Title Title

- Nitrogen fixation rates increase with diazotroph richness in the global ocean

Authors

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

Affiliations

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

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

[3]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, 8093 Zürich, Switzerland.

*Corresponding authors: deriksson@ethz.ch, ssunagawa@ethz.ch, meike.vogt@env.ethz.ch

Abstract

Marine diazotrophs convert atmospheric nitrogen gas into bioavailable forms of nitrogen and are thus critical to maintaining the productivity of the ocean. However, little is known about the link between global-scale diazotroph diversity and marine biological nitrogen fixation rates. Here, we address this question by integrating 22,000 sequencing and microscopy-based observations for 15 diazotroph taxa into an ensemble of 90 species distribution models. Our ensemble of models predicts a strong latitudinal gradient in diazotroph species richness from the tropics to the poles, driven by environmental factors such as temperature, nitrate, phosphate, and silicate concentrations. Non-cyanobacterial diazotrophs show a higher richness in upwelling regions compared to their cyanobacterial counterparts, which exhibit richness hotspots within oligotrophic gyres. Diazotroph richness

is positively correlated with nitrogen fixation rates derived from various independent observational/observation-based estimates. Analyses of community composition suggest a selection driven effect underlying the positive biodiversity-ecosystem function relationship where cyanobacterial diazotrophs are the major drivers of the observed positive relationship. Our findings underscore the pivotal role of diazotroph richness in alleviating nitrogen limitation in marine ecosystems through enhanced marine biological nitrogen fixation rates, thus potentially mitigating adverse climate change impacts on primary production in tropical and temperate ocean regions. Overall, our study supports a biodiversity-ecosystem functioning relationship crucial to the global marine nitrogen cycle.

MAIN TEXT

Introduction

Nitrogen-fixing microorganisms, collectively termed diazotrophs, convert atmospheric dinitrogen gas into fixed forms of bioavailable nitrogen. In this way, they supply a substantial fraction of the nitrogen needed to support primary production in many oligotrophic regions of the tropics and subtropics (*1*, *2*). Marine biological nitrogen fixation (BNF) also resupplies most of the fixed nitrogen that is lost from the ocean as a consequence of denitrification processes (*3*, *4*).

Marine BNF is a highly specialized process that only a handful of bacterial and archaeal groups can perform (*5*). Therefore, the overall richness of diazotrophic species is relatively low compared to the richness of species that provide more widespread ecosystem functions such as photosynthesis. However, recent molecular evaluations of up to 29000 unique sequences associated with the *nifH* gene (*6*), which is critically involved in the marine BNF process, revealed several new species groups capable of marine BNF, substantially

expanding the diversity of known diazotrophs. First global scale collections of diazotroph occurrence and abundance have been compiled (*7*, *8*), and are continuously updated, with recent databases describing cyanobacterial and non-cyanobacterial organism groups at 55286 sampling locations (*9*). The fact that novel diazotroph groups are still being discovered and that global-scale observations remain scarce and biased (*7–9*) highlights our limited understanding of the diversity and biogeographic distribution of this important taxonomic group and its impact on ecosystem functions associated with the global nitrogen and carbon cycle. This knowledge gap limits not only our ability to model the global distribution of diazotrophs in space and time (*7*), but also our ability to assess how the diversity of diazotrophs and the marine BNF will respond to future climate change (*10*).

Historically, it was believed that marine BNF in the global ocean was primarily driven by one group of cyanobacteria, that is, colony-forming species in the genus *Trichodesmium* (*11*, *12*). *Trichodesmium spp.* can be easily recognized using microscopy-based sampling strategies and has been studied for decades. However, a second group of diazotrophs among the genera *Richelia* and *Calothrix* has been discovered, which live in symbiosis with diatoms of the genera *Chaetoceros*, *Hemiaulus* and *Rhizosolenia* (*13*). Their contribution to global marine BNF was believed to be much smaller than that of *Trichodesmium*, mainly due to the strong limitation of silicic acid of host diatoms that prevents them from growing in many low-nutrient regions (*14*).

The advent of culture-independent DNA sequencing methods allowed for exploring the diversity of organisms using the *nifH* gene encoded enzyme nitrogenase that catalyzes the splitting of the nitrogen molecule (*5*). Amplification of the *nifH* gene by PCR and shotgun DNA sequencing using environmental community DNA (metagenomics) revolutionized the identification of diazotrophs and led in the past two decades to the discovery of additional

groups of diazotrophs (*15*, *16*). The first newly discovered group of marine diazotrophs were unicellular cyanobacteria collectively referred to as *UCYN*, with three known subgroups *UCYN-A*, *UCYN-B*, and *UCYN-C (17–19)*. In addition to free-living cells, this group also contains species that are found in symbiosis with photosynthetic eukaryotes or as aggregates such as the colony-forming *Trichodesmiums spp.* (*19*, *20*). Within the Candidatus species *Atelocyanobacterium thalassa* (*UCYN-A*), several strains have been recognized (*A1* to *A6*) (*18*). The strain *UCYN-A2* has been described as a symbiont of the prymnesiophyte algae *Braarudosphaera bigelowii*, and the smaller *UCYN-A1* as an associate of a yet unidentified relative of *B. bigelowii* (*20*). Representatives of the genus *Crocosphaera* (*UCYN-B*) have been found to be free-living, aggregate-forming, or living in symbiosis with the genus *Climacodium,* depending on the specific strain (*19*, *21*). Cyanothece-like *UCYN-C* diazotrophs are presumably small free-living diazotrophs that can form aggregates of up to 500 μm that contribute to the rapid sinking of particulate organic carbon (*17*, *22*). The second group of newly recognized diazotrophs consists of potentially heterotrophic bacteria and archaea, referred to as non-cyanobacterial diazotrophs, which are capable of marine BNF, possibly due to the presence of the *nifH* operon in their genomes (*15*). Although the presence of nitrogen-fixing genes does not necessarily indicate activity, the detection of transcripts suggests additional evidence for the role of non-cyanobacterial diazotrophs in marine BNF (*23*). Although gene transcription indicates gene expression, it does not provide quantitative data. This limitation prevents us from fully assessing the global significance of non-cyanobacterial diazotrophs in marine BNF.

Regarding their impact on global nitrogen cycling, our knowledge of the global magnitude and biogeographic pattern of marine BNF has continuously increased in recent decades (*5*, *8*, *9*, *24*). Although initial estimates were based nearly exclusively on Trichodesmium incubation (*12*), whole community assays and geochemical approaches have expanded our

104 understanding of marine BNF by revealing the contribution of a wider range of diazotrophic
105 organisms (*25–28*). Yet, *in situ* marine BNF measurements remain sparse and extrapolations
106 tentative, with current best estimates of global marine BNF ranging between 85 and 238 Tg
107 N yr^{-1} (*5*, *29*, *30*). Although work has been conducted to quantify the overall diversity of
108 diazotrophs and the gross marine BNF rates (*8*), the relationship between the global
109 diversity of diazotroph species and marine BNF has not been investigated so far. Identifying
110 such a connection would provide evidence for an important biodiversity-ecosystem
111 functioning (BEF) relationship associated with global biogeochemical cycling (*31–33*).
112 Together with the identification of its underlying drivers, the existence of such a BEF
113 relationship may have far-reaching implications for our ability to anticipate and project the
114 response of marine plankton ecosystems and associated global biogeochemical cycles to
115 climate change, particularly through changes in the distribution of diazotrophs, which will
116 alter nitrogen fixation regions and thereby impact marine productivity. Recent modeling
117 work suggests a key control of marine diazotrophy on future marine primary productivity,
118 with the potential to reverse the projected negative trend in NPP to the middle and high
119 latitudes in current climate models (*3*).
120 Biodiversity plays a crucial role in shaping the functioning of ecosystems through its
121 influence on processes such as productivity, stability, and nutrient dynamics (*31*, *34*). In
122 aquatic systems, experimental studies have shown that mixtures of species tend to increase
123 the rate of ecosystem functions, such as productivity, relative to species in monoculture
124 (*32*). Positive BEF relationships can emerge through ecological processes such as niche
125 complementarity (*35*) or selection effects (*36–38*). Through niche complementarity, diverse
126 communities can encompass a wider range of resource utilization strategies compared to
127 less diverse communities, weakening competition through the asynchronous use of
128 resources and maintaining high levels of ecosystem function performance over time (*39–*

41). Through selection effects, higher rates of ecosystem functions could be achieved when a high-performing species dominates community abundance due to its fitter combination of traits (*38*). Understanding which of these processes dominates can be crucial for predicting the impacts of changing biodiversity on ecological processes and nutrient cycling in various ecosystems.

If such a BEF relationship was identified for marine diazotrophy, it would be essential to understand how taxonomic and physiological trait diversity contributes to marine BNF. Cyanobacterial diazotrophs exhibit various ecophysiological strategies to achieve nitrogen fixation, including diurnal and nocturnal fixation patterns and the formation of oxygen-protective heterocysts (*5*). Less is known about the ecophysiological strategies of non-cyanobacterial diazotrophs (*42*). The latter group is distributed globally within the ocean and can be found from surface to depth and therefore shows a variety of strategies such as utilizing different organic carbon sources, forming symbiotic relationships with other organisms, and adapting to various oxygen levels (*42*). Given the substantial diversity of traits in diazotrophs, we hypothesize that this could lead to an increase in the overall marine BNF, as varying optimal environmental conditions allow different taxa to fix nitrogen (*5*). The role of non-cyanobacterial diazotrophs in BEF relationships has remained unclear. Nitrogen fixation rates from non-cyanobacterial diazotrophs are often close to detection limits and, therefore, several orders of magnitude lower when compared to their cyanobacterial counterparts (*12*, *43*). Current evidence suggests that cyanobacterial diazotrophs contribute disproportionately to nitrogen fixation rates, and the principal ecological mechanism underlying the BEF relationship is likely to be driven by a selection effect (*7*, *9*, *12*). Understanding such mechanisms helps us to predict how climate change can impact marine BNF dynamics, potentially altering the biogeography of diazotroph taxa and consequently marine ecosystem productivity and biogeochemical cycling.

Here, we analyze the biogeographic distribution of marine diazotrophs, including cyanobacterial and non-cyanobacterial taxa, by integrating both traditional (microscopy) and sequence-based (qPCR and metagenomics) field records into ensembles of species distribution models (SDMs) to project diazotroph biogeography in space and time. To enhance the reliability of our model projections, we applied an ensemble strategy by averaging predictions from multiple SDMs to address uncertainties related to environmental predictor selections, algorithm choice, and background selection of individual models (see Material and Methods). The latter are hereby referred to as 'ensemble members'. We aggregate the monthly taxon-specific spatial distribution (i.e., maps of presence-absence) predicted from our models to estimate annual global diazotroph richness and analyze its emergent correlation against independent estimates of global marine BNF on an annual mean scale (*9*, *30*). To diagnose spatial changes in community composition underlying species richness patterns and their emergent links with marine BNF, we computed beta diversity indices (i.e., species turnover and nestedness; (*44*). We assume that nestedness is indicative for selection effects, as specific taxa of a community can disproportionately influence high rates of ecosystem functioning and hypothesize that certain species are key to maximize marine BNF. Using this information, we address three questions: (i) What is the global richness pattern of pelagic diazotrophs? (ii) How does this pattern co-vary with marine BNF? (iii) Which ecological mechanisms may help explain the emergent BEF relationship of cyanobacterial, non-cyanobacterial and total diazotroph diversity?

Results

Biogeographic patterns of diazotroph richness

To investigate the global annual diazotroph richness, we computed the annual species richness normalized by the number of species modeled based on their species-specific

presence-absence maps from an ensemble of 90 SDMs. A species was considered as annually present within one grid cell if it was modeled as present for at least one month out of the 12 possible (i.e., annual presence = 1). Diazotroph richness is calculated as the sum of taxa (n = 15) successfully modeled as present by all 90 ensemble members (Fig. S1). We find a strong latitudinal gradient in the global diazotroph species richness (Fig. 1A and B). Analyzed by latitude, diazotroph richness reaches its maximum at ~15.5° north (mean +/- sd = 0.54 +/- 0.09), and ~16.5° south of the equator (mean = 0.54 +/- 0.07) (Fig. 1B). At the equator, the richness drops to a longitudinally integrated mean value of 0.41 (+/-0.07). This drop results from the low annual diazotroph richness predicted for the upwelling region in the equatorial Pacific Ocean. This latitudinal gradient in diazotroph richness is consistent across all ensemble members (Fig. 1B). Considering the fraction of grid cells where the mean diazotroph richness exceeds 0.5, the Indian Ocean ranks highest with 38% of grid cells, followed by the Pacific Ocean (24%) and the Atlantic (8%), highlighting hotspots of diazotroph richness in the central Indian Ocean and within the North Pacific Gyre. The emergent diazotroph richness distribution is robust to differences in sampling methodology, as shown in Fig. S2, where the input data for the modeling pipeline has been re-run using only microscopy-based and sequencing-based observations (Pearson $r = 0.98$, $p < 0.001$). The ensemble spread, a measure of incongruencies across the ensemble members, is generally higher at high latitudes, lowest within intermediate latitudes, and increases again towards the tropics (Fig. S3A). The spread is highest at ~72.5° northern and ~76.5° southern latitudes, and smallest at ~38.5° and ~36.5° northern and southern latitudes. The tropical regions of the northern hemisphere show a higher ensemble spread than the tropical regions of the southern hemisphere. The global median of the diazotroph richness across ensemble members ranges from 0.23 (IQR = 0.14) to 0.44 (IQR = 0.17) (Fig. S4A).

The global richness distribution patterns differed between cyanobacterial (n = 9) and non-cyanobacterial diazotrophs (n = 6) in several ocean areas (Fig. 1C and E). Cyanobacterial diazotroph richness aligns closely with the latitudinal gradient observed for the total community (Spearman's $\rho = 0.89$, $p < 0.001$), with notable declines in upwelling-influenced regions (Fig. 1D). In contrast, non-cyanobacterial diazotroph richness exhibits a latitudinal gradient increasing from poles to tropics without a significant drop at the equator, with these differences being particularly pronounced in the Pacific Ocean (Fig. 1E and F).

The ensemble spread analysis reveals distinct latitudinal patterns in projections uncertainty (Fig. S3B and C). Cyanobacterial diazotroph richness shows a low spread at high latitudes and an increasing spread toward the tropics (Fig. S3B), while non-cyanobacterial diazotroph richness exhibits a higher spread in higher latitudes and a lower spread in mid- and low latitudes (Fig. S3C). These findings underscore the importance of considering spatial variability when assessing diazotroph community dynamics.

Environmental drivers of diazotroph biogeography and diversity

To understand which environmental drivers may govern diazotroph biogeography and diversity, we quantified predictor ranking for the biogeographies of the modeled taxa and assessed correlation of these with diazotroph richness, given that this is an emergent property of our study. To identify the environmental predictors found to be most important for modeling the diazotroph biogeographies by our SDM, we calculated the mean rank based on the adjusted D-squared for General Linear Model (GLM), adjusted R-squared for General Additive Model (GAM) and Out-Of-Bag Error for Random Forest (RF) for each predictor, thus identifying ecologically relevant predictor candidates for inclusion in our models (Fig. S5A; Table S1). Our analysis highlights sea surface temperature as the most

influential predictor for diazotroph biogeography, with nitrate and phosphate concentrations following in second and third rank, respectively (Fig. S5A). Specifically, cyanobacterial taxa exhibited stronger associations with N* (excess concentration of nitrate in relation to the redfield ratio), Si* (the ratio of nitrate to silicic acid) and photosynthetic active radiation (PAR), while non-cyanobacterial diazotrophs showed closer ties to chlorophyll-a, phosphate, and nitrogen concentrations (Fig. S5B), indicating a preference for more productive regions. As to derived diazotroph richness, sea surface temperature explains most of the variation in total diazotroph richness (Adj. R-squared = 0.87), with nutrients such as nitrate (Adj. R-squared = 0.78) and phosphate (Adj. R-squared = 0.68) ranking second and third. This suggests that diazotroph diversity is governed mostly by interactions between temperature, nitrate, and phosphate concentrations.

Diazotroph richness and marine BNF rates

To test whether a BEF relationship exists between diazotroph richness and marine BNF, we analyze the correlation between different estimates of global marine BNF rates and our diazotroph richness for total, cyanobacterial, and non-cyanobacterial diazotroph richness (Fig. 2; Fig. S6).

First, we matched our modeled richness estimates with the model-based marine BNF rates from Wang et al. (*30*). We find an increase in marine BNF rates with increasing total diazotroph richness (Spearman's $\rho = 0.84$, $p < 0.001$; Fig. 2A). Marine BNF rates are higher in locations of higher total diazotroph richness. The positive BEF relationship holds for cyanobacterial and non-cyanobacterial diazotrophs (Spearman's $\rho = 0.86$, $p < 0.001$, Spearman's $\rho = 0.46$, $p < 0.001$; Fig 2B and C). Yet, a notable contrast emerges in the strength of the BEF relationship when the richness of the two groups is analyzed separately. As diazotroph richness increases, marine BNF for cyanobacterial diazotrophs

does not reach a plateau (Fig. 2B), whereas non-cyanobacterial diazotrophs exhibit a plateau at high richness (Fig. 2C), suggesting a potentially reduced impact on BEF dynamics for the non-cyanobacterial community compared to cyanobacterial diazotrophs. To assess whether the diagnosed BEF relationship is independent of the marine BNF rate data product used, we further examine the correlation between our annual total, cyanobacterial, and non-cyanobacterial richness against marine BNF rates from *in situ* measurements (Shao et al. 2023). While the variability is substantially higher in the *in situ* data, we still confirm the positive correlation between total (Fig. S6A; Spearman's $\rho = 0.15$, $p < 0.001$) and cyanobacterial richness (Fig. S6B; Spearman's $\rho = 0.19$, $p < 0.001$). However, we cannot confirm the positive BEF relationship between *in situ* measured marine BNF rates and non-cyanobacterial diazotroph richness (Spearman's $\rho = 0.02$, $p > 0.05$; Fig. S6C). This suggests that the positive BEF relationship is more strongly tied to cyanobacterial diazotrophs than to non-cyanobacterial ones. The scales covered by the two types of marine BNF estimates differ in terms of spatiotemporal resolution, which has implications for the robustness and interpretation of the BEF relationships. While the *in situ* measured marine BNF rates correspond to local discrete measurements integrated over 24 hours influenced by submeso- and mesoscale processes, the estimate from Wang et al. (*30*) is derived from an inverse biogeochemical model that is representative of 1x1° grid boxes and annual mean scales. This result supports that the positive BEF may be stable across spatiotemporal scales.

Ecological mechanisms underlying the BEF between diazotroph richness and global marine BNF rates

To disentangle which ecological mechanisms may underlie the positive BEF relationship above between diazotroph richness and marine BNF rates, we first diagnose how

differences in diazotroph community composition structure the global diazotroph richness gradient. To this end, we computed the beta ratio as the ratio between nestedness and the Jaccard dissimilarity index (Fig. S7A and B). A beta ratio of > 0.5 indicates that the change in community composition is dominated by nestedness rather than species turnover.

On a global scale, the beta ratio is above 0.5 for most ocean basins except the North Atlantic (Fig. 3A). This result indicates that nestedness (Fig. S7B) contributes more to the richness gradient than species turnover (Fig. S7C), indicating that the decrease in richness from the tropical to the polar oceans primarily stems from the loss of certain taxa with increasing latitude, as evidenced by a decreasing beta ratio with decreasing richness, rather than their replacement within the total diazotroph community (Fig. 3B). To further visualize which species are lost as we move from high to low diazotroph richness across different regimes of marine BNF, we show the relative averaged habitat suitability index across each nitrogen fixation regime for each individual taxon (Fig. 3C). The relative average habitat suitability index for all cyanobacterial diazotrophs, except the genus Calothrix, is lower in regions with low annual marine BNF rates (Fig. 3C), whereas non-cyanobacterial diazotrophs exhibit higher habitat suitability in these regions. This further supports the hypothesis of a selection driven effect as cyanobacterial richness is highest within regions of high marine BNF rates.

The beta ratio estimates exhibit much greater variability across ensemble members in comparison to species richness (Fig. S8). However, the majority of the ensemble members (79 out of 90) agree on nestedness being the major factor underlying the global richness gradient. The robustness of this result across ensemble members indicates that variations in diazotroph communities at spatial scales primarily involve the occurrence of subsets of richer diazotroph communities, while species turnover plays a minor role.

We acknowledge that a part of this pattern could be due to species succession in time, but we have checked the change in community composition at monthly scale and found that within tropical regions where we find highest richness patterns, communities are rather stable on a monthly resolution (Fig. S9).

Discussion

This work provides the first observation-based estimate of the relationship between global diazotroph diversity distribution and marine BNF rates. The integration of classical microscopy-based and sequence-based (qPCR and metagenomic) data sources allowed us to model the biogeography of 15 diazotroph taxa (Fig. S1) spanning multiple life history strategies and including non-cyanobacterial diazotrophs for which global biogeographies have not yet been described (Fig. S10). In accordance with previous SDM-based studies based on plankton occurrences (*45–48*), as well as studies based on metagenomic surveys (*6*, *9*, *49*), we found diazotroph diversity to be highest in tropical and subtropical regions and to decrease towards the poles (Fig. 1). This pattern is in line with the diversity distribution of other plankton functional groups and higher trophic levels (*47*, *50*), implying that the environmental drivers of species diversity are common to widely different marine clades on a global scale (*45*, *47*, *51*).

Our analysis further unveiled complementary patterns of cyanobacterial and non-cyanobacterial diazotroph richness with some overlap in richness hotspots in the Indian and Pacific Ocean (Fig. 1C and E). This is further supported by a study by Weiyi Tang and Nicolas Cassar (*8*) who found a high degree of niche overlap for cyanobacterial diazotrophs in those regions. Less is known about the global richness patterns of non-cyanobacterial diazotrophs (*42*). However, our results showing increased richness in tropical upwelling influenced regions are in line with recent evidence of samples taken during the *Tara* Oceans expedition that identified non-cyanobacterial diazotrophs within

the free living and particle-attached fraction of nutrient rich upwelling influenced waters (*6*). The differences between the biogeographies of these two groups suggest that non-cyanobacterial diazotrophs are the dominant taxa in nutrient-rich upwelling regions where high primary production leads to higher amounts of particulate organic matter (*52*), a carbon source that may create oxygen-depleted microniches that could be occupied by this group. This conclusion is corroborated by abundance-based modeling of individual diazotroph taxa (*53*), although existing data on diazotroph biomass remain too scarce and heterogeneous to allow for comprehensive abundance-based modeling.

We found a positive relationship between marine BNF and diazotroph richness (Fig. 2A) that is robust to bin-size choice (Fig. S11), independent of the temporal integration scale (Fig. S9), independent of data type (traditional versus metagenomic, see Fig. S2), and independent of the global marine BNF data source chosen (model-based (Fig. 2) and *in situ* measurements (Fig. S6)). However, our correlative methods preclude the identification of causal links between marine BNF and diazotroph richness, which means further laboratory or model-based work is required to shed light on the underlying physiological and ecological mechanisms. However, the current literature on positive BEF relationships provides us with ample evidence to formulate plausible hypotheses for the underlying ecological mechanisms that may be tested in future mechanistic or experimental work regarding diazotrophs (*5*). Temporal differences in marine BNF activities can lead to enhanced co-existence and increased resource use efficiency, and thus enhanced community-level marine BNF rates, in particular under stable environmental conditions such as those prevailing in the tropics (*54*), where our SDMs have projected the highest diazotroph richness estimates (Fig. 1A). The high energetic costs and oxygen sensitivity of the nitrogen fixing enzyme led to a variety of strategies enabling marine BNF (*5*). Depending on the prevailing environmental conditions, some

nitrogen fixing strategies are favored over others, which can be understood as a form of temporal niche complementarity among diazotrophic species. While our data suggests a limited role of succession at the monthly scale (Fig. S9), especially in the tropical ocean, this may play a role in temperate oceans or different seasons (*55*). Due to the temporal variability in environmental conditions, several strategies may be expressed throughout the diurnal cycle, at daily or weekly scales, which may lead to a cumulative effect of submonthly succession patterns on total annual and monthly mean marine BNF rates that are not currently resolved in our approach, which may allow one to maintain efficient marine BNF through temporal niche partitioning (*56*). This is further supported by the differing environmental optima of individual response curves fitted by our SDMs across the different taxa (Fig. S12). Temporal niche partitioning supports the insurance hypothesis of biodiversity (57), suggesting that higher species richness increases the likelihood of maintaining ecosystem functions like marine BNF, which may explain why regions with higher marine BNF coincide with regions of higher diazotroph species richness on an annual scale.

To diagnose what ecological process underlies the positive BEF, we computed the ratio of the nestedness component to the total dissimilarity based on the Jaccard's index. We found that the loss of cyanobacterial diazotrophs from the community contributes most to the changes in community composition along a gradient of marine BNF (Fig. 3). When using *in situ* marine BNF rates representative of finer scales, cyanobacterial richness was positively correlated, while non-cyanobacterial richness showed no statistically significant correlation to marine BNF (Fig. S6B and C). This suggests that cyanobacterial diazotroph diversity potentially controls global marine BNF rates, while non-cyanobacterial diazotrophs may play a minor role. Such a selection driven effect has been found in another model-based study by Pedro Cermeño et al. (*38*), although not explicitly

for diazotrophs, but marine phytoplankton. The study found a general positive BEF relationship between phytoplankton richness and marine primary productivity. The growth model that best described the observed productivities, was the model where dominant species of the community approached their maximum growth rates. Given that a positive BEF is identified between annual diazotroph richness and global marine BNF rates, the analogy between productivity and marine BNF suggests that maintaining high diazotroph diversity seems crucial for optimizing marine BNF in marine environments via selection effects, where species richness increases the likelihood of communities to harbor highly productive species such as *Trichodesmium* (*12*). *In situ* measured non-cyanobacterial BNF rates have been close to the detection limit and therefore several orders of magnitude smaller when compared to cyanobacterial marine BNF rates, supporting the view that cyanobacterial diazotrophs contribute relatively more to marine BNF (*42*).

Our findings highlight the role of temperature and nutrient concentrations in driving diazotroph richness (Fig. S5) and are in line with other studies (*8*). Temperature exhibits a first order control on the distribution of diazotrophs species richness and lend support to the kinetic energy hypothesis, where higher temperatures promote higher species diversity through increased speciation rates and the selection of warm-water-tolerant species (45–47, 50). As anthropogenic climate change continues to raise ocean temperatures and alter nutrient distributions through ocean stratification (58), our results suggest that the diversity of diazotroph communities could increase substantially, particularly in the South Pacific and the Central Atlantic, as discussed in Dutkiewicz et al. (*59*) and thus increase the magnitude of marine BNF in the tropical and temperate ocean. This potential increase in marine BNF rates could constitute a negative feedback to climate change impacts on surface ocean nutrient availability and stimulate ocean

productivity, potentially counteracting the increasing nitrogen limitation caused by ocean stratification and mitigating the negative effects on primary production (*60*). Since diazotrophy is a major source of uncertainty in NPP projections (*3*), understanding these dynamics is crucial for predicting how climate change will affect ocean ecosystems and their capacity to support marine life and global nutrient cycles in the future.

In addition to the valuable insights gained from this study, several limitations should be acknowledged. Our study is based on a limited set of 15 diazotroph taxa (Fig. S1). Although we provide first estimates for global diazotroph diversity, these likely do not capture the full spectrum of diazotroph diversity existing. While the number of marine diazotrophs has increased with the advent of sequence-based surveys, current estimates based on *nifH* sequences mapped against a gene catalog of unique *nifH* sequences counted up to 762 nifH sequences that mapped with at least 80% similarity in Tara Oceans datasets (6). Based on a current estimate identifying 34 operational taxonomic units (OTUs) of non-cyanobacterial diazotrophs (42), we were able to model 18% of these OTUs, which is similar to the percentage of diversity of marine plankton captured by other recent modeling work (10-30%; (*45*, *47*)). A successive removal of one species at a time implies that up to the removal of nine species, the clustering quality remains relatively stable (range Mean Silhouette Index: 0.43 to 0.45; Fig. S13). This indicates that the clustering structure is robust to the removal of up to 60% of species.

The findings of this study open avenues for valuable contributions to future research in marine ecology and biogeochemistry. The identification of a positive relationship between diazotroph richness, especially cyanobacterial diazotroph richness, and marine BNF provides a foundation for deeper exploration of the ecological mechanisms driving these relationships. The recognition of non-cyanobacterial diazotrophs in nutrient-rich upwelling regions raises further questions on their activity within such marine regions

since it is not yet clear if they are capable of marine BNF in the environment, although genetic elements for marine BNF are present within their genome**.**

Importantly, our study's approach of integrating diverse data sources, including microscopy and sequence-based observations, provides a methodological framework for addressing challenges associated with data scarcity in microbial plankton research. This approach allowed us to maximize the number of observations and to show that observations either retrieved from microscopy or sequence-based methodologies lead to similar global richness patterns of marine diazotrophs (Fig. S2), indicating that the origin of the data source has no effect that substantially biases our SDM results. 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.

Our work shows a positive relationship between diazotroph diversity and global marine BNF, primarily driven by cyanobacterial diazotrophs. This underscores the importance of understanding how increased marine BNF could influence global biogeochemical cycles, counteract nitrogen limitation from ocean stratification, and mitigate impacts on primary production, which is crucial for predicting climate change effects on ocean ecosystems.

Materials and Methods

Experimental Design

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), (*7*), (*8*), (*25*), (*27*), (*28*), Phytobase (*61*), (*6*) and (*62*) (Fig. S14).

We included further quality controls on observations that have been retrieved via public databases such as GBIF or OBIS. We used an ocean mask (*63*) 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 (*15*). We used the metagenomic assembled genomes (MAGs) computed by (*15*) and screened the Ocean

Microbiomics Database (*62*) which compiles data from (*64*), (*23*), (*65*), (*66*), (*15*), (*67*) and (*68*) for matching metagenomic operational taxonomic units (mOTUs) to retrieve a taxonomic annotation of those genomes using the Genome Taxonomy Database (*69*). 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 S2) (*45*, *70*, *71*). 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 (*72*) and from regions characterized by climatological surface salinities below 20 (*73*).

Species distribution models

SDMs fit statistical associations between species' observed occurrences and environmental variables; i.e., they estimate a species' realized environmental niche of a taxon by fitting a response curve between the distributions of occurrence data and variables (Fig. S12; (*70*). 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 (*74*) 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 (*74*),

and rapidly respond when conditions turn suitable (*33*). Since the distribution patterns of diazotrophic taxa are likely to change seasonally (*75*), 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. (*45*) and Benedetti et al. (*46*) 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 (*45*, *76*). 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. (*77*).

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 (*61*) 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 (*45*) 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.

Statistical analysis

All statistical analysis was performed in R version 4.2.2.

Algorithm and complexity choice

Statistical algorithm choice represents a main source of uncertainty in studies relying on SDMs (*78*). 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. (*79*), 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. (*77*). 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 D2 for GLM, adjusted R2 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 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 S3). 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 (*80*). 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. S2 shows a heatmap of TSS scores of each successfully modeled diazotroph across all 90 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 (*81*).

Diazotroph diversity

Richness is the simplest measure of diversity and corresponds to the number of species present within one assemblage (i.e., grid cell). We stacked the monthly SDM projections from each ensemble member of the successfully modeled taxa (n = 15) within each grid cell. For each model, we then calculated the annual richness estimate by summing up the annual average species-level presence-absence maps. The number of species per area was then computed as the mean richness of each grid cell across all 90 ensemble members. As an annual presence, we assigned a presence when the taxon was present at least once in twelve month. The final map on global annual diazotroph richness distribution is therefore

computed as an ensemble of 90 models each including 15 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), six different background selection strategies (total target-group, group-specific, cruise-specific with overlapping and non-overlapping options) and five predictor settings per taxon. 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 types of SDMs covers the main source of uncertainty in SDMs based studies and increases the generalization of the main diversity patterns we estimate (Diniz-Filho et al., 2009; Thuiller et al., 2019).

To estimate beta diversity and investigate the ecological processes underlying 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 *beta.pair* function from the *betapart* package in R (*82*) 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 calculate the pairwise dissimilarity between all site pairs and then average the results for each grid cell across all other grid cells. In this calculation, '*a*' represents the number of shared species between two cells, '*b*' stands for the number of species unique to the poorer site, and '*c*' denotes the number of species unique to the richer site. All beta diversity components are based on the annual presence maps (as defined above). Moreover, the beta ratio was computed as $ß_{ratio} = ß_{jne}/ß_{jac}$ to identify which component has the highest contribution to total dissimilarity. Here, 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 and diazotroph richness, we used the global marine BNF estimates originating from two independent studies. Wang et al. (*30*) 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 marine BNF compiled by Shao et al. (*9*). To make both marine 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 marine 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). To assess the effect of removing species on clustering quality on the global richness projection, we implemented a function to calculate the Silhouette Index ((*83*); Fig. S13), using hierarchical clustering. We performed a permutation test to calculate the Mean Silhouette Index after sequentially removing 1 to 14 species from the dataset. For each number of removed species, 100 iterations were conducted to ensure robustness, with the mean Silhouette Index recorded for each removal scenario.

References

1. T. Shiozaki, D. Bombar, L. Riemann, M. Sato, F. Hashihama, T. Kodama, Tanita, Iwan, Takeda, Shigenobu, Saito, Hiraoki, Hamasaki, Koji, Furuya, Ken, Linkage Between Dinitrogen Fixation and Primary Production in the Oligotrophic South Pacific Ocean. *Global Biogeochem. Cycles* **32**, 1028–1044 (2018).

2. D. Karl, R. Letelier, L. Tupas, J. Dore, J. Christian, D. Hebel, The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* **Nature**, 533–538 (1997).

3. L. Bopp, O. Aumont, L. Kwiatkowski, C. Clerc, L. Dupont, C. Ethé, T. Gorgues, R. Séférian, A. Tagliabue, Diazotrophy as a key driver of the response of marine net primary productivity to climate change. *Biogeosciences* **19**, 4267–4285 (2022).

4. N. Gruber, J. N. Galloway, An Earth-system perspective of the global nitrogen cycle. *Nature* **451**, 293–296 (2008).

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

6. J. J. Pierella Karlusich, E. Pelletier, F. Lombard, M. Carsique, E. Dvorak, S. Colin, M. Picheral, F. M. Cornejo-Castillo, S. G. Acinas, R. Pepperkok, E. Karsenti, C. de Vargas, P. Wincker, C. Bowler, R. A. Foster, Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat. Commun.* **12**, 4160 (2021).

7. Y.-W. Luo, S. C. Doney, L. A. Anderson, M. Benavides, I. Berman-Frank, A. Bode, S. Bonnet, K. H. Boström, D. Böttjer, D. G. Capone, E. J. Carpenter, Y. L. Chen, M. J. Church, J. E. Dore, L. I. Falcón, A. Fernández, R. A. Foster, K. Furuya, F. Gómez, K. Gundersen, A. M. Hynes, D. M. Karl, S. Kitajima, R. J. Langlois, J. LaRoche, R. M. Letelier, E. Marañón, D. J. McGillicuddy Jr, P. H. Moisander, C. M. Moore, B. Mouriño-Carballido, M. R. Mulholland, J. A. Needoba, K. M. Orcutt, A. J. Poulton, E. Rahav, P. Raimbault, A. P. Rees, L. Riemann, T. Shiozaki, A. Subramaniam, T. Tyrrell, K. A. Turk-Kubo, M. Varela, T. A. Villareal, E. A. Webb,

A. E. White, J. Wu, J. P. Zehr, Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst. Sci. Data* **4**, 47–73 (2012).

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

9. Z. Shao, Y. Xu, H. Wang, W. Luo, L. Wang, Y. Huang, Y.-W. Luo, Version 2 of the global oceanic diazotroph database. *Earth Syst. Sci. Data Discuss.*, doi: 10.5194/essd-2023-13 (2023).

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

11. 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).

12. D. G. Capone, J. A. Burns, J. P. Montoya, A. Subramaniam, C. Mahaffey, T. Gunderson, A. F. Michaels, E. J. Carpenter, Nitrogen fixation byTrichodesmiumspp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochem. Cycles* **19** (2005).

13. 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).

14. 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).

15. T. O. Delmont, C. Quince, A. Shaiber, Ö. C. Esen, S. T. Lee, M. S. Rappé, S. L. McLellan, S. Lücker, A. M. Eren, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).

16. H. Farnelid, A. F. Andersson, S. Bertilsson, W. A. Al-Soud, L. H. Hansen, S. Sørensen, G. F. Steward, Å. Hagström, L. Riemann, Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* **6**, e19223 (2011).

17. A. N. Knapp, S. E. Fawcett, A. Martínez-Garcia, N. Leblond, T. Moutin, S. Bonnet, Nitrogen isotopic evidence for a shift from nitrate- to diazotroph-fueled export production in the VAHINE mesocosm experiments. *Biogeosciences* **13**, 4645–4657 (2016).

18. 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).

19. 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).

20. A. Thompson, B. J. Carter, K. Turk-Kubo, F. Malfatti, F. Azam, J. P. Zehr, Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ. Microbiol.* **16**, 3238–3249 (2014).

21. R. A. Foster, M. M. M. Kuypers, T. Vagner, R. W. Paerl, N. Musat, J. P. Zehr, Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. *ISME J.* **5**, 1484–1493 (2011).

22. S. Bonnet, M. Benavides, F. A. C. Le Moigne, M. Camps, A. Torremocha, O. Grosso, C. Dimier, D. Spungin, I. Berman-Frank, L. Garczarek, F. M. Cornejo-Castillo, Diazotrophs are overlooked contributors to carbon and nitrogen export to the deep ocean. *ISME J.* **17**, 47–58 (2023).

23. G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, Tara Oceans Coordinators, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).

24. 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).

25. M. R. Gradoville, H. Farnelid, A. E. White, K. A. Turk-Kubo, B. Stewart, F. Ribalet, S. Ferrón, P. Pinedo-Gonzalez, E. V. Armbrust, D. M. Karl, S. John, J. P. Zehr, 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).

26. M. Chen, Y. Lu, N. Jiao, J. Tian, S.-J. Kao, Y. Zhang, Biogeographic drivers of diazotrophs in the western Pacific Ocean. *Limnol. Oceanogr.* **64**, 1403–1421 (2019).

27. 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).

28. C. Martínez-Pérez, W. Mohr, C. R. Löscher, J. Dekaezemacker, S. Littmann, P. Yilmaz, N. Lehnen, B. M. Fuchs, G. Lavik, R. A. Schmitz, J. LaRoche, M. M. M. Kuypers, The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**, 16163 (2016).

29. J. N. Galloway, F. J. Dentener, D. G. Capone, E. W. Boyer, R. W. Howarth, S. P. Seitzinger, G. P. Asner, C. C. Cleveland, P. A. Green, E. A. Holland, D. M. Karl, A. F. Michaels, J. H. Porter, A. R. Townsend, C. J. Vöosmarty, Nitrogen Cycles: Past, Present, and Future. *Biogeochemistry* **70**, 153–226 (2004).

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

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

32. L. Gamfeldt, J. S. Lefcheck, J. E. K. Byrnes, B. J. Cardinale, J. E. Duffy, J. N. Griffin, 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. F. van der Plas, Biodiversity and ecosystem functioning in naturally assembled communities. *Biol. Rev. Camb. Philos. Soc.* **94**, 1220–1245 (2019).

35. L. W. Aarssen, High Productivity in Grassland Ecosystems: Effected by Species Diversity or Productive Species? *Oikos* **80**, 183–184 (1997).

36. J. J. Stachowicz, J. F. Bruno, J. E. Duffy, Understanding the Effects of Marine Biodiversity on Communities and Ecosystems. *Annu. Rev. Ecol. Evol. Syst.* **38**, 739–766 (2007).

37. P. B. Reich, J. Knops, D. Tilman, J. Craine, D. Ellsworth, M. Tjoelker, T. Lee, D. Wedin, S. Naeem, D. Bahauddin, G. Hendrey, S. Jose, K. Wrage, J. Goth, W. Bengston, correction: Plant diversity enhances ecosystem responses to elevated CO2 and nitrogen deposition. *Nature* **411**, 824 (2001).

38. P. Cermeño, P. Chouciño, B. Fernández-Castro, F. G. Figueiras, E. Marañón, C. Marrasé, B. Mouriño-Carballido, M. Pérez-Lorenzo, T. Rodríguez-Ramos, I. G. Teixeira, S. M. Vallina, Marine Primary Productivity Is Driven by a Selection Effect. *Frontiers in Marine Science* **3** (2016).

39. M. E. S. Bracken, Complementarity in spatial subsidies of carbon associated with resource partitioning along multiple niche axes. *Oecologia* **193**, 425–436 (2020).

40. J. Otero, X. A. Álvarez-Salgado, A. Bode, Phytoplankton Diversity Effect on Ecosystem Functioning in a Coastal Upwelling System. *Frontiers in Marine Science* **7** (2020).

41. S. Mensah, B. du Toit, T. Seifert, Diversity-biomass relationship across forest layers: implications for niche complementarity and selection effects. *Oecologia* **187**, 783–795 (2018).

42. K. A. Turk-Kubo, M. R. Gradoville, S. Cheung, F. Cornejo-Castillo, K. J. Harding, M. Morando, M. Mills, J. P. Zehr, Non-cyanobacterial diazotrophs: Global diversity, distribution, ecophysiology, and activity in marine waters. *FEMS Microbiol. Rev.*, doi: 10.1093/femsre/fuac046 (2022).

43. P. H. Moisander, M. Benavides, S. Bonnet, I. Berman-Frank, A. E. White, L. Riemann, Chasing after Non-cyanobacterial Nitrogen Fixation in Marine Pelagic Environments. *Front. Microbiol.* **8**, 1736 (2017).

44. R. Omidipour, P. Tahmasebi, M. Faal Faizabadi, M. Faramarzi, A. Ebrahimi, Does β diversity predict ecosystem productivity better than species diversity? *Ecol. Indic.* **122**, 107212 (2021).

45. 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).

46. F. Benedetti, M. Vogt, U. H. Elizondo, D. Righetti, N. E. Zimmermann, N. Gruber, Major restructuring of marine plankton assemblages under global warming. *Nat. Commun.* **12**, 5226 (2021).

47. F. Benedetti, N. Gruber, M. Vogt, Global gradients in species richness of marine plankton functional groups. *J. Plankton Res.* **45**, 832–852 (2023).

48. F. M. Ibarbalz, N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J.

Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, Tara Oceans Coordinators, M. Babin, E. Boss, D. Iudicone, O. Jaillon, S. G. Acinas, H. Ogata, E. Pelletier, L. Stemmann, M. B. Sullivan, S. Sunagawa, L. Bopp, C. de Vargas, L. Karp-Boss, P. Wincker, F. Lombard, C. Bowler, L. Zinger, Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **179**, 1084–1097.e21 (2019).

49. S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, Tara Oceans Coordinators, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard, H. Ogata, S. Pesant, M. B. Sullivan, P. Wincker, C. de Vargas, Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).

50. D. P. Tittensor, C. Mora, W. Jetz, H. K. Lotze, D. Ricard, E. V. Berghe, B. Worm, Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010).

51. T. O. Gagné, G. Reygondeau, C. N. Jenkins, J. O. Sexton, S. J. Bograd, E. L. Hazen, K. S. Van Houtan, Towards a global understanding of the drivers of marine and terrestrial biodiversity. *PLoS One* **15**, e0228065 (2020).

52. I. Stuhldreier, C. Sánchez-Noguera, F. Roth, J. Cortés, T. Rixen, C. Wild, Upwelling Increases Net Primary Production of Corals and Reef-Wide Gross Primary Production Along the Pacific Coast of Costa Rica. *Frontiers in Marine Science* **2** (2015).

53. N. S. Knecht, F. Benedetti, U. Hofmann Elizondo, N. Bednaršek, S. Chaabane, C. de Weerd, K. T. C. A. Peijnenburg, R. Schiebel, M. Vogt, The impact of zooplankton calcifiers on the marine carbon cycle. *Global Biogeochem. Cycles* **37** (2023).

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

55. A. J. Schlenger, S. Libralato, L. T. Ballance, Temporal Variability of Primary Production Explains Marine Ecosystem Structure and Function. *Ecosystems* **22**, 331–345 (2019).

56. K. S. McCann, The diversity-stability debate. *Nature* **405**, 228–233 (2000).

57. M. Loreau, *From Populations to Ecosystems: Theoretical Foundations for a New Ecological Synthesis (MPB-46)* (Princeton University Press, 2010).

58. G. Li, L. Cheng, J. Zhu, K. E. Trenberth, M. E. Mann, J. P. Abraham, Increasing ocean stratification over the past half-century. *Nat. Clim. Chang.* **10**, 1116–1123 (2020).

59. S. Dutkiewicz, B. A. Ward, J. R. Scott, M. J. Follows, Understanding predicted shifts in diazotroph biogeography using resource competition theory. *Biogeosciences* **11**, 5445–5461 (2014).

60. W. Fu, J. T. Randerson, J. K. Moore, Climate change impacts on net primary production (NPP) and export production (EP) regulated by increasing stratification and phytoplankton community structure in the CMIP5 models. *Biogeosciences* **13**, 5151–5170 (2016).

61. 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).

62. L. Paoli, H.-J. Ruscheweyh, C. C. Forneris, F. Hubrich, S. Kautsar, A. Bhushan, A. Lotti, Q. Clayssen, G. Salazar, A. Milanese, C. I. Carlström, C. Papadopoulou, D. Gehrig, M.

Karasikov, H. Mustafa, M. Larralde, L. M. Carroll, P. Sánchez, A. A. Zayed, D. R. Cronin, S. G. Acinas, P. Bork, C. Bowler, T. O. Delmont, J. M. Gasol, A. D. Gossert, A. Kahles, M. B. Sullivan, P. Wincker, G. Zeller, S. L. Robinson, J. Piel, S. Sunagawa, Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).

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

64. S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, Tara Oceans coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

65. S. J. Biller, P. M. Berube, K. Dooley, M. Williams, B. M. Satinsky, T. Hackl, S. L. Hogle, A. Coe, K. Bergauer, H. A. Bouman, T. J. Browning, D. De Corte, C. Hassler, D. Hulston, J. E. Jacquot, E. W. Maas, T. Reinthaler, E. Sintes, T. Yokokawa, S. W. Chisholm, Marine microbial metagenomes sampled across space and time. *Sci Data* **5**, 180176 (2018).

66. S. G. Acinas, P. Sánchez, G. Salazar, F. M. Cornejo-Castillo, M. Sebastián, R. Logares, M. Royo-Llonch, L. Paoli, S. Sunagawa, P. Hingamp, H. Ogata, G. Lima-Mendez, S. Roux, J. M. González, J. M. Arrieta, I. S. Alam, A. Kamau, C. Bowler, J. Raes, S. Pesant, P. Bork, S. Agustí, T. Gojobori, D. Vaqué, M. B. Sullivan, C. Pedrós-Alió, R. Massana, C. M. Duarte, J. M. Gasol, Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol* **4**, 604 (2021).

67. T. Klemetsen, I. A. Raknes, J. Fu, A. Agafonov, S. V. Balasundaram, G. Tartari, E. Robertsen, N. P. Willassen, The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.* **46**, D692–D699 (2018).

68. M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, R. Stepanauskas, Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623–1635.e11 (2019).

69. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

70. P. Brun, M. Vogt, M. R. Payne, N. Gruber, C. J. O'Brien, E. T. Buitenhuis, C. Le Quéré, K. Leblanc, Y.-W. Luo, Ecological niches of open ocean phytoplankton taxa. *Limnol. Oceanogr.* **60**, 1020–1038 (2015).

71. 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).

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

73. M. M. Zweng, D. Seidov, T. P. Boyer, M. Locarnini, H. E. Garcia, A. V. Mishonov, O. K. Baranova, K. Weathers, C. R. Paver, I. Smolyar, Others, World ocean atlas 2018, volume 2: Salinity. (2019).

74. 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).

75. J. Ladau, T. J. Sharpton, M. M. Finucane, G. Jospin, S. W. Kembel, J. O'Dwyer, A. F. Koeppel, J. L. Green, K. S. Pollard, Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* **7**, 1669–1677 (2013).

76. S. J. Phillips, M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, S. Ferrier, Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* **19**, 181–197 (2009).

77. 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).

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

79. P. Brun, W. Thuiller, Y. Chauvier, L. Pellissier, R. O. Wüest, Z. Wang, N. E. Zimmermann, Model complexity affects species distribution projections under climate change. *J. Biogeogr.* **47**, 130–142 (2020).

80. C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, S. Lautenbach, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46 (2013).

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

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

83. A. Starczewski, A. Krzyżak, "Performance Evaluation of the Silhouette Index" in *Artificial Intelligence and Soft Computing* (Springer International Publishing, Cham, 2015)*Lecture notes in computer science*, pp. 49–58.

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. MV and MB acknowledge funding through BiOcean5D, which was co-funded by the European Union (GA#101059915 - BIOcean5D) and has received funding from the Swiss State Secretariat

for Education, Research and Innovation (SERI) under contract No. 22.00255. DR acknowledges funding from the Swiss National Science Foundation, grant no. P500PB_203241. The authors thank Mar Benavides for inputs on our analysis within 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. We thank Mridul Thomas and Alexandre Schickele for fruitful discussion and ideas. AI-assisted have been used as aids in writing.

Funding:

EC Horizon 2020 Framework Programme (H2020) 862923

BNP Paribas, Fondation BNP Paribas (BNP Paribas Foundation) 1-006245-000

Swiss National Science Foundation (SNSF) grant no. 205321_184955 and P500PB_203241.

BiOcean5D European Union (BIOcean5D) grant no. GA#101059915

Author contributions:

Conceptualization: DE, NG, DR, FB, GS, SS, MV

Data Curation: DE & DR, LP

Methodology: DE, DR, FB

Software: DE, DR

Formal analysis: DE, FB

Visualization: DE

Writing - Original Draft: DE based on a report by DR and DE

Writing - Review & Editing: DE, NG, FB, DR, LP, SS, MV

Supervision: NG, SS, MV

Funding acquisition: NG, SS, MV

Competing interests:

The authors have no conflict of interest to report.

Data and materials availability:

All codes and data used for this analysis are publicly available. The dataset is stored at the ETH Zurich Research collection with the doi: 10.3929/ethz-b-000635803 and codes are publicly available at the GitHub repository https://github.com/Domle/Notion.git.

Figures and Tables

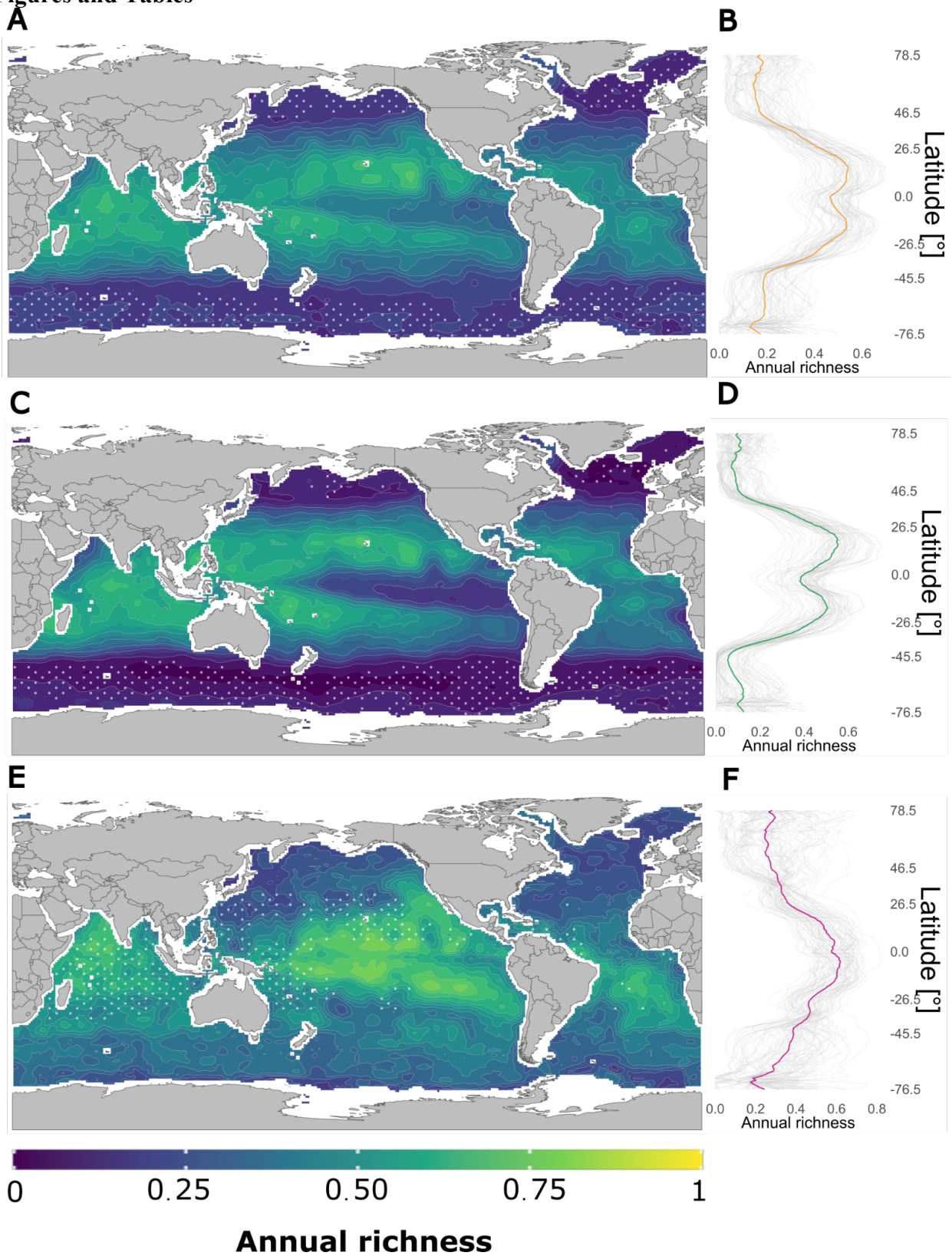

Fig. 1. Distribution of global diazotroph richness. Global maps of the annual diazotroph richness computed across 90 ensemble members, with each grid cell covering 1° longitude by 1° latitude for **(A)** the total (n = 15; orange line), **(C)** the cyanobacterial (n = 9; green line) and **(E)** the non-cyanobacterial (n = 6; purple line) diazotroph community. We established the criterion for annual presence as the occurrence of at least one species within a grid cell for a minimum of one month over a twelve-month period. Thus, a value of 1 indicates an annual presence of all diazotrophs at the specified locations for at least one month annually. White stipples indicate areas where the coefficient of variation was above the 70th percentile, marking greater differences between model projections. Line plots showing the global 1° binned latitudinal annual diazotroph richness gradients for each of the 90 members (grey lines) for **(B)** the total, **(D)** the cyanobacterial and **(F)** the non-cyanobacterial diazotroph community. The black line is the mean across all 90 models.

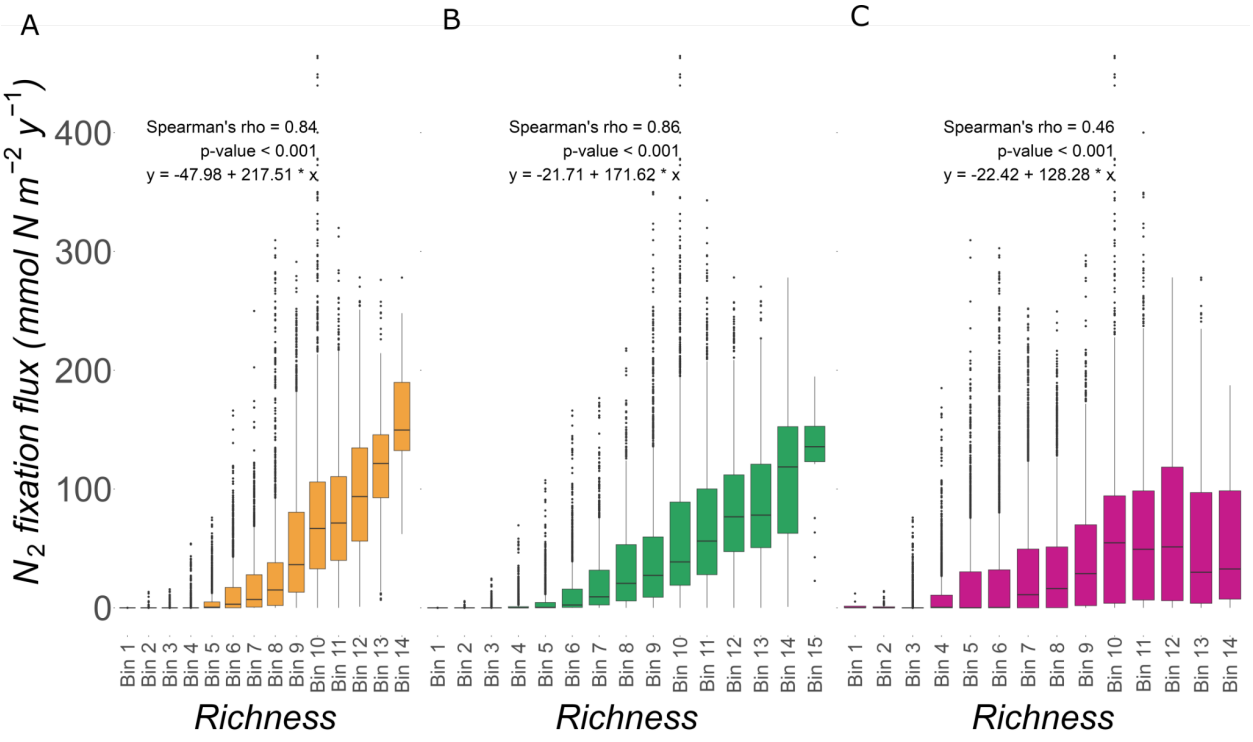

Fig. 2. Relationship between diazotroph richness and biological nitrogen fixation.

Boxplots between colocated nitrogen fixation rates (mmol N m^{-2} y^{-1}) (Wang et al. 2019) and annual diazotroph richness estimates for **(A)** the total (n = 15; orange boxplots), **(B)** cyanobacterial (n = 9; green boxplots) and **(C)** non-cyanobacterial diazotroph community (n = 6; purple boxplots). The annual diazotroph richness is computed across 90 ensemble members with each grid cell covering 1° longitude by 1° latitude and has been normalized by the number of taxa. An annual presence has been assigned when a taxon was present at least once in twelve month. Therefore a value of 1 indicates that all taxa have been present at least one month of the year across all ensemble members. Spearman's correlation coefficient, associated p-values and linear fit are given in the upper left corner. The bins illustrate normalized richness intervals, each beginning from 0.05 and spanning a width of 0.05.

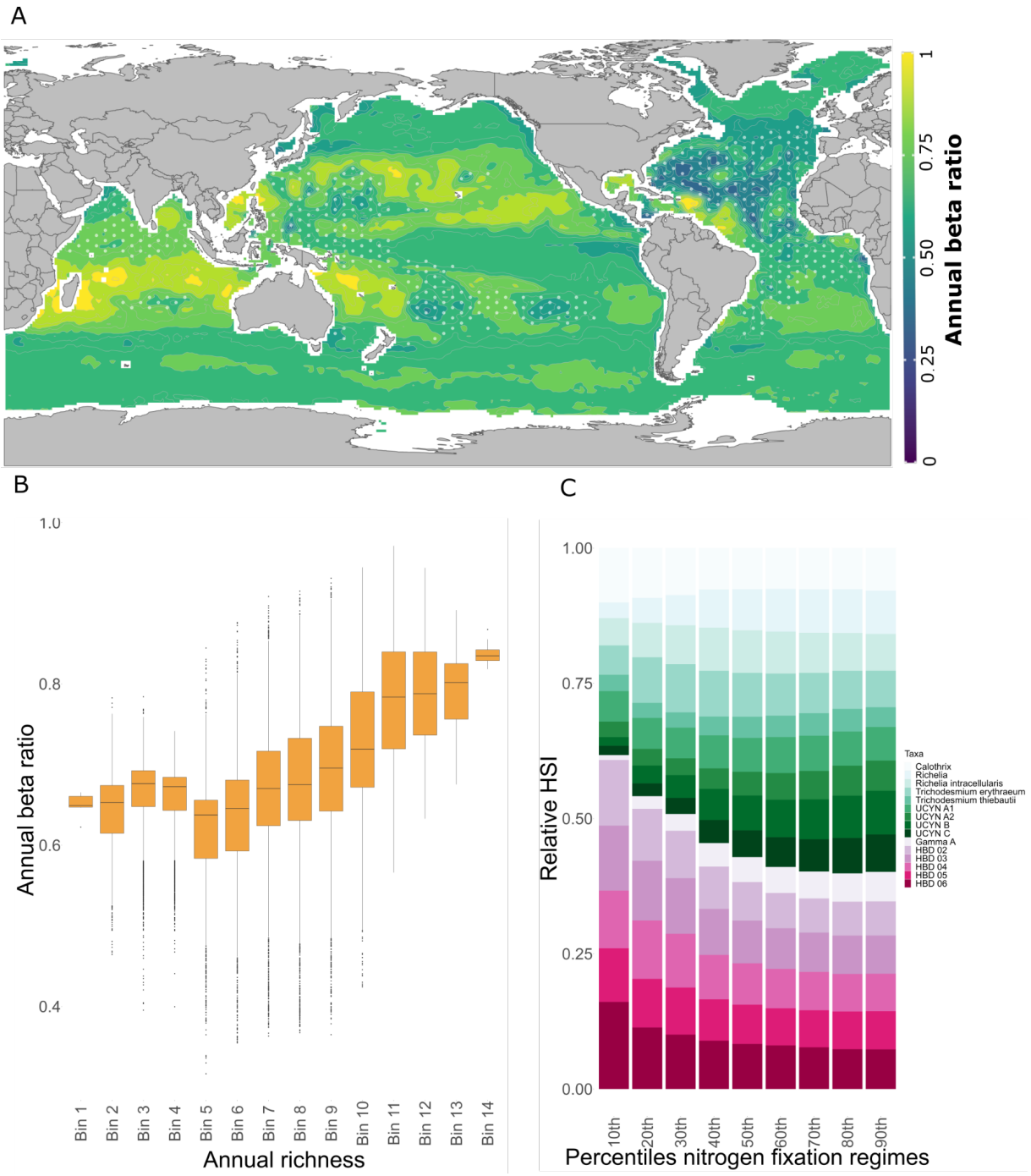

Fig. 3. Global annual beta ratio. (A) A map of the global annual beta ratio (nestedness/jaccard) computed across 90 ensemble members , with each grid cell covering 1° longitude by 1° latitude. A value of > 0.5 indicating that the nestedness is the major contributor underlying the change in community composition. **(B)** Boxplots of the annual beta ratio (y-axis) and the annual richness (x-axis). The bins illustrate normalized richness intervals, each beginning from 0.05 and spanning a width of 0.05. **(C)** The stacked bar chart depicts the average habitat suitability

indices for individual diazotroph taxa, with cyanobacterial taxa represented in green and non-cyanobacterial taxa in purple. The y-axis illustrates the relative proportion of mean annual habitat suitability. The x-axis represents locations ranked from highest (90th percentile) to lowest (10th percentile) annual nitrogen fixation rates Wang et al. (2019).

Supplementary Materials for

Nitrogen fixation rates increase with diazotroph richness in the global ocean

Dominic Eriksson* *et al.*

*Corresponding author. Email:deriksson@ethz.ch, ssunagawa@ethz.ch, meike.vogt@env.ethz.ch

This PDF file includes:

Figs. S1 to S14
Tables S1 to S3

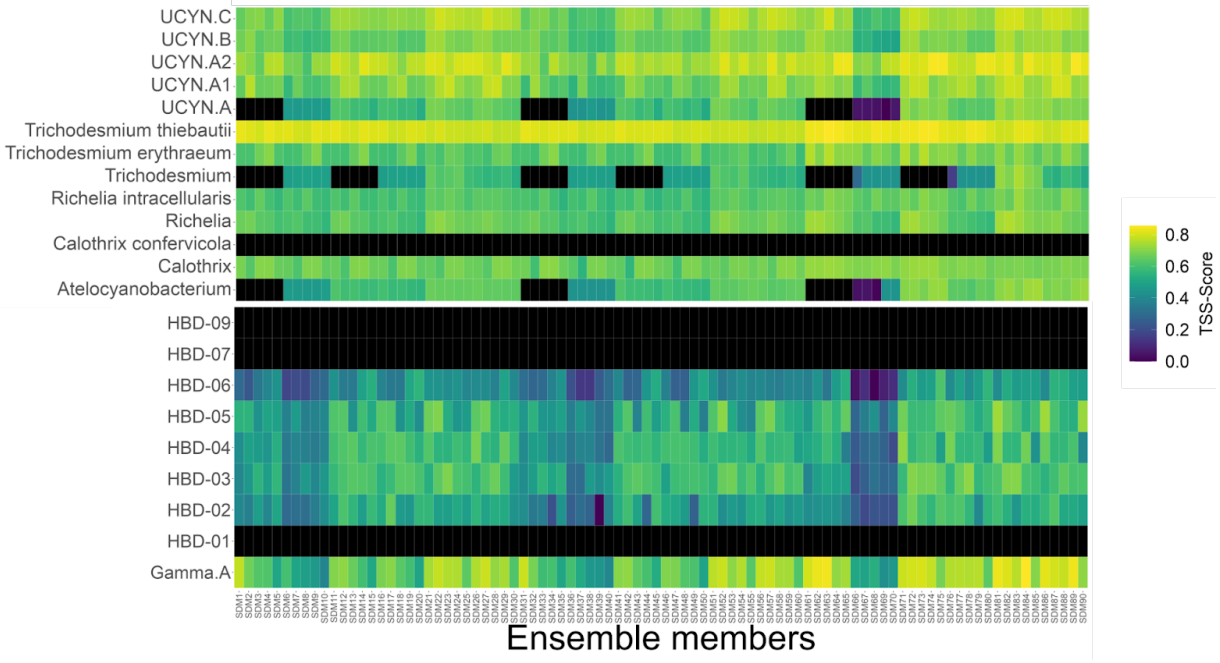

Fig. S1.

Heatmap of True Skill Statistic (TSS) scores for each modelled diazotrophic taxa (y-axis) and all 90 ensemble members (x-axis). Theoretical TSS scores range from minus one to one, and black color indicates unsuccessful member models at TSS scores below 0.3. For our analysis we only kept taxa that were modeled successfully by all 90 ensemble members to maintain consistency across taxa.

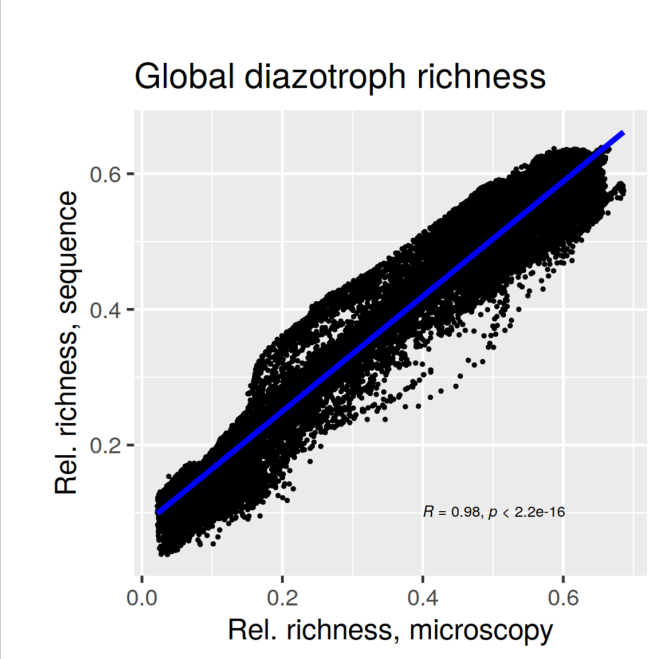

Fig. S2.

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 General Additive Model. Both axes show the normalized richness for each location (i.e., global ocean grid cell).

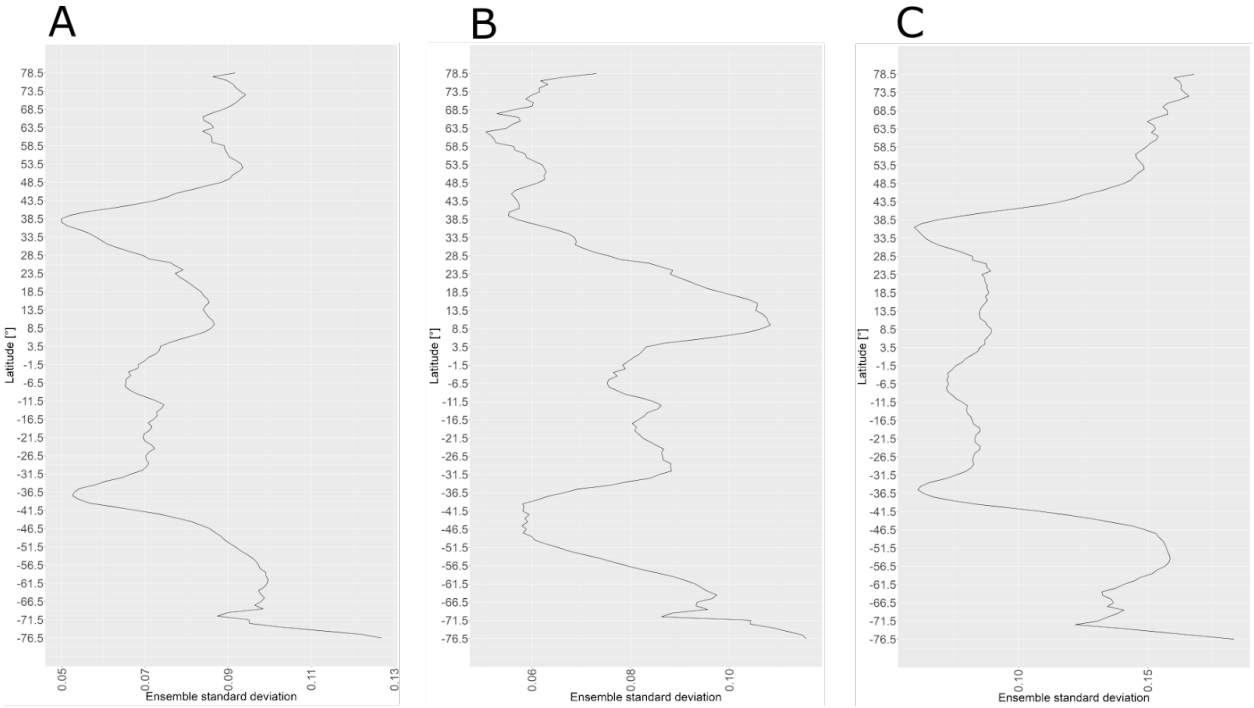

Fig. S3.

Standard deviations of yearly averaged diazotroph richness estimates across latitudinal 1° bins for **(A)** total, **(B)** cyanobacterial, and **(C)** non-cyanobacterial diazotroph richness. Diazotroph richness is computed as the time-averaged species richness normalized by the number of species modelled using presence-absence maps, based on an ensemble of 90 species distribution models (SDMs). An annual presence is defined as a species occurring at least once in twelve months within a grid cell. Richness values are based on the sum of taxa modelled successfully by all 90 ensemble members (n = 15).

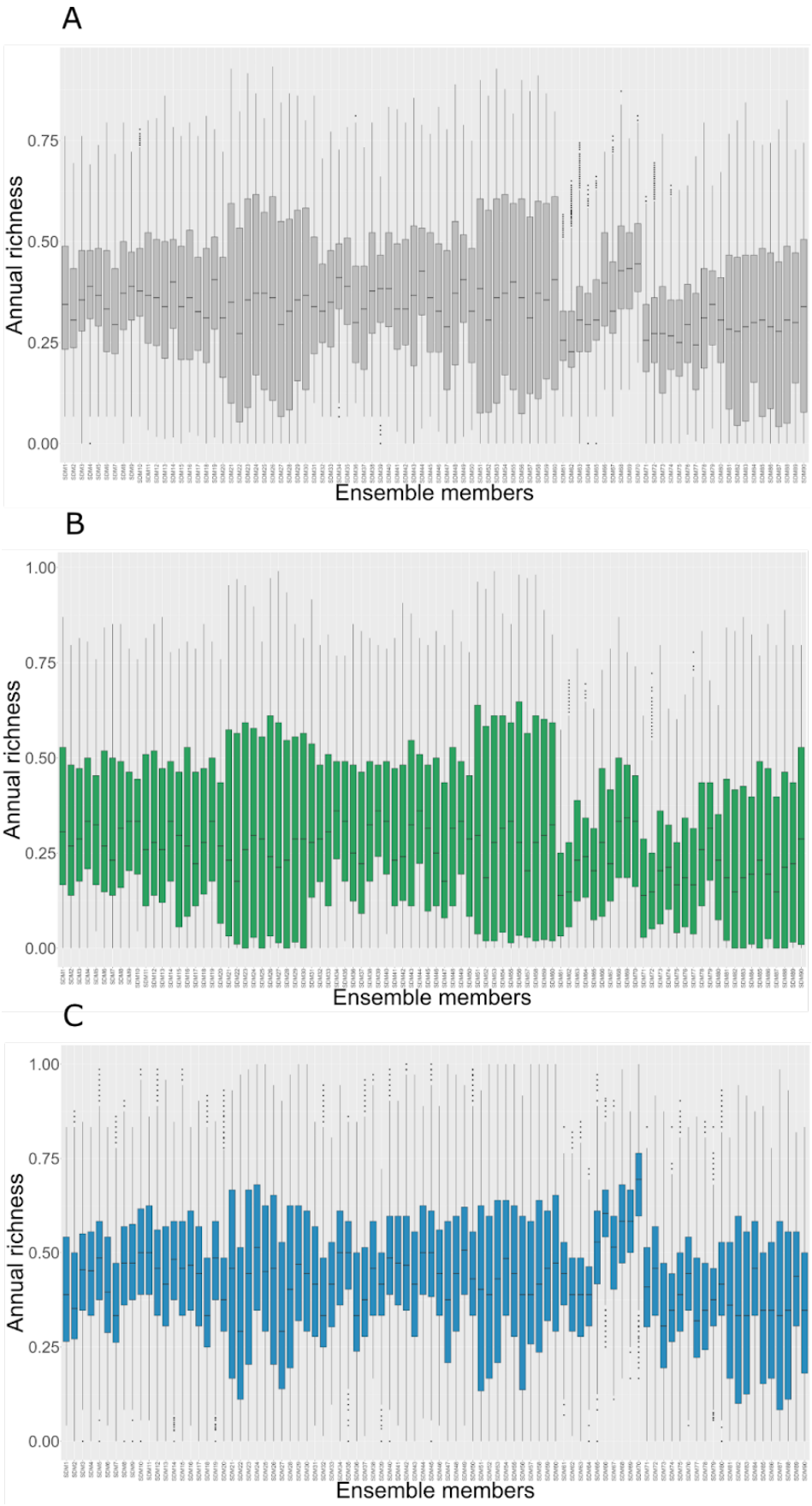
A
Annual richness
0.00
0.25
0.50
0.75
Ensemble members
B
Annual richness
0.00
0.25
0.50
0.75
1.00
Ensemble members
C
Annual richness
0.00
0.25
0.50
0.75
1.00
Ensemble members

Fig. S4.

Boxplots showing the range of annual diazotroph richness estimates (y-axis) across each one of the 90 ensemble members (x-axis) for **(A)** total (n = 15), **(B)** cyanobacterial (n = 9), and **(C)** non-cyanobacterial (n = 6) diazotroph community. The annual diazotroph richness is computed with each grid cell covering 1° longitude by 1° latitude and has been normalized by the number of taxa. An annual presence has been assigned when a taxon was present at least once in twelve month. Therefore a value of 1 indicates that all taxa have been present at least one month of the year across all ensemble members.

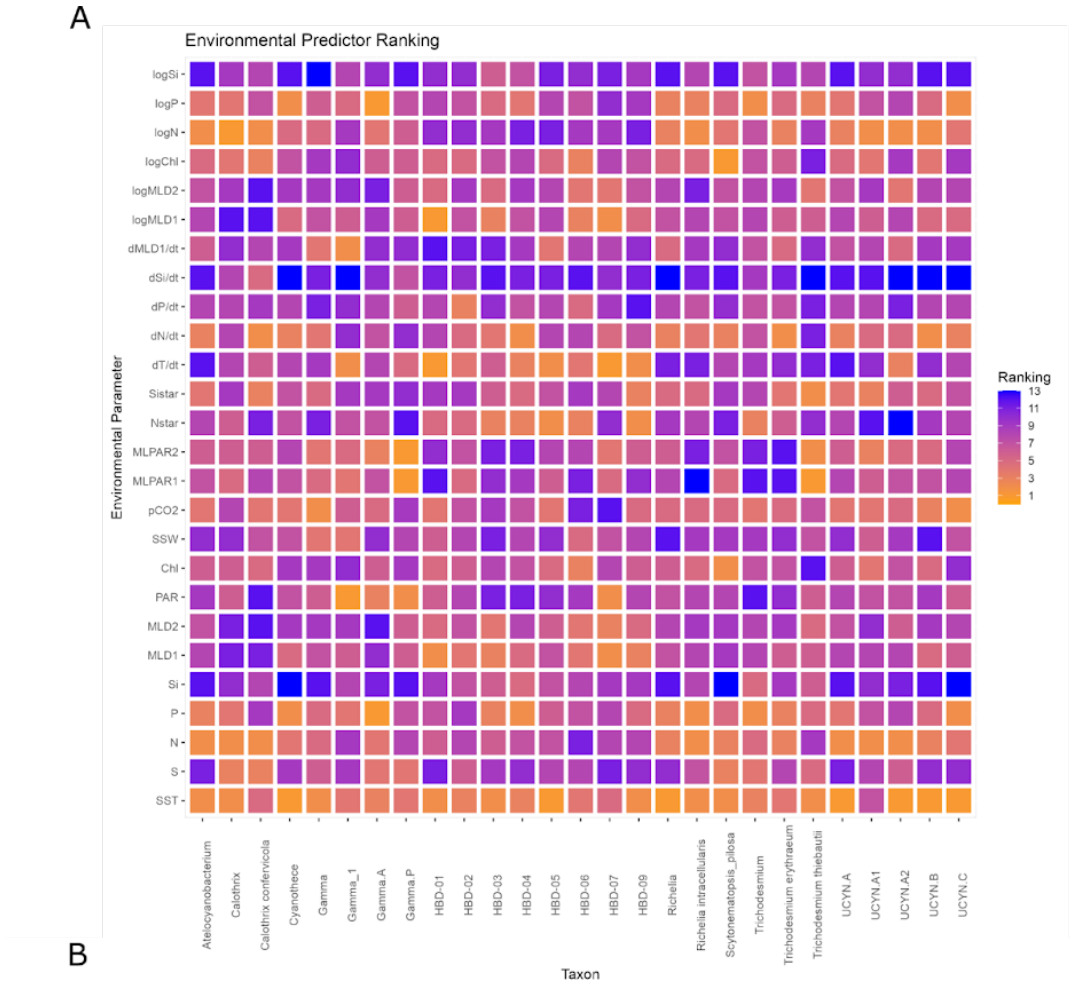
A
Environmental Predictor Ranking
Environmental Parameter
logSi
logP
logN
logChl
logMLD2
logMLD1
dMLD1/dt
dSi/dt
dP/dt
dN/dt
dT/dt
Sistar
Nstar
MLPAR2
MLPAR1
pCO2
SSW
Chl
PAR
MLD2
MLD1
Si
P
N
S
SST
Ranking
13
11
9
7
5
3
1
Ateleocyanobacterium
Calothrix
Calothrix confervicola
Cyanothece
Gamma
Gamma_1
Gamma-A
Gamma-P
HBD-01
HBD-02
HBD-03
HBD-04
HBD-05
HBD-06
HBD-07
HBD-09
Richelia
Richelia intracellularis
Scytonematopsis_pilosa
Trichodesmium
Trichodesmium erythraeum
Trichodesmium thiebautii
UCYN-A
UCYN-A1
UCYN-A2
UCYN-B
UCYN-C
Taxon
B
Wind.CCMP
MLD1
Sal
Sistar
Nstar
Gamma-A
Chl
N
P
Calothrix
HBD-05
MLPAR1
PAR
T
pCO2
CCA1
CCA2
-1.0
-0.5
0.0
0.5
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8

Fig. S5.

(A) Mean ranking of environmental parameters based on single factor analysis on each diazotroph species applying three different algorithms namely General Linear Model General Additive Model, and a Random Forest. 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 $m^{-2}s^{-1}$), Chl (chlorophyll, µg l^{-1}), Wind.CCMP (sea surface wind stress, m s^{-1}), pCO2 (carbon dioxide partial pressure in the surface sea, µatm), MLPAR1 (photosynthetically available radiation over the mixed layer depth, µmol $m^{-2}s^{-1}$), 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 l^{-1}), logN (logarithmic nitrate concentration, µM), logP (logarithmic phosphate concentration, µM), logSi (logarithmic silicic acid concentration, µM). **(B)** Non-metric multidimensional scaling for the main environmental predictors ranked highest. Centroids have been computed on the annually averaged projected habitat suitability indices projected by our SDM on a 1° longitude by 1° latitude spatial resolution and is an average across 90 ensemble members. Environmental variables were matched up with each co-located annual habitat suitability index for each taxon.

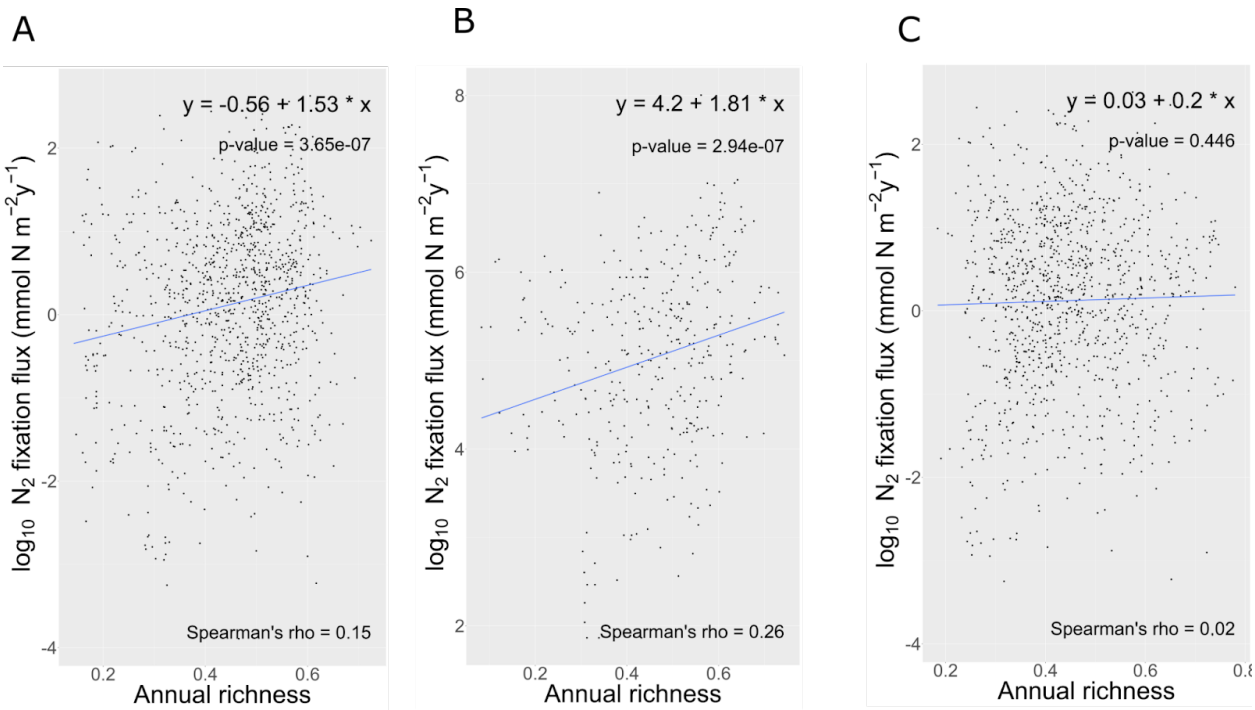

Fig.S6.

Correlation between grid cells that contain *in situ* nitrogen fixation measurements compiled by from Luo et al. (2014), Bonnet et al. (2017) and Shao et al. (2023) colocated modeled annual diazotroph richness of **(A)** total (n = 15), **(B)** cyanobacterial (n = 9) and **(C)** non-cyanobacterial (n = 6) diazotrophs. Spearman rank correlation coefficient is given in the lower right corner along with the linear fit and p-value in the top right. The annual diazotroph richness is computed with each grid cell covering 1° longitude by 1° latitude and has been normalized by the number of taxa and averaged across 90 ensemble members. Within each ensemble an annual presence has been assigned when a taxon was present at least once in twelve month. Therefore a value of 1 indicates that all taxa has been present at least one month of the year across all ensemble members.

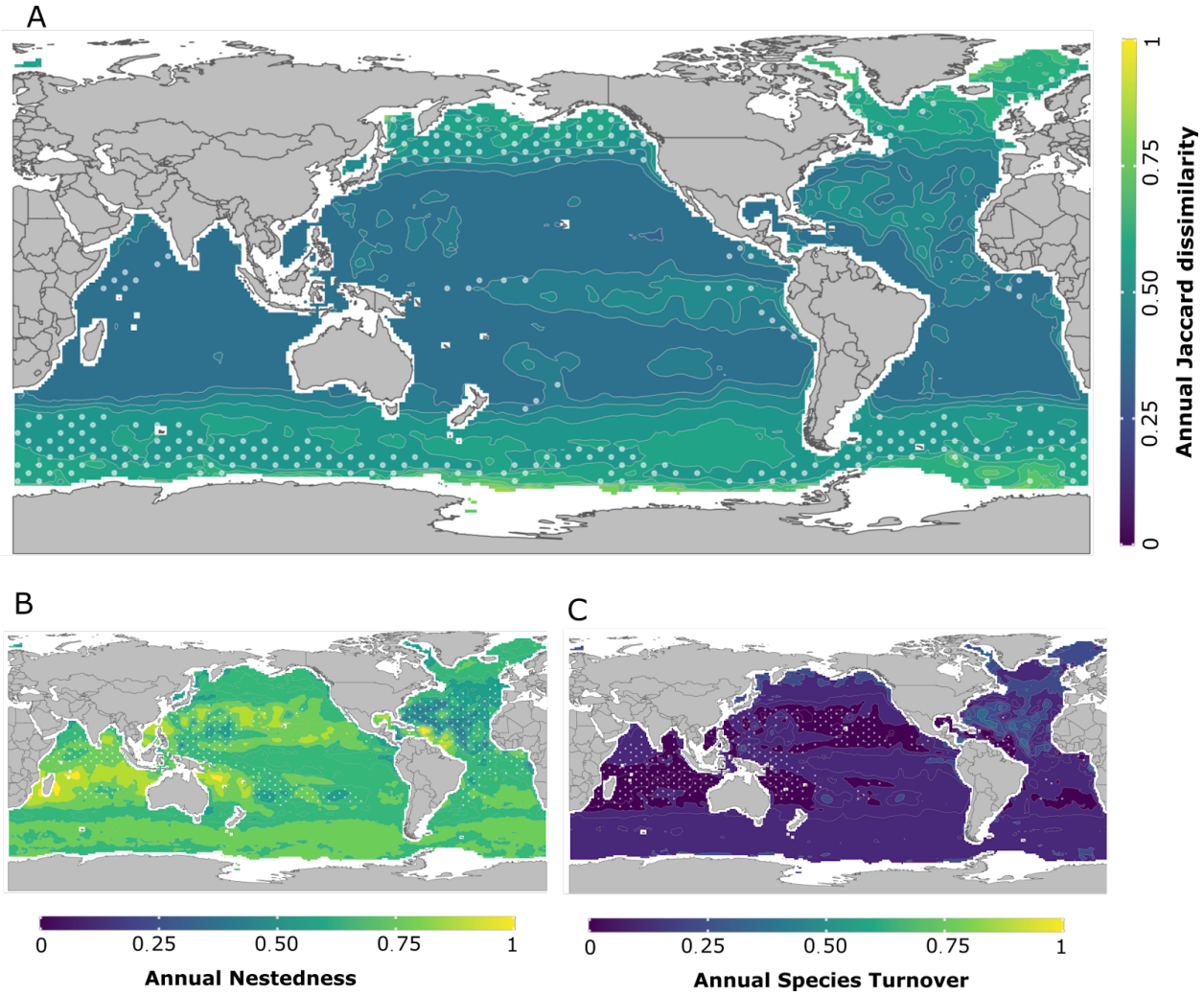

Fig.S7.

Global maps of beta diversity. **A)** The annual Jaccard dissimilarity, **B)** annual nestedness and **C)** annual species turnover. All three beta diversity indices were computed across 90 ensemble members, with each grid cell covering 1° longitude by 1° latitude for the total diazotroph community (n = 15). We established the criterion for annual presence as the occurrence of at least one species within a grid cell for a minimum of one month over a twelve-month period. Thus, a value of 1 indicates the presence of all diazotrophs at the specified locations for at least one month annually. White stipples indicate areas where the coefficient of variation was above the 70th percentile, marking greater differences between model projections.

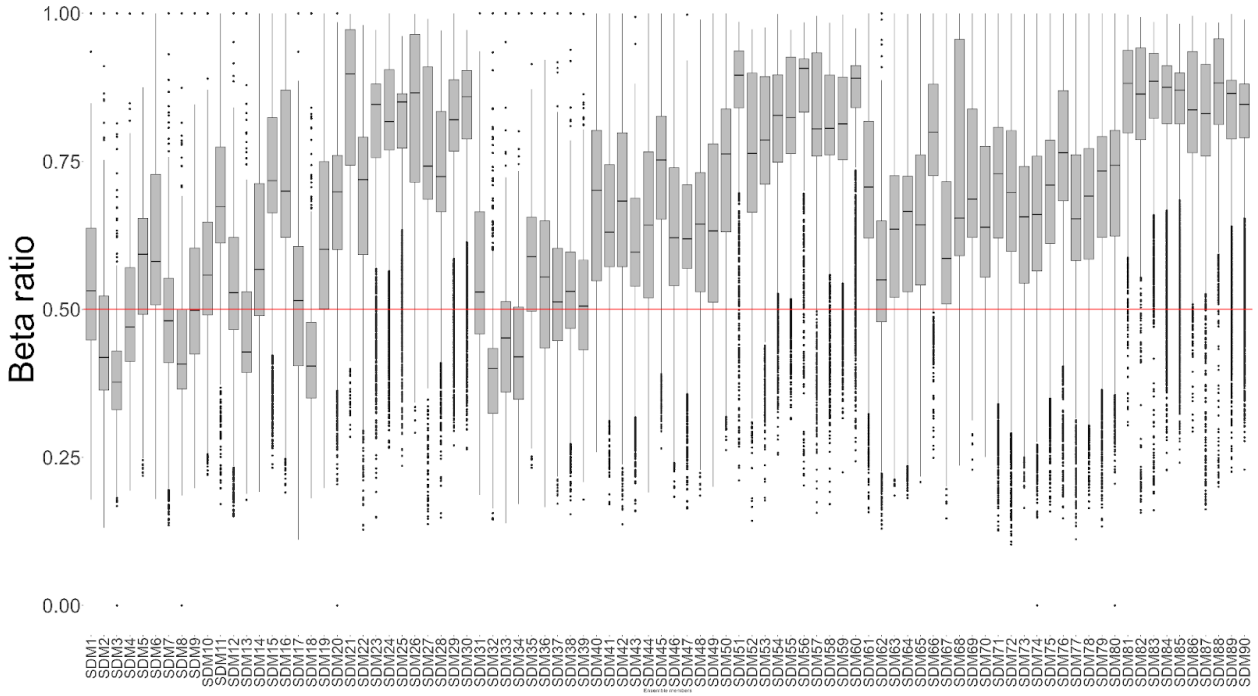

Fig. S8.

Boxplots showing the range of annual diazotroph beta ratio estimates (y-axis) across each one of the 90 ensemble members (x-axis) for a total of 15 modeled diazotrophs. The annual diazotroph beta ratio is computed as the ratio of nestedness over Jaccard dissimilariy with each grid cell covering 1° longitude by 1° latitude. The red line has been placed to visualize the threshold of either nestedness (> 0.5) or species turnover (< 0.5) being the dominant driver underlying the global richness pattern.

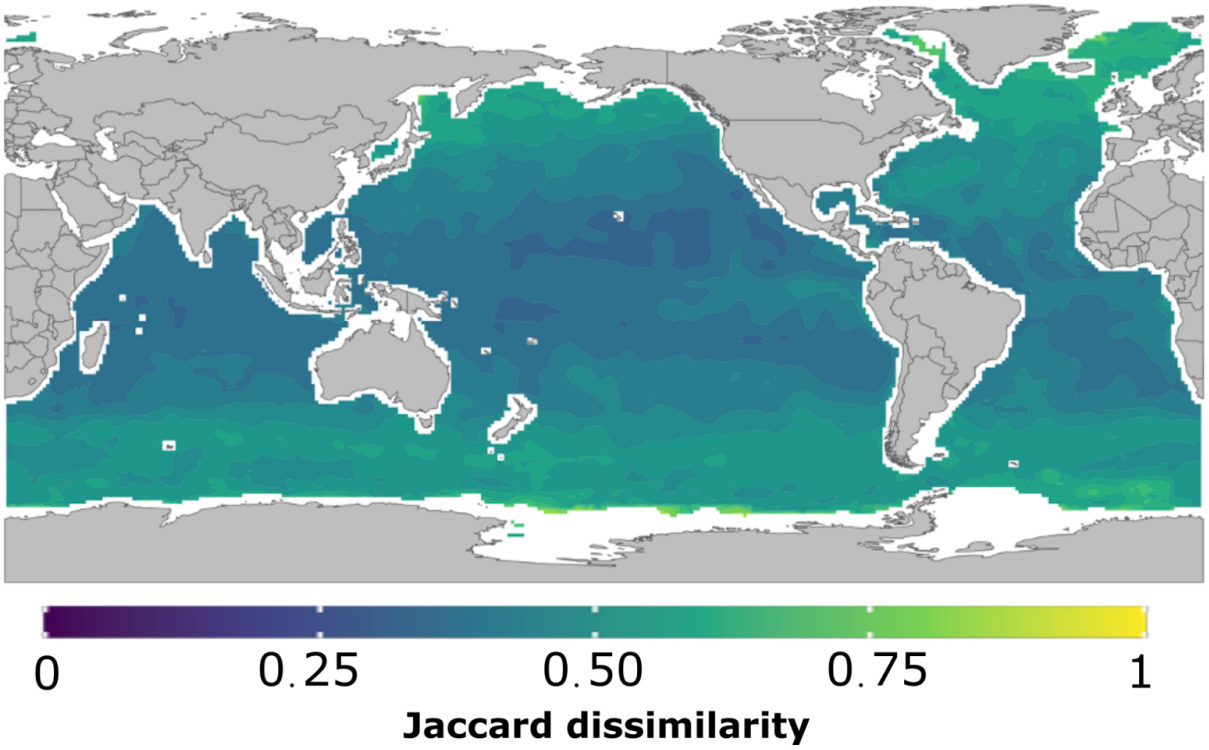

Fig. S9.

Global maps of Jaccard dissimilarity. The Jaccard dissimilarity for each grid cell computed as the mean across all 12 monthly pairs. The Jaccard dissimilarity is computed as an ensemble across 90 ensemble members, with each grid cell covering 1° longitude by 1° latitude for the total diazotroph community (n = 15). We established the criterion for annual presence as the occurrence of at least one species within a grid cell for a minimum of one month over a twelve-month period. A value of zero indicates that the communities remain stable all year around, showing no change in community composition on a temporal (monthly) scale.

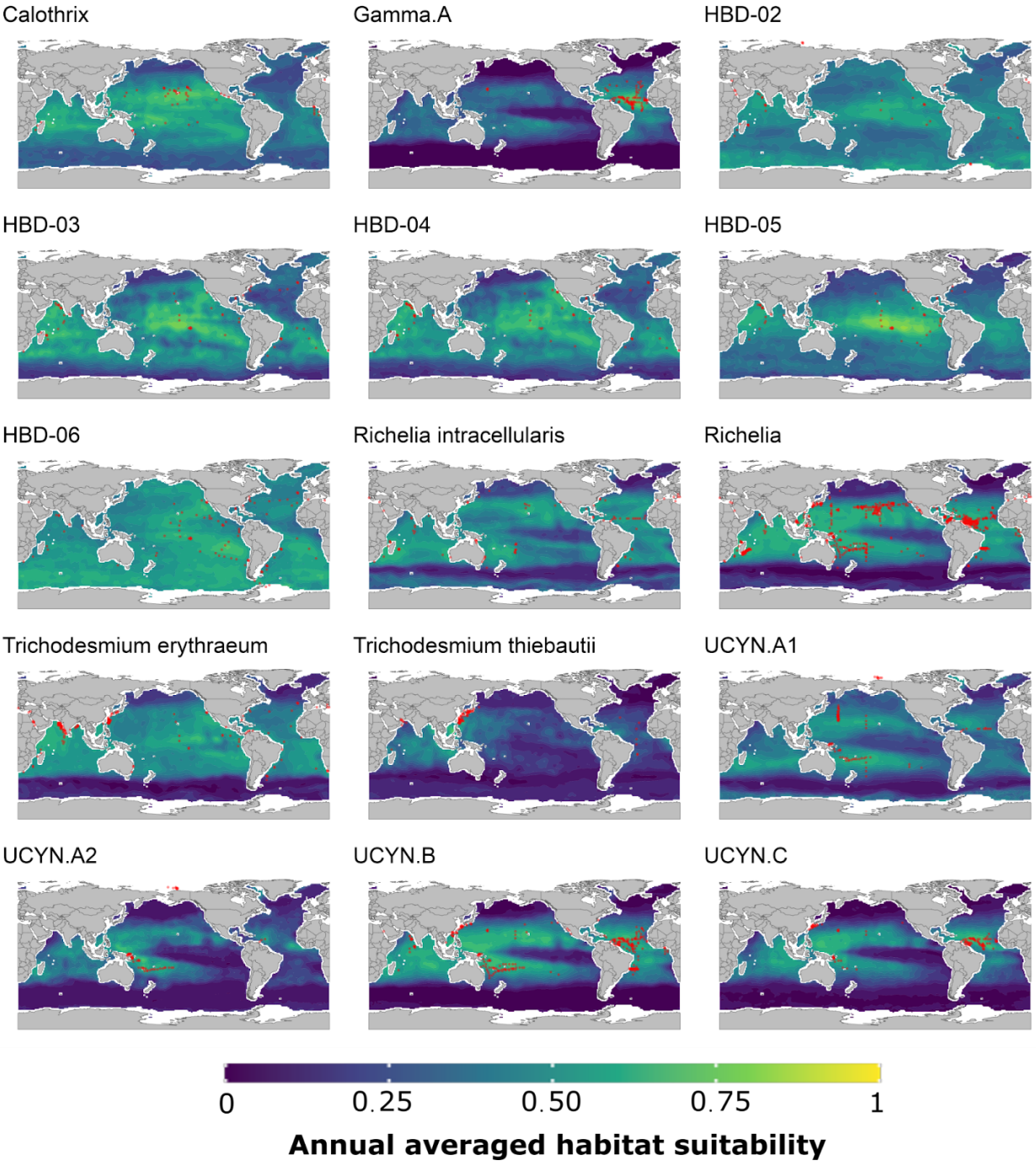

Fig. S10.

Species or genus-level spatial distribution of diazotrophs computed as the time-averaged habitat suitability per taxon per area across 12 months and 90 ensemble members. Each taxon was modeled separately for each month on a 1° longitude 1° latitude spatial resolution. Red points indicate the gridded presences that were available for the modeling pipeline.

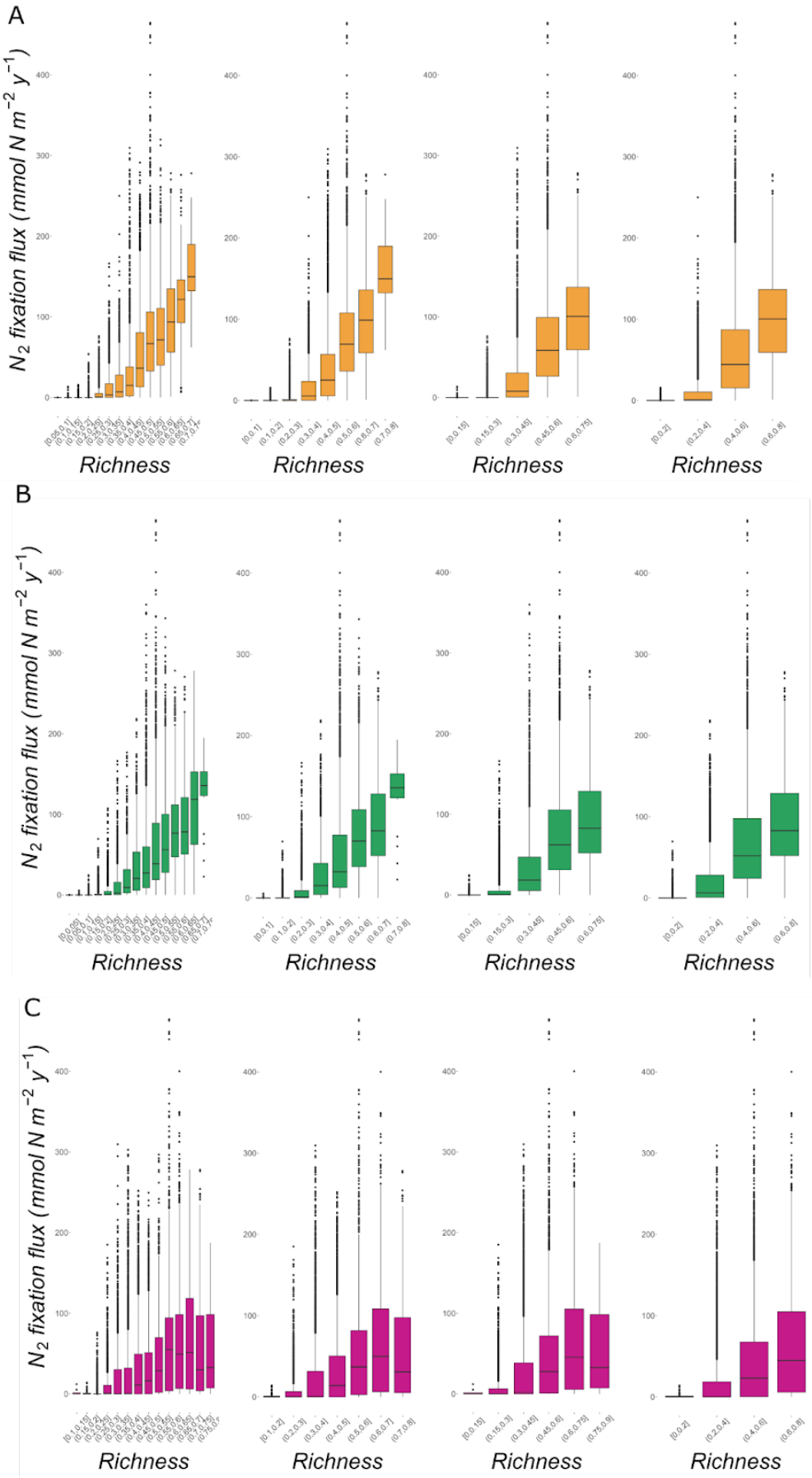

Fig. S11.

Boxplots showing increasing annual diazotroph richness binned according to bin sizes on the x-Axis and nitrogen fixation rates (Wang et al. 2019) on the y-Axis for **(A)** total, **(B)** cyanobacterial and **(C)** non-cyanobacterial diazotroph community.

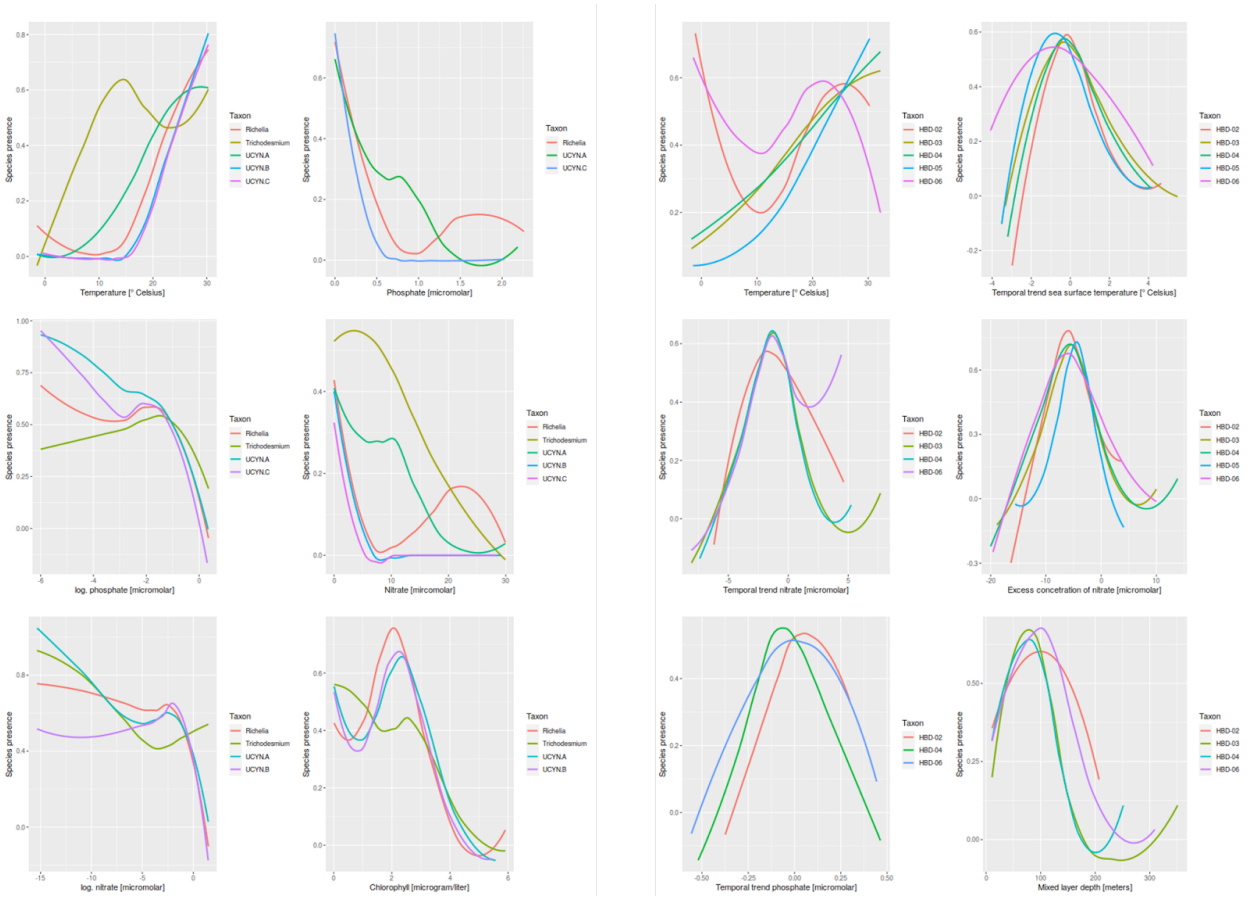

Fig. S12.

Response curves fitted for diazotroph taxa and top ranked environmental predictors, based on the General Additive Model. Response curves have been averaged according to different background selection strategies.

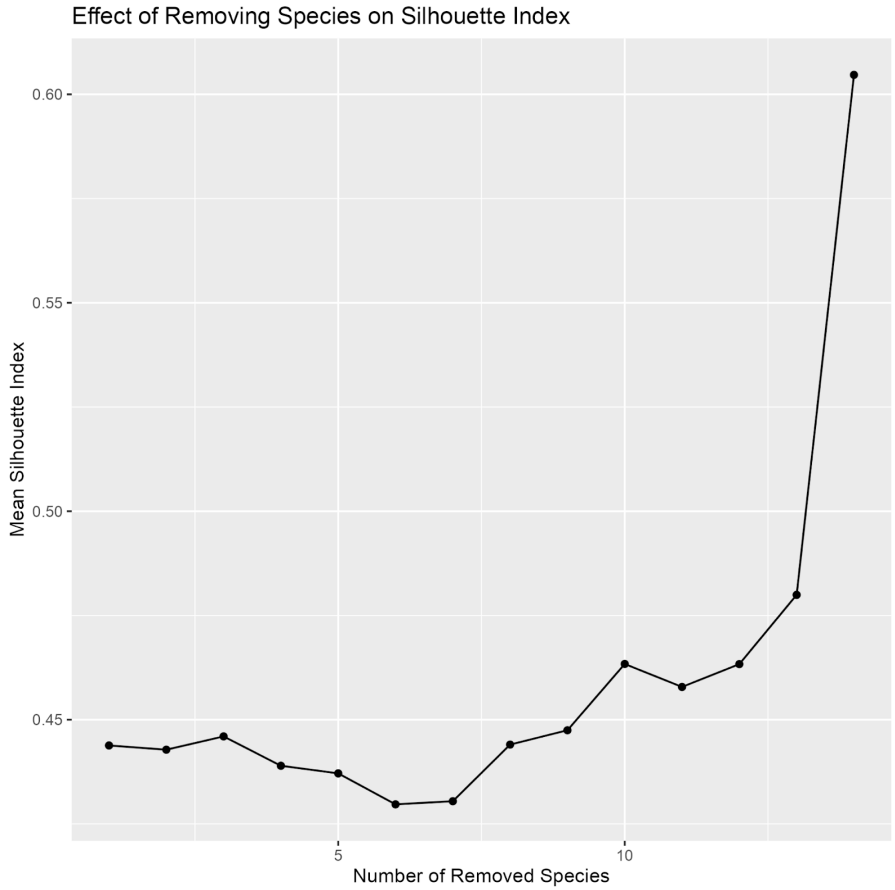

Fig. S13

The plot depicts the Mean Silhouette Index calculated after sequentially removing 1 to 14 species from the dataset. The x-axis represents the number of species removed, ranging from 1 to 14, and the y-axis indicates the corresponding Mean Silhouette Index values.

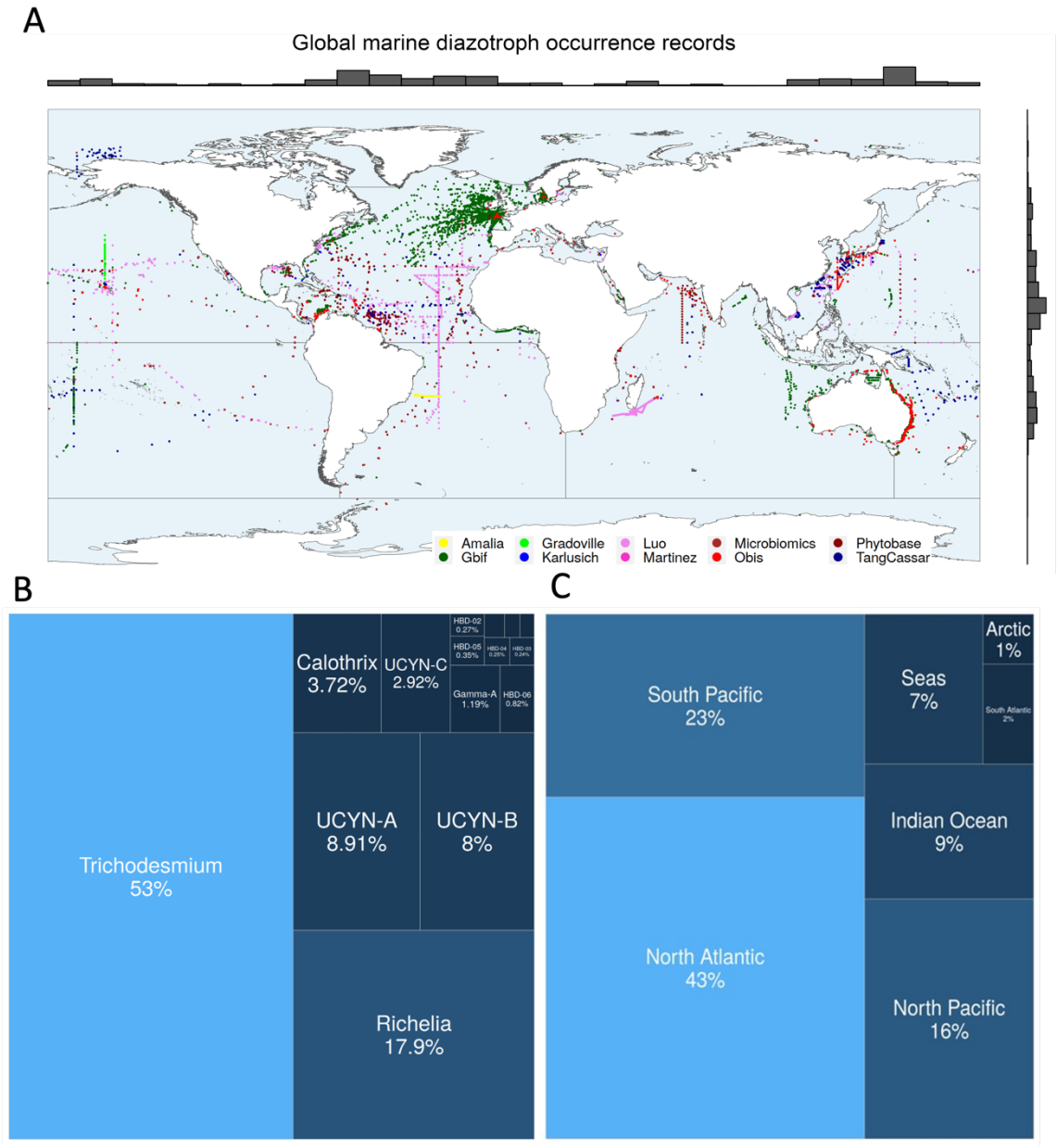

Fig. S14.

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 gradients.

Table S1.

Table showing mean ± sd explanatory skill used for the environmental predictor ranking for GLM (D2), GAM (adjusted R^2), and RF (Out-Of-Bag Error) across all six background selection strategies. 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 $m^{-2}s^{-1}$), Chl (chlorophyll, µg l^{-1}), Wind.CCMP (sea surface wind stress, m s^{-1}), pCO2 (carbon dioxide partial pressure in the surface sea, µatm), MLPAR1 (photosynthetically available radiation over the mixed layer depth, µmol $m^{-2}s^{-1}$), 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), dP_dt (temporal trends of phosphate, µM), dSi_dt (temporal trends of silicate, µM), dMLD1_dt (temporal trends of mixed layer depth, meters)

Taxon	Statistic	T	Sal	N	P	Si	MLD1	PAR	Chl	Wind.CCMP	pCO2	MLPAR1	Nstar	Sistar	dT_dt	dN_dt	dP_dt	dSi_dt	dMLD1_dt
Calothrix	1-OOB.error	0.69 ± 0.02	0.69 ± 0.03	0.68 ± 0.02	0.65 ± 0.01	0.63 ± 0.02	0.67 ± 0.03	0.67 ± 0.01	0.67 ± 0.02	0.66 ± 0.03	0.63 ± 0.02	0.65 ± 0.01	0.66 ± 0.02	0.67 ± 0.02	0.67 ± 0.02	0.66 ± 0.01	0.64 ± 0.02	0.64 ± 0.02	0.66 ± 0.02
Calothrix	adj.Dsq	0.07 ± 0.03	0.11 ± 0.01	0.06 ± 0.03	0.05 ± 0.04	0.02 ± 0.02	0.03 ± 0	0.05 ± 0	0.02 ± 0.01	0.03 ± 0.01	0.03 ± 0.03	0.06 ± 0.01	0.1 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.03 ± 0.02	0.04 ± 0.01	0.05 ± 0.01
Calothrix	adj.Rsq	0.12 ± 0.01	0.15 ± 0.02	0.06 ± 0.03	0.05 ± 0.03	0.03 ± 0.02	0.03 ± 0	0.05 ± 0	0.07 ± 0.01	0.06 ± 0.01	0.04 ± 0.03	0.07 ± 0.01	0.1 ± 0.02	0.05 ± 0.01	0.04 ± 0.01	0.05 ± 0	0.04 ± 0.02	0.04 ± 0.01	0.06 ± 0.01
Cyanothece	1-OOB.error	0.67 ± 0.03	0.61 ± 0.03	0.62 ± 0.03	0.65 ± 0.03	0.59 ± 0.02	0.66 ± 0.02	0.68 ± 0.02	0.65 ± 0.03	0.65 ± 0.01	0.67 ± 0.02	0.64 ± 0.02	0.6 ± 0.03	0.63 ± 0.02	0.64 ± 0.02	0.65 ± 0.03	0.63 ± 0.02	0.61 ± 0.02	0.66 ± 0.02
Cyanothece	adj.Dsq	0.2 ± 0.04	0.03 ± 0.03	0.16 ± 0.05	0.15 ± 0.07	0.01 ± 0.02	0.14 ± 0.01	0.06 ± 0.01	0.02 ± 0.01	0.03 ± 0.01	0.09 ± 0.04	0.11 ± 0.01	0.06 ± 0.05	0.11 ± 0.02	0.04 ± 0.01	0.13 ± 0.03	0.03 ± 0.03	0.01 ± 0	0.04 ± 0.01
Cyanothece	adj.Rsq	0.18 ± 0.04	0.03 ± 0.02	0.13 ± 0.04	0.13 ± 0.06	0.01 ± 0.01	0.13 ± 0.01	0.06 ± 0.01	0.03 ± 0.01	0.14 ± 0.01	0.12 ± 0.03	0.14 ± 0.02	0.07 ± 0.05	0.1 ± 0.02	0.06 ± 0.01	0.13 ± 0.03	0.04 ± 0.03	0.03 ± 0.02	0.07 ± 0.02
Gamma.A	1-OOB.error	0.62 ± 0.07	0.65 ± 0.06	0.6 ± 0.08	0.64 ± 0.08	0.59 ± 0.06	0.63 ± 0.06	0.66 ± 0.07	0.63 ± 0.07	0.6 ± 0.07	0.62 ± 0.08	0.61 ± 0.07	0.59 ± 0.06	0.62 ± 0.06	0.61 ± 0.06	0.59 ± 0.09	0.58 ± 0.07	0.6 ± 0.06	0.65 ± 0.06
Gamma.A	adj.Dsq	0.15 ± 0.08	0.18 ± 0.05	0.14 ± 0.09	0.21 ± 0.08	0.04 ± 0.03	0.04 ± 0.01	0.16 ± 0.05	0.06 ± 0.04	0.04 ± 0.03	0.12 ± 0.06	0.14 ± 0.03	0.13 ± 0.05	0.04 ± 0.02	0.11 ± 0.02	0.09 ± 0.07	0.09 ± 0.05	0.07 ± 0.02	0.03 ± 0.02
Gamma.A	adj.Rsq	0.13 ± 0.07	0.18 ± 0.06	0.12 ± 0.07	0.2 ± 0.07	0.05 ± 0.03	0.04 ± 0.01	0.18 ± 0.03	0.06 ± 0.04	0.07 ± 0.03	0.12 ± 0.05	0.17 ± 0.03	0.13 ± 0.05	0.09 ± 0.03	0.12 ± 0.02	0.09 ± 0.07	0.09 ± 0.04	0.1 ± 0.02	0.04 ± 0.03
HBD-02	1-OOB.error	0.64 ± 0.03	0.67 ± 0.05	0.6 ± 0.05	0.61 ± 0.03	0.66 ± 0.03	0.64 ± 0.05	0.6 ± 0.06	0.57 ± 0.05	0.57 ± 0.06	0.64 ± 0.06	0.58 ± 0.06	0.61 ± 0.06	0.65 ± 0.04	0.6 ± 0.04	0.62 ± 0.04	0.6 ± 0.04	0.63 ± 0.04	0.6 ± 0.02
HBD-02	adj.Dsq	0.03 ± 0.02	0.03 ± 0.01	0.02 ± 0	0.05 ± 0.03	0.02 ± 0.01	0.02 ± 0.02	0.03 ± 0.02	0.02 ± 0.03	0.03 ± 0.02	0.03 ± 0.02	0.05 ± 0.02	0.07 ± 0.03	0.01 ± 0.01	0.04 ± 0.03	0.02 ± 0.02	0.03 ± 0.04	0.01 ± 0.01	0.01 ± 0.01
HBD-02	adj.Rsq	0.09 ± 0.04	0.04 ± 0.02	0.01 ± 0.01	0.07 ± 0.03	0.04 ± 0.03	0.03 ± 0.02	0.04 ± 0.02	0.02 ± 0.03	0.03 ± 0.01	0.03 ± 0.02	0.09 ± 0.05	0.09 ± 0.04	0.01 ± 0.01	0.04 ± 0.03	0.02 ± 0.02	0.05 ± 0.04	0.02 ± 0.02	0.01 ± 0.01
HBD-03	1-OOB.error	0.65 ± 0.06	0.63 ± 0.04	0.64 ± 0.04	0.62 ± 0.06	0.61 ± 0.05	0.62 ± 0.08	0.59 ± 0.08	0.54 ± 0.08	0.57 ± 0.06	0.61 ± 0.08	0.55 ± 0.06	0.63 ± 0.06	0.65 ± 0.05	0.62 ± 0.06	0.62 ± 0.05	0.59 ± 0.06	0.63 ± 0.04	0.6 ± 0.04
HBD-03	adj.Dsq	0.05 ± 0.05	0.02 ± 0.01	0.04 ± 0.02	0.1 ± 0.02	0.08 ± 0.02	0.06 ± 0.03	0.01 ± 0.01	0.02 ± 0.03	0 ± 0.01	0.03 ± 0.01	0.03 ± 0.02	0.1 ± 0.04	0.03 ± 0.02	0.04 ± 0.02	0.06 ± 0.02	0.01 ± 0.01	0 ± 0	0.03 ± 0.01
HBD-03	adj.Rsq	0.04 ± 0.04	0.03 ± 0.02	0.04 ± 0.02	0.1 ± 0.02	0.08 ± 0.03	0.06 ± 0.04	0.01 ± 0.01	0.02 ± 0.03	0.03 ± 0.03	0.02 ± 0.01	0.06 ± 0.02	0.11 ± 0.04	0.05 ± 0.05	0.05 ± 0.02	0.06 ± 0.02	0.01 ± 0.01	0 ± 0	0.03 ± 0.02
HBD-04	1-OOB.error	0.64 ± 0.04	0.6 ± 0.05	0.61 ± 0.04	0.61 ± 0.05	0.62 ± 0.05	0.63 ± 0.08	0.56 ± 0.08	0.56 ± 0.08	0.58 ± 0.08	0.61 ± 0.08	0.56 ± 0.06	0.63 ± 0.06	0.63 ± 0.05	0.63 ± 0.06	0.64 ± 0.04	0.6 ± 0.04	0.62 ± 0.04	0.61 ± 0.05
HBD-04	adj.Dsq	0.05 ± 0.05	0.02 ± 0.01	0.04 ± 0.02	0.09 ± 0.02	0.05 ± 0.02	0.03 ± 0.02	0.01 ± 0.01	0.02 ± 0.03	0.03 ± 0.01	0.04 ± 0.02	0.03 ± 0.02	0.09 ± 0.04	0.04 ± 0.02	0.05 ± 0.02	0.07 ± 0.02	0.02 ± 0.02	0 ± 0	0.03 ± 0.01
HBD-04	adj.Rsq	0.04 ± 0.04	0.04 ± 0.04	0.04 ± 0.01	0.09 ± 0.01	0.05 ± 0.03	0.03 ± 0.02	0.02 ± 0.02	0.02 ± 0.02	0.05 ± 0.03	0.06 ± 0.01	0.06 ± 0.01	0.09 ± 0.04	0.03 ± 0.02	0.06 ± 0.02	0.07 ± 0.02	0.03 ± 0.02	0 ± 0	0.03 ± 0.01
HBD-05	1-OOB.error	0.67 ± 0.04	0.63 ± 0.03	0.61 ± 0.03	0.62 ± 0.03	0.6 ± 0.04	0.63 ± 0.03	0.58 ± 0.05	0.57 ± 0.05	0.61 ± 0.06	0.65 ± 0.07	0.62 ± 0.06	0.66 ± 0.04	0.63 ± 0.04	0.65 ± 0.08	0.62 ± 0.04	0.62 ± 0.04	0.64 ± 0.04	0.65 ± 0.04
HBD-05	adj.Dsq	0.11 ± 0.06	0.03 ± 0.01	0.04 ± 0.02	0.09 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.03 ± 0.01	0.06 ± 0.05	0.02 ± 0.01	0.12 ± 0.03	0.04 ± 0.01	0.16 ± 0.05	0.03 ± 0.02	0.06 ± 0.03	0.02 ± 0.01	0.01 ± 0.02	0 ± 0	0.05 ± 0.03
HBD-05	adj.Rsq	0.11 ± 0.05	0.04 ± 0.03	0.04 ± 0.02	0.09 ± 0.02	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.05 ± 0.04	0.02 ± 0.01	0.12 ± 0.03	0.07 ± 0.03	0.16 ± 0.04	0.02 ± 0.02	0.08 ± 0.03	0.02 ± 0.01	0.01 ± 0.02	0 ± 0	0.04 ± 0.02
HBD-06	1-OOB.error	0.6 ± 0.1	0.58 ± 0.08	0.61 ± 0.07	0.62 ± 0.08	0.55 ± 0.11	0.61 ± 0.09	0.56 ± 0.1	0.55 ± 0.1	0.56 ± 0.11	0.57 ± 0.13	0.53 ± 0.1	0.6 ± 0.1	0.59 ± 0.09	0.61 ± 0.1	0.59 ± 0.08	0.57 ± 0.11	0.56 ± 0.09	0.6 ± 0.08
HBD-06	adj.Dsq	0.01 ± 0.01	0 ± 0.01	0.02 ± 0.03	0.04 ± 0.03	0.01 ± 0.01	0.03 ± 0.02	0.02 ± 0.01	0.03 ± 0.04	0.03 ± 0.01	0.01 ± 0.02	0.01 ± 0.01	0.05 ± 0.02	0.02 ± 0.02	0.01 ± 0.01	0.02 ± 0.02	0.01 ± 0.02	0.01 ± 0	0.02 ± 0.01
HBD-06	adj.Rsq	0.04 ± 0.02	0.05 ± 0.04	0.03 ± 0.04	0.05 ± 0.03	0.02 ± 0.02	0.04 ± 0.02	0.02 ± 0.01	0.03 ± 0.03	0.04 ± 0.02	0.02 ± 0.02	0.01 ± 0.01	0.06 ± 0.03	0.04 ± 0.03	0.02 ± 0.02	0.04 ± 0.02	0.01 ± 0.02	0.01 ± 0.01	0.02 ± 0.02
Richelia intracellularis	1-OOB.error	0.68 ± 0.02	0.69 ± 0.03	0.65 ± 0.01	0.66 ± 0.02	0.66 ± 0.02	0.65 ± 0.02	0.67 ± 0.02	0.64 ± 0.02	0.63 ± 0.01	0.67 ± 0.02	0.63 ± 0.02	0.63 ± 0.02	0.66 ± 0.03	0.64 ± 0.02	0.66 ± 0.02	0.63 ± 0.01	0.63 ± 0.02	0.64 ± 0.03
Richelia intracellularis	adj.Dsq	0.09 ± 0.03	0.01 ± 0.01	0.1 ± 0.03	0.06 ± 0.05	0.01 ± 0.02	0.04 ± 0	0 ± 0	0.05 ± 0.03	0.02 ± 0	0.02 ± 0.02	0.02 ± 0	0.03 ± 0.03	0.09 ± 0.01	0.01 ± 0.01	0.05 ± 0.02	0.02 ± 0.02	0.01 ± 0.01	0.05 ± 0.01
Richelia intracellularis	adj.Rsq	0.11 ± 0.02	0.01 ± 0.01	0.09 ± 0.03	0.07 ± 0.05	0.02 ± 0.02	0.04 ± 0	0.06 ± 0	0.04 ± 0.03	0.06 ± 0	0.03 ± 0.02	0.02 ± 0	0.03 ± 0.02	0.08 ± 0.01	0.03 ± 0	0.05 ± 0.02	0.03 ± 0.02	0.02 ± 0.01	0.05 ± 0.01
Richelia	1-OOB.error	0.63 ± 0.07	0.59 ± 0.06	0.58 ± 0.07	0.57 ± 0.07	0.55 ± 0.06	0.62 ± 0.06	0.59 ± 0.07	0.58 ± 0.07	0.55 ± 0.06	0.59 ± 0.07	0.56 ± 0.07	0.56 ± 0.06	0.59 ± 0.07	0.56 ± 0.06	0.59 ± 0.07	0.56 ± 0.07	0.55 ± 0.06	0.62 ± 0.06
Richelia	adj.Dsq	0.18 ± 0.04	0.01 ± 0.01	0.14 ± 0.04	0.08 ± 0.06	0.01 ± 0.01	0.05 ± 0.01	0.07 ± 0.01	0.05 ± 0.02	0.02 ± 0.01	0.07 ± 0.03	0.07 ± 0.01	0.02 ± 0.03	0.11 ± 0.02	0.01 ± 0.01	0.09 ± 0.03	0.03 ± 0.02	0 ± 0	0.08 ± 0.01
Richelia	adj.Rsq	0.19 ± 0.04	0.04 ± 0.01	0.14 ± 0.04	0.08 ± 0.05	0.01 ± 0.01	0.05 ± 0.01	0.09 ± 0.02	0.08 ± 0.02	0.04 ± 0.01	0.1 ± 0.02	0.07 ± 0.01	0.02 ± 0.03	0.12 ± 0.02	0.03 ± 0.01	0.1 ± 0.03	0.03 ± 0.02	0 ± 0	0.08 ± 0.02
Trichodesmium erythraeum	1-OOB.error	0.67 ± 0.04	0.68 ± 0.03	0.66 ± 0.03	0.67 ± 0.03	0.65 ± 0.04	0.71 ± 0.04	0.67 ± 0.04	0.66 ± 0.04	0.66 ± 0.05	0.64 ± 0.01	0.66 ± 0.04	0.67 ± 0.04	0.67 ± 0.03	0.65 ± 0.04	0.68 ± 0.04	0.67 ± 0.04	0.65 ± 0.04	0.69 ± 0.04
Trichodesmium erythraeum	adj.Dsq	0.12 ± 0.02	0 ± 0.01	0.07 ± 0.03	0.03 ± 0.04	0.01 ± 0.01	0.03 ± 0.01	0.01 ± 0	0.03 ± 0.01	0.04 ± 0.02	0.06 ± 0	0 ± 0	0.05 ± 0.01	0.06 ± 0.02	0.01 ± 0.01	0.07 ± 0.02	0.02 ± 0.01	0.01 ± 0	0.05 ± 0.01
Trichodesmium erythraeum	adj.Rsq	0.13 ± 0.01	0.01 ± 0	0.06 ± 0.02	0.04 ± 0.03	0.01 ± 0.01	0.03 ± 0.01	0.03 ± 0	0.05 ± 0.01	0.05 ± 0.01	0.05 ± 0.01	0.01 ± 0.01	0.05 ± 0.01	0.07 ± 0.01	0.01 ± 0.01	0.08 ± 0.02	0.03 ± 0.01	0.01 ± 0	0.06 ± 0.02
Trichodesmium thiebautii	1-OOB.error	0.76 ± 0.02	0.79 ± 0.03	0.74 ± 0.03	0.76 ± 0.01	0.76 ± 0.04	0.79 ± 0.05	0.74 ± 0.02	0.69 ± 0.04	0.77 ± 0.05	0.74 ± 0.02	0.78 ± 0.01	0.73 ± 0.02	0.8 ± 0.03	0.77 ± 0.04	0.73 ± 0.03	0.72 ± 0.03	0.68 ± 0.06	0.76 ± 0.04
Trichodesmium thiebautii	adj.Dsq	0.3 ± 0.04	0.15 ± 0.06	0.1 ± 0.03	0.17 ± 0.01	0.28 ± 0.06	0.24 ± 0.07	0.19 ± 0.02	0.07 ± 0.01	0.24 ± 0.06	0.19 ± 0.01	0.34 ± 0.03	0.05 ± 0.03	0.31 ± 0.05	0.09 ± 0.05	0.03 ± 0.01	0.01 ± 0	0.02 ± 0.01	0.13 ± 0.02
Trichodesmium thiebautii	adj.Rsq	0.28 ± 0.03	0.38 ± 0.08	0.11 ± 0.03	0.21 ± 0.01	0.3 ± 0.07	0.32 ± 0.08	0.26 ± 0.02	0.09 ± 0.01	0.31 ± 0.07	0.2 ± 0.02	0.36 ± 0.01	0.14 ± 0.03	0.4 ± 0.06	0.11 ± 0.06	0.12 ± 0.02	0.14 ± 0.01	0.05 ± 0.02	0.14 ± 0.02
UCYN.A1	1-OOB.error	0.63 ± 0.07	0.64 ± 0.06	0.61 ± 0.09	0.59 ± 0.06	0.64 ± 0.06	0.64 ± 0.07	0.62 ± 0.08	0.6 ± 0.06	0.6 ± 0.05	0.65 ± 0.05	0.61 ± 0.07	0.58 ± 0.06	0.63 ± 0.07	0.62 ± 0.05	0.61 ± 0.08	0.63 ± 0.07	0.6 ± 0.07	0.62 ± 0.06
UCYN.A1	adj.Dsq	0.03 ± 0.02	0.01 ± 0.01	0.05 ± 0.05	0.04 ± 0.05	0.02 ± 0.01	0.06 ± 0.03	0.05 ± 0.02	0.07 ± 0.04	0.08 ± 0.03	0.08 ± 0.03	0.06 ± 0.03	0.01 ± 0.01	0.09 ± 0.05	0.07 ± 0.02	0.05 ± 0.03	0.02 ± 0.02	0.03 ± 0.02	0.08 ± 0.03
UCYN.A1	adj.Rsq	0.06 ± 0.04	0.1 ± 0.02	0.06 ± 0.06	0.03 ± 0.05	0.04 ± 0.02	0.06 ± 0.02	0.06 ± 0.02	0.1 ± 0.04	0.1 ± 0.02	0.09 ± 0.05	0.06 ± 0.03	0.01 ± 0.01	0.11 ± 0.04	0.09 ± 0.02	0.07 ± 0.04	0.05 ± 0.02	0.03 ± 0.01	0.07 ± 0.03
UCYN.A2	1-OOB.error	0.72 ± 0.05	0.69 ± 0.04	0.64 ± 0.08	0.61 ± 0.05	0.65 ± 0.04	0.67 ± 0.06	0.7 ± 0.05	0.64 ± 0.03	0.65 ± 0.04	0.65 ± 0.04	0.66 ± 0.04	0.63 ± 0.05	0.67 ± 0.05	0.71 ± 0.04	0.65 ± 0.06	0.63 ± 0.05	0.6 ± 0.04	0.66 ± 0.04
UCYN.A2	adj.Dsq	0.18 ± 0.09	0.04 ± 0.02	0.19 ± 0.1	0.05 ± 0.05	0.01 ± 0.01	0.07 ± 0.03	0.01 ± 0.01	0.09 ± 0.03	0.08 ± 0.04	0.14 ± 0.04	0.05 ± 0.03	0.01 ± 0.01	0.09 ± 0.06	0.15 ± 0.04	0.08 ± 0.06	0.01 ± 0.02	0.01 ± 0.01	0.1 ± 0.03
UCYN.A2	adj.Rsq	0.2 ± 0.09	0.24 ± 0.02	0.18 ± 0.09	0.05 ± 0.05	0.01 ± 0.01	0.09 ± 0.01	0.08 ± 0.04	0.1 ± 0.02	0.12 ± 0.02	0.16 ± 0.03	0.1 ± 0.02	0.02 ± 0.01	0.14 ± 0.04	0.14 ± 0.04	0.08 ± 0.06	0.02 ± 0.02	0.01 ± 0.01	0.1 ± 0.03
UCYN.B	1-OOB.error	0.65 ± 0.11	0.59 ± 0.11	0.59 ± 0.1	0.56 ± 0.1	0.54 ± 0.1	0.62 ± 0.1	0.58 ± 0.1	0.59 ± 0.11	0.55 ± 0.12	0.61 ± 0.12	0.58 ± 0.1	0.57 ± 0.1	0.59 ± 0.12	0.57 ± 0.1	0.6 ± 0.11	0.57 ± 0.1	0.54 ± 0.1	0.59 ± 0.09
UCYN.B	adj.Dsq	0.24 ± 0.07	0.02 ± 0.01	0.13 ± 0.07	0.09 ± 0.07	0.02 ± 0.02	0.06 ± 0.02	0.07 ± 0.02	0.05 ± 0.03	0.02 ± 0.01	0.1 ± 0.04	0.08 ± 0.02	0.03 ± 0.03	0.08 ± 0.04	0.03 ± 0.02	0.1 ± 0.06	0.05 ± 0.03	0 ± 0	0.05 ± 0.02
UCYN.B	adj.Rsq	0.24 ± 0.06	0.03 ± 0.01	0.13 ± 0.06	0.09 ± 0.06	0.01 ± 0.01	0.08 ± 0.02	0.07 ± 0.01	0.08 ± 0.04	0.02 ± 0.01	0.13 ± 0.04	0.09 ± 0.02	0.03 ± 0.02	0.09 ± 0.03	0.04 ± 0.01	0.11 ± 0.05	0.05 ± 0.03	0 ± 0	0.05 ± 0.02

Table S2.

Environmental parameters that have been chosen in regard to reflect oceanic conditions that shape species' distributions via effects on physiology, growth, or species competition. 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 S3.

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