

Machine learning maps the “dark matter” of the global marine fish food web

Jorge Assis

jorgemfa@gmail.com

Centro de Ciências do Mar do Algarve (CCMAR/CIMAR LA) <https://orcid.org/0000-0002-6624-4820>

Maria Arredondo

Centro de Ciências do Mar do Algarve (CCMAR/CIMAR LA) <https://orcid.org/0009-0000-5186-6654>

Frederico Mestre

Centro de Ciências do Mar do Algarve (CCMAR/CIMAR LA) <https://orcid.org/0000-0002-7390-1120>

Miguel Araújo

Museo Nacional de Ciencias Naturales <https://orcid.org/0000-0002-5107-7265>

Eliza Fragkopoulou

Centro de Ciências do Mar do Algarve (CCMAR/CIMAR LA) <https://orcid.org/0000-0002-0557-3954>

Article

Keywords:

Posted Date: July 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10194574/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

Marine food webs structure the flow of energy, nutrients, and biomass through the ocean, yet our understanding of their global organization is severely limited by the scarcity and geographical bias of documented species interactions. To address this gap, we constructed a global probabilistic metaweb for 2,454 marine fish species using a positive-unlabeled machine-learning framework. By integrating traits, taxonomic hierarchies, and environmental niches with 10,549 verified interactions, we identified 50,725 novel, high-probability feeding links, raising global connectance from an empirical baseline of 0.18% to 0.98%. We demonstrate that these missing interactions are highly structured, with the largest gaps occurring in hyper-diverse tropical realms, including the Central Indo-Pacific, the High Seas, and upper pelagic zones between 50 and 500 m. Structurally, the undocumented component of the global web is overwhelmingly wider rather than taller. Predicted local connectance and diet breadth increase sharply, whereas mean trophic level and food-chain length remain largely unchanged. This lateral expansion shifts our view of marine trophic architecture from sparse and fragmented to densely connected and trophically redundant, implying empirical records substantially underestimate ecosystem resilience. Ultimately, this baseline provides a crucial trophic prior for modelling extinction risks and forecasting food-web rewiring under climate-driven range shifts.

Introduction

Marine food webs structure the flow of energy, nutrients, and biomass through the global ocean, governing biogeochemical cycling, fisheries productivity, and the resilience of marine ecosystems (Albouy et al., 2019; Caron et al., 2022; du Pontavice et al., 2021; Galiana & Araújo, 2026). Network ecology represents these systems through the metaweb, the complete pool of biologically possible interactions among a species pool, i.e., the upper bound of trophic capacity from which local food webs are filtered (Banville et al., 2025a; González-Trujillo et al., 2025; Mestre & Bastazini, 2026; Poisot et al., 2016). Mapping the metaweb is a prerequisite for forecasting how marine ecosystems will reorganize under climate change and fisheries pressure, and for anticipating the trophic cascades these stressors may propagate (Albouy et al., 2014), complementing trophic-guild approaches that reveal regularities in marine community organisation under environmental gradients (González-Trujillo et al., 2025). Yet the central obstacle to mapping it is the severe incompleteness of documented species interactions (Morales-Castilla et al., 2015), often referred to as the Eltonian shortfall (Peterson et al., 2011). Trophic interactions are ephemeral, context-dependent, and rarely observed directly, so empirical records often reflect research effort as much as network topology (Brimacombe et al., 2023; Menegotto & Rangel, 2018). The bias is not evenly distributed. Tropical assemblages are thought to harbour more dietary specialists, whose narrow and cryptic feeding links are intrinsically harder to document than the broader diets of high-latitude consumers (Hayden et al., 2019). As a result, the regions where interaction data are most needed, including the hyper-diverse tropical Indo-Pacific, are often those where any given trophic link is least likely to have been observed. The ocean's trophic architecture is largely there but invisible to us, a massive and consequential reservoir detectable only indirectly, the ecological equivalent of "dark matter".

Overcoming this limitation requires shifting from mapping only observed interactions to predicting plausible but undocumented ones (Caron et al., 2022; Strydom et al., 2022). The theoretical potential of this shift is substantial. A model that learns the rules governing who eats whom can fill gaps left by uneven sampling, assign probabilities to candidate interactions rather than treating them as simply present or absent, and project how networks may reassemble as species shift their ranges (Morales Castilla et al. 2015). A series of metaweb reconstructions has shown this approach can be scaled to macroecological extents, including global terrestrial vertebrates (Hao et al., 2025), European tetrapods (Maiorano et al., 2020), and marine fish (Albouy et al., 2019). These efforts have established that the possibility space of interactions can be inferred from traits, co-occurrence, and biomechanics. However, by delineating a deterministic interaction space, they do not directly quantify the gap between documented and likely interactions. They do not distinguish observed links from unrecorded but probable ones, nor do they provide the calibrated probabilities and pairwise uncertainty required for downstream forecasting. Without quantifying the magnitude and spatial structure of this missing interaction data, models of future ecosystem change risk propagating historical observation biases into projections of trophic cascades and ecosystem reorganization under global change.

Generating such a probabilistic baseline exposes a fundamental limitation of standard binary classification. Conventional classifiers require negative examples, yet in sparse interaction datasets the only available negatives are unobserved species pairs. Treating these pairs as true absences is problematic because a missing record may reflect incomplete sampling rather than biological impossibility (Banville et al., 2025a; Habedank et al., 2025). Positive-unlabeled learning resolves this problem by treating the unobserved background as unlabeled rather than absent, mirroring the asymmetry of ecological knowledge: presences can be confirmed, whereas absences are rarely certain (Fernández & Nakamura, 2015), much as presence-only species-distribution models treat the environmental background (Strydom et al., 2021). By doing so, positive-unlabeled learning converts the metaweb from a deterministic catalogue of possible interactions into a probability-weighted representation of both known and likely undocumented links (Strydom et al., 2022). This directly responds to recent calls for probabilistic frameworks that fuse empirical observation with mechanistic inference, allowing interaction uncertainty to be propagated into downstream analyses (Mestre & Bastazini, 2026; Poisot et al., 2016).

Despite growing recognition that interaction data are incomplete and unevenly sampled, the magnitude and geography of this incompleteness remain poorly quantified. Here, we address this gap for marine fish, the best-documented marine taxon for both trophic interactions and functional traits, and, therefore, a natural proof-of-concept for quantifying missing trophic information at macroecological scale. We construct a probabilistic metaweb for 2,454 species using a positive-unlabeled machine-learning framework that integrates 10,549 empirical trophic interactions with morphological traits, taxonomic and trophic hierarchies, and environmental niches across the global ocean. We use this framework to (i) quantify the relative contributions of biomechanical, habitat, taxonomic, and macroecological filters in determining realized interactions, (ii) map the magnitude and spatial geography of undocumented but probable feeding links at global resolution, and (iii) determine whether recovering these links widens

local food webs, through increased connectance and diet breadth, or deepens them, through changes in trophic level and food-chain length. We recover tens of thousands of probable but undocumented feeding pathways and expose macroecological gradients in food-web structure that sparse databases cannot detect, opening the way to anticipating how marine food webs may rewire as climate-driven range shifts force species into novel co-occurrence.

Results

A calibrated probabilistic framework for the global marine fish metaweb

The ensemble achieved high and stable predictive performance on the strictly held-out test set. Averaged across the 99 independent training rounds, PR-AUC was 0.945 ± 0.002 and F1 score was 0.879 ± 0.004 , with mean precision at 0.836 ± 0.014 and mean recall at 0.927 ± 0.015 . The complementary ROC-AUC of 0.947 ± 0.002 confirms strong overall discrimination. The narrow standard deviations indicate that the predictive signal is driven by stable ecological structure rather than proxy-negative sampling stochasticity (Fig. S1). The ensemble was also highly resilient to all data perturbations. PR-AUC remained between 0.942 and 0.948 (Table S1), and feature-importance rank order was preserved (Spearman $\rho = 1.0$) under observational dropout, trait degradation, and spatial boundary shifts. Isotonic recalibration successfully corrected the prevalence-induced miscalibration, dropping the Brier score from 0.240 to 0.091. The calibrated ensemble assigned a mean probability of 0.928 to empirically observed interactions and 0.191 to the unobserved background, confirming clean segregation of the trophic signal.

Hierarchy of ecological drivers

SHAP-based attribution revealed a clear hierarchy of interaction predictors (Fig. 1). The two genus-level neighbourhood features dominated (prey_coverage mean SHAP = 2.121 and predator_pressure mean SHAP = 1.055). Environmental niche overlap ranked third (niche_overlap = 0.419). By contrast, biomechanical and habitat features contributed at a markedly lower magnitude (size_ratio = 0.082; depth_overlap = 0.051), while trophic index differences made the smallest overall contributions. Stratifying test-set performance clarified the roles of these features.

Different-genus pairs maintained a strong PR-AUC of 0.942 (Table S2), showing that prediction does not collapse to within-genus signal. In a direct ablation, removing the two genus-level features left balanced discrimination essentially intact (PR-AUC from 0.945 to 0.927; F1 from 0.879 to 0.875) but inflated the shortfall in almost tenfold (predicted interactions from 59,190 to 554,059) and degraded specificity at matched recall (from 99.2% to 90.9%). Across size-ratio quartiles, positive prevalence rose from 17.4% to 74.1% (Table S2), confirming the allometric expectation that asymmetric pairs interact more frequently. Crucially, in the lowest size-ratio quartile, where gape limitation is weakest, a PR-AUC of 0.750 shows

that macroecological features retain strong independent discriminatory power (Table S2). The ensemble also learned trophic directionality. Whereas 4.9% of empirical links were bidirectional, the model predicted bidirectional rates below this baseline (1.5% under the 80%-sensitivity threshold; 0.98% under $K = 9$; 0.75% under $K = 38$), a suppression driven by the directional size_ratio and trophic levels features.

By comparing feature-space distributions across empirically verified interactions, predicted hidden links, and confidently rejected pairs, we verify that on all biomechanical features, hidden links closely matched known interactions while remaining sharply separated from rejected pairs (Table S3). The genus-level features showed an informative divergence of hidden links occupying an intermediate feature-space position between known and rejected pairs ($K-S$ distance ≈ 0.17), demonstrating that the model extrapolates into undersampled biological space rather than merely echoing empirical priors (Table S3).

The hidden global metaweb

Predicting the metaweb to the full pool of 6,019,662 directed pairs revealed a trophic architecture vastly larger than current empirical data suggests (Fig. 2). Under the 80%-sensitivity threshold, the model projected 59,190 plausible interactions, uncovering 50,725 novel feeding pathways, while maintaining 99.2% specificity against the unlabeled background. The Top-K constraints bracketed this space: the specialist ceiling ($K = 9$) yielded 19,758 novel links, while the opportunistic ceiling ($K = 38$) projected 87,913 novel links. Global network connectance rose from a baseline of 0.18% to 0.98% under the 80%-sensitivity threshold, and to between 0.37% and 1.55% across the Top-K ceilings.

The spatial projections exposed the geographic signature of the shortfall. Species richness peaks in the tropical and temperate Indo-West Pacific, the Caribbean Sea, and the tropical Brazilian coast (Fig. 2); empirical interactions, however, are concentrated in temperate–tropical Atlantic waters and, among the high-richness regions, only in the Caribbean and the Brazilian coast. The predicted metaweb at the 80%-sensitivity threshold fills this gap (Fig. 2), and it does so unevenly. While the Southern Ocean, the Arctic, and the Temperate Northern Atlantic expanded interactions by less than 3.17-fold, the Central Indo-Pacific, the Eastern Indo-Pacific, and the Tropical Eastern Pacific expanded by more than 15-fold, with some areas gaining over 10,000 novel interactions (Table S5; Fig. 2). Globally, the High Seas emerged as a major trophic reservoir, gaining interactions by a factor of ~ 5 (2,513 novel interactions; Table S5). The shortfall is also bathymetric: higher documentation gaps occur within the shallow-to-upper-twilight realms, peaking in the 50–200 m band (~ 7.5 -fold), while the 200–500 m band harboured the highest absolute volume of novel pathways (13,922; Table S6).

Connectance declined steeply with species richness in both the empirical and the predicted networks, peaking in the species-poor polar and subpolar realms (Table S5; Fig. 2). The Southern Ocean reached the highest connectance of any realm (empirical mean 0.131; 80%-sensitivity mean 0.235), far above the hyper-diverse Central Indo-Pacific (0.012; 0.044) and Tropical Atlantic (0.011; 0.033) despite the latter two holding roughly eight times as many species (Table S5). These patterns remained consistent across the 80%-sensitivity baseline and both Top-K constraints (Fig. S2).

Relative to the empirical baseline, the 80%-sensitivity reconstruction raised mean trophic level only 1.10-fold and left mean food-chain length virtually unchanged (Fig. 3). Across every threshold, the undocumented web is lateral, not vertical. By adding prey to existing trophic levels rather than creating new ones, the undocumented interactions widen local networks without deepening them, an expansion that scales inversely with local species richness. Spatially, mean trophic level peaks at high latitudes, particularly the Southern Ocean (Fig. 3). Mean food-chain length shows the opposite gradient, rising toward the equator; this pattern is robust, eroding only under the most permissive assumption ($K = 38$), where link saturation collapses the chains.

Empirically, redundancy (i.e., diet breadth), the clearest spatial fingerprint of the lateral expansion, peaked in temperate and polar realms (the Southern Ocean, 2.54; the Temperate Northern Atlantic, 2.44) and collapsed across the hyper-diverse Central Indo-Pacific (1.10) and Western Indo-Pacific (1.04; Table S5). This is a direct artifact of the shortfall. Recovering the hidden links largely realigns diet breadth with diversity. Under the 80%-sensitivity reconstruction it rose in every realm, but most steeply where documentation was poorer: ~5.9-fold in the Central Indo-Pacific, ~4.1-fold in the Western Indo-Pacific, versus ~2.0-fold in the Southern Ocean (Fig. 3; Table S5). The highest predicted diet breadth thus concentrates in productive, species-rich regions, i.e., the Caribbean, the tropical Brazilian shelf, and the Indo-Pacific, with a further peak in the temperate NE Pacific, where the per-cell maximum expanded more than threefold (from ~7.5 to ~25 prey per predator; Fig. 3).

Discussion

By reconstructing a probabilistic metaweb for marine fish at global scale, we show that the ocean may be more tightly connected and functionally integrated than current interaction databases suggest. Missing interaction data thin local networks unevenly, with the greatest gaps where biodiversity and sampling effort are most mismatched. Geographically, these gaps are largest in the hyper-diverse tropics, including the Central Indo-Pacific and the Coral Triangle, where the richest marine fish faunas coincide with sparse records of feeding interactions. Vertically, they peak between 50 and 500 m, spanning the productive epipelagic layer and the upper twilight zone. The undocumented links we recover are overwhelmingly lateral: they increase local connectance and diet breadth several-fold while leaving mean trophic level and food-chain length largely unchanged. In other words, the undocumented component of the network is wider, not taller. Because these lateral links provide alternative feeding pathways that can buffer networks against secondary extinctions and perturbations, their absence from empirical databases implies that observed food webs may substantially underestimate the redundancy (Landi et al., 2018), and potentially the resilience, of marine ecosystems (Heymans et al., 2014).

The strength of the ensemble predictions rests on clear ecological structure. The two strongest predictors, the genus-level neighbourhood features, capture the taxonomic conservatism of diet. That is, closely related predators tend to consume closely related prey (Munoz et al., 2014; Wiens & Graham, 2005). Their importance draws predictions towards patterns already present among related lineages. Although the dominance of these features could raise concerns about circularity, a direct ablation

experiment suggests otherwise. Removing them caused only negligible loss of discriminatory performance, but inflated the predicted metaweb almost tenfold. If these features were merely generating spurious links, their removal should have reduced the number of predicted interactions, not multiplied it. Consistent with this interpretation, the predicted undocumented links occupy regions adjacent to verified interactions within the genus-level feature space, while retaining a distinct statistical signature ($K-S \approx 0.17$). This indicates that the model extrapolates into undersampled biological space rather than merely reproducing the training prior.

Environmental niche compatibility is also a strong predictor of interaction probability: predators rarely consume prey with which they share no environmental space (Astarloa et al., 2019; Olivier et al., 2024a). Biomechanical features ranked lower, but remain ecologically essential. Their role is best understood as setting boundaries rather than driving most realized interactions (Van Kleunen et al., 2026; Weigel & Bonsdorff, 2018). They exclude, for instance, predator-prey pairs in which prey body size exceeds plausible gape constraints, or pairs that imply incompatible directions of energy flow (Cuff et al., 2024; Strydom et al., 2022). This directional prior fits the majority of feeding relationships, but it inevitably suppresses a minority of genuinely reciprocal or stage-dependent interactions, such as intraguild predation, mutual predation, and ontogenetic role reversals. The most directly comparable study, Albouy et al. (2019), constructed a deterministic fish metaweb primarily from an allometric niche rule based on body-size relationships across ~ 225 unique predator-prey pairs (far fewer than the $\sim 10,500$ underlying our ensemble), which was then filtered by bathymetry, diet, and co-occurrence. Our framework formalizes an inverse distinction: biomechanics define the primary filter, or the possibility space, while taxonomic and environmental context determine where within that space interactions are most likely to occur.

The positive-unlabeled framework addresses a common limitation of standard evaluation in sparse networks: high-confidence predictions for unobserved pairs are often scored as errors, even when they are biologically plausible (Habedank et al., 2025; Van Kleunen et al., 2026). In our case, high-probability predictions absent from empirical databases (e.g., *Myoxocephalus stelleri* feeding on *Enophris lucasi*, $p = 0.984$, size ratio 3.10; or *Lutjanus argentiventris* feeding on *Selene peruviana*, $p = 0.980$, size ratio 1.77; Table S7), satisfy the ecological constraints imposed by the ensemble. These cases are, therefore, better interpreted not as the upper tail of model error, but as the lower-bound estimate of undocumented trophic interactions and a prioritized agenda for empirical validation (Banville et al., 2025a; Marquez et al., 2022). Conversely, verified interactions assigned low probabilities fall into two main classes. Some are biological exceptions, often associated with severely inverted size ratios, indicating ontogenetic feeding, scavenging, or other context-dependent processes, as in *Scomber japonicus* feeding on the larger *Ophiodon elongatus* ($p < 0.05$, size ratio 0.42; Table S7). Others appear to reflect spatial-mapping artefacts, where verified pairs receive near-zero environmental niche overlap, as in boreal *Reinhardtius hippoglossoides* on tropical *Caranx rhonchus* ($p < 0.05$, niche overlap 0; Table S7). These discrepancies point to the need for improved trait, distribution, and interaction data, rather than to a failure of the modeling framework itself.

The geography of undocumented interactions is starkly uneven. It is smallest in the comparatively well-sampled temperate Atlantic and an order of magnitude larger in the hyper-diverse tropics. This confirms and extends the tropical complexity gradient reported by Albouy et al. (2019), showing that much of that complexity remains undocumented rather than absent. The Coral Triangle, the global epicentre of marine fish diversity (Tittensor et al., 2010), is also among the poorest regions in documented interactions, with important implications for modelling the planet's most biodiverse marine ecosystems. The same pattern emerges vertically and jurisdictionally: undocumented interactions are especially concentrated in the twilight zone and the High Seas, two of the least-documented ocean compartments and, at the same time, frontiers of emerging exploitation and management, including under the 2023 Biodiversity Beyond National Jurisdiction Agreement.

The latitudinal behaviour of connectance, one of the most discussed patterns in marine food web structure, is consistent across empirical and predicted networks. It remains relatively flat across the species-rich tropical–temperate band and peaks at high latitudes, particularly in the Southern Ocean, in line with Albouy et al. (2019). Diet breadth, by contrast, shows a different pattern. Once undocumented interactions are included, diet breadth peaks in the species-rich tropics, suggesting that its empirical decline with richness is largely an artefact of undersampling. Connectance and diet breadth thus describe different aspects of trophic organization. A polar predator may consume a large fraction of a small prey pool, yielding high connectance, while still feeding on relatively few prey species in absolute terms. A tropical predator, by contrast, may consume many prey species, yielding high diet breadth, while exploiting only a vanishing fraction of an enormous pool. This distinction qualifies the common characterization of tropical assemblages as dominant dietary specialists (Hayden et al., 2019). Tropical consumers may be specialists in proportional terms, but not necessarily in absolute diet breadth. Polar networks, conversely, may be densely connected but depend on relatively few prey species, making them potentially tolerant of random species loss but vulnerable to the targeted removal of highly connected prey, such as the Antarctic silverfish *Pleuragramma antarcticum* (La Mesa & Eastman, 2012).

Recovering undocumented interactions expands the volume of local food webs while leaving their vertical organization almost unchanged. This carries a broader lesson for ecological inference under data incompleteness: descriptors that scale directly with link count (e.g., connectance and diet breadth) are most distorted by missing interactions, whereas height-based descriptors (mean trophic level, food-chain length) remain comparatively robust. In our networks, mean food-chain length increases toward warmer and more diverse realms. Mean trophic level, however, peaks at higher latitudes in both empirical and predicted networks, sharply departing from Albouy et al. (2019). This discrepancy reflects a methodological difference. Albouy and colleagues included primary and secondary producers, allowing trophic levels to be measured from the base of the food web upwards. Our network contains only fish–fish interactions, so species are positioned near the upper part of the trophic system and lack basal taxa (producers) to anchor the scale. Our trophic levels, therefore, track relative position among fish, rather than the full basal-anchored depth of the marine food web described by Albouy and colleagues' analysis.

Despite the strong performance of the ensemble and its immediate applications, three limitations deserve explicit framing. First, the restriction to marine fish is a principled starting point. Fish are the best-documented marine taxon for both trophic interactions and functional traits, making them appropriate proof-of-concept for a framework of this scale. However, extending the approach to invertebrates, marine mammals, seabirds, and foundation taxa (corals, macroalgae, seagrasses) will be essential as their trait, distribution, and interaction databases mature. Second, the metaweb predicts the probability that two species interact, but not interaction strength, frequency, or the fine-scale dynamics that determine realized energy flux (Banville et al., 2025a; Toscano et al., 2014). Extending the framework from interaction presence to interaction strength is the next frontier, and the calibrated probability surface is a natural starting point: interaction probabilities can serve as priors for link strengths in dynamic food-web models (Olivier et al., 2024b). Third, the current framework collapses complex life histories into single, static species nodes. By relying on aggregate species-level traits, such as maximum body length and mean trophic indices, the model inherently smooths over ontogenetic dietary shifts. In marine systems, larval and juvenile stages frequently occupy entirely different trophic niches than their adult counterparts (De Roos & Persson, 2013), engaging in reciprocal or stage-structured predation that single-node architectures cannot fully resolve.

By treating missing records as uncertainty rather than evidence of absence, calibrating predictions to empirical priors, and validating against independent data, machine learning can recover components of trophic architecture that observational data alone cannot resolve (Banville et al., 2025a; Habedank et al., 2025b). The undocumented marine fish food web is thus not unknowable, but increasingly inferable and mappable. As marine biodiversity redistributes under warming oceans (Pech et al., 2017), species will track suitable conditions into novel assemblages, generating new trophic interactions and weakening historical food-web configurations (Bartley et al., 2019; Kortsch et al., 2015). Integrated with next-generation species distribution modelling, the probabilistic metaweb presented here lays the foundation for anticipating this reorganisation, assessing its extinction risks, and supporting conservation and management decisions before ecological change becomes irreversible.

Online methods

Empirical data

We assembled three aligned data domains for the global marine fish metaweb: empirical trophic interactions, species-level functional traits, and global environmental niche suitability maps. Empirical predator-prey interactions were compiled from FishBase, the Global Biotic Interactions database (GloBI), (Barnes et al., 2008), (Silva et al., 2022), the Mangal ecological interactions database, and the EcoBase repository of Ecopath with Ecosim models. These sources initially contributed 3,362, 10,152, 151, 230, 40, and 29 predator-prey records, respectively. After taxonomic harmonization to accepted scientific names using the World Register of Marine Species, removal of duplicated records across databases, and the exclusion of non-feeding interactions, including parasitic and commensal associations, the final empirical dataset comprised 10,549 unique predator-prey pairs among 2,454 marine fish species.

Species-level traits were compiled from FishBase. These included minimum and maximum habitat depth, maximum recorded body length, used as a proxy for gape limitation, two complementary trophic indices (TrophicDiet, estimated from quantitative diet compositions; TrophicFood, estimated from food-item occurrences), categorical habitat affiliations (pelagic, demersal, benthic, bathypelagic and reef-associated), and categorical feeding guilds, including carnivores, herbivores, filter feeders, scavengers, and omnivores. This guild-based representation follows recent global approaches to marine trophic biogeography based on dietary preferences and resource-use categories (González-Trujillo et al., 2025). We also compiled species-specific environmental suitability maps from AquaMaps (Kesner-Reyes et al., 2016) on a 0.5° global resolution grid, corresponding to a 360 × 720-cell grid. Following Klein et al. (Klein et al., 2015), cells were classified as environmentally suitable for a species when the mapped suitability exceeded 0.5.

Feature engineering

For every predator–prey pair, we engineered nine pairwise features that capture four tiers of interaction constraint (Olivier et al., 2024a; Van Kleunen et al., 2026; Weigel & Bonsdorff, 2018).

Biomechanical tier: *size_ratio* (predator's maximum body length divided by prey's, encoding gape limitation) (Gravel et al., 2013); *TrophicDiet_difference* and *TrophicFood_difference* (absolute differences in the two trophic indices, encoding the trophic-level gap).

Habitat tier: *depth_overlap* (proportion of shared depth range); *feeding_compatibility* (an ecology-informed guild filter assigning 1 exclusively when the predator is a carnivore and the prey belongs to a compatible guild, thereby eliminating implausible roles such as herbivores acting as predators); *habitat_overlap* (Jaccard similarity across categorical habitats).

Taxonomic tier: *prey_coverage* (proportion of species within the prey's genus consumed by any member of the predator's genus; *predator_pressure* (proportion of species within the predator's genus consuming any member of the prey's genus). These genus-level features, linked to niche conservatism (Wiens & Graham, 2005), were computed strictly within the training partition to prevent data leakage.

Macroecological tier: *niche_overlap* (Jaccard similarity of shared suitable grid cells derived from AquaMaps environmental envelopes, i.e., modeled environmental-suitability overlap rather than observed geographic co-occurrence or range overlap).

Positive-unlabeled learning framework and ensemble architecture

Because interaction databases document only confirmed feeding observations (Banville et al., 2025a; Habedank et al., 2025), we implemented a framework (Bekker & Davis, 2020; Elkan & Noto, 2008) in which the model is trained on verified positive interactions, while unobserved pairs are treated as an unlabeled background from which proxy-negatives are drawn stochastically (Habedank et al., 2025a). We

embedded this sampling within a 99-round ensemble to isolate macroecological signals from noise: each round drew an independent 1:1 proxy-negative sample and trained an independent classifier. The 10,549 positive interactions were partitioned once into training (70%), validation (15%), and test (15%) sets by stratified random sampling. The validation set was used for hyperparameter tuning, while the test set was reserved for final evaluation, and was never exposed to model fitting. Proxy-negative sampling was performed independently within each partition, so that validation and test negatives were drawn only from their respective unlabeled backgrounds. Each ensemble round trained an Extreme Gradient Boosting classifier (XGBoost; (Chen & Guestrin, 2016)), chosen for its ability to capture non-linear threshold effects and natively handle missing trait values. Hyperparameters were tuned by grid search over maximum tree depth $\in \{2, 4, 6, 8\}$, learning rate $\in \{0.01, 0.03, 0.05, 0.08\}$, and L2 regularization (reg_lambda) $\in \{1.0, 2.0, 3.0\}$, maximizing the Precision–Recall Area Under the Curve (PR-AUC) on the validation set. PR-AUC was prioritized over ROC-AUC because the latter is inflated by the abundance of true negatives inherent to sparse interaction data. The maximum F1 score was also computed to confirm precision–recall balance at the operating threshold used for network inference. Because models were trained on a 1:1 ratio, raw scores reflected an artificial 50% prevalence. We corrected this using isotonic regression (Zadrozny & Elkan, 2002), reweighting the validation set to the empirical prevalence ($\sim 0.18\%$; 10,549 interactions across 6,019,662 directed pairs) before generating final probabilities.

Feature importance and global metaweb inference

We computed exact tree-based SHapley Additive exPlanations (SHAP; (Lundberg et al., 2017)) for each of the 99 trained classifiers. SHAP matrices were averaged across all rounds so that final attributions reflect stable, sampling-independent feature contributions.

All 6,019,662 directed species pairs were scored by the ensemble. Because no single decision threshold is uniquely defensible for an inherently probabilistic surface (Poisot, 2023), and because our objective is to recover undocumented rather than to reject implausible interactions, we adopted a fixed-sensitivity cutoff retaining 80% of verified interactions. Fixing a threshold to a target sensitivity is an established alternative to a default 0.5 cutoff, which performs poorly under the extreme class imbalance of sparse interaction data (Liu et al., 2013). At this operating point the ensemble correctly excluded 99.2% of the unlabeled background pairs, while keeping false positives low (see Results). Because the recovered shortfall nonetheless remains conditional on the binarization rule, we report it as a range, bracketing the 80%-sensitivity cutoff with two predator-centric ceilings, evaluating $K = 9$ (75th percentile of observed diet breadths, a specialist ceiling) and $K = 38$ (95th percentile, an opportunistic ceiling), reading the spread among them as the operative range of the shortfall.

Binarized species-to-species links were assigned to the grid cells where both species were environmentally suitable. Network topology metrics were computed per cell for both the documented and the predicted local food webs: (i) interaction count; (ii) connectance (realized interactions relative to all possible local interactions); (iii) mean diet breadth (i.e., average number of prey per predator); and (iv) mean trophic level and mean food-chain length (respectively, shortest-path height from basal

components and shortest basal-to-apex path length, on the same backbone), describing vertical structure.

Sensitivity analysis

To verify that the ensemble's predictions reflect fundamental macroecological rules rather than data-limitation artifacts, we independently perturbed each data domain and reran the full pipeline under each scenario: (i) observational dropout, withholding 10% of positive interactions at random from training and validation to simulate database incompleteness; (ii) trait degradation, injecting noise to 10% of each species-level traits before feature engineering; and (iii) spatial boundary shifts, resetting the AquaMaps suitability threshold to a lenient (0.3) and a strict (0.7) value. For each scenario, predictive stability was assessed via PR-AUC and F1 on the held-out test set, and interpretive stability via the Spearman rank correlation of ensemble mean SHAP values against the unperturbed baseline. Finally, we ran an ablated model that removed the taxonomic prey_coverage and predator_pressure features, to test whether they spuriously generate hidden links via circular reasoning; we report its discrimination, the size of the inferred shortfall, and specificity at matched recall.

Declarations

Data availability

The species interaction and trait data are available at <https://doi.org/10.6084/m9.figshare.32802599>, while the environmental niche maps are available at <https://www.aquamaps.org/>.

Code availability

The python code used to produce the results is available at <https://doi.org/10.6084/m9.figshare.32802599>.

Competing interests

The authors declare no competing interests.

Author contributions

J.A. coordinated and produced the results, developed the code, and wrote the original draft. F.M. and M.A. co-wrote the original draft. E.F. supervised all phases of production and article writing.

Acknowledgements

This work was funded by FCT - Foundation for Science and Technology through contracts UID/04326/2025 (<https://doi.org/10.54499/UID/04326/2025>), UID/PRR/04326/2025 (<https://doi.org/10.54499/UID/PRR/04326/2025>), LA/P/0101/2020 (<https://doi.org/10.54499/LA/P/0101/2020>) and the Individual Call to Scientific Employment Stimulus 2022.00861.CEE-CIND/CP1729/CT0003 (<https://doi.org/10.54499/2022.00861.CEECIND/CP1729/CT0003>). MBA's research is funded by the Spanish 'Ministerio de Ciencia, Innovación y Universidades' project PID2024-159356NB-I00 and by the CSIC Intramural project 202630E50; his contribution to this research was conducted while visiting the German Centre for Integrative Biodiversity Research (iDiv).

References

1. Albouy C, Archambault P, Appeltans W, Araújo MB, Beauchesne D, Cazelles K, Cirtwill AR, Fortin MJ, Galiana N, Leroux SJ, Pellissier L, Poisot T, Stouffer DB, Wood SA, Gravel D (2019) The marine fish food web is globally connected. *Nat Ecol Evol* 3(8). <https://doi.org/10.1038/s41559-019-0950-y>
2. Albouy C, Velez L, Coll M, Colloca F, Loc'h L, Mouillot F, D., Gravel D (2014) From projected species distribution to food-web structure under climate change. *Glob Change Biol* 20(3). <https://doi.org/10.1111/gcb.12467>
3. Astarloa A, Louzao M, Boyra G, Martinez U, Rubio A, Irigoien X, Hui FKC, Chust G (2019) Identifying main interactions in marine predator–prey networks of the Bay of Biscay. *ICES J Mar Sci* 76(7):2247–2259. <https://doi.org/10.1093/icesjms/fsz140>
4. Banville F, Strydom T, Blyth PSA, Brimacombe C, Catchen MD, Dansereau G, Higino G, Malpas T, Mayall H, Norman K, Gravel D, Poisot T (2025a) Deciphering Probabilistic Species Interaction Networks. In *Ecology Letters* (Vol. 28, Number 6). <https://doi.org/10.1111/ele.70161>
5. Banville F, Strydom T, Blyth PSA, Brimacombe C, Catchen MD, Dansereau G, Higino G, Malpas T, Mayall H, Norman K, Gravel D, Poisot T (2025b) Deciphering Probabilistic Species Interaction Networks. *Ecol Lett* 28(6). <https://doi.org/10.1111/ele.70161>
6. Barnes C, Bethea DM, Brodeur RD, Spitz J, Ridoux V, Pusineri C, Chase BC, Hunsicker ME, Juanes F, Kellermann A, Lancaster J, Ménard F, Bard FX, Munk P, Pinnegar JK, Scharf FS, Rountree RA, Stergiou KI, Sassa C, Jennings S (2008) Predator and prey body sizes in marine food webs. *Ecology* 89(3). <https://doi.org/10.1890/07-1551.1>
7. Bartley TJ, McCann KS, Bieg C, Cazelles K, Granados M, Guzzo MM, MacDougall AS, Tunney TD, McMeans BC (2019) Food web rewiring in a changing world. In *Nature Ecology and Evolution* (Vol. 3, Number 3). <https://doi.org/10.1038/s41559-018-0772-3>
8. Bekker J, Davis J (2020) Learning from positive and unlabeled data: a survey. *Mach Learn* 109(4):719–760. <https://doi.org/10.1007/s10994-020-05877-5>
9. Brimacombe C, Bodner K, Michalska-Smith M, Poisot T, Fortin M-J (2023) Shortcomings of reusing species interaction networks created by different sets of researchers. *PLoS Biol* 21(4):e3002068. <https://doi.org/10.1371/journal.pbio.3002068>

10. Caron D, Maiorano L, Thuiller W, Pollock LJ (2022) Addressing the Eltonian shortfall with trait-based interaction models. In *Ecology Letters* (Vol. 25, Number 4, pp. 889–899). John Wiley and Sons Inc. <https://doi.org/10.1111/ele.13966>
11. Chen T, Guestrin C (2016) XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 785–794. <https://doi.org/10.1145/2939672.2939785>
12. Cuff JP, Labonte D, Windsor FM (2024) Understanding Trophic Interactions in a Warming World by Bridging Foraging Ecology and Biomechanics with Network Science. *Integr Comp Biol* 64(2):306–321. <https://doi.org/10.1093/icb/icae070>
13. De Roos AM, Persson L (2013) Population and community ecology of ontogenetic development. In *Population and Community Ecology of Ontogenetic Development*. <https://doi.org/10.23943/princeton/9780691137575.001.0001>
14. du Pontavice H, Gascuel D, Reygondeau G, Stock C, Cheung WWL (2021) Climate-induced decrease in biomass flow in marine food webs may severely affect predators and ecosystem production. *Glob Change Biol* 27(11). <https://doi.org/10.1111/gcb.15576>
15. Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220. <https://doi.org/10.1145/1401890.1401920>
16. Fernández D, Nakamura M (2015) Estimation of spatial sampling effort based on presence-only data and accessibility. *Ecological Modelling*, 299. <https://doi.org/10.1016/j.ecolmodel.2014.12.017>
17. Galiana N, Araújo MB (2026) Biotic interactions biogeography: A framework for understanding how species interactions shape biodiversity patterns across scales. *PLoS Biol* 24(6):e3003813. <https://doi.org/10.1371/journal.pbio.3003813>
18. González-Trujillo JD, Assis J, Serrão E, Costello MJ, Fragkopoulou E, Mendoza M, Araújo MB (2025) Trophic convergence of marine vertebrate communities worldwide. *Proceedings of the National Academy of Sciences*, 122(17). <https://doi.org/10.1073/pnas.2403754122>
19. Gravel D, Poisot T, Albouy C, Velez L, Mouillot D (2013) Inferring food web structure from predator-prey body size relationships. *Methods Ecol Evol* 4(11). <https://doi.org/10.1111/2041-210X.12103>
20. Habedank J, Horn S, Lemke P, Sidorenko V, Wiltshire KH (2025a) A new perspective for revealing ‘hidden’ interactions in ecological networks. *Methods Ecol Evol* 16(7):1314–1321. <https://doi.org/10.1111/2041-210X.70077>
21. Habedank J, Horn S, Lemke P, Sidorenko V, Wiltshire KH (2025b) A new perspective for revealing ‘hidden’ interactions in ecological networks. *Methods Ecol Evol* 16(7). <https://doi.org/10.1111/2041-210X.70077>
22. Hao X, Holyoak M, Zhang Z, Yan C (2025) Global Projection of Terrestrial Vertebrate Food Webs Under Future Climate and Land-Use Changes. *Glob Change Biol* 31(2). <https://doi.org/10.1111/gcb.70061>

23. Hayden B, Palomares MLD, Smith BE, Poelen JH (2019) Biological and environmental drivers of trophic ecology in marine fishes - a global perspective. *Sci Rep* 9(1).
<https://doi.org/10.1038/s41598-019-47618-2>
24. Heymans JJ, Coll M, Libralato S, Morissette L, Christensen V (2014) Global patterns in ecological indicators of marine food webs: A modelling approach. *PLoS ONE* 9(4).
<https://doi.org/10.1371/journal.pone.0095845>
25. Kesner-Reyes K, Kaschner K, Kullander S, Garilao C, Barile J, Froese R (2016) AquaMaps: Algorithm and Data Sources for Aquatic Organisms AquaMaps for Marine Species. *Froese, R. and D. Pauly. Editors. 2012. FishBase. World Wide Web Electronic Publication. Www.Fishbase.Org, Version (04/2012), 1–23*
26. Klein CJ, Brown CJ, Halpern BS, Segan DB, McGowan J, Beger M, Watson JEM (2015) Shortfalls in the global protected area network at representing marine biodiversity. *Scientific Reports*, 5.
<https://doi.org/10.1038/srep17539>
27. Kortsch S, Primicerio R, Fossheim M, Dolgov AV, Aschan M (2015) Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists. *Proceedings of the Royal Society B: Biological Sciences*, 282(1814). <https://doi.org/10.1098/rspb.2015.1546>
28. La Mesa M, Eastman JT (2012) Antarctic silverfish: Life strategies of a key species in the high-Antarctic ecosystem. *Fish Fish* 13(3). <https://doi.org/10.1111/j.1467-2979.2011.00427.x>
29. Landi P, Minoarivelo HO, Brännström Å, Hui C, Dieckmann U (2018) Complexity and stability of ecological networks: a review of the theory. In *Population Ecology* (Vol. 60, Number 4).
<https://doi.org/10.1007/s10144-018-0628-3>
30. Liu C, White M, Newell G (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *J Biogeogr* 40(4). <https://doi.org/10.1111/jbi.12058>
31. Lundberg SM, Allen PG, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, (30). <https://github.com/slundberg/shap>
32. Maiorano L, Montemaggiori A, Ficetola GF, O'Connor L, Thuiller W (2020) TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Glob Ecol Biogeogr* 29(9):1452–1457.
<https://doi.org/10.1111/geb.13138>
33. Marquez L, Fragkopoulou E, Cavanaugh KC, Houskeeper HF, Assis J (2022) Artificial intelligence convolutional neural networks map giant kelp forests from satellite imagery. *Sci Rep* 12(1):22196.
<https://doi.org/10.1038/s41598-022-26439-w>
34. Menegotto A, Rangel TF (2018) Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. *Nat Commun* 9(1):4713. <https://doi.org/10.1038/s41467-018-07217-7>
35. Mestre F, Bastazini VAG (2026) Metawebs as a tool to connect ecology and biogeography: A critical review. *Ecol Model* 519:111674. <https://doi.org/https://doi.org/10.1016/j.ecolmodel.2026.111674>
36. Morales-Castilla I, Matias MG, Gravel D, Araújo MB (2015) Inferring biotic interactions from proxies. *Trends Ecol Evol* 30(6):347–356. <https://doi.org/https://doi.org/10.1016/j.tree.2015.03.014>

37. Munoz F, Ramesh BR, Couteron P (2014) How do habitat filtering and niche conservatism affect community composition at different taxonomic resolutions? *Ecology* 95(8).
<https://doi.org/10.1890/13-0064.1>
38. Olivier PEN, Lindegren M, Bonsdorff E, Nordström MC (2024a) A network of biological traits: Profiling consumer-resource interactions. *Food Webs* 38:e00333.
<https://doi.org/10.1016/j.fooweb.2023.e00333>
39. Olivier PEN, Lindegren M, Bonsdorff E, Nordström MC (2024b) A network of biological traits: Profiling consumer-resource interactions. *Food Webs*, 38.
<https://doi.org/10.1016/j.fooweb.2023.e00333>
40. Pecl GT, Araújo MB, Bell JD, Blanchard J, Bonebrake TC, Chen I, Clark TD, Colwell RK, Danielsen F, Evengård B, Falconi L, Ferrier S, Frusher S, Garcia RA, Griffis RB, Hobday AJ, Janion-scheepers C, Jarzyna MA, Jennings S, Williams SE (2017) Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science* 355(6332):eaai9214.
<https://doi.org/10.1126/science.aai9214>
41. Peterson ATT, Soberón J, Pearson RGRG, Anderson RPRP, Martínez-Meyer E, Nakamura M, Bastos Araújo M (2011) Ecological Niches and Geographic Distributions. In S. E. Simon A. Levin and Henry S. Horn (Ed.), *Choice Reviews Online* (Monographs, Vol. 49, Number 11). Princeton University Press.
<https://doi.org/10.5860/CHOICE.49-6266>
42. Poisot T (2023) Guidelines for the prediction of species interactions through binary classification. *Methods Ecol Evol* 14(5). <https://doi.org/10.1111/2041-210X.14071>
43. Poisot T, Cirtwill AR, Cazelles K, Gravel D, Fortin MJ, Stouffer DB (2016) The structure of probabilistic networks. *Methods Ecol Evol* 7(3). <https://doi.org/10.1111/2041-210X.12468>
44. Silva MA, Fonseca CT, Olivar MP, Bernal A, Spitz J, Chouvelon T, Jónasdóttir SH, Colaço A, Carmo V, Sutton TT, Menezes GM, Falkenhaug T, Bergstad OA, Pérez-Jorge S (2022) *Compilation of data to assess trophic interactions in mesopelagic food webs in the central and Northeast Atlantic, and in the Mediterranean Sea*. PANGAEA. <https://doi.org/10.1594/PANGAEA.946020>
45. Strydom T, Bouskila S, Banville F, Barros C, Caron D, Farrell MJ, Fortin MJ, Hemming V, Mercier B, Pollock LJ, Runghen R, Riva D, G. V., Poisot T (2022) Food web reconstruction through phylogenetic transfer of low-rank network representation. *Methods Ecol Evol* 13(12).
<https://doi.org/10.1111/2041-210X.13835>
46. Strydom T, Catchen MD, Banville F, Caron D, Dansereau G, Desjardins-Proulx P, Forero-Muñoz NR, Higino G, Mercier B, Gonzalez A, Gravel D, Pollock L, Poisot T (2021) A roadmap towards predicting species interaction networks (across space and time). In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 376, Number 1837). Royal Society Publishing.
<https://doi.org/10.1098/rstb.2021.0063>
47. Tittensor DP, Mora C, Jetz W, Lotze HK, Ricard D, Berghe E, Vanden, Worm B (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature* 466(7310):1098–1101.
<https://doi.org/10.1038/nature09329>

48. Toscano BJ, Gatto J, Griffen BD (2014) Effect of predation threat on repeatability of individual crab behavior revealed by mark-recapture. *Behav Ecol Sociobiol* 68(3). <https://doi.org/10.1007/s00265-013-1666-7>
49. Van Kleunen LB, Dee LE, Wootton KL, Massol F, Clauset A (2026) Predicting missing links in food webs using stacked models and species traits. *Nat Commun* 17(1):2298. <https://doi.org/10.1038/s41467-026-68769-7>
50. Weigel B, Bonsdorff E (2018) Trait-based predation suitability offers insight into effects of changing prey communities. *PeerJ* 6(11):e5899. <https://doi.org/10.7717/peerj.5899>
51. Wiens JJ, Graham CH (2005) NICHE CONSERVATISM: Integrating Evolution, Ecology, and Conservation Biology. *Annu Rev Ecol Evol Syst* 36:519–539. <https://doi.org/10.1146/annurev.ecolsys.36.102803.095431>
52. Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/775047.775151>

Figures

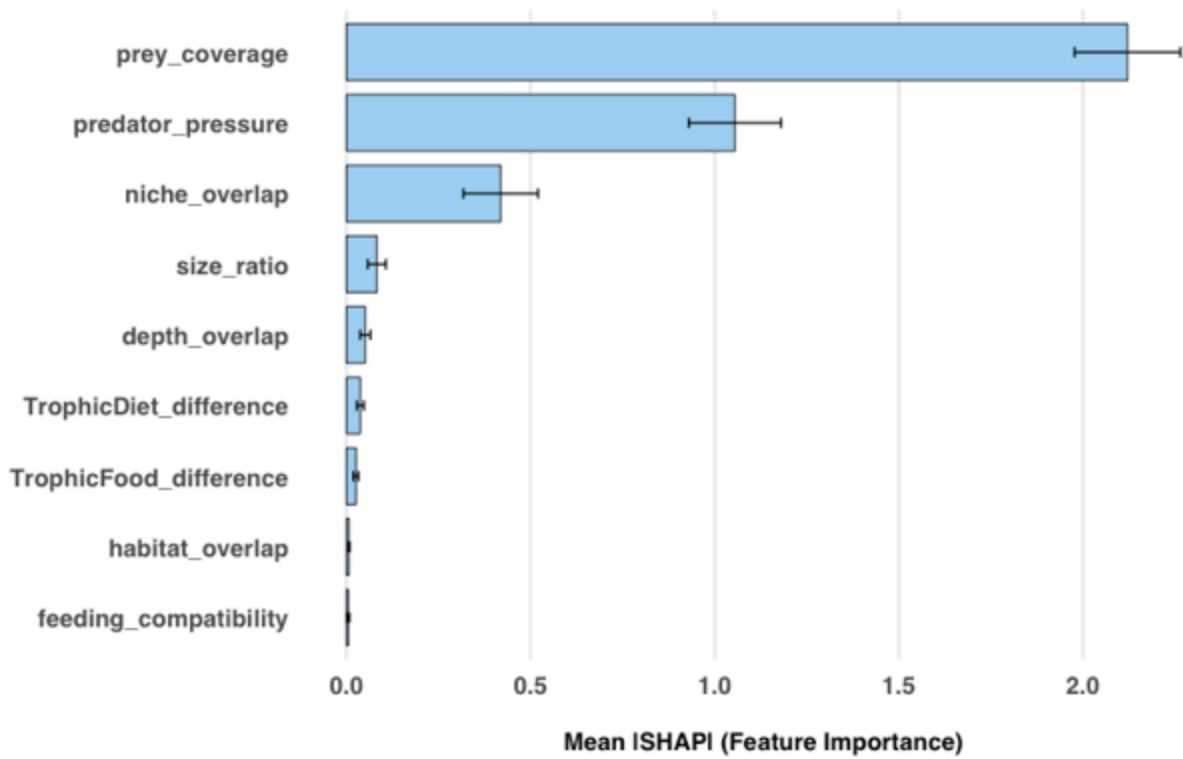


Figure 1

Relative importance of model predictors of marine trophic interactions. Mean absolute SHapley Additive exPlanations (SHAP) values for the nine features, ranked by their marginal contribution to the prediction of a positive interaction. Bars represent ensemble means across all 99 trained classifiers.

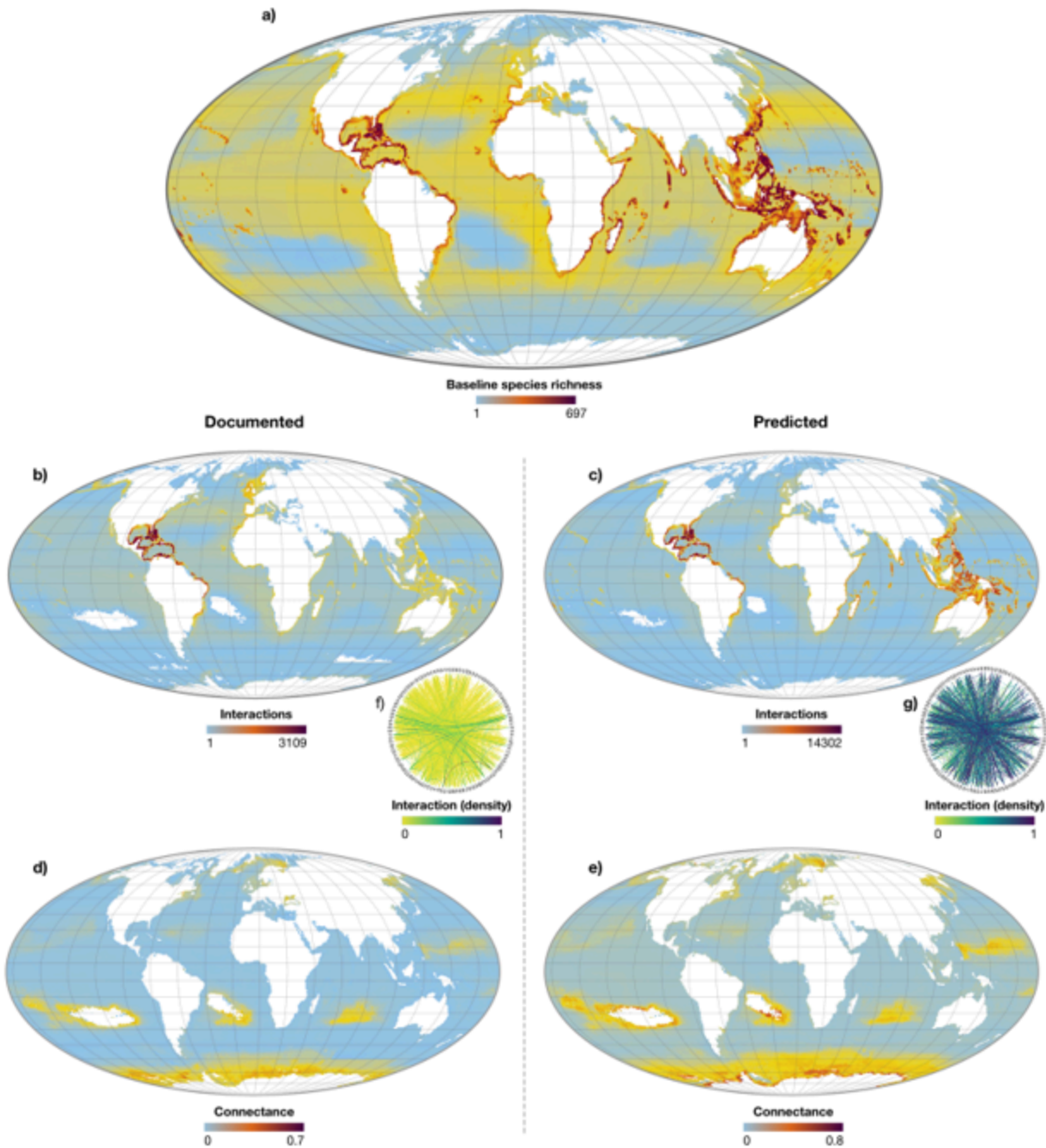


Figure 2

Spatial projection of the global marine fish metaweb. (a) Baseline species richness across the 2,454 species evaluated. (b) Spatial distribution of the 10,549 empirically observed interactions and the (c) predicted interactions recovered under the 80% sensitivity threshold, highlighting the biodiverse regions where the Eltonian shortfall obscures trophic structure. Spatial distribution of connectance, shown for (d) the empirically observed network and (e) the predicted network under the 80% sensitivity threshold. Topologies of the (f) observed and the (g) predicted global marine fish metawebs at the family level under the 80% sensitivity threshold. For clarity, panels show families covering >85% of binarized predicted links. Nodes are labelled with unique three-letter taxonomic codes; full mapping in Supplementary Table S4.

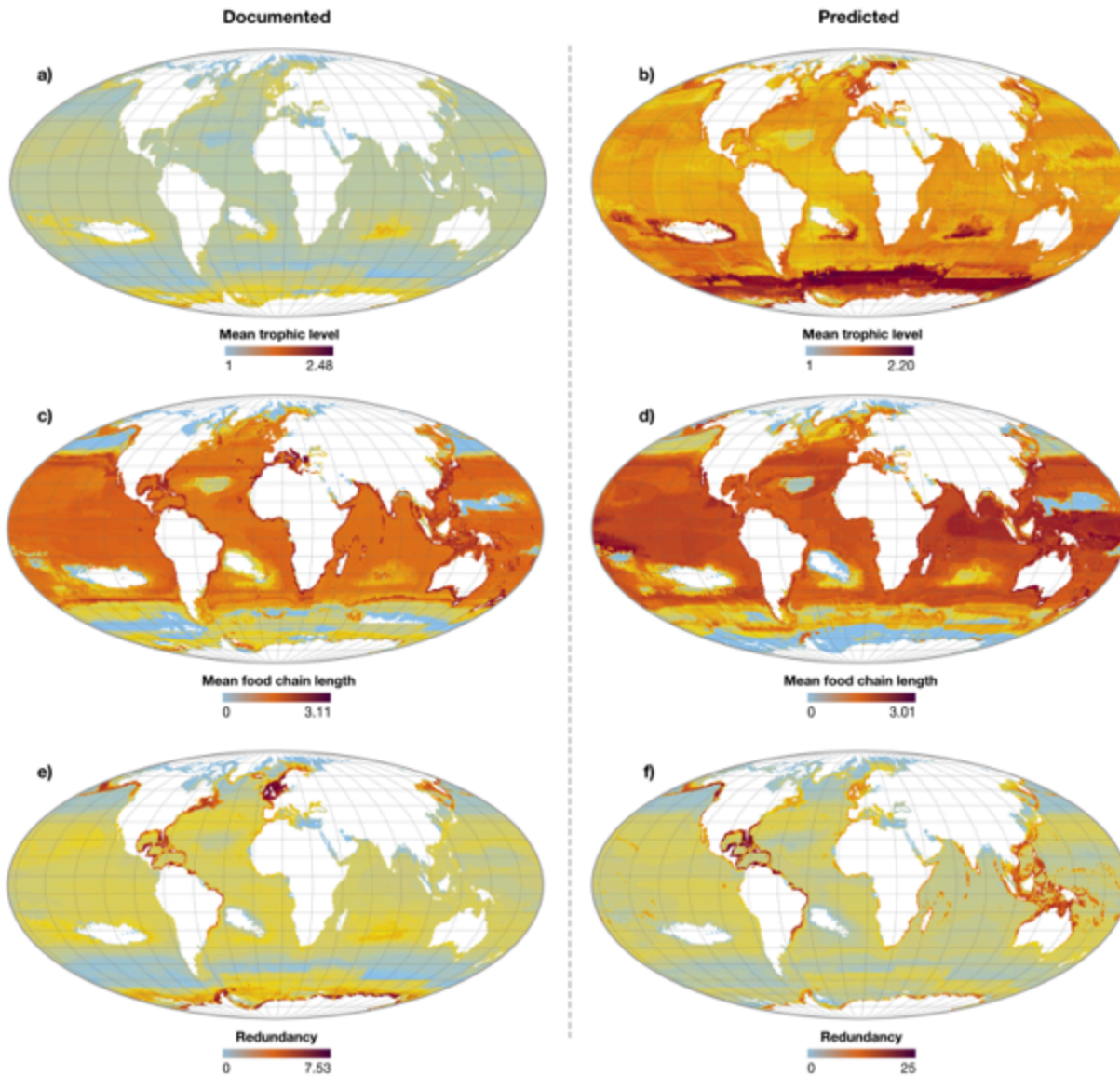


Figure 3

Spatial structure of vertical trophic organization and trophic redundancy across the global ocean. Mean trophic level (shortest-path height from basal components) for (a) the empirically observed network and (b) the predicted network. Mean food-chain length (shortest basal-to-apex path length) for (c) the observed network and (d) the predicted network. Mean redundancy (mean number of prey per predator) for (e) the observed network and (f) the predicted network. Grid cells containing fewer than five species were excluded from the analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS1.xlsx](#)

- [TableS2.xlsx](#)
- [TableS3.xlsx](#)
- [TableS4.xlsx](#)
- [TableS5.xlsx](#)
- [TableS6.xlsx](#)
- [TableS7.xlsx](#)
- [FigureS1.png](#)
- [FigureS2.png](#)