UC Merced

UC Merced Previously Published Works

Title

A genomic catalog of Earth's bacterial and archaeal symbionts

Permalink

https://escholarship.org/uc/item/7ng0f1x8

Journal

Nature Biotechnology

ISSN

1087-0156

Authors

Villada, Juan C
Vasquez, Yumary M
Szabó, Gitta
et al.

Publication Date

2026-08-31

DOI

10.1038/s41587-026-03213-1

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

nature biotechnology

Resource

https://doi.org/10.1038/s41587-026-03213-1

A genomic catalog of Earth's bacterial and archaeal symbionts

Received: 2 June 2025

Accepted: 10 June 2026

Published online: 31 August 2026

Check for updates

Juan C. Villada[1] ✉, Yumary M. Vasquez[1,2], Gitta Szabó[1], Ewan Whittaker-Walker[1,3], Miguel F. Romero[1], Sarina Qin[4], Neha Varghese[1], Emiley A. Eloe-Fadrosh[1], Nikos C. Kyrpides[1], SymGs data consortium*, Axel Visel[1], Tanja Woyke[1] & Frederik Schulz[1] ✉

Microbial symbiosis drives the functional and phylogenomic diversification of life on Earth yet remains underexplored because of culturing challenges. This study used machine learning (ML) to predict symbiotic lifestyles in more than a hundred thousand microbial genomes from diverse environmental metagenome samples and reference genomes. Predictions were performed using `symclatron`, an ML framework developed to identify genomic signatures of symbionts. Predictions were deposited in a catalog we established called Symbiont Genomes (SymGs). The results indicate that 15–23% of uncultivated microorganisms likely engage in symbiotic relationships with other organisms, categorized as host-associated or obligate intracellular lifestyles, and are present in half of all known bacterial and archaeal phyla. We also identify genomic signatures of symbiotic lifestyles, including the loss of certain metabolic functions and the differential presence of metabolic modules that may enable host-dependent living. The `symclatron` software and the SymGs catalog represent valuable resources for studying symbioses, potentially facilitating future mechanistic investigations and engineering of host–microorganism associations.

The intricate and persistent relationships between microorganisms and their hosts, known as microbial symbioses, are fundamental to the functioning of life on Earth[1–3]. These dynamic relationships span a spectrum from mutualism to parasitism[4] and are integral to the health and evolution of all partners involved[5]. On the one hand, microbial symbionts have critical roles in processes such as nutrient cycling[6,7], transfer of energy in the form of ATP[8], enhancement of host stress resistance[9] and disease suppression and protection from pathogens[10–12]. On the other hand, microbial pathogens infect eukaryotic cells and adversely affect their host fitness[13]. The genetic makeup of symbionts often determines host range and interaction[4,14]; however, the availability of symbiont genomes is skewed toward those that are animal or human associated or amenable to laboratory cultivation[15–17]. This bias has caused researchers to overlook the true breadth of symbiont diversity in nature, potentially missing novel and globally important symbiont clades yet to be identified.

Recent advances in metagenomics have provided the tools to capture and analyze the vast array of genetic material present in environmental samples, offering unprecedented insights into the microbial world[18]. The ability to reconstruct population genomes, so-called metagenome-assembled genomes (MAGs) and single amplified genomes (SAGs), from environmental sequencing data has revolutionized our understanding of microbial diversity, functional capabilities and ecology[19–24] and the evolution of the eukaryotic cell[25,26]. To date, hundreds of thousands of MAGs have been generated[27–29]. While MAG taxonomy can be linked to their predicted metabolic traits, the lifestyles of these organisms remain largely unknown. Taxonomy alone is often insufficient to predict lifestyle because of variation within related

[1]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]Department of Life and Environmental Sciences, University of California Merced, Merced, CA, USA. [3]Department of Microbiology and Environmental Toxicology, University of California Santa Cruz, Santa Cruz, CA, USA. [4]School of Natural Sciences, University of California Merced, Merced, CA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: jvillada@lbl.gov; fschulz@lbl.gov

groups and the prevalence of novel organisms in MAG datasets. Analyzing genomic features related to function and dependence offers a more direct approach[30]. In particular, genomes of symbiotic clades have been disproportionately underrepresented in large-scale MAG studies. This bias occurs because estimates of genome completeness often fall below the medium-quality threshold for these genomes[31–33]. The concept of the symbiome encompasses the entire assemblage of colocalized and coevolving organisms, including both host and symbionts[34]. However, a linkage between symbiont and host can often not be established in complex metagenomic datasets. Yet, it is plausible that many MAGs and SAGs represent uncultivated microbial symbionts; therefore, a symbiont-centric view of the global microbiome data could represent a powerful approach to uncover the full extent of the genomic space of bacterial and archaeal symbionts.

Understanding microbial symbiosis has exciting implications for biotechnology. Symbiotic systems represent naturally evolved solutions for metabolic division of labor, offering blueprints for engineering synthetic microbial consortia for industrial applications[35]. The streamlined genomes of obligate symbionts provide insights into minimal genome design for synthetic biology applications, while their specialized metabolic capabilities suggest enzymatic tools for biotechnology[36]. Furthermore, symbiont–host metabolic complementarity can guide the design of engineered microbial communities for bioproduction, bioremediation and agricultural applications[37,38]. A comprehensive catalog of symbiont genomes would accelerate the discovery of additional biosynthetic pathways, specialized enzymes, metabolic division of labor and metabolic interactions that can be harnessed for biotechnological innovation in the engineering of host–microorganism systems.

Here, we leverage machine learning (ML) models trained on large-scale genomic data to map putative microbial symbionts across the globe. For simplicity, we henceforth refer to these putative symbionts as 'symbionts', recognizing that their status remains provisional, being predicted from genomic data. Operationally, in this study, we define a symbiont as an organism that maintains a spatially close and temporally prolonged association with a host organism of a different species, regardless of the function of the symbiont, which can range from mutualistic to parasitic[16]. Microbial symbionts can be associated with host organisms by attaching to external surfaces or host cells or occurring intracellularly at least during part of their life cycle. Some microorganisms are obligate intracellular and cannot replicate outside a host cell. To account for these differences in host dependence, we broadly categorized microbial symbionts as 'host-associated' or, more specifically, as 'obligate intracellular'. Host-associated symbionts encompass both facultatively intracellular microorganisms, which can replicate inside a host cell but also independently of a host, and ectosymbionts (or epibionts), which inhabit the exterior of their host, often attached to the host's surface. In contrast, obligate intracellular symbionts represent a subset of host-associated symbionts that thrive and replicate exclusively inside a host cell. We investigate the prevalence of predicted symbionts across the microbial tree of life, systematically detect the taxonomic extent of symbiont-exclusive clades and analyze the functional and metabolic patterns that characterize host-associated and intracellular lifestyles. The results of our large-scale approach enrich our understanding of microbial symbiont prevalence, diversity and metabolic capacity, providing global insights into microbial symbiosis and establishing a resource for potential biotechnological applications ranging from synthetic consortium design to minimal genome engineering and enzyme discovery.

Results

An artificial intelligence (AI)/ML framework (`symclatron`) for accurately classifying symbiotic lifestyles

Symbiont genome dataset and feature extraction. We constructed a comprehensive symbiont proteome database and extracted genomic features to enable lifestyle classification (Fig. 1). The first part included a reference set of 792 symbiont proteomes from across major clades (Fig. 1a) that was used to capture functional signals of symbiosis. The manually curated origin of the 792 symbiont proteomes and the expert-guided labeling criteria for the 6,751 genomes are described in the Methods. We performed orthogroup inference on these symbiont proteomes, from which we built 20,063 high-confidence profile hidden Markov models (HMMs) for orthogroups that contained ≥5 member proteins. The HMMs serve as potential markers of symbiotic gene content (Fig. 1a). Next, we built a dataset comprising 6,751 labeled microbial genomes (Fig. 1b and Supplementary Table 1) available in the Integrated Microbial Genomes and Metagenomes (IMG/M) database[29], spanning three lifestyle classes: free-living (n = 5,959), host-associated (n = 409) and obligately intracellular (n = 383). Detailed definitions of these lifestyle labels are provided in the Methods. With the aim of accounting for the effect of genome incompleteness in the downstream training of our models, all the genomes were artificially reduced to different levels of genome completeness and fragmentation lengths (Fig. 1b). We then applied uniform genome gene calling and annotation across the entire labeled dataset, yielding ~244 million proteins (Fig. 1b). Each protein was scanned against the symbiont HMM library (Fig. 1a) to determine orthogroup memberships. As a result, each genome was represented as a 20,063-dimensional feature vector reflecting the bitscore of symbiont-related orthogroups.

Genome labeling was performed following our conceptual demarcation of the symbiosis continuum (Fig. 1c). Then, the 20,063-dimensional encoding captured each genome's potential host-dependence traits (Fig. 1c). Accordingly, we used the encodings to train a symbiont classifier model (`symcla`) on the basis of discrete class labels and to train a regressor model (`symreg`) derived from the continuous labels (Fig. 1d). To optimize computational efficiency, the top 1,000 most relevant features were selected for each model.

Calibrating predictions with clade-level cross-validations and the neural network `symclatron`. To benchmark the model's generalizability to novel taxa, we implemented a clade-level leave-one-clade-out cross-validation strategy (Fig. 1e). In total, 11,025 unique clades (across species, genus, family, order, class and phylum levels) were represented in the labeled dataset (Fig. 1e, left). For each validation run, we withheld all genomes from one clade (for example, all members of a given family) from training and then tested the model on those held-out genomes (Fig. 1e, right). This process was repeated such that each clade (at each taxonomic rank) was left out in turn, yielding 11,025 clade-specific validations. By removing entire lineages during training, this rigorous scheme evaluates how well the classifier can predict lifestyles for previously unseen lineages.

We developed a feedforward neural network, termed '`symclatron`', to provide a final prediction of each genome's lifestyle on the basis of the clade-level cross-validation data and other interpretable genomic features (Fig. 1f). Rather than using thousands of raw ortholog features directly, the `symclatron` model integrates seven informative features per genome that summarize its position in symbiotic gene-content space. These features capture both the categorical and the continuous signals of symbiosis, as well as genome quality, and are defined below (Fig. 1f).

Symcla class probabilities (three features): the probabilities of the genome being free-living, host-associated or obligate intracellular, as predicted by an initial gradient-boosted trees classifier trained on the 20,063-dimensional feature matrix.

Symreg score (one feature): a continuous symbiosis score ranging from 0.0 (free-living) to 2.0 (obligate intracellular) produced by a gradient-boosted regression model (`symreg`). This models the genome's position along the host-dependence continuum (Fig. 1c).

Distance to training datasets (two features): for both the symcla classifier feature space and the symreg feature space, we computed the genome's weighted Euclidean distance to the nearest

https://doi.org/10.1038/s41587-026-03213-1

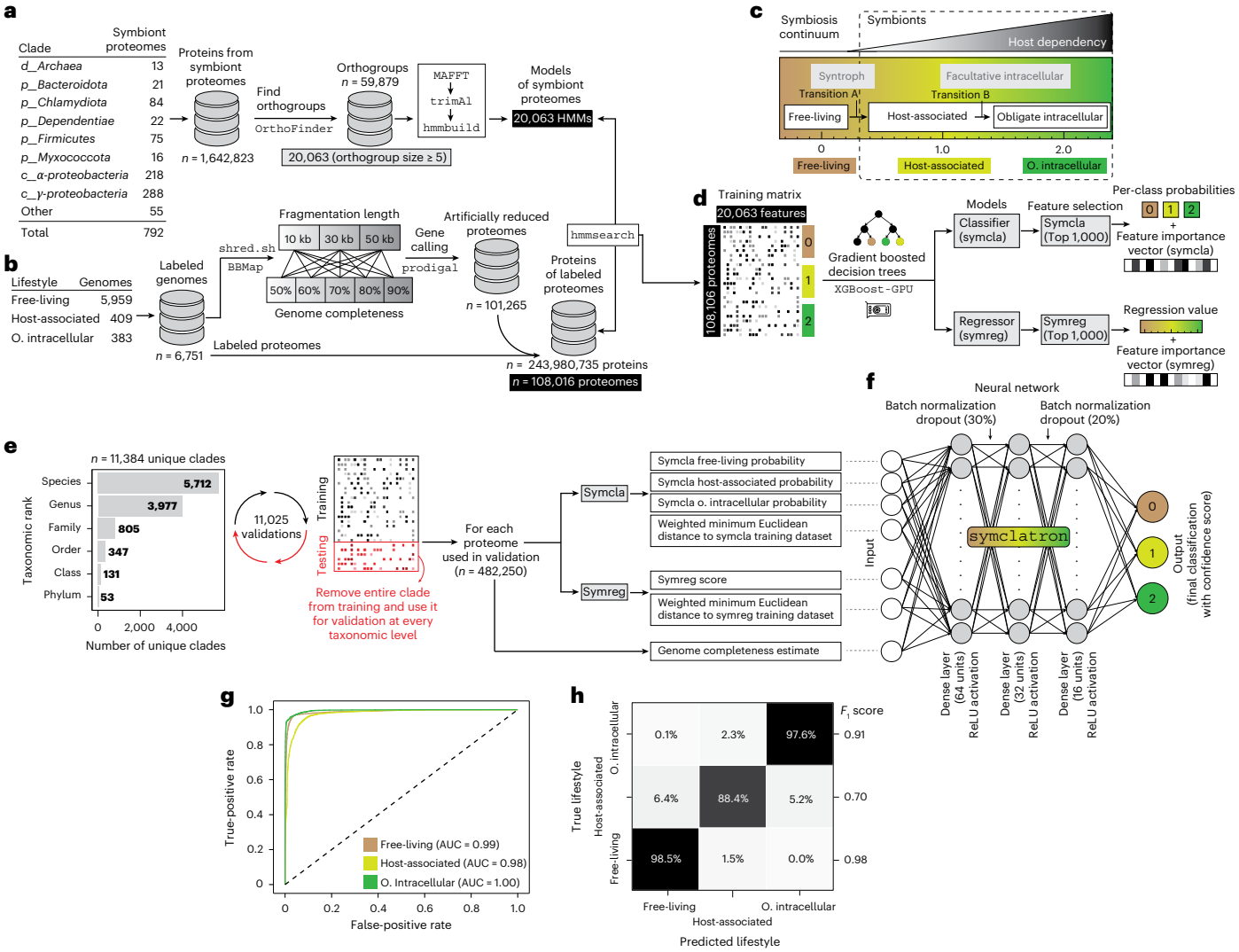


Fig. 1 | Developing an AI/ML framework for symbiotic lifestyle prediction in microbial genomes. a, Feature engineering workflow for detecting the genomic features specific to symbiont genomes. **b**, Construction of a genus-level representative database of lifestyle-labeled genomes accounting for artificial genome reductions. **c**, Schematic of the symbiosis continuum concept applied in the labeling strategy for model training. **d**, Construction of the training matrix for the symbiont classifier (symcla) and symbiont regressor (symreg) using XGBoost. The final models were optimized for the most relevant 1,000 features. **e**, Strategy implemented to calibrate the final lifestyle prediction by training a neural network on the outputs of the out-of-fold validations. Right, the neural network was trained on 11,025 rounds of benchmarking and validations of the classifier and regressor models by blinding the models to entire clades from the training data. Left, bar plot showing the total number of unique clades analyzed at each taxonomic rank. **f**, Architecture of the neural network featuring batch normalization and dropout layers (30% and 20%), with three dense layers using ReLU activation. The output layer also yields confidence scores for each lifestyle class (0, free-living; 1, host-associated; 2, obligate intracellular). **g**, ROC curves showing classification performance for each lifestyle category (one versus rest). ROC curves are shown for the final neural network model evaluated on the stratified 80:20 split of the integrated out-of-fold benchmarking dataset used for neural network model selection. **h**, Confusion matrix at 0.725 confidence threshold depicting correct classifications for each category. The data split for the neural network validation was applied at 80% for training (n = 385,800 proteomes) and 20% for testing (n = 96,450 proteomes).

training-set genome. These distance metrics (weighted by feature importances from the symcla/symreg models) quantify how atypical or novel the genome's gene content is relative to known symbionts and free-living organisms.

Genome completeness (one feature): an estimate of genome completeness (assembly quality) for the proteome, included to ensure that partial genomes are appropriately handled.

All seven features were scaled and passed as input to the neural network. The `symclatron` architecture consists of fully connected dense layers with batch normalization and dropout regularization (Fig. 1f). Specifically, three dense layers (each with ReLU activation and dropout rates) transform the input features, followed by an output layer with a softmax activation that yields probability scores for each of the three lifestyle classes (0, free-living; 1, host-associated; 2, obligate intracellular). This design leverages biologically interpretable inputs, derived from gene-content patterns and host-dependence signals, while using the neural network to capture nonlinear combinations of these inputs for final prediction. By building on the intermediate symcla and symreg models (which themselves provide feature importance insights), the `symclatron` predictions remain grounded in tangible biological features rather than opaque latent variables.

Classification performance across lifestyles. The `symclatron` model achieved high classification accuracy for all three lifestyle categories in cross-validation (Fig. 1g,h). Receiver operating characteristic (ROC) analysis (Fig. 1g) showed that the model discriminates lifestyles with excellent performance; the one-versus-all AUC was 0.99 for free-living, 0.98 for host-associated and 1.00 for obligate intracellular genomes.

These high AUC values indicate that symclatron consistently ranks the true class higher than the alternatives across diverse genomes. Correspondingly, the classifier's confusion matrix confirms accurate assignment of lifestyles (Fig. 1h). Overall accuracy was 96%, with most genomes falling on the diagonal of the confusion matrix. Free-living bacteria were almost always correctly identified (98.5% of free-living genomes predicted as free-living, Fig. 1h), with only minimal misclassification (1.5% predicted as host-associated and effectively none as intracellular). Obligate intracellular symbionts were also recognized with high precision (97.6% correctly classified; Fig. 1h). The model exhibited a slight tendency to confuse the two symbiotic classes in a minority of cases; about 5.2% of host-associated genomes were mispredicted as obligately intracellular and 2.3% of obligate intracellular genomes were predicted as host-associated (Fig. 1h). This minor class mixing reflects the biological reality that some host-associated bacteria approach an intracellular lifestyle but, importantly, very few symbionts were mistaken for free-living organisms or vice versa (Fig. 1h). Given that the benchmark set is class imbalanced, we emphasize balanced metrics in addition to accuracy. On the held-out validation set, symclatron achieved class-specific F_1 values of 0.982 for free-living, 0.699 for host-associated and 0.908 for obligate intracellular genomes. The strong performance on held-out clades underscores that the thoroughly calibrated symclatron can generalize to novel taxa, even when evaluating genomes from entirely untrained lineages, the model maintained robust accuracy with only rare errors. Our results demonstrate that the neural network, informed by independent models built on gene-content features, can reliably distinguish free-living versus symbiotic lifestyles and even delineate the degree of host dependence among symbionts.

Prediction confidence. To translate symclatron's predictions into actionable biological insights, we calibrated the model's probability outputs to assess prediction confidence (Extended Data Fig. 1). We first examined the distribution of prediction confidence scores (the highest class probability for each prediction) for correct versus incorrect assignments (Extended Data Fig. 1a). This analysis revealed a clear separation: the vast majority of true-positive predictions were made with high confidence, whereas false predictions tended to have lower confidence scores (usually <0.6). On the basis of this distribution, we selected a confidence threshold of 0.725, above which misclassifications are exceedingly rare (Extended Data Fig. 1a). At this cutoff, the false-positive rate is only 1.5% (that is, only 1.5% of incorrect predictions have a confidence ≥ 0.725), indicating that a prediction with ≥72.5% confidence is highly likely to be correct. Users of symclatron can, thus, choose to accept only predictions above this probability to minimize errors. Applying a confidence threshold allows a tunable trade-off between accuracy and data coverage (Extended Data Fig. 1b). As the minimum required probability is raised, the model's accuracy on retained predictions increases, while the fraction of genomes that meet the threshold (coverage) decreases.

Lastly, we assessed model calibration across the full range of confidence scores by stratifying predictions into 40 bins and computing performance within each bin (Extended Data Fig. 1c). This analysis showed that symclatron's reported probabilities are well aligned with actual outcomes. Higher confidence scores yield higher precision in lifestyle assignment, as reflected in the confusion matrices for each bin (Extended Data Fig. 1c). In the lowest-confidence bins (for example, <0.5), the model's predictions include a mix of errors, especially for the difficult host-associated category. In contrast, in the highest-confidence bin (≥0.95 probability), predictions are almost always correct; free-living and obligate intracellular calls were >99% accurate and even host-associated calls were ~92% accurate. Even at moderately high confidence levels (for instance, ≥0.70), the precision for free-living and intracellular predictions already exceeds 97–98%, with host-associated predictions around ~88% accurate (Extended Data Fig. 1c). These results confirm that the model's confidence scores are meaningful measures of confidence for practical applications of symclatron.

Contribution of symclatron features. To assess feature redundancy and contribution in the final symclatron model, we compared the performance of symcla alone against the full seven-feature neural network on the held-out 20% split of the integrated benchmark dataset (Extended Data Fig. 1d) and then performed grouped permutation importance on the final neural network (Extended Data Fig. 1e). Symcla alone achieved high overall accuracy (95.34%), indicating that it captures much of the dominant signal in the easier regions of the host-dependence continuum. However, because the evaluation set is dominated by free-living genomes, accuracy alone underestimated the contribution of the full model. The complete symclatron model increased macro-F_1 from 0.769 to 0.863, increased host-associated recall and F_1 from 32.7% and 0.467 to 83.3% and 0.697 and increased intracellular recall/F_1 from 83.3% and 0.861 to 93.3% and 0.911. Permutation analysis showed that, although the symcla class probabilities contributed the strongest overall signal, the symreg score and the distance-to-training-set features provided nonredundant information, particularly for host-associated predictions and, to a lesser degree, for intracellular predictions. These results indicate that symcla is an informative coarse classifier, whereas the full symclatron model is required for balanced three-class prediction across the symbiosis continuum.

Predictions as a function of distance from training data. To evaluate whether prediction accuracy declines for genomes that are farther from the training data, we analyzed the leave-one-clade-out benchmark as a function of a conservative combined weighted distance to the remaining training set, defined as the maximum of the symreg and symcla nearest-neighbor distances (Extended Data Fig. 1f–h). As expected, broader held-out clades were farther from the training data, with the median combined distance decreasing from 61.98 at phylum to 36.84 at species (Extended Data Fig. 1f). Genome-weighted performance remained strong through the first four distance quintiles but declined in the most distant quintile, where accuracy dropped from 94.1% to 72.7% (Extended Data Fig. 1g). This decline was strongest for host-associated genomes, whose recall fell from 73.6% in the closest quintile to 7.2% in the most distant quintile; intracellular recall fell from 70.8% to 20.3%. When the same analysis was repeated with equal weighting of held-out clades rather than genome count, the same pattern remained, with the decrease concentrated in the most distant quintile (Extended Data Fig. 1h). These results show that symclatron generalizes well across broad phylogenetic diversity but predictions for the most phylogenetically novel clades, especially within the host-associated category, should be interpreted with greater caution.

Symclatron availability. Our symbionts prediction framework uses biologically interpretable features of genomes to accurately predict microbial lifestyle with high confidence (Fig. 1g,h and Extended Data Fig. 1c). The model is robust to taxonomic novelty (validated across thousands of clades) and provides probabilistic predictions that are well calibrated, enabling flexible, threshold-tuned deployment. These characteristics make symclatron a powerful tool for characterizing the lifestyle of microorganisms, even in the face of incomplete genomes or previously unseen lineages, while retaining transparency and control over prediction confidence. Symclatron is publicly available from GitHub (https://github.com/NeLLi-team/symclatron).

SymGs: a global-scale genomes catalog of microbial symbionts

Genome catalog construction and symbiont classification. We started by binning a comprehensive set of 31,152 public environmental sequencing data projects from the IMG/M database to construct a catalog of MAGs referred to as 'NeLLi2023'. We also collected reference genomes from IMG/M and Genome Taxonomy Database (GTDB). After

https://doi.org/10.1038/s41587-026-03213-1

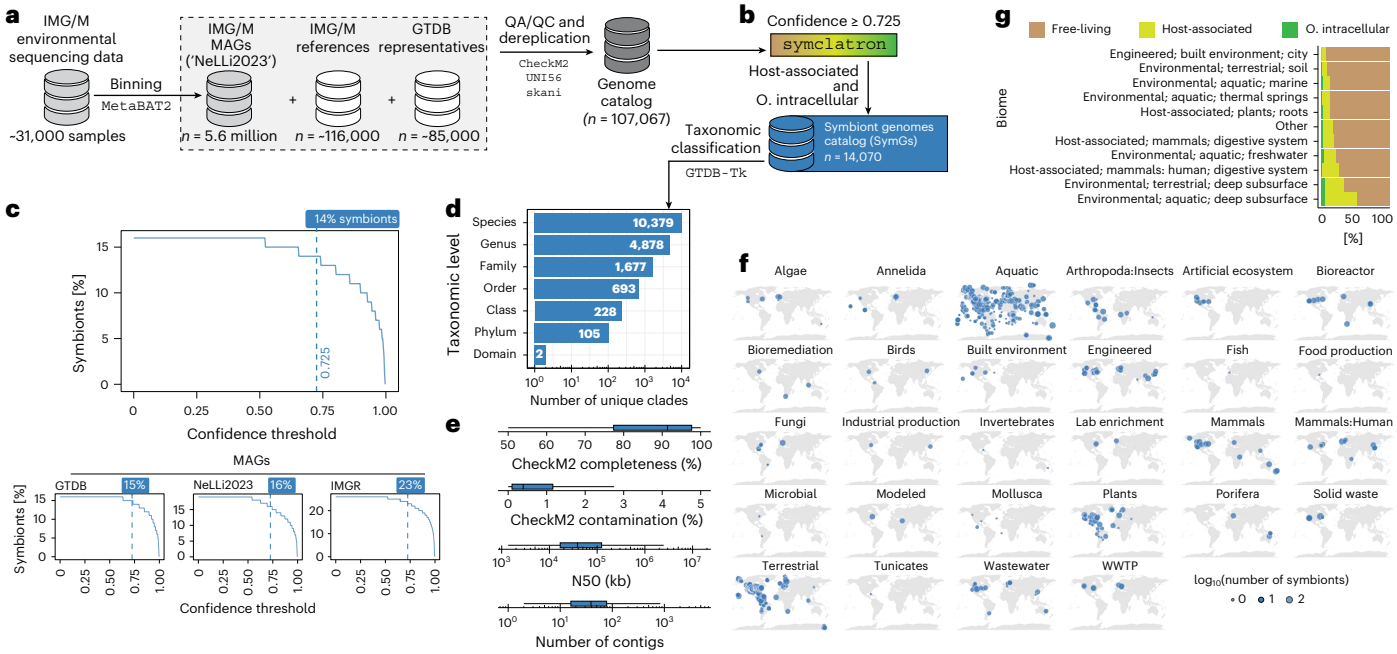


Fig. 2 | Cataloging the taxonomic and global environmental distribution of microbial symbionts. a, Overview of the pipeline for creating the genome catalog. QA/QC, quality assurance and control. **b**, The genome catalog was subjected to `symclatron` to predict lifestyles of Bacteria and Archaea; only high-confidence predictions are reported in all downstream results. **c**, Percentage of predicted symbionts as the added percentage of the categories host-associated and obligate intracellular across varying confidence thresholds. Bottom, percentages specified only for the three databases of MAGs used in this study. **d**, Taxonomic breakdown of the clades identified in the SymGs catalog. **e**, Quality assessment metrics of the SymGs catalog (each box plot is derived from the total genomes in the catalog; $n = 14{,}070$). The box plot shows the median and interquartile range; whiskers represent the minimum and maximum values. **f**, Global distribution of genomes in the SymGs catalog across different biomes. Geographical locations and environment data are only shown for metagenomes with available metadata in the IMG/M. **g**, Percentage of predicted lifestyles (only MAGs) across the top ten biomes with most symbiont prevalence. Remaining biomes were collapsed into the 'other' category. Bars show the relative frequency of each category in that biome.

quality assurance, quality control and dereplication at 95% average nucleotide identity (ANI), we obtained a genome catalog encompassing 107,067 genomes of Bacteria and Archaea (Fig. 2a). To predict symbiotic lineages within our catalog, we applied `symclatron`. Using the stringent confidence threshold 0.725, we identified 14,070 high-confidence symbiont genomes in the catalog (Supplementary Table 2), equating to 14% of all genomes in the catalog; we define the collection of high-confidence symbiont genomes as SymGs (Fig. 2b). In other words, roughly one in seven genomes in the assembled dataset was predicted to represent a symbiont (Fig. 2c). The symbiont classification rate was even higher when computed only for the MAGs fraction of the catalog, ranging from 15% up to 23% symbionts (Fig. 2c); roughly one in five MAGs was predicted to be a symbiont.

Taxonomic diversity and genome quality of SymGs. The SymGs catalog with 14,070 symbiont genomes spans a wide phylogenetic range across the prokaryotic tree of life, highlighting the broad taxonomic diversity of microbial symbionts (Fig. 2d). SymGs encompassed representatives from all prokaryotic ranks, including dozens of bacterial phyla (Fig. 2d). This diversity indicates that symbiotic lifestyles have arisen in many branches of Bacteria and Archaea. Indeed, the overall taxonomic profile of the SymGs closely mirrored the relative composition of the full 107,067-genome catalog (Extended Data Fig. 2a), implying that most high-level clades contain at least some symbiotic members. As expected, lineages well known for symbiosis were prominent within SymGs (for example Pseudomonadota, Proteobacteria, Bacteroidota and Firmicutes), reflecting an enrichment of symbiont-rich clades in the catalog. At the same time, SymGs also included members of less-characterized phyla (including candidate phyla), underscoring that symbiotic associations are phylogenetically widespread and not limited to traditionally studied groups. In addition to their broad taxonomic spread, the SymGs are distinguished by high genome quality and characteristic features (Fig. 2e). Most symbiont genomes were high-quality MAGs or nearly complete genomes with minimal assembly fragmentation (Fig. 2e). `CheckM2` quality assessment indicated that SymGs had a median completeness higher than 90% and low contamination (typically <0.2%), comparable to the quality metrics observed for genomes in the overall catalog (Extended Data Fig. 2b). Many symbiont MAGs are relatively small, streamlined genomes, which contributed to high assembly contiguity (high N50 values) and low contig counts per genome (often only a few contigs; Fig. 2e). These statistics are consistent with the biology of well-studied obligate symbionts, which often possess reduced, gene-dense genomes that could be fully or nearly fully recovered from metagenomes. In summary, SymGs represents a set of genomes that not only spans diverse taxa but also meets high standards of completeness and purity, ensuring that downstream analyses of symbiont biology rest on a solid genomic foundation.

Ecological and geographic distribution of symbiont genomes. SymGs exhibits broad ecological and geographic coverage, demonstrating that symbiotic microorganisms are ubiquitous across global environments (Fig. 2f). SymGs was recovered from environmental samples spanning all major biome categories and continents or oceans (Fig. 2f), highlighting a near-global distribution of these genomes. Notably, symbiont MAGs were disproportionately derived from host-associated and aquatic biomes, which together accounted for the majority of SymG source environments (Fig. 2f). Host-associated samples (such as animal microbiomes, insect guts and plant rhizospheres) and aquatic ecosystems (marine and freshwater environments) each contributed large numbers of symbiont genomes. Overall, the global and ecological spread of SymGs parallels the broad sampling of the entire genome dataset (Extended Data Fig. 2c) but

with a clear emphasis on environments involving host organisms or aquatic environments where symbiotic interactions are expected to be especially common.

To further characterize the environmental contexts of the symbiont genomes, we examined the most common biome sources represented in SymGs (Fig. 2g). The top contributing biomes included the deep subsurface of terrestrial and aquatic samples. Each of these biomes contributed hundreds to thousands of symbiont genomes, underlining that symbiotic microorganisms thrive in a variety of environments but are especially prevalent in association with animal hosts and aquatic ecosystems. On the contrary, as expected, obligate intracellular microorganisms were not detected in built environments (Fig. 2g). Nonhuman animal digestive systems showed presence of predicted obligate intracellular symbionts, while host-associated (and not obligate intracellular) lifestyles were predominant in human digestive systems (Fig. 2g), which may represent more facultative or environmentally transmitted symbionts.

To evaluate how representative the labeled training set is relative to the IMG/M-derived NeLLi2023 catalog, we compared GTDB taxonomic composition across the 6,751 labeled training genomes, the full 107,067-genome NeLLi2023 catalog and the 14,070 high-confidence SymGs subset (Extended Data Fig. 2d,e). The training set spans both Bacteria and Archaea domains and includes 53 phyla, 131 classes, 347 orders, 805 families, 3,977 genera and 5,712 species, indicating broad phylogenetic coverage but not a compositional mirror of the MAG-derived catalog. At the phylum level, the training set was moderately concordant with the full NeLLi2023 catalog (Spearman $\rho = 0.57$; Jensen–Shannon distance = 0.36) and with SymGs (Spearman $\rho = 0.51$; Jensen–Shannon distance = 0.66). This pattern is expected because the labeled set was built from high-quality, manually curated reference genomes, whereas NeLLi2023 and SymGs derive largely from MAGs reconstructed from environmental projects and, therefore, extend much further into uncultivated diversity. Thus, the training set captures many of the major clades present in the catalog while remaining intentionally biased toward better-characterized reference lineages. This ecological resemblance is shown explicitly in Extended Data Fig. 2f, where the relative representation of IMG ecosystem categories in SymGs closely tracks that of the full NeLLi2023 catalog overall (Spearman $\rho = 0.84$), while revealing enrichment of SymGs in several host-associated and aquatic environments. Host-associated datasets constituted a minority of the MAG input data given that, among the NeLLi2023 MAGs with environment metadata, 13.2% derived from IMG datasets categorized as host-associated, compared to 12% of the high-confidence SymGs subset, indicating that the prevalence of predicted symbionts is not driven simply by overrepresentation of host-associated metagenomes.

Collectively, the construction and characterization of SymGs reveal a remarkable breadth of microbial symbionts. The SymGs comprise over 14,000 genomes spanning myriad taxa and including high-quality representatives of known and novel symbiont lineages. By contrasting the SymGs with the broader genome catalog, we observe that symbiotic lineages make up roughly 15–23% of the sampled microbial diversity, a substantial and previously underappreciated fraction.

Functional traits that differentiate microbial lifestyles and allow accurate prediction of symbionts

To identify functional traits that differentiate microbial lifestyles, we performed SHAP (Shapley additive explanations) analysis on the annotated features used by the symcla and symreg models of `symclatron`, which stratified features into quadrants defined by SHAP contribution and HMM bit score (Extended Data Fig. 3a for symcla; Extended Data Fig. 4a for symreg). Analysis of Gene Ontology (GO) term mapping resolved two major sets of lifestyle-associated traits in the NeLLi2023 catalog. In the symcla analysis, low detection of Q3 features contributed negatively to symbiotic classification and was therefore more indicative of a free-living lifestyle (Extended Data Fig. 3b, Q3). In contrast, high detection of Q2 features contributed positively to symbiotic classification and was therefore more indicative of a symbiotic lifestyle; the corresponding GO terms were enriched in transport, membrane and other host-dependence-related functions (Extended Data Fig. 3b, Q2). The top 35 features of symcla Q2 accounted for 25.7% of the cumulative mean SHAP signal (Extended Data Fig. 3c). These high-contributing features sat on top of a strongly long-tailed distribution rather than a small discrete set, indicating that symbiont prediction draws on a broader continuum of lower-contributing domains (Extended Data Fig. 3d).

The same quadrant contrast was recovered when the analysis was repeated on the labeled training genomes. Low detection of Q3 features was again more indicative of a free-living lifestyle, whereas high detection of Q2 features was more indicative of a symbiotic lifestyle; the corresponding Q2 terms emphasized transport and membrane-related functions (Extended Data Fig. 3e). The highest-ranking symcla Q2 features in the training data closely resembled those identified in the environmental catalog (Extended Data Fig. 3f) and the full ranked SHAP distribution retained the same long-tailed structure (Extended Data Fig. 3g). Across the shared Q2 feature set, mean absolute SHAP values were strongly correlated between the catalog and the labeled training genomes (Spearman $\rho = 0.976$) and 31 of the top 35 features were shared between the two rankings (Extended Data Fig. 3h). These results indicate that the functional signals highlighted by the symcla model are not unique to the environmental catalog but are also present in the curated reference set used for training.

The parallel symreg analysis (Extended Data Fig. 4a) recovered the same broad biological distinction while emphasizing a complementary subset of predictive features. In the NeLLi2023 catalog, low detection of Q3 features was more indicative of a free-living lifestyle and the corresponding terms were enriched in biosynthetic functions, whereas high detection of Q2 features was more indicative of a symbiotic lifestyle and the corresponding terms were enriched in transport and membrane-associated functions (Extended Data Fig. 4b); the top 35 symreg Q2 features accounted for 52.2% of the cumulative mean SHAP signal (Extended Data Fig. 4c). The corresponding ranked distribution was long-tailed in both the catalog and the labeled training genomes (Extended Data Fig. 4d,g) and the same GO term contrast was recapitulated in the training set (Extended Data Fig. 4e). The highest-ranking training features closely matched those from the catalog (Extended Data Fig. 4f) and catalog and training symreg estimates were strongly concordant across shared Q2 features (Spearman $\rho = 0.951$), with 28 of the top 35 features shared between datasets (Extended Data Fig. 4h). However, the catalog-level symcla and symreg Q2 rankings shared only three of their respective top 35 features and showed low overall concordance across the shared Q2 set (Spearman $\rho = 0.096$), indicating that the classifier and regression branches capture complementary aspects of host dependence while retaining a limited shared core of symbiosis-associated markers. Feature rank discrepancies between the models arise from distinct mathematical optimization objectives. The classifier (symcla) establishes discrete boundaries between the three lifestyle classes by prioritizing category-separating attributes. Conversely, the regressor (symreg) optimizes for a continuous score along the host-dependence continuum to capture progressive functional contraction trends. Permutation analysis (Extended Data Fig. 1e) confirmed that both models provide nonredundant information that the final neural network integrates to improve classification accuracy.

`Symclatron` allows identification of bacterial and archaeal clades dominated by symbionts

Widespread distribution of symbionts across Bacteria and Archaea. Microbial symbionts are phylogenetically widespread, appearing in numerous branches of the bacterial (Fig. 3) and archaeal (Fig. 4) trees of life. Nearly all major phyla include some host-associated members,

https://doi.org/10.1038/s41587-026-03213-1

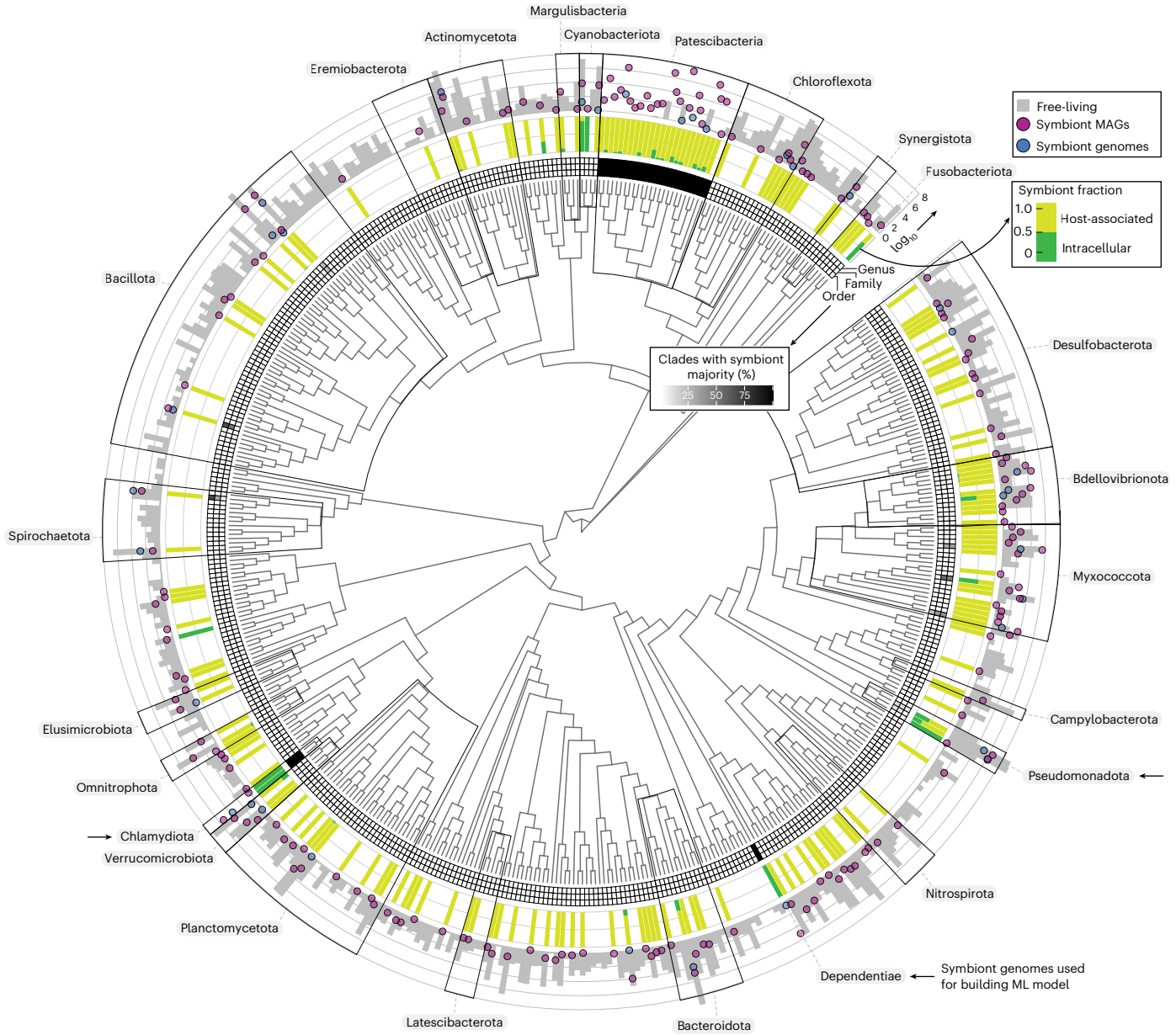


Fig. 3 | Phylogenetic distribution of the bacterial symbionts. The tree encompasses a representative genome from each class spanning various phyla. The tree was rooted using Fusobacteriota. From the innermost to the outermost rings, the following information is displayed: (1) the heat map ring shows the percentage of symbiont-exclusive clades per class found at the order, family and genus levels; (2) the bars represent the specific lifestyle of the symbiont fraction within the class; (3) the filled dots and the gray bars show the $\log_{10}$-scaled total number of genomes analyzed in each class; (4) clades from which genomes were used in building the classifier and regressor models are designated by black arrows. Additional details on genomes of predicted symbionts and symbiont-exclusive clades are provided in Supplementary Tables 2 and 3. Tree reconstruction is described in the Methods.

indicating that transitions to symbiosis have occurred repeatedly in diverse lineages. However, the prevalence of symbionts varies greatly among clades, with certain groups especially enriched for symbiotic lifestyles (Figs. 3 and 5). For example, the Candidate Phyla Radiation of bacteria (this entire lineage is now classified as a phylum-level group; analogous to the Patescibacterota, it has been formally designated as Minisyncoccota[39]) and the DPANN superphylum of archaea (for example, Nanoarchaeota and Aenigmatarchaeota) consist almost entirely of host-dependent organisms. By contrast, in large phyla such as Pseudomonadota or Actinomycetota, symbionts represent a smaller subset of the total diversity, yet, even in these groups, multiple subclades (for instance, the order Rickettsiales in Alphaproteobacteria or family Mycobacteriaceae in Actinobacteria) have evolved host-associated lifestyles (Fig. 3).

Our analysis also shows that many symbiont-exclusive clades are dominated by intracellular symbionts (detailed metrics of every clade analyzed in this study in Supplementary Table 3). For instance, the phylum Chlamydiota (Fig. 3) is composed entirely of obligate intracellular bacteria (most are expected to be pathogens of animals according to the current understanding of this clade) and the Dependentiae (formerly TM6) are similarly almost exclusively intracellular symbionts (most are expected to be endosymbionts of protists according to the current understanding of this clade). In other clades, the symbionts are predominantly host-associated but not obligate intracellular, as in the case of certain bacterial clades within the Bacteroidota and Bacillota (formerly known as Firmicutes) phyla that are traditionally known to colonize the gut of animal hosts without residing inside host cells. In sum, the broad phylogenetic survey of symbiont genomes shows that

https://doi.org/10.1038/s41587-026-03213-1

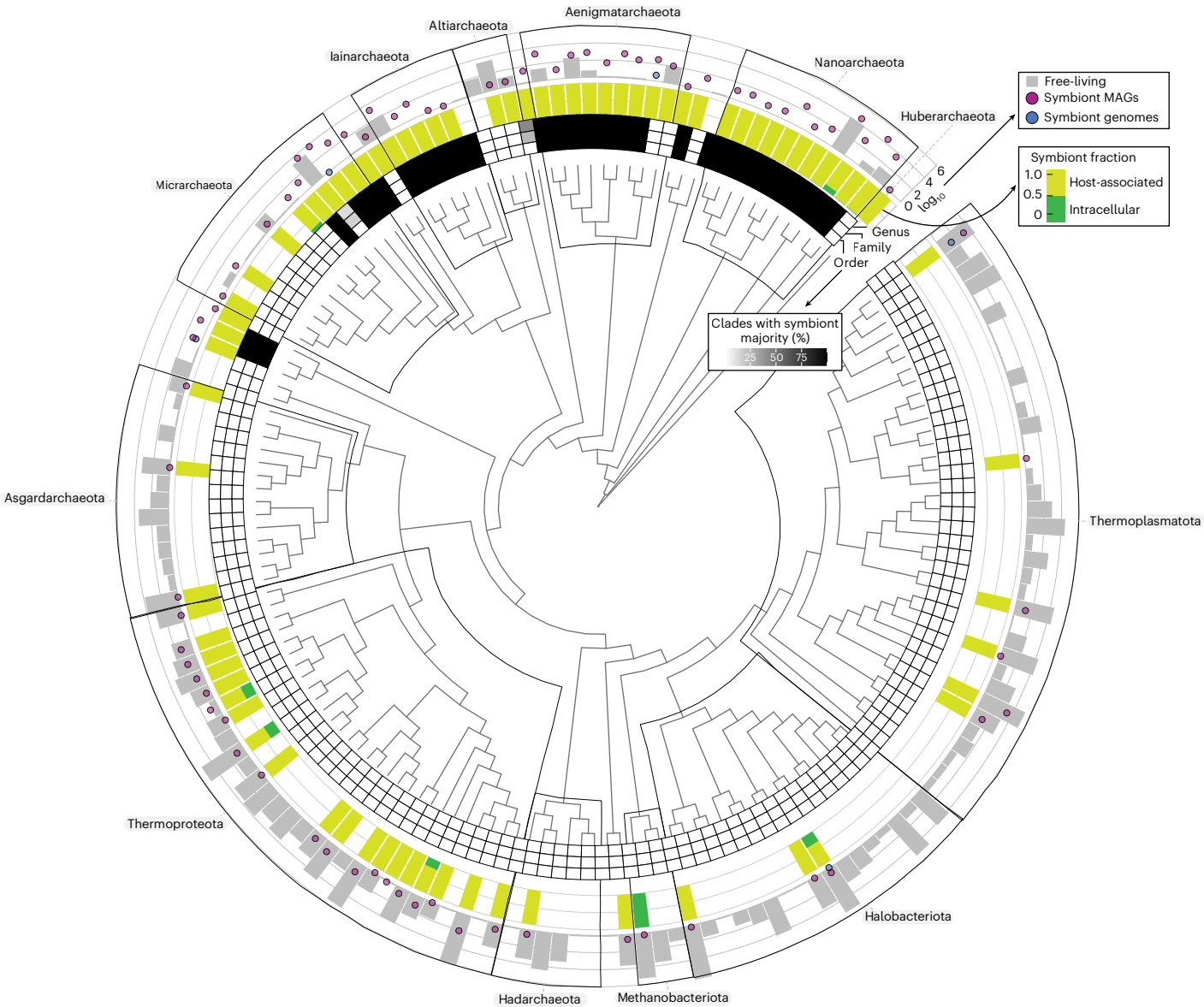


Fig. 4 | Phylogenetic distribution of the archaeal symbionts. The tree encompasses a representative genome from each order spanning various phyla. The tree was rooted using bacterial outgroups that were removed from the visualization. From the innermost to the outermost rings, the following information is displayed: (1) the heatmap ring shows the percentage of symbiont-exclusive clades per order found at the order, family and genus levels; (2) the bars represent the specific lifestyle of the symbiont fraction within the order; (3) the filled dots and the gray bars show the $\log_{10}$-scaled total number of genomes analyzed in each order. Additional details on genomes of predicted symbionts and symbiont-exclusive clades are provided in Supplementary Tables 2 and 3. Tree reconstruction is described in the Methods.

symbiotic relationships have evolved in diverse taxa of the microbial world, with some lineages undergoing especially pronounced shifts to host-dependent, often intracellular, existence in Bacteria (Fig. 3) and Archaea (Fig. 4).

Symbiont-exclusive clades. Using the above phylogenomic framework, we identified numerous 'symbiont-exclusive' clades, defined as taxonomic groups in which ≥90% of the member genomes are predicted to be symbionts with high-confidence (Fig. 5a). As expected, because of microbial diversification, the majority of symbiont-exclusive clades are found at the lower taxonomic ranks; over 1,000 genera consist almost entirely of symbionts. At progressively higher ranks, fewer clades meet the ≥90% symbiont criterion. We detected on the order of tens of classes and hundreds of orders and families that are symbiont dominated but only a handful of phyla (Fig. 5a). The few top-level groups with pervasive symbionts include well-known obligate symbiont lineages such as Chlamydiota (phylum of obligate intracellular bacteria), as well as several candidate phyla in bacteria and archaea that lack free-living representatives. This pattern indicates that completely host-dependent groups exist at all taxonomic scales but broad, phylum-level specializations to symbiosis are rare. More commonly, symbiosis evolves as a specialized subset within larger lineages. The prevalence of so many symbiont-exclusive genera and families suggests multiple independent acquisitions of a host-associated lifestyle within each major bacterial phylum. Furthermore, many of these clades have only recently been recognized through metagenomic sequencing (Fig. 5a), implying that our knowledge of microbial diversity has expanded to include previously undiscovered symbiont lineages.

Indeed, a notable fraction of our high-confidence symbiont clades appear to be MAG-exclusive (known solely from MAGs), highlighting how cultivation-independent approaches are revealing symbionts that were missed in cultivation-based surveys (Fig. 5a). Genome

https://doi.org/10.1038/s41587-026-03213-1

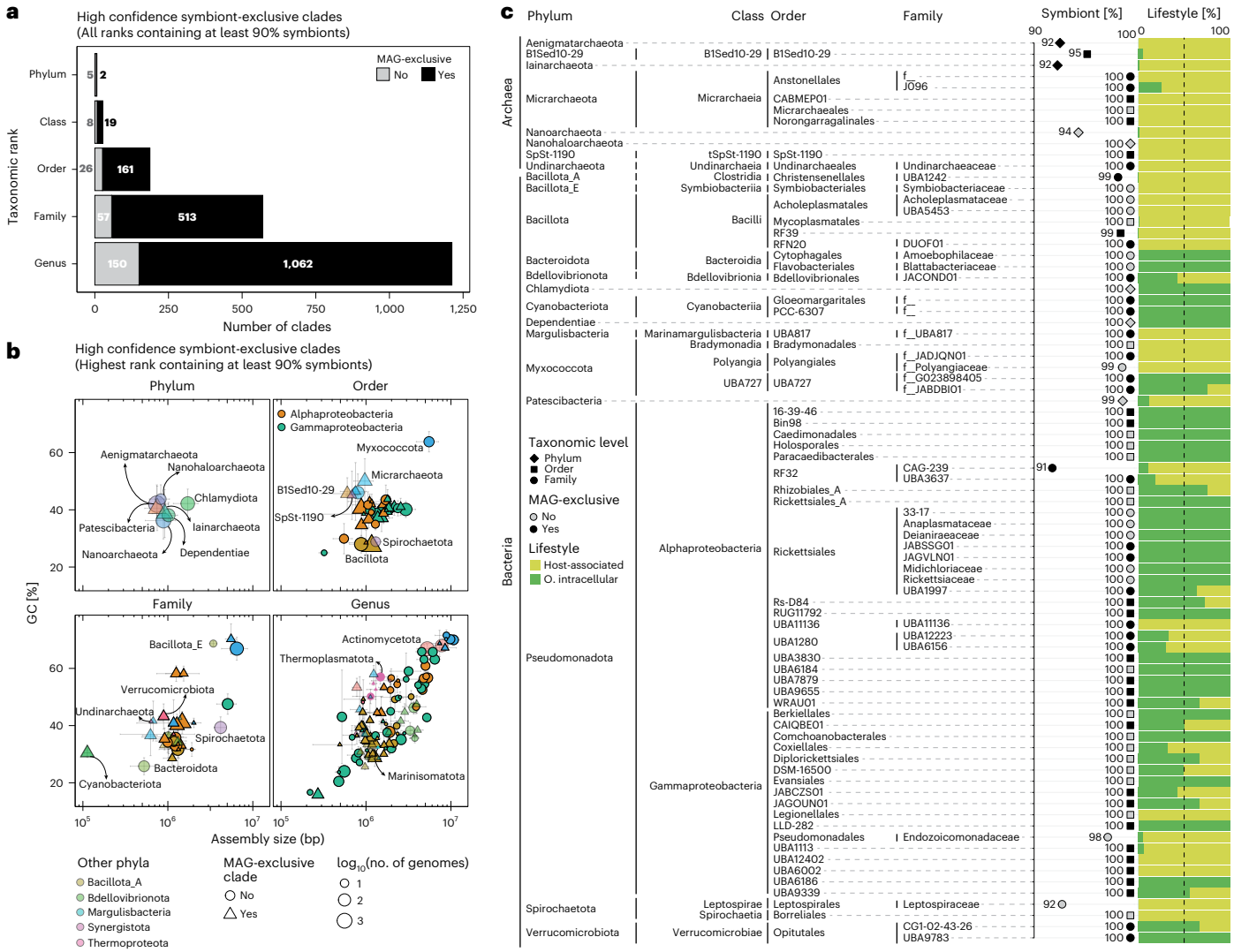


Fig. 5 | Genomic characterization of symbiont-exclusive clades. a, Distribution of symbiont-exclusive clades across different taxonomic ranks. A clade is depicted as symbiont-exclusive at any taxonomic rank if ≥90% of genomes were classified as symbionts with mean confidence of the predictions of the clade ≥ 0.725 and s.d. ≤ 0.2. A clade is depicted as 'MAG-exclusive' if it was not previously represented by reference isolates. **b**, G+C content and assembly size of symbiont-exclusive clades across different taxonomic ranks, from phylum to genus. Dot color indicates the phylum in which we found symbiont-exclusive clades; only Alphaproteobacteria and Gammaproteobacteria are shown at the class level. Each point represents a clade. **c**, Lifestyle composition of symbiont-exclusive clades down to the family taxonomic rank. Unknown taxonomic classifications are depicted by the first letter of the taxonomic rank and two underscore signs. The two rightmost columns indicate the fraction of genomes in that clade classified as symbionts and its percentage distribution of symbiotic lifestyles.

characteristics of many symbiont-exclusive clades reflect reductive evolution but some have larger genomes and high G+C content (Fig. 5b). There is a pronounced trend toward genome miniaturization; many obligate symbionts in our dataset have genome sizes well under 1.5 Mbp and some approach the theoretical lower size limit for cellular life (on the order of only a few hundred kilobases). These reduced genomes also tend to have biased nucleotide compositions, often <40% G+C. For instance, several known insect endosymbiont clades (Fig. 5b) harbor genomes around ~0.2–0.7 Mb with extreme A+T bias (for example, one lineage at ~13–20% G+C). Such values are in line with the most extreme cases previously reported in endosymbiotic bacteria such as *Candidatus Zinderia insecticola* with a 208-kb, 13.5% G+C genome[40].

To examine the nature of these symbiont-dominated clades more concretely, we broke down a selection of high-confidence symbiont-exclusive clades (down to family level for clarity), detailing their taxonomy, symbiont percentage and lifestyle composition (Fig. 5c). All of the highlighted clades have at least 90% of members identified as symbiotic, with many at a full 100%. The proportion of host-associated versus intracellular symbionts is indicated for each clade, illustrating different degrees of integration with host organisms. Some clades are composed almost entirely of obligate intracellular symbionts. For example, the family Blattabacteriaceae (order Flavobacteriales) is 100% symbiotic, consisting of obligate intracellular (traditionally known to be endosymbionts of cockroaches) bacteria that live within host fat-body cells[41]. Similarly, the Amoebophilaceae (order Cytophagales) are intracellular symbionts traditionally known to live inside amoeba hosts[42]. In these groups, virtually all symbionts inhabit host cells, indicating an extreme specialization and dependence on the host environment.

In contrast, other clades, while still exclusively host-associated, predominantly contain nonobligate intracellular symbionts (Fig. 5c). An example is the order Acholeplasmatales (class Bacilli (formerly Mollicutes), phylum Bacillota), which includes *Mycoplasma*-like bacteria found on animal and plant hosts; these symbionts require a host

to survive but typically reside on cell surfaces or in host fluids rather than intracellularly invading host cells. As a result, Acholeplasmatales genomes, although reduced, are not as minimal as those of strictly intracellular bacteria (Bacillota dots in Fig. 5b). Likewise, certain gut bacterial clades like Christensenellales (former Christensenellaceae) within the class Clostridia and phylum Bacillota_A (former Firmicutes) are known to be almost entirely restricted to animal hosts yet remain outside host cells[43] as predicted by us (Fig. 5c).

Among the Pseudomonadota symbionts that were used for building the classifier, only Rickettsiales did not meet the criteria for being considered a symbiont-exclusive clade despite its known occurrence as symbionts of a wide variety of eukaryotic species (Fig. 5c). Both orders Holosporales and Legionellales did form symbiont-exclusive clades (Fig. 5c). This is consistent with recent genome-based analyses suggesting free-living or possibly facultative host-associated lifestyles for certain, basal lineages of Rickettsiales[44]. Within Rickettsiales, only Rickettsiaceae, Midichloriaceae, JAGVLN01, JABSSG01, Deianiraeaceae, Anaplasmataceae and 33-17 formed symbiont-exclusive families of obligate intracellular symbionts (Fig. 5c). In contrast, some members of the Rickettsiales family UBA1997 were classified as host-associated. Alternatively, these could also be ectosymbiotic, a lifestyle that has been reported in the Rickettsiales for *Candidatus Deianiraea vastatrix*[45]. Genome-based inferences proposed that intracellularity likely evolved multiple times in the evolution of Rickettsiales, which is consistent with the prediction of host-associated, not necessarily obligate intracellular lifestyles in certain lineages of this order[44,46].

Overall, our data indicate that most clades dominated by symbionts have gravitated toward intracellular niches (Fig. 5c), although there are puzzling cases such as the MAG-exclusive unknown families of Gloeomargaritales and PCC-6307 within Cyanobacteriota (Fig. 5c). Nevertheless, in other clades, a notable subset maintain a less intimate, nonobligate intracellular association with their hosts. This variation suggests that, while host dependence is the common thread uniting these clades, the mode of symbiosis (intracellular endosymbiosis versus host association) can differ, likely influenced by the specific ecological context and evolutionary history of each lineage.

Differential metabolic module completeness found in symbiont-exclusive clades compared to their free-living relatives

Widespread loss of biosynthetic pathways in symbiont genomes. We calculated metabolic module completeness for each clade and used bootstrap resampling to detect major differences between symbiont and free-living groups. For each symbiont-exclusive clade, bootstrap comparisons were performed against the corresponding free-living sister set from the immediately higher parent clade after excluding the focal symbiont clade itself; 100 bootstrap iterations were then generated by sampling 100 genomes with replacement from each lifestyle group per iteration and clades were retained only when at least five free-living sister genomes were available. The 100 random subsamples of genomes were used to generate confidence intervals for each module's completeness in symbiont versus free-living cohorts. This analysis ensured that observed differences (for example, loss of amino acid synthesis or enrichment of nitrogen metabolism) are not because of random genome content variation but reflect true adaptive divergences.

We observed that, across diverse lineages, many core metabolic pathways are incomplete or entirely absent in symbiont-exclusive clades (Fig. 6). For example, symbionts in the families CG1-02-43-26 and UBA9783 within the order Opitutales (phylum Verrucomicrobiota) and the order Borreliales within the phylum Spirochaetota show pervasive loss of amino acid biosynthesis pathways that are present in related free-living bacteria. Modules for the production of branched-chain amino acids (valine, isoleucine and leucine), as well as other essential amino acids such as ornithine, arginine and histidine, are largely missing or pseudogenized in the symbionts (Fig. 6). This trend is echoed in the symbiont-exclusive order Borreliales within the phylum Spirochaetota, which lack key enzymes for serine, threonine and arginine biosynthesis that their free-living relatives retain (Fig. 6). We generally observed that symbionts show pervasive loss of amino acid biosynthesis pathways that are present in related free-living bacteria. Similarly, symbionts rely on salvage pathways for nucleotides; we found de novo purine and pyrimidine biosynthesis modules (for example, M00048 and M00051) to be partially or wholly absent in symbiont clades, whereas free-living strains encode these pathways completely (Fig. 6).

The two symbiont genera *UMGS1491* and *CAG-557* within the family Acutalibacteraceae (class Clostridia and phylum Bacillota_A) were characterized by having also shed parts of central metabolic pathways such as the citrate cycle (TCA cycle), being absent in all examined symbionts within these two clades (Fig. 6). Their absent citrate cycle indicates a reduced capacity for aerobic respiration in these symbionts. Nevertheless, they do show a more complete module for V-type ATPase metabolism.

In a contrasting example of related symbiont clades with potentially different hosts, we observed that the symbiont-exclusive order UBA1113 within Gammaproteobacteria retained a complete module cytochrome *o* ubiquinol oxidase but the same module was not consistently detected in the related symbiont-exclusive order UBA11136 and was completely undetectable in free-living counterparts (Fig. 6). This differential presence of the cytochrome *o* ubiquinol oxidase module indicates that UBA1113 families are still occupying niches with oxygen flux while UBA11136 families are most likely adapting to oxygen-poor host environments.

Our data show that the loss of energy-intensive biosynthetic functions indicates that symbionts frequently obtain amino acids, nucleotides, cofactors, vitamins and other key metabolites (Extended Data Fig. 5b) from their hosts (or cosymbionts), a hallmark of metabolic dependence in host-associated bacteria[47–49]. However, some symbiont-exclusive clades can also show more complete modules for these biosynthetic functions (Extended Data Fig. 5b), indicating that the selective pressure of the specific host or niche environment is a differential factor in the evolution of metabolic capabilities of symbionts and a reduction in metabolic function is not the only adaptation mechanism of host-associated and intracellular microbial symbionts.

Retention of niche-specific metabolism in symbiont genomes. We observe that, despite overall metabolic contraction, symbionts commonly retain or even enrich certain pathways that may confer a selective advantage within the host environment. Some biosynthetic modules are more complete in symbionts than in related free-living taxa (Fig. 6). For example, the modules for C5 and C10–C20 isoprenoid biosynthesis and pentose phosphate pathway in the symbiotic order Borreliales within Spirochaetota are consistently more complete.

Nitrogen metabolism modules in symbiont-exclusive genera within the family Thioglobaceae (class Gammaproteobacteria) are more complete than in free-living relatives (Fig. 6), indicating potentially nitrogen-poor and sulfur-poor diets of their hosts or that the symbionts' retention in nitrogen acquisition modules likely alleviates their hosts' dietary deficiencies, providing a source of usable nitrogen. Likewise, we found that sulfur metabolic modules are retained in these symbiont clades, probably reflecting adaptation to anoxic, sulfate-rich gut environments. Most likely, the nitrogen and sulfur chemoautotrophic metabolism that is fully present enables the host to exploit inorganic energy sources in harsh environments like deep-sea vents. In the symbiotic genus *CAG-557* within the class Clostridia (Fig. 6), the dTDP-L-rhamnose biosynthesis module, often critical for the virulence of human pathogens[50], is consistently complete in contrast to other symbiont clades and free-living relatives.

https://doi.org/10.1038/s41587-026-03213-1

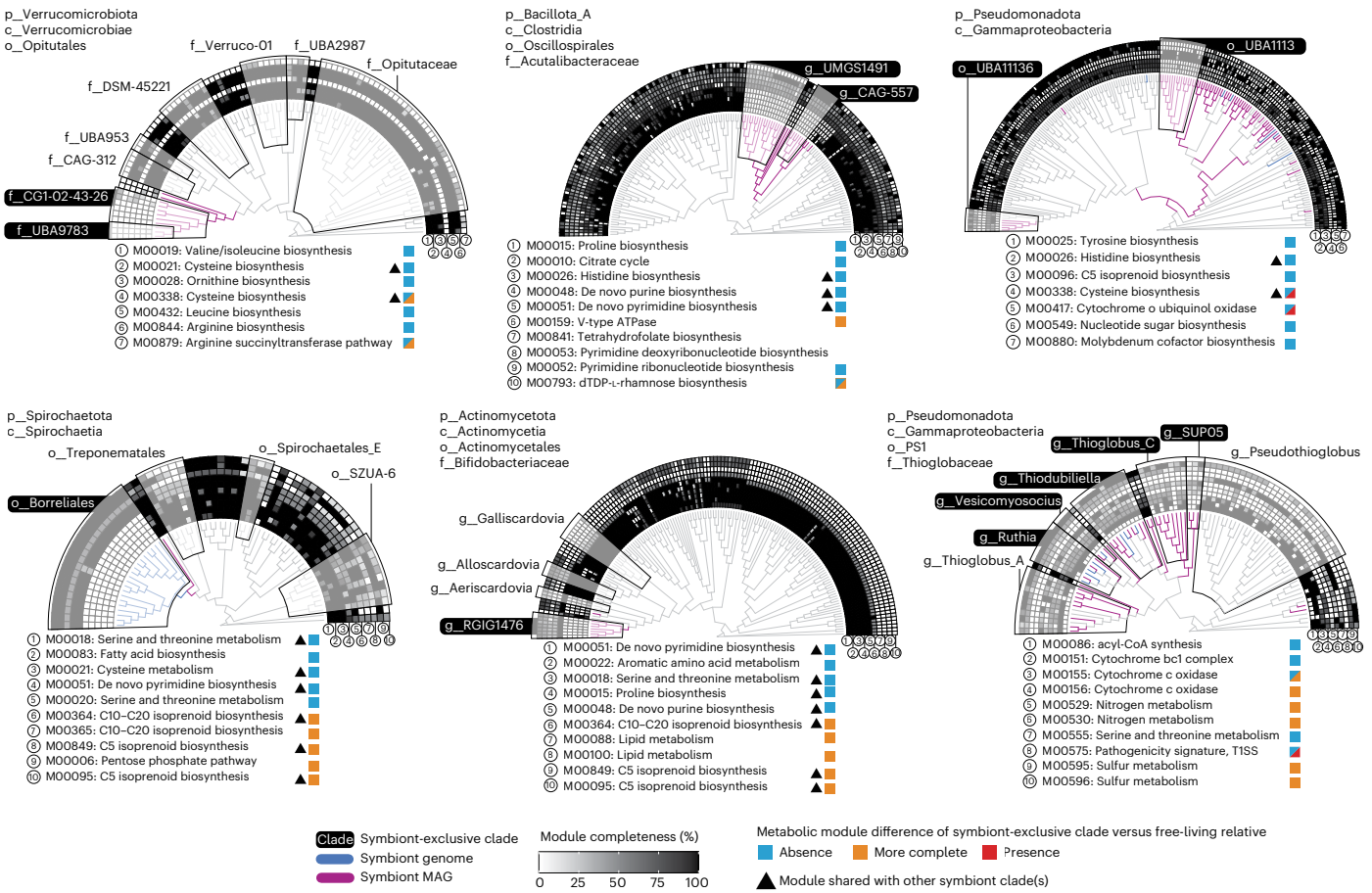


Fig. 6 | Differential metabolic module completeness between genomes of symbionts and their free-living sister clades. Phylogenomics trees showing clades identified as symbiont-exclusive alongside related taxa. Key metabolic modules were automatically selected using the bootstrapping approach detailed in Extended Data Fig. 5a. Modules that are shared between two clades are denoted by triangles. Module identification was based on the KEGG database. Each ring of cells around the tree represents a KEGG metabolic module, with fill color indicating the completeness or the presence or absence of that module in each genome.

Overall, we observe that symbionts may allow hosts to live in habitats or use resources that would otherwise be inaccessible, illustrating an adaptive advantage of maintaining certain metabolic modules. In general, the genes that remain in reduced symbiont genomes (Extended Data Fig. 5b) tend to be those directly involved in metabolic exchanges with the host, such as nutrient provisioning or waste recycling, highlighting the tight metabolic interdependence of host and symbiont.

Discussion

Our study leverages large-scale genome-resolved metagenomics and ML to explore the global diversity of microbial symbionts. By compiling and analyzing a comprehensive catalog of genomes primarily composed of MAGs, here, we show the extensive diversity and pervasiveness of bacterial and archaeal symbionts across Earth's biomes, an idea long hypothesized[16,51] but not documented until now.

We developed the `symclatron` software designed for the classification of genomes. One of the main criteria for `symclatron` development and optimization was to minimize the false-positive rate for symbiont prediction, which provides a conservative trustworthy ground for the host-associated and obligate intracellular predictions made here. Nevertheless, `symclatron` showed remarkable accuracy by correctly classifying symbiotic lifestyles in clades such as Bipolaricaulota and Omnitrophota, consistent with the recent literature[52,53]. Another example includes *Candidatus Azoamicus ciliaticola* (GTDB: f__UBA6186, o__UBA6186), a Gammaproteobacterial endosymbiont that uses denitrification to generate ATP for its ciliate host[8], which was the first described member of a much larger clade of protist endosymbionts[54]. In our analysis, we were able to identify 21 genomes from the order UBA6186 that are part of our SymGs catalog. `Symclatron` predicted all the symbionts within the UBA6186 order as intracellular (Supplementary Table 3). Similarly, the model correctly predicted recently described obligate intracellular endosymbionts of dinoflagellates, demonstrating robust generalization to novel lineages explicitly outside the main training bias. For example, the bacterial genomes RS3 and XS4 (endosymbionts of *Citharistes regius*) were accurately classified despite their closest relatives in the training data being free-living[55]. Furthermore, `symclatron` correctly identified *Candidatus Sukunaarchaeum mirabile*, a newly discovered deeply branching archaeal lineage recovered from a single sorted dinoflagellate, as an obligate intracellular symbiont[56].

Although the training data of `symclatron` were skewed toward bacterial symbionts of Eukaryotes and involved only a few archaeal genomes, the model correctly predicted several lineages of archaeal symbionts affiliated with DPANN lineages that have previously been characterized as episymbionts of other archaea. One such example is Nanoarchaeota[57,58], families of which associate obligately with other Archaea within the Thermoproteota. Many members of Thermoproteota were also identified as symbionts by `symclatron`; however, our classifier predicted the known Thermoproteota hosts *Ignicoccus hospitalis*[59] and Acd1 (ref. 60) to be correctly free-living. Furthermore, it correctly identified the symbiotic and free-living lifestyle of the Huberarchaeota–Altiarchaeota symbiosis[61]. Lastly, it predicted symbiont-exclusive clades for Micrarchaeota, which were recently

described as an example for Archaea–Archaea interactions[62]. `Symclatron`'s capability in predicting prokaryote–prokaryote associations is potentially supported further by the high prevalence of predicted symbionts in deep subsurface samples. This biome is known to harbor eukaryotic organisms such as fungi and protists but typically in very low abundance[63,64]. The proportion of predicted symbionts becomes more plausible when considering the likelihood of prokaryote–prokaryote associations in this environment. Similarly, within the phylum Patescibacteriota[19], our classifier predicted most (99% of genomes) as symbionts. Members of this lineage have previously been described as associated with other bacteria and archaea[65,66]. Additionally, similarly to DPANN archaea, they have been detected less frequently in samples from animal microbiomes[22]. However a small fraction of Patescibacteriota was predicted as free-living by our model, which suggests that, despite their limited metabolic capacities and likely dependence, some members of this phylum might not require other organisms to thrive, which would support recent findings[67,68].

We detected multiple other host-associated clades, such as in the case of the MAG-exclusive Bipolaricaulota (Fig. 3). Although MAGs from this phylum have been recently sequenced from coral hosts[52], it is important to note that they are phylogenetically distant from the clades used for feature extraction in our classifier. In agreement with recent results[53], our classifier also predicted both host-associated and obligate intracellular clades within the Omnitrophota. A detailed comparison of the ecological genomics of animal-associated and aquatic-enriched predicted symbiont lineages remains an important next step made possible by the SymGs catalog and the symclatron framework.

Because of the capability of microbial symbionts to be a key driver of complex life evolution by influencing the adaptive radiation of host lineages to a variety of environments, understanding symbiont diversity and pervasiveness is essential[69]. Some symbioses, especially the transition from free-living to obligate intracellular lifestyles, lead to an irreversible genome reduction and metabolic streamlining of the symbiont genome[31,70–73]. However, although it is clear symbiont genomes are reduced during coevolution with their host, there is evidence that gene gain may enable the initial symbiotic interaction that leads to prolonged association. For example, a recent discovery indicated gene gain[74] as an evolutionary mechanism in the transition to symbiosis for Chlamydiota. Our study also reveals a streamlining of certain biochemical networks in symbiont genomes, such as central metabolic pathways and biosynthesis of nucleotides and amino acids, presumably sustained by hosts. Yet, we also report here differential function presence in symbiont-exclusive clades compared to their free-living relatives across several pathways and from multiple clades (Fig. 6). One clear example is the increased completeness found in the metabolic module M00417, cytochrome *o* ubiquinol oxidase (part of the bacterial electron transport chain complex IV), in the symbiont-exclusive order UBA1113 within Gammaproteobacteria, linking our results to recent findings of gain of function by Chlamydiota in their transition to endosymbiotic lifestyle[74]. These findings were further extended through the detection of presence or more complete metabolic modules, in several symbiont-exclusive clades, involved in pathways related to cofactor, vitamin and ATP metabolism.

The reported differential presence of metabolic capacity challenges the traditional paradigm of genome streamlining in host-associated and intracellular bacteria; hence, we provide extensive evidence of evolution toward selective retention of metabolic capabilities in symbionts. These findings extend our understanding of the adaptive outcomes that genomes of symbionts underwent (or are currently undergoing) to sustain ecological complexity[75] and thrive in their specific hosts and their niches. Different mechanisms of gene transfer have been recently described in bacterial symbionts, such as gene transfer agents in endosymbionts within the Alphaproteobacteria[76], ancestral vertical inheritance accompanied by gene innovations in Rickettsiales[44] and gene gain in Chlamydiota[74]. The evolutionary mechanisms of metabolic retention and potential gain of functions found in symbiont lineages in our collection remain to be investigated.

We collectively demonstrate that microbial symbionts have evolved streamlined genomes that sacrifice broad metabolic independence in favor of a pared-down set of functions tailored to the host environment. Key biosynthetic pathways have been lost in multiple symbiont lineages, indicating a convergent evolutionary response to the host-provided niche. Conversely, symbionts preferentially retain or enhance pathways that fulfill host needs or exploit host habitat resources, underscoring a fine-tuned metabolic complementarity between host and symbiont (Fig. 6 and Extended Data Fig. 5). These findings highlight how intimate symbiotic associations drive metabolic specialization and codependence, ultimately enabling hosts to occupy ecological niches that would otherwise be unattainable.

Our work also contributes to the rapidly advancing field of engineered endosymbiosis, a paradigm with transformative potential for biotechnology. Applications range from regenerative medicine, where endosymbionts can act as 'biological remote controls' for cellular function[77], and immunotherapy using nonreplicating 'cyborg pathogens' (ref. 78) to sustainable bioproduction, where photosynthetic endosymbiotic systems convert CO_2 into valuable chemicals[79,80]. The success of a host–symbiont pairing is governed by complex and often unpredictable molecular interactions, as evidenced by the complex phenotypic outcomes of engineered auxotrophies for minimal genome construction of synthetic symbiosis[81]. The reliance on experimental screening creates an important bottleneck that limits the translation of promising concepts into functional technologies. The computational framework presented here, comprising the `symclatron` software and the SymGs genomic catalog, addresses this information bottleneck. By cataloging the global diversity of symbionts and their key genomic features derived from the molecular hallmarks of symbiosis, such as symbiont-specific and symbiont-exclusive clade metabolic capability, our work could aid in directing the engineering for scoped symbiont selection and bioenergetic exchange with hosts[82]. Our approach will potentially enable the future rational, hypothesis-driven design of novel host–symbiont pairs for biotechnological applications. This capacity to computationally screen and prioritize candidates could accelerate the design–build–test–learn cycle and lead the way for the efficient and scalable development of next-generation symbiotic systems tailored for specific and diverse biotech industries.

Our findings address a gap in cataloging and better understanding of the bacterial and archaeal symbionts and offer a global perspective on the ecological, evolutionary and biochemical implications of microbial symbiosis, which may assist future studies aimed at engineering artificially induced symbioses and discovering the novel functional roles and mechanisms of symbionts in understudied hosts from various difficult-to-access ecosystems.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-026-03213-1.

References

1. Douglas, A. E. Lessons from studying insect symbioses. *Cell Host Microbe* **10**, 359–367 (2011).
2. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
3. Douglas, A. E. Symbiosis as a general principle in eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* **6**, a016113 (2014).

4. Drew, G. C., Stevens, E. J. & King, K. C. Microbial evolution and transitions along the parasite-mutualist continuum. *Nat. Rev. Microbiol.* **19**, 623–638 (2021).
5. Bordenstein, S. R. & Theis, K. R. Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLoS Biol.* **13**, e1002226 (2015).
6. Kipp, M. A. et al. Nitrogen isotopes reveal independent origins of N_2-fixing symbiosis in extant cycad lineages. *Nat. Ecol. Evol.* **8**, 57–69 (2024).
7. Rädecker, N., Escrig, S., Spangenberg, J. E., Voolstra, C. R. & Meibom, A. Coupled carbon and nitrogen cycling regulates the cnidarian-algal symbiosis. *Nat. Commun.* **14**, 6948 (2023).
8. Graf, J. S. et al. Anaerobic endosymbiont generates energy for ciliate host by denitrification. *Nature* **591**, 445–450 (2021).
9. Hector, T. E., Hoang, K. L., Li, J. & King, K. C. Symbiosis and host responses to heating. *Trends Ecol. Evol.* **37**, 611–624 (2022).
10. Daisley, B. A., Chmiel, J. A., Pitek, A. P., Thompson, G. J. & Reid, G. Missing microbes in bees: how systematic depletion of key symbionts erodes immunity. *Trends Microbiol* **28**, 1010–1021 (2020).
11. Villada, J. C. & Schulz, F. The endosymbiotic box of protective tricks. *Nat. Rev. Microbiol.* **20**, 255 (2022).
12. Zhao, D., Zhang, Z., Niu, H. & Guo, H. Pathogens are an important driving force for the rapid spread of symbionts in an insect host. *Nat. Ecol. Evol.* **7**, 1667–1681 (2023).
13. Salje, J. Cells within cells: Rickettsiales and the obligate intracellular bacterial lifestyle. *Nat. Rev. Microbiol.* **19**, 375–390 (2021).
14. Perreau, J. & Moran, N. A. Genetic innovations in animal–microbe symbioses. *Nat. Rev. Genet.* **23**, 23–39 (2022).
15. Relman, D. A. Microbial genomics and infectious diseases. *N. Engl. J. Med.* **365**, 347–357 (2011).
16. Husnik, F. et al. Bacterial and archaeal symbioses with protists. *Curr. Biol.* **31**, R862–R877 (2021).
17. Delaux, P.-M. & Schornack, S. Plant evolution driven by interactions with symbiotic and pathogenic microbes. *Science* **371**, eaba6605 (2021).
18. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolution of microbial genomics: conceptual shifts over a quarter century. *Trends Microbiol.* **29**, 582–592 (2021).
19. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
20. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
21. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
22. Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
23. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
24. Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
25. Stairs, C. W. et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci. Adv.* **6**, eabb7258 (2020).
26. Eme, L. et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
27. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
28. Gurbich, T. A. et al. MGnify Genomes: a resource for biome-specific microbial genome catalogues. *J. Mol. Biol.* **435**, 168016 (2023).
29. Chen, I.-M. A. et al. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
30. Guerrero-Egido, G. et al. bacLIFE: a user-friendly computational workflow for genome analysis and prediction of lifestyle-associated genes in bacteria. *Nat. Commun.* **15**, 1–18 (2024).
31. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).
32. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
33. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
34. Tripp, E. A. et al. Reshaping Darwin's tree: impact of the symbiome. *Trends Ecol. Evol.* **32**, 552–555 (2017).
35. Lindemann, S. R. et al. Engineering microbial consortia for controllable outputs. *ISME J.* **10**, 2077–2084 (2016).
36. Moya, A., Peretó, J., Gil, R. & Latorre, A. Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat. Rev. Genet.* **9**, 218–229 (2008).
37. Hays, S. G., Patrick, W. G., Ziesack, M., Oxman, N. & Silver, P. A. Better together: engineering and application of microbial symbioses. *Curr. Opin. Biotechnol.* **36**, 40–49 (2015).
38. Kenny, D. J. & Balskus, E. P. Engineering chemical interactions in microbial communities. *Chem. Soc. Rev.* **47**, 1705–1729 (2018).
39. Nakajima, M. et al. *Minisyncoccus archaeiphilus* gen. nov., sp. nov., a mesophilic, obligate parasitic bacterium and proposal of Minisyncoccaceae fam. nov., Minisyncoccales ord. nov., Minisyncoccia class. nov. and Minisyncoccota phyl. nov. formerly referred to as Candidatus Patescibacteria or candidate phyla radiation. *Int. J. Syst. Evol. Microbiol.* **75**, 006668 (2025).
40. McCutcheon, J. P. & Moran, N. A. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol. Evol.* **2**, 708–718 (2010).
41. Vicente, C. S. L. et al. Genome analysis of new *Blattabacterium* spp., obligatory endosymbionts of *Periplaneta fuliginosa* and *P. japonica*. *PLoS ONE* **13**, e0200512 (2018).
42. Selberherr, E. et al. The life cycle-dependent transcriptional profile of the obligate intracellular amoeba symbiont *Amoebophilus asiaticus*. *FEMS Microbiol. Ecol.* **98**, fiac001 (2022).
43. Waters, J. L. & Ley, R. E. The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health. *BMC Biol.* **17**, 83 (2019).
44. Schön, M. E., Martijn, J., Vosseberg, J., Köstlbacher, S. & Ettema, T. J. G. The evolutionary origin of host association in the Rickettsiales. *Nat. Microbiol.* **7**, 1189–1199 (2022).
45. Castelli, M. et al. *Deianiraea*, an extracellular bacterium associated with the ciliate *Paramecium*, suggests an alternative scenario for the evolution of *Rickettsiales*. *ISME J.* **13**, 2280–2294 https://doi.org/10.1038/s41396-019-0433-9 (2019).
46. Castelli, M. et al. Host association and intracellularity evolved multiple times independently in the Rickettsiales. *Nat. Commun.* **15**, 1093 (2024).
47. Boscaro, V. et al. Parallel genome reduction in symbionts descended from closely related free-living bacteria. *Nat. Ecol. Evol.* **1**, 1160–1167 (2017).
48. Nechitaylo, T. Y. et al. Incipient genome erosion and metabolic streamlining for antibiotic production in a defensive symbiont. *Proc. Natl Acad. Sci. USA* **118**, e2023047118 (2021).
49. Osvatic, J. T. et al. Gene loss and symbiont switching during adaptation to the deep sea in a globally distributed symbiosis. *ISME J.* **17**, 453–466 (2023).

50. Mistou, M.-Y., Sutcliffe, I. C. & van Sorge, N. M. Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria. *FEMS Microbiol. Rev.* **40**, 464–479 (2016).
51. Margulis, L. *Symbiotic Planet: A New Look At Evolution* (Basic Books, 2008).
52. Tandon, K. et al. Genomic view of the diversity and functional role of archaea and bacteria in the skeleton of the reef-building corals *Porites lutea* and *Isopora palifera*. *GigaScience* **12**, giac127 (2022).
53. Seymour, C. O. et al. Hyperactive nanobacteria with host-dependent traits pervade Omnitrophota. *Nat. Microbiol.* **8**, 727–744 (2023).
54. Speth, D. R. et al. Genetic potential for aerobic respiration and denitrification in globally distributed respiratory endosymbionts. *Nat. Commun.* **15**, 9682 (2024).
55. Nakayama, T. et al. Marked genome reduction driven by a parasitic lifestyle: two complete genomes of endosymbiotic bacteria possibly hosted by a dinoflagellate. *Microbes Environ.* **40**, ME25005 (2025).
56. Harada, R. et al. A cellular entity retaining only its replicative core: Hidden archaeal lineage with an ultra-reduced genome. Preprint at *bioRxiv* https://doi.org/10.1101/2025.05.02.651781 (2025).
57. Wurch, L. et al. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* **7**, 12115 (2016).
58. Jarett, J. K. et al. Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. *Microbiome* **6**, 161 (2018).
59. Paper, W. et al. *Ignicoccus hospitalis* sp. nov., the host of '*Nanoarchaeum equitans*'. *Int. J. Syst. Evol. Microbiol.* **57**, 803–808 (2007).
60. Podar, M. et al. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* **8**, 9 (2013).
61. Schwank, K. et al. An archaeal symbiont-host association from the deep terrestrial subsurface. *ISME J.* **13**, 2135–2139 (2019).
62. Krause, S. et al. The importance of biofilm formation for cultivation of a Micrarchaeon and its interactions with its Thermoplasmatales host. *Nat. Commun.* **13**, 1735 (2022).
63. Borgonie, G. et al. Eukaryotic opportunists dominate the deep-subsurface biosphere in South Africa. *Nat. Commun.* **6**, 8952 (2015).
64. Bhattacharjee, A. S., Schulz, F., Woyke, T., Orcutt, B. N. & Martínez Martínez, J. Genomics discovery of giant fungal viruses from subsurface oceanic crustal fluids. *ISME Commun.* **3**, 10 (2023).
65. Utter, D. R., He, X., Cavanaugh, C. M., McLean, J. S. & Bor, B. The saccharibacterium TM7x elicits differential responses across its host range. *ISME J.* **14**, 3054–3067 (2020).
66. Kuroda, K. et al. Microscopic and metatranscriptomic analyses revealed unique cross-domain parasitism between phylum Candidatus Patescibacteria/candidate phyla radiation and methanogenic archaea in anaerobic ecosystems. *MBio* **15**, e0310223 (2024).
67. Beam, J. P. et al. Ancestral absence of electron transport chains in Patescibacteria and DPANN. *Front. Microbiol.* **11**, 1848 (2020).
68. Chaudhari, N. M. et al. The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. *Environ. Microbiome* **16**, 24 (2021).
69. McCutcheon, J. P. The genomics and cell biology of host-beneficial intracellular infections. *Annu. Rev. Cell Dev. Biol.* **37**, 115–142 (2021).
70. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
71. Wernegreen, J. J. For better or worse: genomic consequences of intracellular mutualism and parasitism. *Curr. Opin. Genet. Dev.* **15**, 572–583 (2005).
72. Toft, C. & Andersson, S. G. E. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* **11**, 465–475 (2010).
73. Bennett, G. M. & Moran, N. A. Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. *Proc. Natl Acad. Sci. USA* **112**, 10169–10176 (2015).
74. Dharamshi, J. E. et al. Gene gain facilitated endosymbiotic evolution of Chlamydiae. *Nat Microbiol* **8**, 40–54 (2023).
75. Muñoz-Gómez, S. A., Kreutz, M. & Hess, S. A microbial eukaryote with a unique combination of purple bacteria and green algae as endosymbionts. *Sci. Adv.* **7**, eabg4102 (2021).
76. George, E. E. et al. Gene transfer agents in bacterial endosymbionts of microbial eukaryotes. *Genome Biol. Evol.* **14**, evac099 (2022).
77. Contag, C. H. Engineered endosymbionts for cellular control. *Open Access Government* https://doi.org/10.56367/OAG-041-11213 (2024).
78. Khan, S., Lin, P.-R. & Tan, C. Engineering Cyborg pathogens through intracellular hydrogelation. *ACS Synth. Biol.* **13**, 3609–3620 (2024).
79. Gao, Y.-L. et al. Introducing carbon assimilation in yeasts using photosynthetic directed endosymbiosis. *Nat. Commun.* **15**, 5947 (2024).
80. Chandra De, B. et al. Photosynthetic directed endosymbiosis to investigate the role of bioenergetics in chloroplast function and evolution. *Nat. Commun.* **15**, 1–14 (2024).
81. Mehta, A. P. et al. Toward a synthetic yeast endosymbiont with a minimal genome. *J. Am. Chem. Soc.* **141**, 13799–13802 (2019).
82. Cournoyer, J. E., De, B. C. & Mehta, A. P. Molecular and biochemical insights from natural and engineered photosynthetic endosymbiotic systems. *Curr. Opin. Chem. Biol.* **87**, 102598 (2025).

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

https://doi.org/10.1038/s41587-026-03213-1

SymGs data consortium

Ramunas Stepanauskas[5], Christopher A. Francis[6], Rick Cavicchioli[7], Katherine Trina McMahon[8], Steven Hallam[9], Wen-Tso Liu[10], Thomas Mock[11], James Tiedje[12] & Jiarong Guo[13]

[5]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA. [6]Earth System Science and Oceans Departments, Stanford Doerr School of Sustainability, Stanford University, Stanford, CA, USA. [7]School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney, New South Wales, Australia. [8]Departments of Civil and Environmental Engineering and Bacteriology, University of Wisconsin-Madison, Madison, WI, USA. [9]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada. [10]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [11]School of Environmental Sciences, University of East Anglia (UEA), Norwich Research Park, Norwich, UK. [12]Center for Microbial Ecology, Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA. [13]Center of Microbiome Science, The Ohio State University, Columbus, OH, USA.

Methods

Definition of symbionts, host-associated and obligate intracellular

In this study, we adopted the definitions recently delineated by Husnik et al.[16] to characterize symbiotic relationships. We define a symbiont as an organism that maintains a spatially close and temporally prolonged association with a host organism of a different species. Symbionts can be broadly categorized as host-associated or, more specifically, as obligate intracellular. Host-associated symbionts encompass both facultative endosymbionts, which can replicate inside a host cell but also independently of a host, and ectosymbionts (or epibionts), which inhabit the exterior of their host, often attached to the host's surface. In contrast, obligate intracellular symbionts represent a subset of host-associated symbionts that thrive and replicate exclusively inside a host cell. To better reflect the continuum of symbiosis and the blurred boundaries of these definitions, we use the term obligate intracellular as a more specific and exclusive categorization of symbionts that replicate exclusively within host cells. A prediction of host-associated could also indicate that a microorganism can facultatively thrive intracellularly but can also thrive outside of host cells. According to our definition, obligate intracellular symbionts can include both mutualists and pathogens, provided that replication is restricted to host cells and no established host-independent replication is known.

Our labeling strategy is designed to represent a continuum of host dependence, which is reflected in the genomic features our model uses for prediction. The three main categories are free-living, host-associated symbiont and obligate intracellular symbiont. We define free-living organisms as those capable of replicating independently of a host. The term symbiont is used here in its broad biological sense, referring to any organism engaged in a specific and persistent association with another, regardless of whether the relationship is neutral, beneficial or detrimental to the host. Our labeling approach allows our ML model to focus on the convergent genomic signatures of host dependence, such as genome reduction and metabolic specialization. The distinction between host-associated and obligate intracellular symbiont labels reflects the evolutionary trajectory of the host-dependence continuum. The symbiont categories are hierarchical, not necessarily mutually exclusive. Host-associated symbiont is a broad classification that includes any organism predicted to have a host-dependent lifestyle. The obligate intracellular symbiont label is a more specific subset of host-associated symbionts, reserved for organisms showing strong genomic evidence of being unable to survive outside of a host cell. Consequently, all organisms classified as obligate intracellular symbiont are also, by evolutionary trajectory, already associated with a host and are, therefore, part of the host-associated symbiont category. This nested approach allows us to conservatively classify organisms with facultative intracellular stages as host-associated symbionts, as their genomes must retain the functions for both extracellular and facultative intracellular survival. This ensures that the obligate intracellular symbiont category remains specific to organisms with the highest degree of host dependence. In consequence, in our training data, we assigned higher scores to known obligate intracellular symbionts (label = 2, obligate intracellular) and lower scores to known facultative endosymbionts and ectosymbionts (label = 1, host-associated). Because the training set for label 2 includes only well-characterized obligate intracellular symbionts, predicted obligate intracellular calls should be interpreted conservatively; we cannot rule out that some obligate ectosymbionts or epibionts with similar extreme host-dependence signatures may also be assigned to this category.

Syntrophic relationships, in which two organisms cooperate to metabolize a compound that neither could break down alone, may fall outside of the current definition of symbiosis if the partners interact only through environmental metabolite exchange, without a specific physical association or long-term coadaptation[83]. High-impact theoretical frameworks emphasize that true symbiosis involves specific host–microorganism associations maintained across generations and shaped by selection[84,85]. In contrast, broader ecological interactions such as syntrophy, while conferring functional advantages, may be better conceptualized as community-level cooperation within a shared environment rather than symbiosis, when the partners are linked only by metabolic exchange[86,87].

In the current scope of the `symclatron` project, host refers to the partner organism in a sustained, spatially close symbiotic association and may be either eukaryotic or prokaryotic; in contrast, we do not classify generic syntrophic or diffusible cross-feeding interactions in microbial communities as symbiosis unless a stable host-associated relationship is known.

Expert workflow for manual lifestyle curation

Lifestyle labels for the 6,751 training genomes were assigned by manual expert curation rather than by automated database annotations. IMG/M was used only as a source of genome sequences and assembly metadata; lifestyle labels were not transferred from IMG/M, GTDB, National Center for Biotechnology Information or any other automated metadata field to avoid circularity and to prevent propagation of inconsistent or erroneous annotations. The curation workflow used a divide-and-conquer strategy in which genomes were partitioned across experts according to their taxonomic and biological familiarity and then proceeded through a fixed sequence of decisions. In step 1, one or two genomes per genus were selected from IMG/M to reduce redundancy while maintaining broad phylogenetic coverage of both bacteria and archaea. In step 2, each genome was assigned to a curator with relevant expertise in microbial symbiosis, who first relied on their own archives of organism-specific publications, experiments and long-term familiarity with the focal symbiotic systems and consulted the primary literature when needed to resolve cases that were not already clear from that expertise. In step 3, because organism names, strain names and higher-rank assignments are often inconsistent across the literature and genome repositories, curators reconciled taxonomic synonyms and nomenclatural changes to ensure that published biological evidence corresponded to the correct genome lineage under current taxonomy. In step 4, labels were assigned according to explicit biological rules; free-living was used for organisms known to replicate independently of a host, host-associated was used for organisms with persistent association with a host, including ectosymbionts, epibionts, gut-associated symbionts and facultative intracellular organisms, when evidence for exclusive intracellular replication was lacking and obligate intracellular was reserved for taxa for which replication only occurs within host cells and no established host-independent replication. In step 5, ambiguous cases were handled conservatively; when the available evidence did not justify a confident obligate intracellular label, the genome was assigned to the broader host-associated category rather than overcalled and generic syntrophy, diffusible cross-feeding or community-level metabolic cooperation without documented stable host association was not considered sufficient to assign a symbiont label. In step 6, when a label could not be resolved confidently by the initial curator, the case was reassigned to a more senior expert or discussed with a senior subject-matter expert with deeper taxonomic or biological familiarity with the lineage. In step 7, labels were frozen before feature extraction, model training and benchmarking, such that no genome-derived prediction, functional annotation summary or model output was used to decide the training labels. This workflow was designed for genome-level biological adjudication rather than for formal systematic-review screening and it is better suited to a literature in which taxonomy, nomenclature and lifestyle terminology are often unstable and in which the evidence relevant to a genome's lifestyle may be distributed across organism descriptions, ecological reports, host biology studies and experimental literature rather than captured in a single standardized annotation resource.

AI/ML-based classification of microbial lifestyles

To construct the reference datasets, we first assembled a manually curated set of 792 symbiont proteomes spanning the major clades shown in Fig. 1a. These proteomes were selected by expert review and were used only for orthogroup inference and HMM construction. The orthogroup and HMM feature space was constructed from curated symbiont proteomes to enrich for symbiosis-associated signal while maintaining computational tractability and software portability; extending orthogroup inference to the full diversity of free-living genomes would greatly expand feature-space size and HMM-scanning cost, with limited expected benefit for the already well-resolved free-living class.

Independently, to build the labeled training set, we downloaded one or two genomes per genus from IMG/M (6,751 genomes total) and then assigned each genome de novo to one of three lifestyle categories by manual expert curation rather than by automated database annotations. This curation was performed by a group of scientists with extensive expertise in microbial symbiosis and was based on established biological knowledge from the primary literature. Obligate intracellular was reserved for taxa for which the literature supports replication only within host cells, whereas host-associated was used for persistent host-dependent taxa, including ectosymbionts, epibionts, gut-associated symbionts and facultative intracellular organisms, for which exclusive intracellular replication was not established. Free-living was reserved for organisms known to replicate independently of a host. When evidence for strict intracellular dependence was incomplete or ambiguous, we used the more conservative host-associated label.

We next developed an ML classifier to systematically classify microbial genomes on the basis of their lifestyles. First, all the proteins (n = 1,642,823) found in known genomes of symbionts within the clades detailed in Fig. 1a were subjected to `OrthoFinder` (version 2.5.4)[88] to detect orthogroups of proteins across the different microbial clades of symbionts (Fig. 1a). Only protein sequences from orthogroups with at least five sequences were retained. Sequence alignment and trimming were performed using `MAFFT` (version 7.481) and `trimAl` (version 1.4), respectively, to prepare for the construction of HMMs. A total of 20,063 HMMs were generated to represent the protein orthogroups (Fig. 1a).

To construct a database of labeled genomes according to lifestyles (Fig. 1b), we downloaded one (or two when available) genomes from each bacterial genus in the IMG/M (accessed in October 2022). Because currently available reference genomes are biased toward better-characterized symbiotic systems, especially bacterial symbionts of eukaryotes, we reduced redundancy by sampling at one or two genomes per genus and evaluated generalization explicitly by leave-one-clade-out benchmarking across multiple taxonomic ranks. The genomes (n = 6,751; Supplementary Table 1) were then labeled (free-living, n = 5,959; host-associated, n = 409, intracellular, n = 383) by scrutinizing for lifestyle categories using manual literature searches.

Next, every labeled genome was artificially fragmented to emulate various levels of genome completeness (50%, 60%, 70%, 80% and 90%) and fragmentation lengths (10 kb, 30 kb and 50 kb) with `shred.sh` from `BBMap` (version 39.01)[89]. Estimated completeness below 50% was treated as a conservative warning flag for reduced confidence rather than as an automatic exclusion or relabeling criterion; this lower bound is aligned with the minimum completeness floor commonly associated with the minimum information about an MAG[32] medium-quality MAG practice, whereas the UNI56 marker score used by `symclatron` serves only as a rapid internal completeness proxy and not as a substitute for full genome quality control, which remains the responsibility of the user. Gene prediction on these fragmented genomes was conducted using gene calling with `Prodigal` (version 2.6.3) to identify proteins after the fragmentation iterations. The predicted proteins from labeled genomes were then searched against the 20,063 HMMs generated from the symbiont orthogroups using `hmmsearch` to obtain bitscores. To keep consistent and reproducible results, all `hmmsearch` runs were with the parameters '`-E 1000`, `--incE 1000`, and `-Z 61295632`' to guarantee that the entire set of results were output independently of potential E-value biases generated by database size. This parameter was chosen intentionally with the goal of maximizing sensitivity. We aimed to capture all potential homologs, including those with only remote similarity, ensuring that potentially relevant but divergent features were not missed. While a high E-value threshold allows many hits, the statistical significance and strength of each hit are still captured by the HMMER bitscore. These E-value parameters were chosen to maximize sensitivity and capture even more remote homologs, allowing the ML step to interpret these features in context and produce reliable predictions. Finally, the 20,063 HMMs deriving from the symbiont genomes were used to search against the proteins of labeled genomes (Fig. 1b) with `hmmsearch` (version 3.3.2)[90].

For the AI/ML component (Fig. 1d), we constructed a training matrix with 20,063 features (from the HMMs), encapsulating the bitscores of every genome and its simulated genomes from the labeled dataset. These features were fed into two gradient-boosted decision tree models. A classifier (named symcla for symbiont classifier) and a regressor (named symreg for symbiont regressor), using the GPU version of `XGBoost` (version 1.7.4)[91] for `Python`. For computational efficiency, we used computing nodes with NVIDIA A100 (40 GB) GPUs on the Perlmutter supercomputer of the National Energy Research Scientific Computing Center (NERSC; https://www.nersc.gov, US Department of Energy (DOE), Office of Science). We did not perform an exhaustive hyperparameter search; instead, we used fixed `XGBoost` settings selected from preliminary training behavior with learning rate = 0.1 and 1,000 estimators across final training and clade-held-out benchmarking, while the neural network was trained on a small fixed architecture, with class weighting and early stopping.

The two models predict where a genome falls along the symbiosis continuum (Fig. 1c), from free-living to host-associated and obligate intracellular lifestyles, encoded as 0, 1 and 2, respectively. Symcla predicts the lifestyle according to the probability assigned to each class, whereas symreg assigns a continuous regression score in the range from 0.0 to 2.0 along the same continuum. The neural network received the full three-class softmax output of symcla as input; although one probability is formally determined by the other two, retaining the complete vector avoids arbitrarily omitting one class and is inconsequential in this small seven-feature model.

The neural-network-based classifier was then built by first extracting a set of numeric features from each genome and its classifications by the classifier and the regressor (for example, predicted scores from regression and classification models, completeness and distance to training data measures). These features were then combined into a feature matrix and the corresponding labels were one-hot-encoded. The dataset was split into training and validation subsets at an 80:20 ratio and a StandardScaler was fitted on the training data to normalize each feature's distribution. This step helped to stabilize the subsequent model training process. Next, a feedforward neural network architecture was defined. It consisted of an input layer feeding into three dense layers with ReLU activation functions, each followed by BatchNormalization (to stabilize gradients) and Dropout (to reduce overfitting). The output layer was a three-neuron softmax layer, producing probability estimates for the three lifestyle classes. To address potential class imbalance, class weights were calculated and applied during training. The model was optimized using Adam (learning rate = 0.001) with categorical cross-entropy loss and training was monitored with early stopping based on validation loss.

The final softmax probability from the neural network serves as an initial confidence score for each prediction. For deployment, this value was conservatively downweighted when predictions were made for genomes that were clearly more distant from the labeled training set or were incomplete. Specifically, we applied a 0.8 penalty factor when the weighted minimum Euclidean distance exceeded 80 and an additional 0.8 penalty factor when estimated completeness was below 50. These heuristic adjustments were used only to temper the reported

confidence score and did not alter the predicted lifestyle class. The final confidence score was calculated by multiplying the base confidence by the penalty factor when applicable, resulting in a metric ranging from 0 to 1 that integrated both statistical certainty and biological context. The final AI/ML framework composed of the classifier, the regressor, the extraction of genome features and the neural network was named `symclatron`.

To avoid information leakage between model stages, we separated deployment training from benchmark data generation. All phylogenomic benchmark outputs used for performance evaluation and for neural network training were generated independently by leave-one-clade-out retraining. Specifically, at each taxonomic rank from phylum to species, one clade was removed in its entirety, a new symreg or symcla model was initialized and fitted from scratch on the remaining clades; predictions were recorded only for the held-out genomes. Across 11,025 unique clades, this produced individual out-of-fold predictions for symreg and symcla. The neural network did not receive the 20,063 ortholog features or any fitted tree parameters; instead, it used only seven scalar inputs derived from these out-of-fold results (the symreg score, the three symcla class probabilities, genome completeness and the two distance-to-training-set features). The distance features were computed after removing the held-out clade from the reference set. For neural network model selection, the integrated out-of-fold table was split 80:20 in a stratified manner.

The culmination of the AI/ML framework efforts was the development of a novel software tool, `symclatron` (https://github.com/NeLLi-team/symclatron). In the `symclatron` software, we developed and included one module for rapid estimation of genome completeness through the detection of UNI56 marker genes as a close but computationally efficient approximation to `CheckM2` completeness estimates (Extended Data Fig. 6).

Systematic cross-validation and calibration of final predictions

To ensure that our predictive model generalizes well across diverse microbial clades, we meticulously cross-validated predictions accounting for taxonomic diversity of the labeled genomes using the classifier and the regressor. We adopted a stringent strategy where clades were entirely omitted from the training dataset to evaluate each model's performance on unseen data. Each of the training and testing matrices includes an iterative process of removing entire clades corresponding to different taxonomic levels, ranging from phylum to species levels (n = 11,025, one per clade found). For example, when benchmarking at the phylum level, we removed the entire set of genomes belonging to the phylum Pseudomonadota from the training dataset; we trained on all other phyla and used all Pseudomonadota only in the testing dataset. The same was applied for all clades at every taxonomic rank. This validation technique ensures that each model's predictive power is not artificially inflated by overrepresented taxonomic groups. Each training iteration of the classifier used a learning rate of 0.1 and 1,000 estimators. The aggregated outputs of each of the 11,025 cross-validation exercises were used as input to the neural network for final classification calibration and confidence estimation (Fig. 1e).

For distance-dependent generalization analysis, we restricted the integrated leave-one-clade-out benchmark to the unshredded genome evaluations to isolate phylogenetic novelty from completeness and fragmentation effects. We defined a conservative combined weighted distance to the remaining training set as the maximum of the symreg and symcla nearest-neighbor-weighted Euclidean distances. Performance was first summarized across genome-weighted quintiles (quintiles containing equal numbers of genome evaluations) of this combined distance using accuracy, macro-F_1 and class-specific recall. To control for potential overrepresentation of many closely related held-out clades, we then summarized each unique held-out clade by its median combined distance and its clade-level accuracy and macro-F_1 and grouped clades into quintiles of median distance for an equal-clade-weight analysis.

Building a nonredundant genome catalog of Earth's microbiomes and reference genomes

Metagenomic bins were obtained through metagenome binning and gene calling on assembled contigs from 31,152 public environmental sequencing data projects through the IMG/M system up to April 2023. Only contigs with length equal to or greater than 5 kbp were processed. Metagenome samples without associated depth data were removed. Binning was performed on each sample with `MetaBAT2` (version 2.15; parameters: '--minContig 5000 --minClsSize 20000 --cvExt')[92], yielding 5,635,662 bins recovered from metagenomes, cell enrichments and SAGs (Fig. 2a), from which 136,784 passed the quality criteria of `CheckM2` (version 1.0.2)[33] estimates for completeness equal or higher than 50% and contamination lower than 5%. Gene calling was performed on each bin with `Prodigal` (version 2.6.3) with the '`-p meta`' parameter[93].

Adding to our catalog, reference genomes and MAGs available through publicly available repositories were downloaded from the IMG isolates repository (n = 115,969)[29] and from the bacterial and archaeal representative datasets of GTDB (n = 85,205)[27]. All genomes and MAGs were subsequently subjected to `CheckM` (version 1.1.3) with the UNI56 markers for detection of universal markers required for further phylogenomics reconstruction, `CheckM2` (version 1.0.2) for quality assessment and `seqkit` (version 2.5.1)[94] for assembly characterization. All genomes and bins with less than 50% UNI56 markers or `CheckM2` completeness lower than 50% or `CheckM2` contamination higher than 5% were removed from the genome catalog. All MAGs included in NeLLi2023 and SymGs passed quality control filtering based on `CheckM2` and UNI56 marker completeness estimates, thereby restricting the catalog to cellular genomes; consequently, phage-associated domains detected in feature analyses likely reflect genes encoded within cellular genomes, including integrated prophage functions, rather than free viral bins.

Cluster representatives were determined on the basis of the highest assembly quality and dereplication through `skani` (version 0.1.4, parameters: `ANI` ≥ 95% `align_fraction_ref` ≥ 50% and `align_fraction_query` ≥ 50%)[95]. A single metric for assembly quality Q was computed as `Q = (0.5 × completeness_checkm2) + [0.2 × (100 − contamination_checkm2)] + (0.3 × completeness_UNI56)` to quantitatively represent the quality criteria applied. Then, the cluster representatives were selected as those with the highest Q and, if ties were found, the assemblies with the highest number of genes predicted and the highest N50 were selected. The dereplicated high-quality genome catalog was finally composed of 107,067 genomes and bins; these were further taxonomically characterized with `GTDB-Tk` (version 2.4.0)[96] from the GTDB Release 220.

To assess the taxonomic representativeness of the labeled training set, we parsed `GTDB-Tk` lineages for the 6,751 manually labeled training genomes, the full 107,067-genome NeLLi2023 catalog and the 14,070 high-confidence SymGs subset. We then compared unique clade counts across domain-to-species ranks and summarized relative phylum composition across the three datasets. Similarity in phylum composition was quantified using Spearman rank correlation and Jensen–Shannon distance.

Feature contribution to microbial lifestyle prediction

The feature contributions to the model's predictions were analyzed using the SHAP methodology, implemented within the `shap Python` library. SHAP values, which provide a breakdown of the impact of each feature on the model's output, were computed for the symcla and symreg predictions. This approach unveils the impact of individual genomic features on the predictive model's output, providing an interpretable mechanistic basis for the classification of microbial lifestyles by decomposing the symcla and symreg predictions into the

sum of effects from each feature, allowing us to quantify the impact of individual genomic attributes. We computed the SHAP values for all the features of each model on each one of the predicted symbionts. Next, all proteins in the 1,000 most relevant features were functionally annotated using `EggNOG-mapper` (version 2.1.10)[97], effectively assigning each feature to Kyoto Encyclopedia of Genes and Genomes (KEGG) ortholog groups (KOs)[98], clusters of orthologous groups[99] and Pfam (protein families)[100], allowing for the biological interpretation of the features and aligning the ML insights with known biological functions. GO terms were mapped from Pfam using `pfam2go` (https://current.geneontology.org/ontology/external2go/pfam2go).

For `symcla`, which predicts three class probabilities, we generated a model-wide feature-attribution matrix by retaining, for each genome–feature pair, the SHAP value of the class probability with the largest absolute contribution after orienting its sign toward symbiotic classification; contributions selected from the free-living class were sign-inverted, whereas contributions selected from the host-associated or obligate intracellular classes retained their sign. Functional summaries were obtained by mapping dominant Pfam annotations to GO terms and aggregating quadrant-specific mean SHAP values across features.

Our framework uses a quadrant-based approach to interpret the contributions of genomic features toward the classification of microbial symbiotic lifestyles. As shown in Extended Data Fig. 3a and Extended Data Fig. 4a, this framework delineates features into four quadrants, each representing a unique combination of low or high feature detection and negative or positive SHAP contribution. Positive SHAP values increase support for a symbiotic lifestyle, whereas negative values decrease it and are therefore more indicative of a free-living lifestyle.

Quadrant 1 (Q1): features with positive SHAP values when feature detection is low (`hmmsearch` bitscore < 20), more indicative of symbiotic lifestyles.

Quadrant 2 (Q2): features with positive SHAP values when feature detection is high (`hmmsearch` bitscore ≥ 20), more indicative of symbiotic lifestyles.

Quadrant 3 (Q3): features with negative SHAP values when feature detection is low (`hmmsearch` bitscore < 20), more indicative of free-living lifestyles.

Quadrant 4 (Q4): features with negative SHAP values when feature detection is high (`hmmsearch` bitscore ≥ 20), more indicative of free-living lifestyles.

The bitscore threshold of ≥20, used to determine the potential presence of a given domain within a protein sequence, was established on the basis of the distribution of the gathering threshold for sequence and domain used by the InterPro Pfam HMM profiles (for example, https://www.ebi.ac.uk/interpro/entry/pfam/PF00010/logo/).

To quantify the relative contribution of the seven neural network inputs, we performed a contribution analysis on the held-out 20% evaluation split of the integrated benchmark dataset used for neural network validation. We first compared the performance of symcla alone with that of the full saved `symclatron` neural network, treating the symcla prediction as the argmax of the three class probabilities. We then assessed feature group importance by grouped permutation analysis of the saved neural network, measuring the mean decrease in macro-F_1, host-associated recall and intracellular recall after permuting each feature group across genomes. Feature groups were defined as the symcla probability triplet, the symreg score, the two distance-to-training-set features and genome completeness.

Phylogenomics of the bacterial and archaeal symbionts

We inferred phylogenomic trees using `NSGTree` from concatenated protein sequences from at least half of the UNI56 universal marker genes. In `NSGTree`, UNI56 markers are identified with `hmmsearch` (version 3.3.2; parameter: '`--cut_ga`'). Next, we removed genomes if more than 50% of the markers were absent, if at least one marker was present in more than four copies or if more than 30% of the markers were present in multiple copies. Sequences were aligned using `MAFFT` (version 7.310) and trimmed with `trimAL` (version 1.4.1) using a mild trimming to minimize the loss of potentially informative sites (parameter: '`-gt 0.1`'). Concatenated protein sequences were then used to reconstruct a species tree using `IQ-TREE` (version 2.0.3) with the LG+C60+R8+F model. Trees were visualized with the `ggtree R` package[101].

Definition of symbiont-exclusive clades

We designated a clade as symbiont-exclusive if at least 90% of its members were finally classified as symbionts, the clade included a minimum of three genomes and the mean confidence of the clade was at least 0.725 with s.d. of the confidence below 0.2. This designation was made using the `symclatron` final classification. We considered clades at all taxonomic levels, as determined by `GTDB-Tk`[96].

When symbiont-exclusive clades were compared against their free-living relatives ('sister clades'), as in the case of the metabolic module completeness analyses, free-living clades were only considered if they included at least five genomes and if clade members, at the same or higher taxonomic levels, had predominantly free-living lifestyle predictions.

Comparative analysis of metabolic module completeness

We assessed the completeness of metabolic modules of symbiont-exclusive clades and their free-living relatives on the basis of the metabolic module definitions of KEGG[98]. The KO functional annotations of every genome were extracted from the results of `EggNOG-mapper` (version 2.1.10)[97]; when more than one KO was assigned to a protein, all the KOs were unnested and considered individual KOs for downstream analyses. The list of KOs per genome was then passed to the script `ko_mapper.py` from `MicrobeAnnotator` (version 2.0.5)[102] to map KOs to their respective modules and compute metabolic module completeness per genome.

We compared the completeness of KEGG metabolic modules between predicted symbiont-exclusive clades (that is, clades with ≥90% symbionts and ≥3 members) and their predicted free-living sister lineages by firstly collecting all available free-living relatives at the same phylogenetic rank (the sister clade). If there were at least five free-living members, 100 bootstrap samples of the same metabolic module were drawn for the symbiont and the free-living groups. Then, for each metabolic module, the average completeness of the bootstrap samples from symbionts was then compared against the average completeness of the bootstrap samples from free-living relatives, yielding a distribution of completeness differences. From this distribution, the mean Δ completeness and 95% confidence intervals were computed for each module.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All FASTA files of genomes included in the SymGs catalog are hosted in its NERSC repository for easy bulk download (https://portal.nersc.gov/cfs/nelli/symgs/) and archived on Zenodo (https://doi.org/10.5281/zenodo.20399071)[103]. All files are publicly available and immediately accessible for download. The `symclatron` database is hosted on GitHub (https://github.com/NeLLi-team/symclatron/releases/tag/db-latest).

We did not generate new sequencing data for this study. All environmental DNA sequencing samples analyzed were obtained from the IMG/M database. Direct download links to the Joint Genome Institute (JGI) Data Portal for each assembly project are provided through a dedicated Zenodo archive (https://doi.org/10.5281/zenodo.20600438)[104]. All the assemblies of microbiome samples, from which we predicted

symbionts, are individually cited in Supplementary Table 4. The JGI is a DOE Office of Science User Facility located at Lawrence Berkeley National Laboratory (Berkeley Lab). To accommodate the massive scale of these datasets, the JGI Data Portal offers highly flexible and scalable retrieval methods. Users can download data manually through a web browser, transfer large-scale datasets efficiently using Globus endpoints or command line tools or access the data programmatically through the JGI API. Data providers were individually contacted in accordance with the JGI Data Policy at least 1 month before manuscript submission. The complete set of symclatron predictions and associated metadata is available in Supplementary Table 5.

Code availability

The code repository for `NSGTree` is publicly available on GitHub (https://github.com/NeLLi-team/nsgtree). The `symclatron` software, along with all accompanying models, databases and comprehensive documentation, is publicly available as an open-source project on GitHub (https://github.com/NeLLi-team/symclatron). To ensure robust reproducibility and ease of use for the community, the tool supports multiple streamlined installation options through modern package managers, including `pixi`, `mamba` and `conda`. The software is cross-platform, natively supporting both Linux and macOS (including Apple Silicon) environments. Strict semantic versioning has been implemented through GitHub Releases (https://github.com/NeLLi-team/symclatron/releases) and system tags (the current being version 0.10.10). The exact code, neural network models and databases used to generate the findings in this study remain permanently archived and readily accessible to users and the broader scientific community.

References

83. Sieber, J. R., McInerney, M. J. & Gunsalus, R. P. Genomic insights into syntrophy: the paradigm for anaerobic metabolic cooperation. *Annu. Rev. Microbiol.* **66**, 429–452 (2012).
84. Bright, M. & Bulgheresi, S. A complex journey: transmission of microbial symbionts. *Nat. Rev. Microbiol.* **8**, 218–230 (2010).
85. López-García, P., Eme, L. & Moreira, D. Symbiosis in eukaryotic evolution. *J. Theor. Biol.* **434**, 20–33 (2017).
86. Douglas, A. E. & Werren, J. H. Holes in the hologenome: why host–microbe symbioses are not holobionts. *MBio* **7**, e02099 (2016).
87. Moran, N. A. & Sloan, D. B. The hologenome concept: helpful or hollow? *PLoS Biol.* **13**, e1002311 (2015).
88. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
89. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. Available from https://www.osti.gov/servlets/purl/1241166 (2014).
90. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
91. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Krishnapuram, B. & Shah, M.) (ACM, 2016).
92. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
93. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
94. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
95. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
96. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database. *Bioinformatics* **38**, 5315–5316 (2022).
97. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
98. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
99. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
100. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
101. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).
102. Ruiz-Perez, C. A., Conrad, R. E. & Konstantinidis, K. T. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC Bioinformatics* **22**, 11 (2021).
103. Villada, J. C. SymGs release 1. *Zenodo* https://doi.org/10.5281/zenodo.20399071 (2026).
104. Villada, J. C. Detailed provenance of genome assemblies, metagenome assemblies and SAM files used in the Symclatron–SymGs project. *Zenodo* https://doi.org/10.5281/zenodo.20600438 (2026).

Acknowledgements

We would like to thank the expert advice of J.-M. Volland regarding the symbiosis continuum and A. Camargo on ML benchmarking. The work conducted by the US Department of Energy Joint Genome Institute (JGI; https://ror.org/04xm1d337) and the National Energy Research Scientific Computing Center (https://ror.org/05v3mvq14) is supported by the US Department of Energy Office of Science user facilities, operated under contract no. DE-AC02-05CH11231.

Author contributions

Conceptualization, F.S. and J.C.V. Software, J.C.V., F.S. and E.W.W. Investigation, J.C.V., F.S., Y.M.V. and S.Q. Writing—original draft, J.C.V., Y.M.V. and F.S. Writing—review and editing, all authors. Visualization, J.C.V., F.S. and Y.M.V. Funding acquisition, F.S. and T.W. Supervision, F.S. Project administration, F.S.

Competing interests

All authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-026-03213-1.

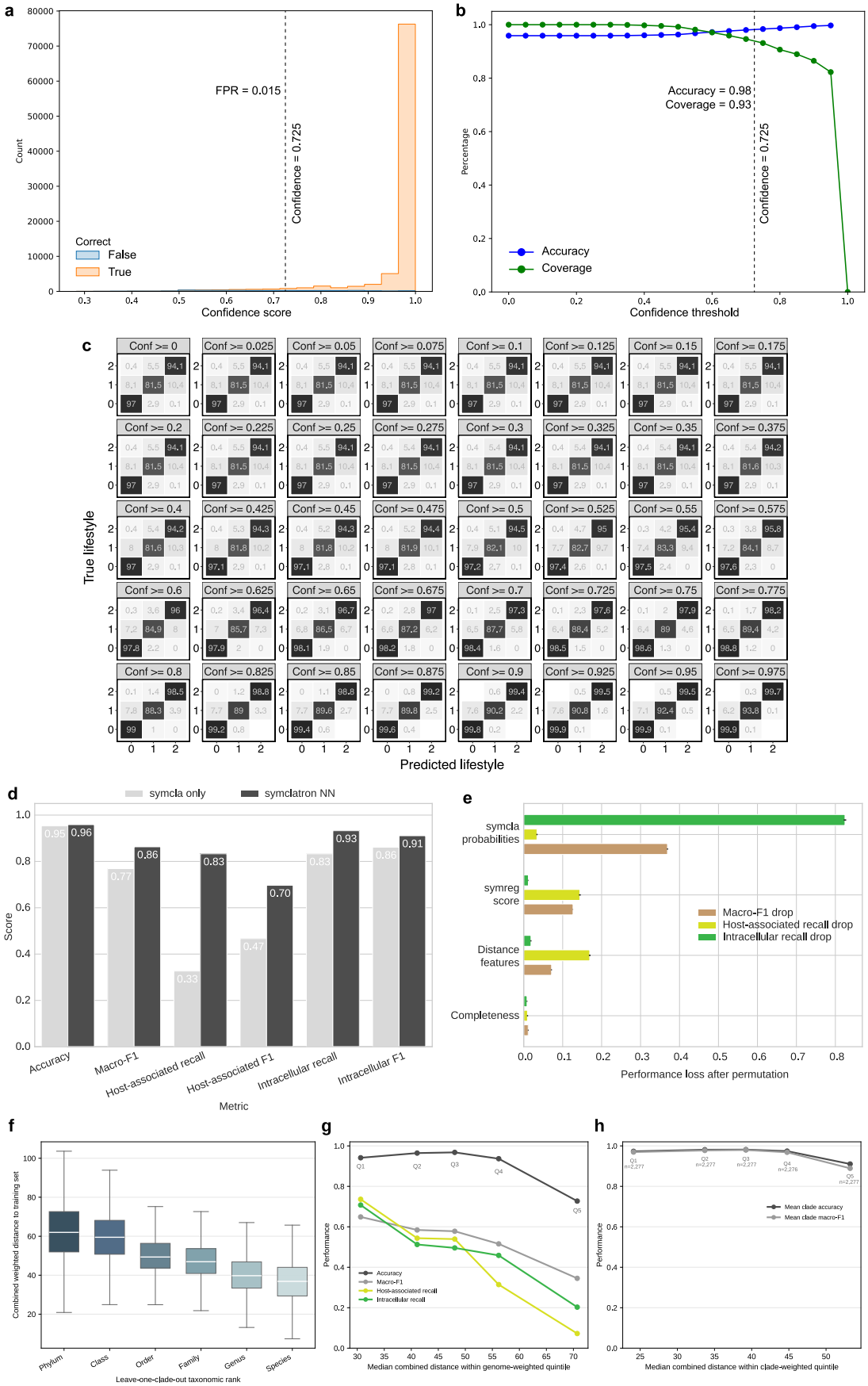
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-026-03213-1.

Correspondence and requests for materials should be addressed to Juan C. Villada or Frederik Schulz.

Peer review information *Nature Biotechnology* thanks Conor Meehan, Maria Rosa Domingo Sananes and Tobias Warnecke for their contribution to the peer review of this work. Peer reviewer reports are available.

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

https://doi.org/10.1038/s41587-026-03213-1

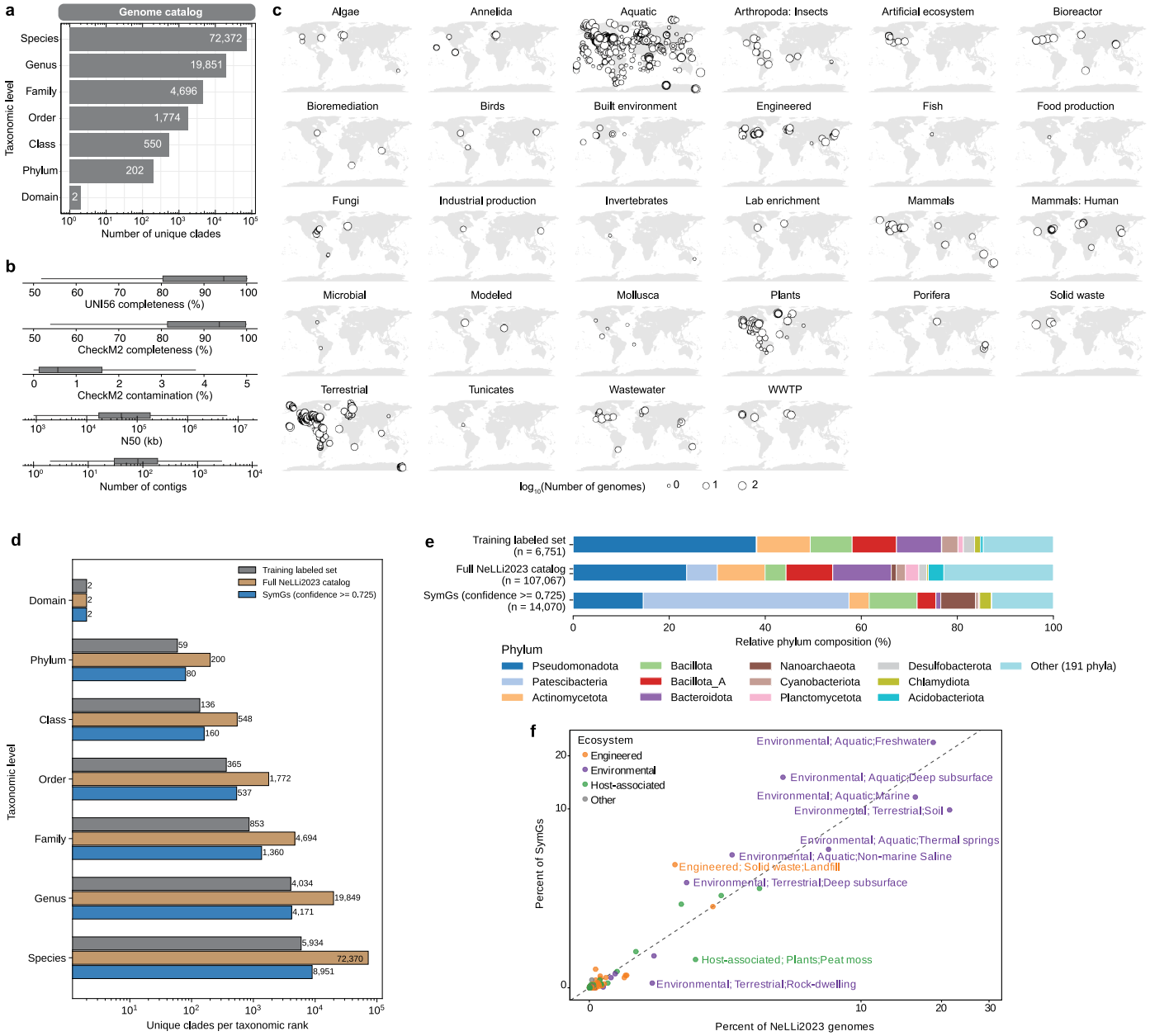


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41587-026-03213-1

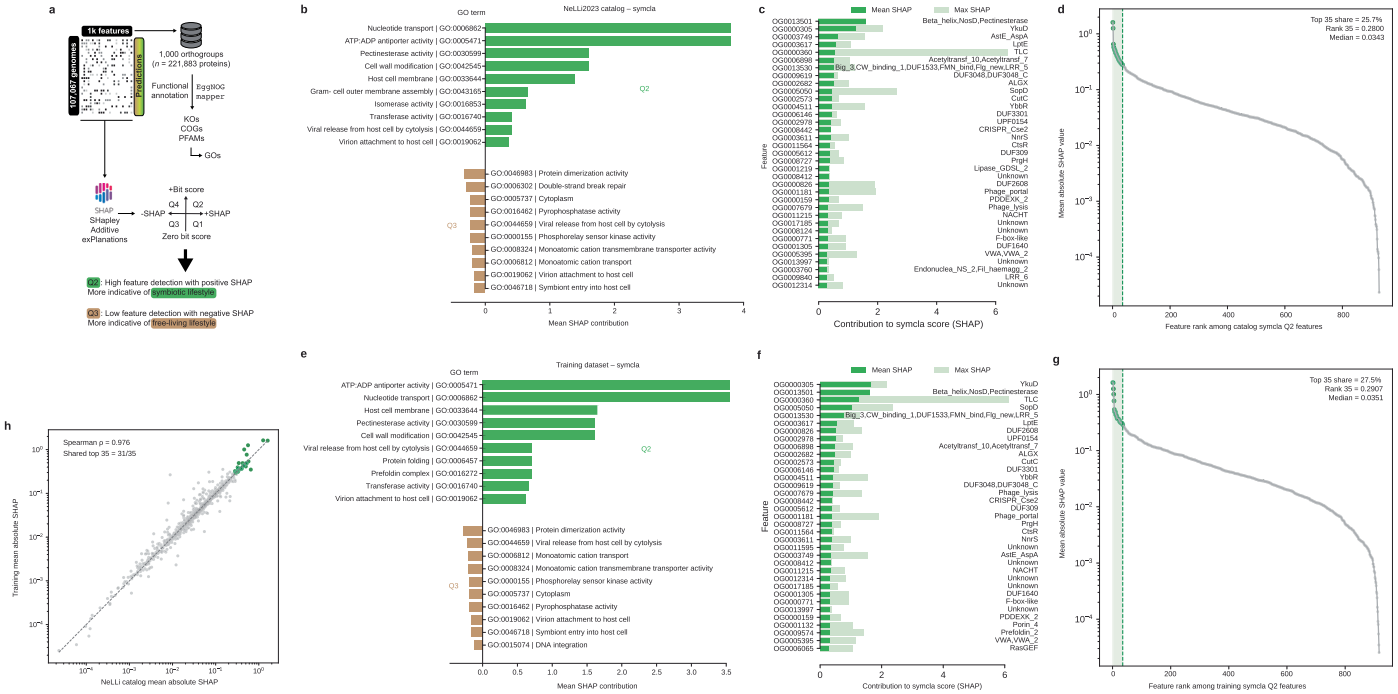
Extended Data Fig. 1 | Performance metrics of the lifestyle prediction model. (**a**) Histogram of the model's confidence scores for all predictions, colored by whether each prediction was correct (orange) or incorrect (blue). The vertical dashed line marks the chosen confidence threshold of 0.725, which corresponds to a false positive rate (FPR) of 0.015. (**b**) Classification accuracy (blue) and coverage (green) as functions of the confidence threshold. The vertical dashed line at 0.725 indicates the operating point where accuracy is 0.98 and coverage is still around 0.95 (exactly 0.9459). (**c**) Confusion matrices for the three lifestyle classes (0 = free-living, 1 = host-associated, 2 = obligate intracellular) at progressively stricter minimum confidence cutoffs (in 0.025 increments). The data shown corresponds to the calibration across all taxonomic ranks. (**d**) Comparison of *symcla* alone versus the full seven-feature `symclatron` neural network on the held-out 20% split of the integrated benchmark dataset. Shown are overall accuracy, macro-F1, and the recall and F1 values for the host-associated and intracellular classes. (**e**) Grouped permutation importance analysis of the `symclatron` neural network on the same held-out split. Bars indicate the mean performance loss after permuting each feature group, measured as the decrease in macro-F1, host-associated recall, or intracellular recall. (**f**) Distribution of the combined weighted distance to the training set for the leave-one-clade-out evaluations, using the maximum of the *symreg* and *symcla* nearest-neighbor distances. (**g**) Genome-count-weighted `symclatron` performance across quintiles of combined distance. Accuracy, macro-F1, host-associated recall, and intracellular recall are shown for each quintile. (**h**) The same analysis shown in panel G after weighting each held-out clade equally rather than weighting by genome count.

https://doi.org/10.1038/s41587-026-03213-1



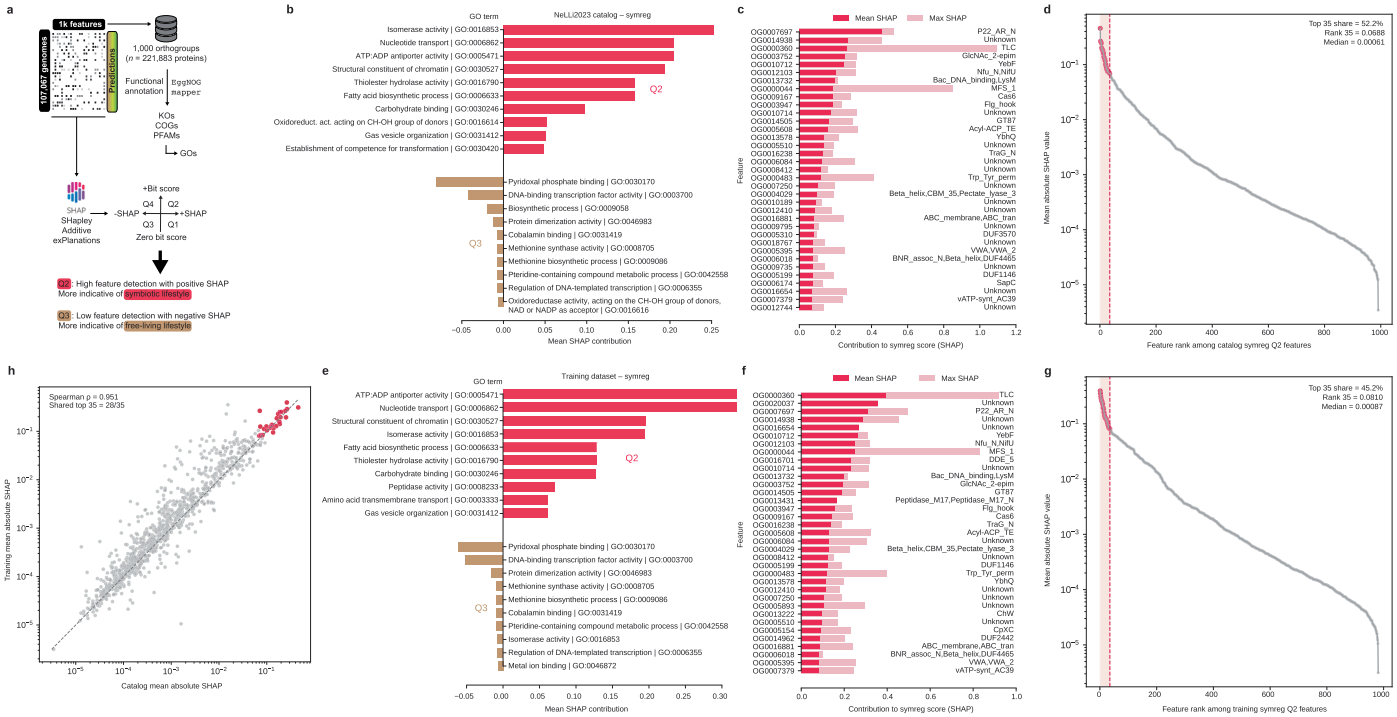
Extended Data Fig. 2 | Overview of the complete genome catalog and its global distribution. (**a**) Taxonomic breakdown of the clades identified in the entire genome catalog. (**b**) Quality assessment metrics of the entire genome catalog. (**c**) Global distribution of genomes in the entire genome catalog across different biomes. Geographical locations and environment data is only shown for metagenomes with available metadata in the IMG/M. (**d**) Unique clade counts across taxonomic ranks for the 6,751 manually labeled training genomes, the full 107,067-genome NeLLi2023 catalog, and the 14,070 high-confidence SymGs subset. (**e**) Relative phylum composition of the labeled training set, the full NeLLi2023 catalog, and SymGs. Bars show the proportion of genomes assigned to the most abundant phyla across the combined datasets, with remaining phyla grouped as Other. (**f**) Comparison of ecological composition between the NeLLi2023 catalog and SymGs. Each point represents one IMG ecosystem category (Ecosystem; Ecosystem Category; Ecosystem Type). The x-axis shows the percentage of NeLLi2023 genomes assigned to that ecosystem, and the y-axis shows the corresponding percentage of high-confidence SymGs. The diagonal marks equal relative representation.

https://doi.org/10.1038/s41587-026-03213-1



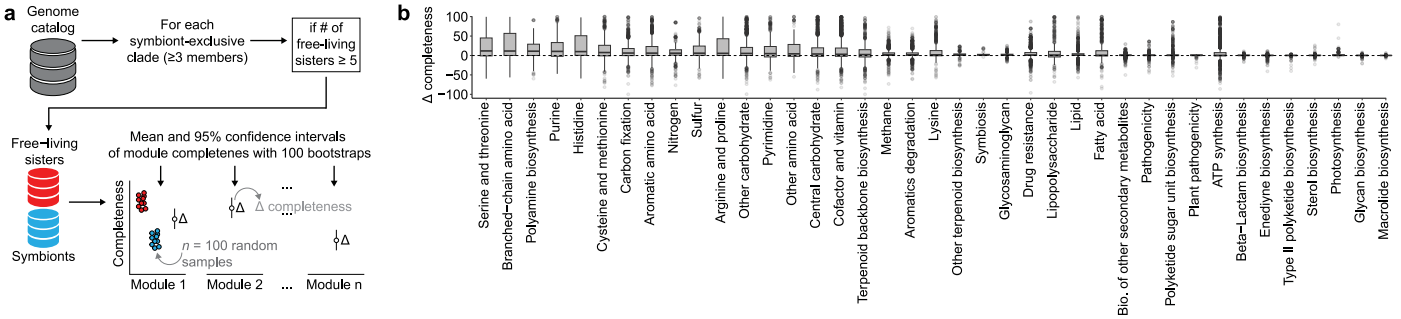
Extended Data Fig. 3 | Different *symcla* functional traits that enable accurate microbial lifestyle predictions by symclatron. (**a**) Method for functional characterization of features and their contribution to lifestyle prediction. (**b**) GO-term summary for the NeLLi2023 catalog for the *symcla* quadrant analysis. Bars in green at the top show the top ten Q2 GO terms (positive SHAP with high feature detection), for which high detection contributes positively to host-associated or obligate intracellular classification. Bars in brown at the bottom show the top ten Q3 GO terms (negative SHAP with low feature detection), for which low detection contributes negatively to these symbiotic classifications and is therefore more indicative of a free-living lifestyle. GO terms were summarized directly from aggregated SHAP values after PFAM-to-GO mapping. (**c**) Top 35 *symcla* Q2 features in the NeLLi2023 catalog. Solid bars show mean SHAP values and translucent bars show maximum SHAP values; PFAM annotations are shown at right where available. (**d**) Ranked distribution of mean absolute SHAP values across all *symcla* Q2 features in the NeLLi2023 catalog. Because *symcla* is multiclass, the underlying SHAP matrix was reduced by retaining, for each genome-feature pair, the class-specific SHAP value with the largest absolute contribution and orienting its sign toward symbiotic classification, as described in the Methods. Q2 therefore reflects positive feature contributions under this symbiont-oriented *symcla* representation. (**e**) Same as panel B but derived from the labeled training genomes. (**f**) Same as panel C but derived from the labeled training genomes. (**g**) Same as panel D but derived from the labeled training genomes. (**h**) Comparison of mean absolute SHAP values for the shared *symcla* Q2 features between the NeLLi2023 catalog and the labeled training genomes.

https://doi.org/10.1038/s41587-026-03213-1



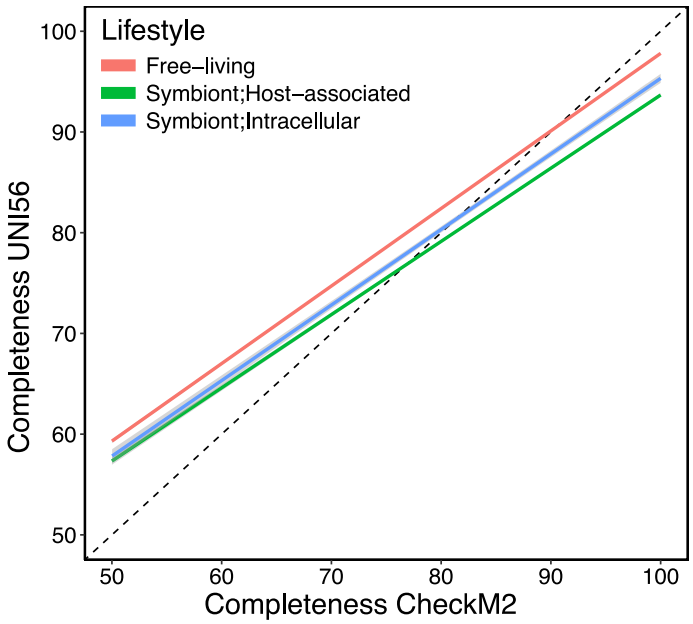
Extended Data Fig. 4 | Different *symreg* functional traits that enable accurate microbial lifestyle predictions by symclatron. (**a**) Method for functional characterization of features and their contribution to lifestyle prediction. (**b**) GO-term summary for the NeLLi2023 catalog for the *symreg* quadrant analysis. Bars in red at the top show the top ten Q2 GO terms (positive SHAP with high feature detection). Bars in brown at the bottom show the top ten Q3 GO terms (negative SHAP with low feature detection). GO terms were summarized directly from aggregated SHAP values after PFAM-to-GO mapping. (**c**) Top 35 *symreg* Q2 features in the NeLLi2023 catalog. Solid bars show mean SHAP values and translucent bars show maximum SHAP values; PFAM annotations are shown at right where available. (**d**) Ranked distribution of mean absolute SHAP values across all *symreg* Q2 features in the NeLLi2023 catalog. (**e**) Same as panel B but derived from the labeled training genomes. (**f**) Same as panel C but derived from the labeled training genomes. (**g**) Same as panel D but derived from the labeled training genomes. (**h**) Comparison of mean absolute SHAP values for the shared *symreg* Q2 features between the NeLLi2023 catalog and the labeled training genomes.

https://doi.org/10.1038/s41587-026-03213-1



Extended Data Fig. 5 | Method developed for automatic detection of differential metabolic module completeness in symbiont-exclusive clades versus their free-living sister lineages. (**a**) Schematic of the bootstrapping method for detection of differential metabolic module completeness between symbionts and free-living sister clades. For each high-confidence symbiont-exclusive clade containing at least five members, a "sister" set of free-living genomes is identified. If at least three free-living genomes exist in the sister clade, the completeness of each module is compared between symbionts and free-living relatives. One hundred bootstrap samples of free-living genomes are used to estimate the mean difference in module completeness (Δ completeness) with 95% confidence intervals. (**b**) Differential metabolic module completeness across major metabolic pathways. Each data point corresponds to a symbiont-exclusive clade's difference from its free-living sister lineage per module, with the horizontal dashed line at 0 indicating no difference. Positive values (above 0) suggest higher module completeness in the free-living set, whereas negative values indicate higher completeness in the symbionts.

https://doi.org/10.1038/s41587-026-03213-1



Extended Data Fig. 6 | Comparison of genome completeness estimates from CheckM2 and UNI56 markers across different predicted lifestyles. The dashed diagonal line indicates a 1:1 relationship. Fitted trend lines for each lifestyle show the correspondence between the two completeness metrics.

nature portfolio

Corresponding author(s): Dr. Juan C. Villada
Dr. Frederik Schulz

Last updated by author(s): 05/25/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For genome labeling: IMG/M (accessed in October 2022) For metagenomic binning: IMG/M (accessed in April 2023) GTDB (Release 220)
Data analysis	OrthoFinder (v2.5.4) MAFFT (v7.481) trimAl (v1.4) BBMap (v39.01) Prodigal (v2.6.3) hmmsearch (v3.3.2) xgboost (v1.7.4) CheckM2 (v1.0.2) MetaBAT2 (v2.15) seqkit (v2.5.1) skani (v0.1.4) GTDB-Tk (v2.4.0) EggNOG-mapper (v2.1.10) IQ-TREE (v.2.0.3) MicrobeAnnotator (v2.0.5)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

DATA AVAILABILITY:

All FASTA files of genomes that are part of the SymGs catalog are hosted at https://portal.nersc.gov/cfs/nelli/symgs/ and are publicly available.

All the environmental DNA sequencing samples from the IMG/M system, in which we found symbionts, are individually cited in Supplemental Table S4. All data can be obtained from the JGI Data Portal at https://data.jgi.doe.gov/. Data providers were individually contacted according to the JGI's Data Policy at least one month before submission of the manuscript.

CODE AVAILABILITY
The symclatron software, and all the accompanying models and data bases, are available at https://github.com/NeLLi-team/symclatron.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	Only public or already published data were used in this study.
Replication	Seed parameter was set for every stochastic method to allow reproducibility.
Randomization	Bootstraping was used to detect metabolic module completeness differences between clades of free-living and symbionts as detailed in Supplemental Figure S4: 100 bootstraps with 100 random samples each.
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A