

Predicting coarse-grained representations of biogeochemical cycles from metabarcoding data

Arnaud Belcour^{1,2,*}, Loris Megy³, Sylvain Stephant⁴, Caroline Michel⁴, Sétareh Rad⁴, Petra Bombach⁵, Nicole Dopffel⁶, Hidde de Jong^{1,2}, Delphine Ropers^{1,2,*}

¹Univ. Grenoble Alpes, Inria, 38000 Grenoble, France

²Université Grenoble Alpes, CNRS, LIPhy, 38000 Grenoble, France

³Gricad, Inria, CNRS, Université Grenoble Alpes, Grenoble INP, 38000 Grenoble, France

⁴French Geological Survey (BRGM), 45060 Orléans, France

⁵Isodetect GmbH, 04103 Leipzig, Germany

⁶NORCE Norwegian Research Center AS, 5008 Bergen, Norway

*Corresponding authors. Arnaud Belcour, Inria—Université Grenoble Alpes, 655 Avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France.

E-mail: arnaud.belcour@inria.fr; Delphine Ropers, Inria—Université Grenoble Alpes, 655 Avenue de l'Europe, Montbonnot, 38334 Saint Ismier CEDEX, France.

E-mail: delphine.ropers@inria.fr.

Abstract

Motivation: Taxonomic analysis of environmental microbial communities is now routinely performed thanks to advances in DNA sequencing. Determining the role of these communities in global biogeochemical cycles requires the identification of their metabolic functions, such as hydrogen oxidation, sulfur reduction, and carbon fixation. These functions can be directly inferred from metagenomics data, but in many environmental applications metabarcoding is still the method of choice. The reconstruction of metabolic functions from metabarcoding data and their integration into coarse-grained representations of biogeochemical cycles remains a difficult bioinformatics problem today.

Results: We developed a pipeline, called Tabigecy, which exploits taxonomic affiliations to predict metabolic functions constituting biogeochemical cycles. In a first step, Tabigecy uses the tool EsMeCaTa to predict consensus proteomes from input affiliations. To optimize this process, we generated a precomputed database containing information about 2404 taxa from UniProt. The consensus proteomes are searched using bigecyhm, a newly developed Python package relying on Hidden Markov Models to identify key enzymes involved in metabolic function of biogeochemical cycles. The metabolic functions are then projected on coarse-grained representation of the cycles. We applied Tabigecy to two salt cavern datasets and validated its predictions with microbial activity and hydrochemistry measurements performed on the samples. The results highlight the utility of the approach to investigate the impact of microbial communities on biogeochemical processes.

Availability and implementation: The Tabigecy pipeline is available at <https://github.com/ArnaudBelcour/tabigecy>. The Python package bigecyhm and the precomputed EsMeCaTa database are also separately available at <https://github.com/ArnaudBelcour/bigecyhm> and <https://doi.org/10.5281/zenodo.13354073>, respectively.

1 Introduction

The characterization of environmental microbial communities has enormously progressed thanks to advances in metabarcoding and metagenomics over the past two decades (Pinto and Bhatt 2024). These approaches are based on high-throughput sequencing and allow the identification of microbial community composition in different environments (Thompson *et al.* 2017, Shaffer *et al.* 2022). Moreover, they enable the analysis of metabolic functions associated with individual microbial species, genera or families in order to understand, e.g. their impact on soil biogeochemistry. One of the ultimate objectives is to elucidate the role these organisms play in major biogeochemical cycles (Sunagawa *et al.* 2020, Sokol *et al.* 2022).

Metagenomics approaches sequence DNA extracted from environmental samples and assemble genomes without the need for lab cultivation (Chen and Pachter 2005, Escobar-Zepeda *et al.* 2015). The metabolic functions associated with the genomic data can be directly inferred by matching DNA sequences with corresponding protein sequences in annotated reference databases. In comparison with metagenomics, metabarcoding approaches are directed against specific

barcode sequences (e.g. 16S rRNA in bacteria). The sequences are amplified through Polymerase Chain Reaction (PCR) and compared with a taxonomic reference database to identify the presence of specific taxa in the sample (Taberlet *et al.* 2012). The association with metabolic functions is then obtained either by literature search or by relying on annotated genomes available in public databases. In metabarcoding, the relation between the DNA extracted from a sample and metabolic functions of the microorganisms is thus more indirect, mediated by taxonomic categories, than in metagenomics. Metabarcoding is still a method of choice in many environmental studies, however, among other things because it (i) requires less DNA for sequencing, (ii) is faster when dealing with a great amount of samples, and (iii) requires less bioinformatics analysis (Srinivas *et al.* 2022).

The mapping of metabarcoding data to networks of metabolic functions interconverting organic and inorganic compounds present in the environment is a complex bioinformatics problem. Several methods have been developed for extracting metabolic functions from metabarcoding data, such as PICRUSt2 (Douglas *et al.* 2020) or Tax4Fun2 (Wemheuer *et al.* 2020). These methods are usually limited to

16S markers and, more fundamentally, they return specific annotation types (EC numbers, KEGG orthologs, ...) that may be difficult to interpret and to integrate into aggregated steps of biogeochemical cycles, such as methanogenesis, denitrification, and sulfate reduction. There exist methods for extracting such coarse-grained metabolic functions from sequencing data of microbial communities, notably METABOLIC (Zhou *et al.* 2022). This method, however, is restricted to metagenomics data and cannot be easily adapted for datasets generated by metabarcoding.

In this work, we propose an integrated approach for the reconstruction of coarse-grained representations of biogeochemical cycles from metabarcoding data which addresses the above problems. The approach takes the form of a bioinformatics pipeline which, in a first step, exploits taxonomic affiliations from metabarcoding sequencing to obtain a list of enzymes (metabolic reactions) occurring in the species associated with the taxa. This step is based on the method EsMeCaTa, which predicts consensus proteomes from taxonomic affiliations (Ruiz *et al.* 2024, Belcour *et al.* 2025), and exploits a precomputed database relating taxa to proteomes to speed up computations. In a second step, the list of enzymes (metabolic reactions) inferred from the metabarcoding data are analyzed by means of Hidden Markov Models (HMMs) linking proteins to coarse-grained metabolic functions associated with biogeochemical cycles. This second step is based on METABOLIC (Zhou *et al.* 2022), while slightly modifying its repertoire of HMMs to account for genes of interest in specific environments. In a third step, the inferred metabolic functions are mapped to carbon, nitrogen, and sulfur cycle diagrams, each consisting of metabolic functions and the compounds they consume and produce. The entire pipeline, called Tabigecy, has been implemented in the workflow language Nextflow (Di Tommaso *et al.* 2017). The code of the pipeline as well as the Python package performing the HMM search of metabolic functions (bigecyhmm) and the precomputed EsMeCaTa database are available.

We show the applicability of the Tabigecy pipeline by analyzing two datasets concerning microbial communities in salt caverns (Bordenave *et al.* 2013, Schwab *et al.* 2022). The analysis reproduces the main features of the interpretation of the data in the source publications, but also refines the conclusions and proposes new insights. Our results notably highlight the diversity of metabolic functions active in different caverns, suggest how this diversity could be related to human activity, and discuss the impact of identified metabolic functions on the capacity for CO₂ and H₂ storage.

The approach presented in this article is generally applicable to metabarcoding datasets in environmental microbiology, possibly after refining the prediction of coarse-grained metabolic functions to the type of communities and biogeochemical cycles studied. The approach is modular in that the prediction of coarse-grained metabolic functions can also start from a list of proteins obtained from the analysis of metagenomics rather than metabarcoding datasets.

2 Materials and methods

2.1 Implementation of the Tabigecy pipeline

The Tabigecy pipeline was developed as a Nextflow workflow (Di Tommaso *et al.* 2017). The pipeline can be directly called at its GitHub repository using `nextflow run ArnaudBelcour/tabigecy ...` in the local command line, as

detailed in the README file of Tabigecy. Tabigecy consists of three processes (Fig. 1) and expects three mandatory inputs: (i) a file with taxonomic affiliations in tabulated format, processed by the ETE3 package (Huerta-Cepas *et al.* 2016), (ii) the precomputed EsMeCaTa database, and (iii) an output folder. There are two optional arguments: (i) a tabulated file listing the abundances associated with the taxonomic affiliations and (ii) the number of cores to be used by bigecyhmm.

2.2 Generation and use of the EsMeCaTa database

To obtain the input for the construction of the EsMeCaTa database, a SPARQL query was performed on the UniProt Endpoint (Bateman *et al.* 2025), resulting in a list of all taxa available for each taxonomic rank ("species," "genus," "family," "order," "class," and "phylum"), corresponding to six tabulated files. For each file, the "esmecata proteomes" command was run against UniProt (version 2024_05; Bateman *et al.* 2025), using the rank limit option associated with the corresponding taxonomic rank. For example, when processing species, the rank limit parameter of EsMeCaTa was set to "species," to avoid EsMeCaTa using higher taxonomic ranks. EsMeCaTa produces a list of taxa, having at least five associated proteomes, and the corresponding proteomes. The latter were then clustered with "esmecata clustering" using 36 CPUs on a HPC cluster by means of MMseqs2 (version 15.6f452; Steinegger and Söding 2017). Next, the consensus proteomes predicted by EsMeCaTa were annotated by means of "esmecata annotation" using eggNOG-mapper (version 2.1.12) and the eggNOG database (version 5.0.2; Huerta-Cepas *et al.* 2019, Cantalapiedra *et al.* 2021). The generation of the annotated consensus proteomes using the HPC cluster took 104 h for "esmecata proteomes," 25 h for "esmecata clustering" (32 cores), and 31 h for "esmecata annotation" (32 cores).

The folders containing the above results were then used as input by the new EsMeCaTa command `esmecata create_db`. The command compiles the information for each taxon into a consensus proteome FASTA file and an annotation file. The annotation file is generated by combining the results from eggNOG-mapper and MMseqs2, listing for each protein cluster the annotation and the included proteins. The new subcommand `esmecata precomputed` extends the standard `esmecata workflow` command to exploit the database thus obtained. It takes as input a tabulated file containing taxonomic affiliations and the compressed zip archive of the database. EsMeCaTa parses the taxonomic affiliations using ETE3 and searches for the presence of corresponding taxa in the database. If it finds a match, the consensus proteome FASTA file and the annotation file are extracted and transferred to the output folder.

The above commands and the precomputed database were included in a new version of EsMeCaTa (version 0.6.0).

2.3 Implementation of the bigecyhmm package

We developed the Python package bigecyhmm, which re-implements the task in METABOLIC (Zhou *et al.* 2022) concerned with the prediction of coarse-grained metabolic functions in biogeochemical cycles from consensus protein sequences.

Function prediction is called by means of the command `bigecyhmm` and takes as input either (i) a FASTA file with protein sequences or (ii) a folder containing multiple such

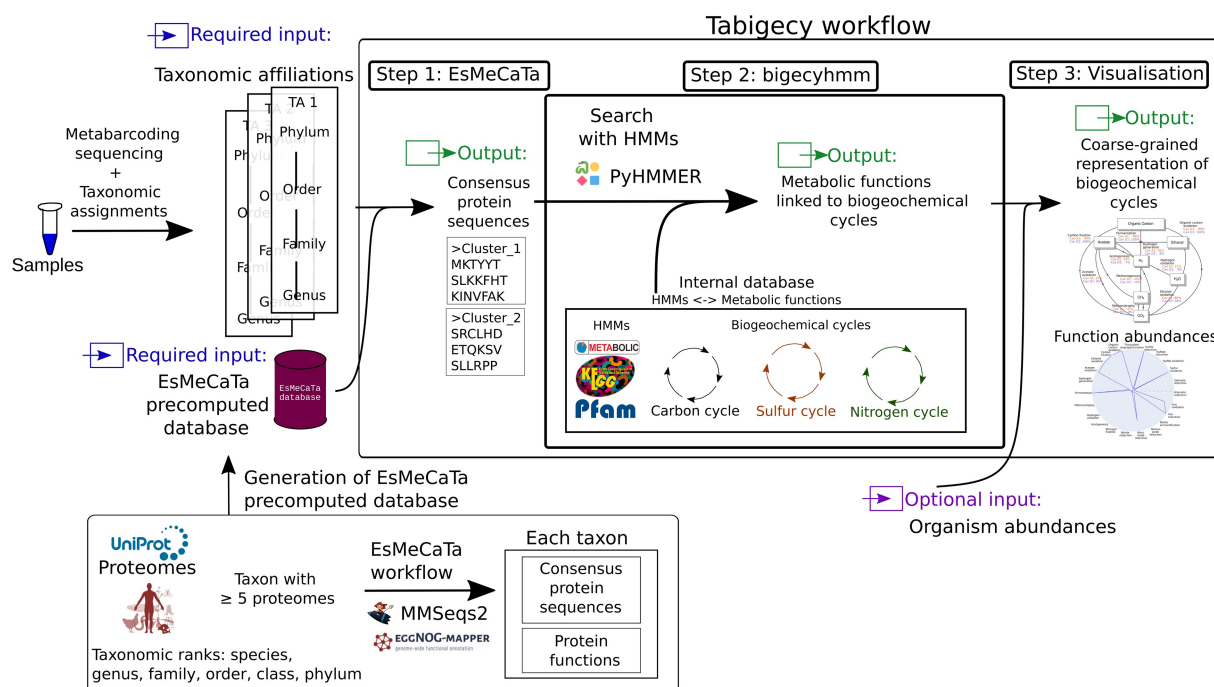


Figure 1. Overview of the Tabigecy pipeline. The input of Tabigecy consists of (i) taxonomic affiliations assigned to metabarcoding data, (ii) the EsMeCaTa database with precomputed mappings of consensus proteomes to taxonomic affiliations, and optionally, (iii) the abundance of the sequenced organisms in the environmental samples. The output of Tabigecy consists of coarse-grained metabolic functions associated with biogeochemical cycles and their visualization. The figure also summarizes the generation of the precomputed EsMeCaTa database. The pipeline draws upon a variety of existing bioinformatics tools, and extends and optimizes them for the purpose of this study, as explained in the main text.

FASTA files. An internal database, contained within the package, includes the HMM profiles for enzymes associated with biogeochemical cycles used by METABOLIC. Compared to the profiles in METABOLIC, several HMM profiles associated with the following genes were added: *hycE* (formate lyase), *acsB* (acetyl-CoA synthase), *fbfS/fbS* (formate-tetrahydrofolate ligase), *mtaB/mtbA* (coenzyme M methyltransferase), and *hydB/hydG* (sulfhydrogenase).

The input protein sequences are searched for matches with the HMM profiles using PyHMMER (Larralde and Zeller 2023), a Python library binding to HMMER (Eddy 2011). Boolean functions consisting of combinations of OR and AND operators between HMM profiles are used to conclude on the presence of specific coarse-grained metabolic functions. When a folder of FASTA files is provided as input to bigecyhm, the number of cores can be specified as an optional argument of the command. bigecyhm then parallelizes the process by distributing the HMM search of individual protein sequence files over the cores.

The biogeochemical cycle diagrams of Zhou *et al.* (2022) were reconstructed as svg templates and the metabolic function predictions are projected on the diagrams using the Pillow library (Murray *et al.* 2025). bigecyhm also computes the occurrence and, when a file with the abundances of microbial species is given as an additional input, the relative abundance of each metabolic function. Occurrences are defined as the number of taxa possessing a function as a fraction of the total number of taxa in the community. Relative abundances correspond to the sum of the abundances of all taxa possessing a function divided by the sum of abundances in the sample. Heatmaps describing either the occurrence or abundance of all metabolic functions are generated using seaborn (Waskom 2021) and matplotlib (Hunter 2007, T. M. D.

Team 2024) packages. The occurrences and abundances of major identified functions in the biogeochemical cycles are also shown in polar plots generated by means of the plotly package (P. T. Inc. 2015).

2.4 Salt cavern datasets

The communities sequenced in the study of Schwab *et al.* (2022) were analyzed by first retrieving the reads available in European Nucleotide Archive (ENA, accession number PRJEB49822). The >500 000 reads were processed by means of the FROGs pipeline (version 4.1.0; Escudié *et al.* 2018, Bernard *et al.* 2021) accessible through Galaxy France (<https://usegalaxy.fr/>; Abueg *et al.* 2024) to infer Amplicon Sequence Variants. In more detail, reads were first merged with VSEARCH (version 2.17.0; Rognes *et al.* 2016) and de-duplicated. They were then clustered using swarm with distance parameter equal to 1 and fastidious option (version 3.0.0; Mahé *et al.* 2021). Chimera were identified by means of VSEARCH and removed. Only clusters comprising at least 0.005% of all sequences were retained and those aligning to contaminants, as determined by means of BLAST (version 2.10; Camacho *et al.* 2009) were removed. Taxonomic assignment was then performed using BLAST and the reference database 16S SILVA 138.1 (Quast *et al.* 2013) with best-hit approach using a 96% identity threshold. The resulting file in Biological Observation Matrix (BIOM) format was converted into a tabulated file.

Table 2 in the study of Bordenave *et al.* (2013) was used to create the Tabigecy input files with taxonomic affiliations. Each genus in the table was processed by the ETE3 package (version 3.1.3; Huerta-Cepas *et al.* 2016) to retrieve the whole lineage (from genus to kingdom). Associating a genus with its complete lineage is necessary to prepare the input file

for EsMeCaTa, in order to allow EsMeCaTa to use a higher taxonomic level when not enough proteomes are found at the genus level.

2.5 Application of the pipeline to salt cavern communities

The two salt cavern datasets were processed using Tabigecy (version 0.1.1). Consensus proteome predictions were performed by EsMeCaTa (version 0.6.4) with the precomputed database (version 1.0.0) generated from UniProt release 2024_05 and the NCBI Taxonomy database with timestamp 2024-10-01. Prediction and visualization of the coarse-grained metabolic functions was achieved by means of bigecyymm (version 0.1.6). On a personal computer, using 5 cores, the whole workflow took 14 min to complete for the Schwab dataset and 3 min for the Bordenave dataset. By comparison, the original EsMeCaTa application without the precomputed database took 12.5 h to process the Schwab dataset and 3 h for the Bordenave dataset on a computer cluster with 10 cores. The PCA biplots in the [Supplementary information](#) were generated using R (version 4.4.1; [R Core Team 2024](#)), factoextra library (version 1.0.7; [Kassambara and Mundt 2020](#)) and the ade4 library (version 1.7–22; [Chessel *et al.* 2004](#), [Dray *et al.* 2007](#), [Dray and Dufour 2007](#), [Thioulouse *et al.* 2018](#)). Correlation plots for PCA dimensions and function abundances were created by means of the R corrplot library (version 0.94; [Wei and Simko 2024](#)).

3 Results

3.1 Pipeline for inferring coarse-grained representations of biogeochemical cycles from taxonomic affiliations

Tabigecy is a workflow which takes as input taxonomic affiliations resulting from the taxonomic assignment of metabarcoding sequences and which produces as output network diagrams of metabolic functions associated with major biogeochemical cycles ([Fig. 1](#)). It accepts the relative abundances of the microorganisms in the community as an additional, optional input to weigh the activity of the metabolic functions. Tabigecy has been implemented in the workflow system Nextflow ([Di Tommaso *et al.* 2017](#)), which allows to construct a computational pipeline from individual tasks defined with input requirements and output declarations. Below we characterize the basic functionality of the pipeline in three steps, putting special emphasis on the novel contributions of this work. For technical details on the implementation of Tabigecy, we refer to the Section 2.

3.1.1 Prediction of consensus proteomes from taxonomic affiliations

The input of the approach consists of the taxonomic affiliations of the microbial species present in the sample, as identified from the metabarcoding data ([Liu *et al.* 2021](#)). We first use EsMeCaTa ([Ruiz *et al.* 2024](#), [Belcour *et al.* 2025](#)) to predict consensus proteomes from the taxonomic affiliations. In summary, EsMeCaTa searches the UniProt database ([Bateman *et al.* 2025](#)) for the proteomes associated with each taxon (species, genus, family, ...). It then uses MMseqs2 ([Steinegger and Söding 2017](#)) to generate consensus sequences for proteins by clustering the UniProt proteomes at a specific taxonomic rank. The resulting consensus proteomes are annotated with eggNOG-mapper ([Huerta-Cepas *et al.* 2019](#),

[Cantalapiedra *et al.* 2021](#)), which assigns functional categories to the protein clusters based on orthology predictions.

The entire process of EsMeCaTa is resource and time-consuming and may take up to tens of hours to complete. In order to overcome this limitation, we generated a precomputed database of EsMeCaTa predictions and exploit this database to optimize the process. The database was obtained by running EsMeCaTa on 25 874 taxa present in UniProt ([Table 1](#)), at different taxonomic ranks, using a computer cluster (Section 2). This greatly reduces the time required by EsMeCaTa to predict consensus proteomes, typically to tens of minutes for the datasets considered below. Moreover, it improves the reproducibility of the results by ensuring that a well-defined database is used in every run of EsMeCaTa.

3.1.2 Inference of coarse-grained metabolic functions from consensus proteomes

In a second step, the workflow uses the consensus proteomes derived in the previous step to predict coarse-grained metabolic functions accomplished by environmental microbial communities. We developed the Python package bigecyymm which implements the search for enzymes involved in metabolic functions using Hidden Markov Models (HMMs), as proposed in METABOLIC ([Zhou *et al.* 2022](#)). The METABOLIC implementation has a complex architecture with multiple dependencies for metabolic function prediction, which does not allow this task to be easily isolated and adapted to metabarcoding data. The bigecyymm package was designed to be portable, self-contained, and easy-to-use.

bigecyymm takes as input protein sequences in FASTA files, which are searched by pyHmmer ([Larralde and Zeller 2023](#)) using an internal database of HMMs. The enzymes thus identified are then mapped to coarse-grained metabolic functions involved in biogeochemical cycles. In comparison with METABOLIC, we extended the repertoire of HMMs with five profiles to account for metabolic specificities of the microbial communities in salt caverns analyzed in this work (Section 2).

3.1.3 Visualization of metabolic functions in biogeochemical cycles

In a third step, Tabigecy visualizes the identified metabolic functions in the form of a polar plot summarizing the major functions. Moreover, it projects the functions on biogeochemical cycle diagrams for carbon, sulfur and nitrogen borrowed from [Zhou *et al.* \(2022\)](#). The diagrams show which organic and inorganic compounds in the environment are consumed and produced by the metabolic functions. If the abundances of the microorganisms in the communities are provided as optional input (e.g. through high-throughput

Table 1. Contents of the precomputed EsMeCaTa database.^a

Taxonomic rank	UniProt taxa	Included taxa	Number of proteomes
Species	13 187	403	5043
Genus	8926	960	11 490
Family	2215	519	9519
Order	899	292	7169
Class	373	139	4298
Phylum	274	91	2942

^a The column “UniProt taxa” refers to the total number of taxa in UniProt and the column “Included taxa” to those having more than five associated proteomes and included in the precomputed EsMeCaTa database.

sequencing of 16S rRNA), then the functions are weighted by the relative abundances. Tabigecy also generates a heatmap showing, for each metabolic function associated with an HMM, the relative abundance of the microorganisms (when available) or the relative occurrence of this function in the taxa identified in the sample (Section 2).

3.1.4 Benchmarking of workflow tasks

While the Tabigecy pipeline performs a task that is, to our knowledge, not covered by other tools in the literature, it is possible to compare subtasks of the pipeline with existing tools. We performed two such benchmarking studies (see [Supplementary Text S1](#) for details). First, we compared the prediction of metabolic functions from reference metabarcoding datasets using both Tabigecy and other tools for this purpose, notably PICRUSt2 and Tax4Fun2 ([Douglas et al. 2020](#), [Wemheuer et al. 2020](#)) ([Supplementary Fig. S1](#)). Second, we compared the predicted coarse-grained representations of biogeochemical cycles obtained by bigecyhmm with the predictions of METABOLIC ([Zhou et al. 2022](#)) using as input several metagenomics datasets available in the literature ([Supplementary Fig. S2](#)). The conclusion of these benchmarking studies is that Tabigecy performs similarly to some of the best available tools on the isolated subtasks.

3.2 Application of the pipeline to the analysis of salt cavern communities

Salt caverns are artificial underground formations that are created by drilling a well into a salt formation and injecting water to dissolve the salt, called solution mining. The resulting brine, consisting of water mixed with salt, is extracted, leaving space for storage. The impermeability of salt caverns prevents gas leakage, making them ideal for gas storage ([Voormeij and Simandl 2004](#)). Despite the high salinity of the environment, microbial life, mostly bacteria and archaea, is abundant in salt caverns. In case of hydrogen storage, microbial activity can change the gas composition and damage the operating infrastructure ([Dopffel et al. 2021](#)). Current knowledge of the microbiology of salt caverns is limited for a number of reasons, including the difficulty to obtain sufficient DNA from samples for characterizing microbial diversity, the under-representation of salt cavern communities in public genome databases, and the difficulty of culturing microbial species present in caverns in the laboratory ([López-López et al. 2013](#), [Słowakiewicz et al. 2024](#)). Predicting the metabolic functions performed by microbial communities in salt caverns is helpful for designing mitigation strategies against harmful microbial reactions. We therefore applied our Tabigecy pipeline to two publicly available datasets (Section 2).

3.2.1 Diversity of acetate metabolism in salt caverns

[Schwab et al. \(2022\)](#) investigated the microbial diversity in five German salt caverns (Cav. A to E) which have been used for natural gas storage. Taxonomic assignment of the metabarcoding reads was performed on duplicate samples from the caverns and the resulting affiliations were used to predict metabolic functions using Tabigecy ([Fig. 1](#)). The results for each replicate and each sample are shown in [Supplementary Figs S3 and S9](#). As observed by [Schwab et al. \(2022\)](#) at the taxonomic level, the replicates of Cav. A, C, and E yield reproducible functional profiles, whereas replicates of Cav. B and D have several differences.

We performed a Principal Component Analysis (PCA) on the predicted metabolic functions in the different samples and found that the caverns separate into two groups ([Supplementary Fig. S4A](#)). We analyzed one representative sample of each group in detail (A1 and D1, the first replicates of caverns A and D). The polar plot in [Fig. 2A](#) shows the main metabolic functions performed by the communities in the two caverns. The plot reveals that many functions are present in both caverns, e.g. generic metabolic functions like fermentation and organic carbon oxidation, but also more specific functions such as iron reduction and sulfur oxidation. However, the analysis also predicts differences, such as the presence of hydrogen generation in A1, but not in D1. One of the major differences concerns acetate metabolism: whereas acetate is produced in A1, it is consumed in D1. The latter observation is confirmed by the PCA in [Supplementary Fig. S4](#), where the analysis of the main principal components shows that acetate oxidation and acetogenesis are seen to be strongly but negatively correlated.

The predicted metabolic functions were mapped to the biogeochemical cycle diagram for carbon, shown in [Fig. 2B](#). The functions were weighted with the measured relative abundances of microorganisms predicted to perform these functions, based on the information provided by the consensus proteomes. When focusing on acetate production and consumption, the diagram indicates that acetogenesis from hydrogen and CO₂, *via* the Wood-Ljungdahl pathway, occurs in A1, whereas acetate oxidation takes place in D1. These two predictions are strongly supported by the abundance data, which suggest that the vast majority of microorganisms in A1 have the capacity for acetogenesis (74%) and in D1 for acetate oxidation (94%). Ethanol oxidation is predicted to be more present in D1 than in A1.

The above predictions can be validated by hydrochemical measurements reported in the same study ([Schwab et al. 2022](#)). Acetate is present at a concentration of 0.5 mM in A1, but is undetectable in D1, consistent with the predominance of acetate production in A1 and of acetate consumption in D1. Ethanol is present in both samples, but at a higher concentration in A1 than in D1 (2.6 versus 7.0 mM), as expected from the abundances. Ethanol is probably artificially introduced in caverns due to injection of workover fluids, showing the possible consequences of human intervention on biogeochemical cycles.

The reconstruction of coarse-grained representations of biogeochemical cycles by means of Tabigecy is thus shown to reproduce and refine the main observations in the original study. In particular, our analysis provides further evidence for the hypothesis of [Schwab et al.](#) that the community is capable of producing acetate from hydrogen and CO₂. While [Fig. 2](#) focuses on the carbon cycle, [Supplementary Fig. S5](#) displays the projection of the metabolic functions on the nitrogen and sulfur cycles.

3.2.2 Impact of microbial activity on CO₂ storage capacity

[Bordenave et al. \(2013\)](#) analyzed four samples from a salt cavern plant in Canada, including brine samples obtained from two salt caverns (SC02 and SC03), in view of the use of these caverns for CO₂ storage. The two caverns differ in that one had been used for oily sand disposal (SC02) whereas the other was unused (SC03). The authors already assigned taxonomic categories to the data and provided the relative abundances of the microorganisms in the samples, which were

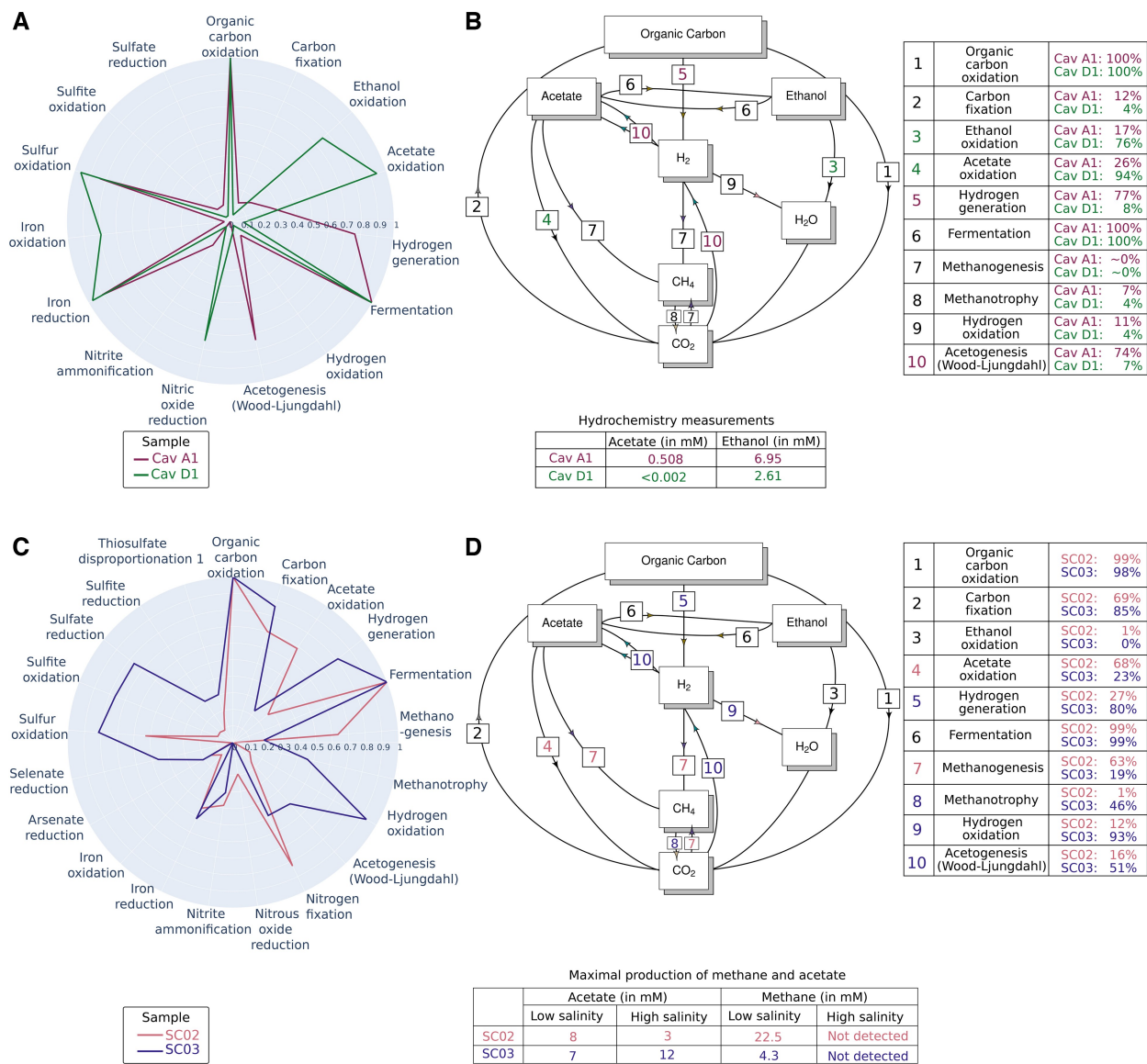


Figure 2. Reconstruction of coarse-grained representations of biogeochemical cycles from the taxonomic affiliations of salt cavern communities. Salt cavern datasets from Schwab *et al.* (2022) and Bordenave *et al.* (2013) were analyzed using the Tabigecy pipeline. (A) Polar plot indicating for each major metabolic function the relative abundance of microorganisms found in two samples (A1 and D1) in the Schwab dataset. Only functions having at least 10% of relative abundance in a sample are shown. (B) Projection of the metabolic functions from panel A on the carbon cycle diagram from Zhou *et al.* (2022), with weights in the functions given by the relative abundance of the organisms in the two samples. Measurements of acetate and ethanol from Schwab *et al.* (2022) are shown in the accompanying table. (C) As in panel A, but for two samples (SC02 and SC03) from the Bordenave dataset. (D) As in panel B, but for the Bordenave dataset, completed with measurements of acetate and methane in high and low salinity conditions (Bordenave *et al.* 2013).

used as input for Tabigecy. The results for all samples are shown in Supplementary Figs S6 and S10.

Like for the previous salt cavern study, we find a global correspondence of metabolic functions in the two samples but also interesting differences (Fig. 2C). A first observation is that in both samples, as observed by Schwab *et al.* (2022), a vast majority of organisms are capable of fermentation and organic carbon oxidation. Moreover, the samples show predicted differences in acetate metabolism, like in the previous study. Whereas acetate oxidation (consumption) is found in SC02, acetate is produced in the cavern corresponding to SC03. Other major differences between the two samples are the capacity for methanogenesis, present in SC02 (63% of

microorganisms) but much less so in SC03 (19%), and the capacity for hydrogen oxidation, present in SC03 (93%) but much less so in SC02 (12%). This observation is confirmed by the differential correlation of the metabolic functions with the first dimension of the PCA, which accounts for almost 80% of the variance in the data (Supplementary Fig. S7).

Mapping the identified functions to the carbon cycle diagram suggests possible explanations for the above-mentioned differences (Fig. 2D). It seems that in SC02, hydrogen accumulating in the caverns is mostly converted into methane, whereas in SC03, hydrogen is utilized both by acetogenesis and hydrogen oxidation. This hypothesis is partially validated by the microbial activity measurements carried out

through enrichment cultures in the laboratory, in the presence of CO₂ and for different levels of salinity (Bordenave *et al.* 2013). A 4-fold higher concentration of methane was found in SC02 as compared to SC03 (at low salinity), while acetate is detected at higher concentrations in SC03 as compared to SC02 (at high salinity). These differences are expected given the higher capacity for methanogenesis in SC02 and for acetogenesis in SC03. Given the high salinity in salt caverns, the absence of methanogenesis in high-salinity conditions suggests little consumption of CO₂, a substrate for methanogenesis together with hydrogen. This implies a lower risk for CO₂ loss in case the caverns are used for CO₂ storage.

The above observations concern the carbon cycle, but analysis of the nitrogen and sulfur cycles shown in [Supplementary Fig. S8](#) may provide further insights. The predicted strong presence of sulfur-related functions producing toxic, corrosive H₂S gas is an interesting observation for further research on the microbiology of salt caverns (Dopffel *et al.* 2021).

4 Discussion

Tabigecy is a workflow to predict coarse-grained metabolic functions accomplished by microbial communities from metabarcoding data and map these functions to global biogeochemical cycles. Tabigecy takes as input taxonomic affiliations inferred from the sequence data and predicts consensus proteomes at the most specific taxonomic ranks possible given the available proteomes in public databases. The consensus proteomes are then analyzed to identify metabolic functions associated with key enzymes and these functions are projected on global cycles for carbon, nitrogen, and sulfur. The identified metabolic functions can be given more or less weight by taking into account, as optional input, the measured abundances of the microorganisms predicted to carry out the functions. The challenge is to make reliable predictions from metabarcoding data, which, while being less informative than metagenomics data, are often found in environmental studies for reasons of costs, speed, and technical constraints (Srinivas *et al.* 2022).

The pipeline is built upon existing bioinformatics tools but adapts and optimizes them for our purpose. Tabigecy uses EsMeCaTa for the prediction of consensus proteomes from taxonomic affiliations (Ruiz *et al.* 2024, Belcour *et al.* 2025), but exploits a newly developed precomputed database which strongly reduces run times (around 60-fold for the datasets considered here, with half the number of the cores, see Section 2). While the idea of inferring metabolic functions from sequence data and the projection of these functions on diagrams of biogeochemical cycles was borrowed from METABOLIC (Zhou *et al.* 2022), Tabigecy adapts the approach both on the conceptual and implementational level. Instead of taking metagenomics data as input, we start from metabarcoding data. The latter are widely used in studies of environmental communities where the DNA yields of samples are often low. Moreover, we reimplemented one of the core tasks of METABOLIC, the HMM search of proteomes to identify enzymes involved in metabolic functions, in a stand-alone Python package, bigecyhmm, allowing this task to be integrated into Tabigecy and other pipelines. We also slightly enlarged the repertoire of HMMs with profiles specific to the communities analyzed in this work.

We tested the capabilities of Tabigecy by analyzing two biogeochemistry datasets obtained in salt caverns. These caverns can be used for the storage of hydrogen and CO₂, which poses the question of the interactions between the stored gases and resident microbial communities. The application of Tabigecy reproduced the observations made in the original source publications (Bordenave *et al.* 2013, Schwab *et al.* 2022), but also provided a more comprehensive and quantitative view. Instead of limiting the analysis to a few *a-priori* selected metabolic functions, we systematically scanned the datasets for evidence of the presence of dozens of functions. Moreover, the integration of measurements of microbial abundances makes it possible to assign a quantitative weight to predicted metabolic functions. It is important to emphasize that the predicted functions correspond to a metabolic potential but not necessarily to a metabolic activity, as the enzymes involved in the functions may be inactive. Some of the predictions of Tabigecy, however, could be validated by relating the microbial functions to reported measurements of substrates and products obtained *in situ* or from laboratory enrichment cultures.

Tabigecy is directly applicable to the study of other environmental microbial communities, for which it is often difficult to extract enough DNA for metagenomic analyses. This generalization of the pipeline to other applications may require the extension of the repertoire of available HMMs and metabolic functions. Moreover, one needs to be aware of possible limitations and biases that were also encountered in the analysis of the salt cavern datasets. First, the functional characterization of communities with little-studied microorganisms may be less reliable. When few relatively close proteomes are available, EsMeCaTa will use more distant organisms to produce consensus proteomes (Belcour *et al.* 2025). The latter are potentially less representative for the microorganisms under study. In addition, due to evolutionary divergence, bigecyhmm may have more difficulty in identifying, in the proteomes, key enzymes of metabolic functions for uncharacterized species. A second bias comes with the use of 16S barcode sequences to quantify abundances of specific microorganisms in a sample (Matchado *et al.* 2024). The copy number of 16S genes varies within and between species (Acinas *et al.* 2004), which may lead to over- or underestimated abundances. While methods have been developed to correct for this bias by predicting 16S gene copy numbers (Douglas *et al.* 2020), they may be difficult to apply in the case of poorly studied microorganisms.

A major strength of Tabigecy is that it integrates the metabolic functions inferred from the sequence data into diagrams that provide a coarse-grained representation of biogeochemical cycles. These diagrams are helpful for understanding the functioning of the communities and identifying possible perturbations due to human interference (Amundson *et al.* 2025), as illustrated by the analysis of the Schwab dataset. The use of the diagrams could be pushed further by realizing that every identified coarse-grained metabolic function corresponds to a macroreaction, e.g. $\text{HCO}_3^- + 4 \text{H}_2 + \text{H}^+ \rightarrow \text{CH}_4 + 3 \text{H}_2\text{O}$ for methanogenesis. Together, the reactions describe the flow of material through carbon, nitrogen, and sulfur cycles. The network of metabolic macroreactions thus obtained could be transformed into a mathematical model enabling the quantitative analysis and simulation of the community dynamics (Widder *et al.* 2016). This opens the perspective of gaining a better, time-resolved understanding of

the interactions between human activities and environmental microbial communities, e.g. in the context of underground gas storage and other industrial applications.

Acknowledgements

Most computations for this work were performed using the GRICAD infrastructure in Grenoble (<https://gricad.univ-grenoble-alpes.fr>).

Author contributions

A.B. conceived the method and implemented the pipeline. A.B., L.M., and D.R. generated the precomputed EsMeCaTa database. A.B., H.d.J., and D.R. conceived the experiments. A.B. conducted the experiments. A.B., S.S., C.M., S.R., P.B., N.D., H.d.J., and D.R. analyzed the results. A.B., H.d.J., and D.R. wrote the manuscript and all authors reviewed the manuscript.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This research was funded by CETPartnership, the European Partnership under Joint Call 2022 for research proposals, co-funded by the European Commission [GA N°101069750] and with the funding organizations of Norway, France, Czech Republic, and Saxon state parliament of Germany listed on the CETPartnership website. In France, this work was supported by the ANR project HyLife [ANR-23-CETP-0002]. This supplement is supported in part by the International Society for Computational Biology (ISCB).

Data availability

The Tabigecy pipeline is available at <https://github.com/ArnaudBelcour/tabigecy>. The Python package bigecyhm and the updated version of EsMeCaTa are also separately available at <https://github.com/ArnaudBelcour/bigecyhm> and <https://github.com/AuReMe/esmecata>, respectively. The precomputed EsMeCaTa database is available at <https://doi.org/10.5281/zenodo.13354073>. Data associated with the experiments presented in this article are available at <https://doi.org/10.5281/zenodo.14762346>. Reads for the Schwab dataset were retrieved from European Nucleotide Archive (ENA, accession number PRJEB49822). Datasets used for the comparison with PICRUSt2 and Tax4Fun2 were retrieved from https://github.com/gavinmdouglas/picrust2_manuscript. Metagenomes used for the METABOLIC comparison were retrieved from NCBI Projects PRJNA390944 and PRJNA449266.

References

Abueg LAL, Afgan E, Allart O *et al.*; The Galaxy Community. The galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res* 2024;52:W83–94.

Acinas SG, Marcelino LA, Klepac-Ceraj V *et al.* Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J Bacteriol* 2004;186:2629–35.

Amundson K, Borton M, Wilkins M. Anthropogenic impacts on the terrestrial subsurface biosphere. *Nat Rev Microbiol* 2025;23:147–61.

Bateman A, Martin M-J, Orchard S *et al.*; The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res* 2025;53:D609–17.

Belcour A, Hamon-Giraud P, Maigne A *et al.* Estimating consensus proteomes and metabolic functions from taxonomic affiliations. *bioRxiv*, 2025, <https://dx.doi.org/10.1101/2022.03.16.484574>, preprint: not peer reviewed.

Bernard M, Rué O, Mariadassou M *et al.* FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. *Brief Bioinform* 2021;22:bbab318.

Bordenave S, Chatterjee I, Voordouw G. Microbial community structure and microbial activities related to CO2 storage capacities of a salt cavern. *Int Biodeterior Biodegradation* 2013;81:82–7.

Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinform* 2009;10:421.

Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 2021;38:5825–9.

Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* 2005;1:e24–112.

Chessel D, Dufour A-B, Thioulouse J. The ade4 package – I: one-table methods. *R News* 2004;4:5–10.

Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–9.

Dopffel N, Jansen S, Gerritse J. Microbial side effects of underground hydrogen storage – knowledge gaps, risks and opportunities for successful implementation. *Int J Hydrogen Energy* 2021;46:8594–606.

Douglas GM, Maffei VJ, Zaneveld JR *et al.* PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* 2020;38:685–8.

Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. *J Stat Soft* 2007;22:1–20.

Dray S, Dufour A-B, Chessel D. The ade4 package – II: two-table and K-table methods. *R News* 2007;7:47–52.

Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.

Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Front Genet* 2015;6:348.

Escudé F, Auer L, Bernard M *et al.* FROGS: find, rapidly, OTUs with galaxy solution. *Bioinformatics* 2018;34:1287–94.

Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 2016;33:1635–8.

Huerta-Cepas J, Szklarczyk D, Heller D *et al.* 0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;5:D309–14.

Hunter JD. Matplotlib: a 2d graphics environment. *Comput Sci Eng* 2007;9:90–5.

Kassambara A, Mundt F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7, 2020. CRAN. <https://dx.doi.org/10.32614/CRAN.package.factoextra>

Larralde M, Zeller G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* 2023;39:btad214.

Liu Y-X, Qin Y, Chen T *et al.* A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* 2021;12:315–30.

López-López A, Richter M, Peña A *et al.* New insights into the archaeal diversity of a hypersaline microbial mat obtained by a metagenomic approach. *Syst Appl Microbiol* 2013;36:205–14.

Mahé F, Czech L, Stamatakis A *et al.* Swarm v3: towards tera-scale amplicon clustering. *Bioinformatics* 2021;38:267–9.

- Matchado MS, Rühlemann M, Reitmeier S *et al.* On the limits of 16S rRNA gene-based metagenome prediction and functional profiling. *Microb Genom* 2024;10:001203.
- Murray A, van Kemenade H, Clark A, *et al.* *python-pillow/pillow: 11.1.0*, 2025. Zenodo. <https://doi.org/10.5281/zenodo.14586647>
- P. T. Inc. Collaborative data science, 2015. <https://plot.ly>
- Pinto Y, Bhatt AS. Sequencing-based analysis of microbiomes. *Nat Rev Genet* 2024;25:829–45.
- Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;41:D590–6.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024.
- Rognes T, Flouri T, Nichols B, *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4: e2584, 2016.
- Ruiz B, Belcour A, Blanquart S *et al.* SPARTA: interpretable functional classification of microbiomes and detection of hidden cumulative effects. *PLoS Comput Biol* 2024;20:e1012577.
- Schwab L, Popp D, Nowack G *et al.* Structural analysis of microbiomes from salt caverns used for underground gas storage. *Int J Hydrogen Energy* 2022;47:20684–94.
- Shaffer JP, Nothias L-F, Thompson LR *et al.*; Earth Microbiome Project 500 (EMP500) Consortium. Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nat Microbiol* 2022;7:2128–50.
- Sokol NW, Slessarev E, Marschmann GL *et al.*; LLNL Soil Microbiome Consortium. Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nat Rev Microbiol* 2022;20:415–30.
- Słowakiewicz M, Goraj W, Segit T *et al.* Hydrochemical gradients driving extremophile distribution in saline and brine groundwater of southern Poland. *Environ Microbiol Rep* 2024;16:e70030.
- Srinivas M, O'Sullivan O, Cotter PD *et al.* The application of metagenomics to study microbial communities and develop desirable traits in fermented foods. *Foods* 2022;11:3297.
- Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8.
- Sunagawa S, Acinas SG, Bork P *et al.*; Tara Oceans Coordinators. Tara oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* 2020;18:428–45.
- Taberlet P, Coissac E, Pompanon F *et al.* Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol* 2012; 21:2045–50.
- T. M. D. Team. *Matplotlib: Visualization with Python*. 2024. Zenodo. <https://doi.org/10.5281/zenodo.13308876>
- Thioulouse J, Dray S, Dufour A-B *et al.* *Multivariate Analysis of Ecological Data with ade4*. New York, NY: Springer, 2018.
- Thompson LR, Sanders JG, McDonald D *et al.*; Earth Microbiome Project Consortium. A communal catalogue reveals Earth's multi-scale microbial diversity. *Nature* 2017;551:457–63.
- Voormeij DA, Simandl GJ. Geological, ocean, and mineral CO₂ sequestration options: a technical review. *Geosci Can* 2004;31:11–22.
- Waskom ML. seaborn: statistical data visualization. *J Open Source Softw* 2021;6:3021.
- Wei T, Simko V. *R package 'corrplot': Visualization of a correlation matrix (Version 0.94)*. 2024. <https://doi.org/10.32614/CRAN.package.corrplot>
- Wemheuer F, Taylor JA, Daniel R *et al.* Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ Microbiome* 2020; 15:11.
- Widder S, Allen R, Pfeiffer T *et al.*; Isaac Newton Institute Fellows. Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J* 2016; 10:2557–68.
- Zhou Z, Tran PQ, Breister AM *et al.* METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* 2022;10:33.