BIOSYSMO

Deliverable D2.2

Report on integrative database construction

Deliverable information	
Responsible Partner:	IDE
Work Package	WP2 Computational tools for biosystems design and implementation
Author(s):	IDE
Contributing Partner(s):	JSI, UBFC, CIIMAR, ICL
Dissemination Level:	Public
Type:	R – Document, report
Due Date:	31/08/2025
Submission Date:	29/08/2025
Version:	1.0

Funded by the European Union under the GA no **101060211**.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or REA. Neither the European Union nor the granting authority can be held responsible for them.

Project Profile

Programme	Horizon Europe
Call	HORIZON-CL6-2021-ZEROPOLLUTION-01
Topic	HORIZON-CL6-2021-ZEROPOLLUTION-01-10
Number	101060211
Acronym	BIOSYSMO
Title	BIOremediation systems exploiting SYnergieS for improved removal of Mixed pOllutants
Start Date	1 September 2022
Duration	48 months

Document History

Version	Date	Author	Remarks
0.1	15/04/2025	IDENER Team	File creation, outline
0.2	06/07/2025	CNRS	Internally reviewed by CNRS
0.3	01/07/2025	IDENER Team	Incorporation of feedback
1.0	29/08/2025	IDENER Team	Final version to submit to EC

Disclaimer

All the contributors to this deliverable declare that they:

- *Are aware that plagiarism and/or literal utilisation (copy) of materials and texts from other Projects, works and deliverables must be avoided and may be subject to disciplinary actions against the related partners and/or the Project consortium by the EU.*
- *Confirm that all their individual contributions to this deliverable are genuine and their own work or the work of their teams working in the Project, except where is explicitly indicated otherwise.*
- *Have followed the required conventions in referencing the thoughts, ideas and texts made outside the Project.*

Table of Contents

Executive Summary

One of the key challenges in developing a computationally framework to support the design of synergistic biosystems for pollutant degradation is the fragmentation and heterogeneity of the underlying biological data. Functional information on chemical compounds, reactions, pathways, enzymes, and organisms involved in pollutant biodegradation is distributed across multiple databases, varies in annotation quality, and often lacks standardized formats that allow seamless integration. This lack of harmonized resources hinders systematic identification of candidate microorganisms with biodegradation potential, reconstruction of degradation pathways, and the implementation of reproducible computational workflows for the assistance in synergistic biosystem design, capable of efficiently degrading and sequestering pollutant mixtures.

To address this gap, the BIOSYSMO project has developed BIOSYSMOdb, a structured, cross-referenced database that consolidates and organizes diverse biological, metabolic and biochemical data relevant to the degradation and sequestration of environmental pollutants. Developed as part of Task 2.1 of the project BIOSYSMOdb has been conceived as a centralized resource to support the identification of organisms with functional traits relevant to pollutant transformation or uptake. It integrates curated information on metabolic pathways, biochemical reactions, enzymatic functions, and associated taxa, enabling the design and evaluation of biosystems for bioremediation applications. Structured as a relational SQL database, BIOSYSMOdb allows integrative querying and consistent cross-referencing across biological levels. It contains data extracted from reference public repositories (e.g., UniProt[1], NCBI[2], RefSeq[3], MetaCyc[4], EAWAG-BBD[5]), complemented with curated literature information and prepared to include experimental results generated within BIOSYSMO.

The resource provides a harmonized, extensible relational design that supports integrative querying, traceability, and scalability, enabling seamless incorporation of that new experimental or curated information throughout the project and after.

Beyond static data storage, BIOSYSMOdb serves as the foundation for key bioinformatic workflows, including the generation of Hidden Markov Models (HMMs) to characterize biodegradation gene families, the annotation of microbial communities (via metabarcoding and 16S rRNA amplicon sequencing), and the simulation of metabolic capabilities through genome-scale metabolic (GEMs) models. It is a central component of the computational infrastructure developed under WP2 and is expected to support both internal project needs and broader research efforts.

The BIOSYSMOdb is being developed under a methodological framework aligned with the FAIR principles, ensuring data are Findable, Accessible, Interoperable, and Reusable within and beyond the consortium through availability on Zenodo[6]. Particular emphasis is placed on enhancing findability through the assignment of persistent identifiers and structured metadata, accessibility via open-access deposition, Interoperability through the use of controlled vocabularies and standard ontologies, and reusability by ensuring transparent data provenance, version control, and clear licensing

This deliverable provides a comprehensive technical overview of BIOSYSMOdb, detailing its design, implementation, content, and practical utilities. Section 1 introduces the database, while Section 2 describes the data processing methodology, including protocols for collection, curation, harmonisation, and integration. Section 3 outlines the content, structure, and accessibility of BIOSYSMOdb,

complemented in Section 4 by additional datasets on bacterial, fungal, and plant metal resistance proteins. Section 5 illustrates the application of BIOSYSMOdb and the complementary datasets through selected case studies, followed by conclusions and references.

Key Insights

- BIOSYSMOdb integrates critical biodegradation knowledge by integrating chemical, metabolic, enzymatic, and organism-level data from multiple heterogeneous sources into a unified, queryable SQL database aligned with FAIR principles.
- Complementary curated protein datasets on metal resistance in bacteria, fungi, and plants expand the functional utility of BIOSYSMOdb beyond organic pollutant degradation, enabling trait screening in diverse biological contexts.
- The database supports both data-driven discovery and applied annotation, serving as the backbone for workflows such as HMM-based model generation, genome-scale metabolic reconstructions, and trait detection in taxonomic or metagenomic datasets.
- Practical case studies confirm the robustness and versatility of the system, showcasing its ability to validate protein clusters, infer functional traits from real environmental samples, and guide experimental prioritization.
- BIOSYSMOdb is a dynamic and extensible platform, designed to accommodate new data and analytical tools over the course of the project and beyond, supporting long-term research in environmental microbiology and bioremediation.

List of Figures

List of Tables

Table of Abbreviations

Abbreviation	Definition
HMM	Hidden Markov Model
EAWAG-BBD	EAWAG Biocatalysis/Biodegradation Database
SQL	Structured-Query Language
NCBI	National Center for Biotechnology Information
GEM	Genome Scale Model
ECHA	European Chemical Agency
KEGG	Kyoto Encyclopedia of Genes and Genomes
GI	GenInfo Identifier
POPs	Persistent Organic Pollutants
WFD	Water Framework Directive
SVHC	Candidate List of Substances of Very High Concern for Authorisation
PAHs	Polycyclic Aromatic Hydrocarbons
CIDs	PubChem Compound Identifiers
WGS	Whole Genome Sequencing
KO	Kegg Orthology Term
MRR	Mean Reciprocal Rank
GA	Gathering Threshold
OTUs	Operational Taxonomic Units

1 BIOSYSMOdb Introduction

Contamination of soil, water and sediment due to industrial, agricultural and urban activities is a growing concern in the European Union, affecting air, water and soil quality, and compromising food security and biodiversity. Pollutants such as heavy metals, persistent organic compounds and other toxic chemicals persist in the environment, bioaccumulating and dispersing globally, negatively affecting the biosphere and human health[7,8]. Bioremediation, which leverages the metabolic potential of microorganisms, plants, and other biological systems to fully or partially degrade, transform, or immobilise environmental contaminants, has emerged as a promising and sustainable remediation strategy. Recent advances in multi-omics technologies, molecular biology, and bioinformatics have significantly enhanced our understanding of the underlying microbial communities, metabolic pathways, and environmental interactions involved in these processes. Nevertheless, relevant data remain fragmented across multiple heterogeneous databases, often lacking standardisation and interoperability, which hampers the integration of knowledge, the development of predictive models, and the translation of research outcomes into practical, scalable bioremediation applications[9].

In line with its objective to support the development of innovative bioremediation strategies, BIOSYSMOdb database integrates and harmonises biodegradation knowledge (Figure 1) from some of the most relevant public repositories in the field: EAWAG-BBD[5], MibPOPdb[10], MetaCyc[4] and KEGG[11]. BIOSYSMOdb compiles and standardises data on metabolic pathways, chemical reactions, associated enzymes and organisms involved in the biodegradation of a wide variety of compounds, including xenobiotics, organic pollutants and metals. This integrated resource provides a valuable knowledge base for broad spectrum of applications that require systematic and unified data on biodegradation processes, including genome-scale modelling of degradation mechanisms and the functional annotation of specific organisms as key activities of WP2. By bridging computational and experimental research, BIOSYSMOdb facilitates a systems biology approach to bioremediation and contributes to accelerating the development of tailored, case-specific remediation strategies providing. By bridging computational and experimental research, BIOSYSMOdb facilitates a systems biology approach to bioremediation and contributes to accelerating the development of tailored, case-specific remediation strategies.

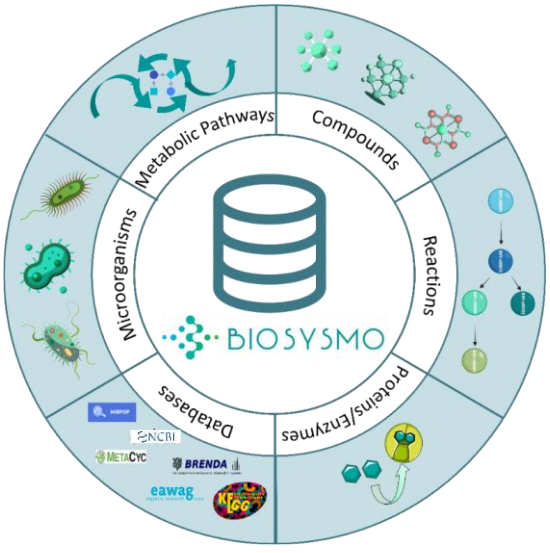


Figure 1. Graphical summary of BIOSYMOdb concept.

2 Data Processing Methodology

A systematic process was implemented to search, extract, and integrate multidimensional data related to target compounds for biological degradation, including metabolic pathways, enzymes and proteins involved, corresponding amino acid sequences and degrading microorganisms for this purpose, specialised databases were accessed through custom-developed web scraping tools designed to ensure efficient and automated data retrieval. The collected data underwent a rigorous curation process, including filtering to eliminate redundancies, standardisation of formats, and the establishment of relational connections to enable integrative analyses. The database architecture was implemented in SQL to ensure data integrity, and efficient query performance. In addition, a graphical user interface has been developed, providing an intuitive and accessible platform for querying and exploring the database content.

2.1 Data collection

Given the multilevel architecture of BIOSYSMOdb, data collection strategies were adapted accordingly. While some protocols focused on retrieving chemical-level data (e.g., chemical compounds, descriptors, regulatory classifications), others targeted metabolic-level information such as reactions, enzymes, and pathways. This complementary approach ensures consistent integration across chemical and biological domains.

Chemical level: At this level, data collection targeted key chemical compounds involved in biodegradation processes and their associated information. The selection of these compounds was based on their relevance within the most recognized health and safety regulations worldwide. In addition, xenobiotic compounds of interest to the objectives of the BIOSYSMO project were included. Data extraction was performed from official websites of regulatory agencies such as European Chemical Agency (ECHA)[12], and CompTox Chemicals Dashboard[13] or consulting other public resources such as PubChem[14] database, through direct downloading or web-scraping techniques. For the collected compounds, physical and chemical properties, biological activities, safety and toxicity information, among others, were extracted. Furthermore, the web tool ClassyFire[15] was used to organize the compounds based on a hierarchical structural classification. Additionally, unique identifiers and other pertinent information were collected from various agencies and databases, thus improving the accessibility and traceability of each compound.

Metabolic level: For this level, four public repositories were consulted: EAWAG-BBD[5], MibPOPdb[10], MetaCyc[4] and KEGG[11]. In all of them, data on degradation reactions and associated proteins and enzymes were collected. Likewise, amino acid sequences corresponding to those proteins and enzymes were collected from UniProt[1] and NCBI-protein[16], where available. In addition, metabolic pathways and the organisms involved were included. For those organisms with sequenced genomes, sequence identifiers (GenInfo (GI) ids and accession numbers) were obtained from NCBI-genome[16].

In order to compile the diverse datasets integrated into BIOSYSMOdb, different data extraction strategies were applied depending on the structure and accessibility of the source. In several cases (e.g., MetaCyc, MibPOPdb, PubChem), data could be downloaded directly in structured formats (e.g., CSV, XML, TSV) via official FTP repositories or API endpoints, allowing efficient parsing and integration. However, for certain key resources such as EAWAG-BBD and MetaCyc, direct access to comprehensive datasets was not available or was limited to user interfaces not optimized for bulk data retrieval. In these cases, custom web scraping solutions were implemented. Specifically, the Web Scraper.io[17] browser extension was used to define navigation schemas and extract structured elements

from dynamic HTML pages, while additional Python scripts using the Beautiful Soup library[18] and Selenium were developed to automate the parsing and cleaning of HTML content. These tools enabled the extraction of data from complex or manually constructed web environments such as EAWAG-BBD and MetaCyc. All scraping procedures were designed to comply with the terms of use of the respective databases and to ensure reproducibility within the project framework.

2.2 Pre-processing and unification of information

Following data extraction, all entries underwent a structured cleaning and unification process aimed at harmonizing formats, ensuring internal consistency, and eliminating redundancies across sources. Given the heterogeneity of the original datasets, a first round of pre-processing was applied independently to the data extracted from each source using custom Python scripts based on the pandas library[19] and related tools. This step included the normalization of field formats, correction of encoding issues, removal of incomplete or inconsistent entries, and harmonization of identifier schemes.

Once source-specific cleaning was completed, a second integration step was performed to merge the datasets into a unified structure aligned with the SQL schema of BIOSYSMOdb. At this stage, unique identifiers were assigned to each chemical compound, reaction, pathway, and enzyme to enable reliable referencing within and across modules. Original database identifiers were systematically retained in auxiliary fields to ensure traceability and facilitate cross-referencing with external resources. To remove redundant entries, particularly at the reaction level, a comparative screening process was implemented (Figure 1). Reactions were compared based on the identity of their reactants and products, as well as their associated enzymes. If all components matched across entries from different sources, the reaction was considered equivalent and merged accordingly.

This strategy allowed the consolidation of information from overlapping databases, ensuring a non-redundant but comprehensive dataset. Figure 2 illustrates an example of the integration of data describing a single metabolic pathway obtained from three different public databases: MetaCyc, MibPOP, and EAWAG-BBD. Each source contains partial and non-overlapping information, some include reactions and enzymes, others provide only compounds or associated microorganisms (e.g., MibPop case, were no reaction are registered as items). The unification process integrates these fragmented entries into a single, harmonized pathway within BIOSYSMOdb. Unique identifiers are assigned to each compound, reaction, enzyme, and organism using standardized BIOSYSMO ID codes (e.g., BSY-c0000X for compounds, BSY-r0000X for reactions) enabling consistent cross-referencing and integrative querying across all biological levels in the database.

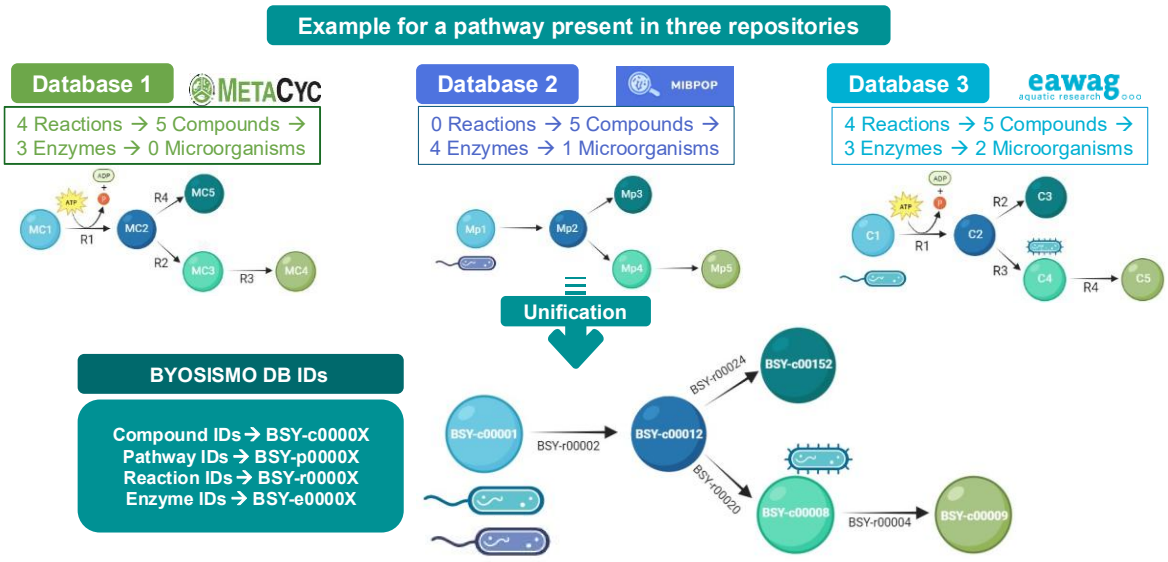


Figure 2. Example of pathway unification from three source databases. Arrows represent biochemical reactions and nodes correspond to compounds. The figure illustrates how a single pathway is reconstructed by integrating overlapping elements (compounds, reactions, enzymes, and organisms) from MetaCyc, MibPOP, and EAWAG-BBD into a unified BIOSYSMOdb entry.

2.3 Implementation of PostgreSQL database

To support the structured storage and retrieval of biodegradation data within the BIOSYSMO project, a PostgreSQL relational database was implemented using a containerised architecture. The development process comprised four main stages: (i) creation of a .NET Core application for schema management and data access, (ii) containerisation of both the application and the PostgreSQL instance with Docker, (iii) automated data ingestion from structured XLSX files, and (iv) system execution and deployment.

(i) The .NET Core application was developed using the Npgsql.EntityFrameworkCore.PostgreSQL package, enabling seamless communication between C# object models and the PostgreSQL database via Entity Framework Core.

(ii) Database tables, relationships, and constraints were defined through C# classes, with migrations generating the corresponding SQL schema. Both the application and database were containerised using Docker, with Visual Studio's Docker support streamlining the creation of the Dockerfile and a docker-compose.yml file orchestrating the environment. The setup ensured the database was initialised prior to application execution.

(iii) Data population was performed through an automated ingestion module within the application, which parsed structured XLSX files, extracted relevant information, and inserted records into the PostgreSQL database. The insertion sequence was controlled to preserve referential integrity.

(iv) Once deployed, the system executed migrations and data population within a reproducible containerised environment. This architecture guarantees consistent alignment between the database schema and application logic, while providing portability, scalability, and a robust foundation for future extensions, querying, and integration with other BIOSYSMO components.

3 Description of BIOSYSOMOdb

3.1 Content of the database

This section describes the scope and composition of, outlining the curated content BIOSYSMOdb on compounds, reactions, pathways, enzymes, and organisms, together with their sources, annotation, and coverage.

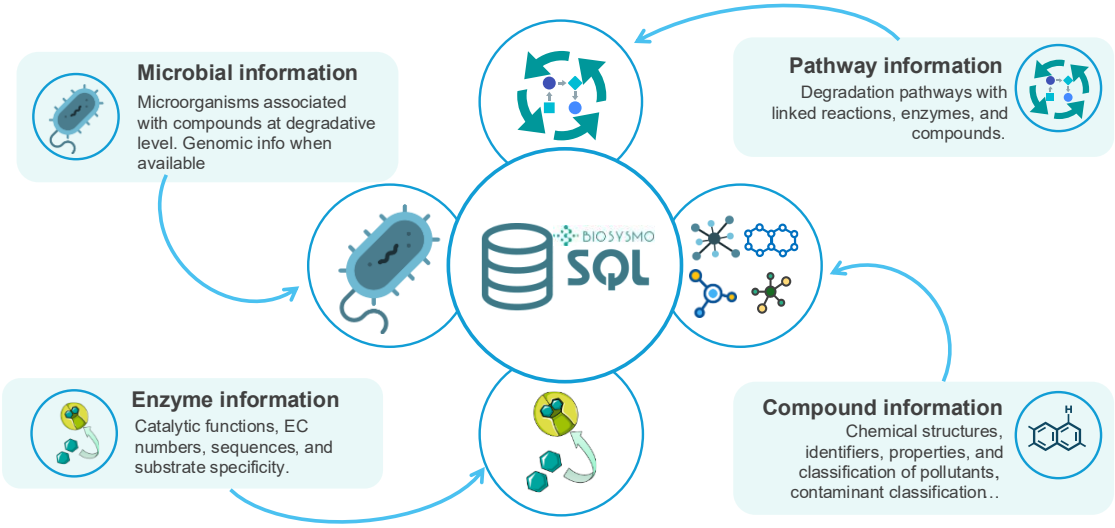


Figure 3. BIOSYSMOdb content diagram

Chemical compounds

The Compounds section in BIOSYSMOdb integrates a wide array of chemical substances relevant to the degradation and transformation of pollutants, including both regulated and unregulated xenobiotics as well as their degradation intermediates and products. The goal of this component is to capture the chemical diversity of BIOSYSMO-relevant pollutants, to support downstream tasks such as target matching, enzyme association, pathway reconstruction, and compound classification.

Content and sources

The compound entries were compiled from a combination of regulatory databases, public chemical repositories, and project-specific sources:

- **Regulatory datasets**: Compound lists from regulatory bodies and frameworks such as European Chemicals Agency (ECHA)[20], the Stockholm Convention on Persistent Organic Pollutants (POPs)[21], and the EU Water Framework Directive (WFD)[22] were integrated. These include databases such as the Candidate List of Substances of Very High Concern for Authorisation (SVHC)[23], and the PBT/vPvB assessment list for persistent, bioaccumulate and toxic substances[24]. Entries were curated and cross-referenced with PubChem[14] and CompTox to obtain chemical identifiers and metadata.
- **Research initiatives and supporting networks**: Additional substances were included from priority pollutant lists developed by collaborative research platforms such as NORMAN[25], which provides expert-curated suspect lists widely used in environmental monitoring and risk assessment. These datasets help capture relevant emerging contaminants.

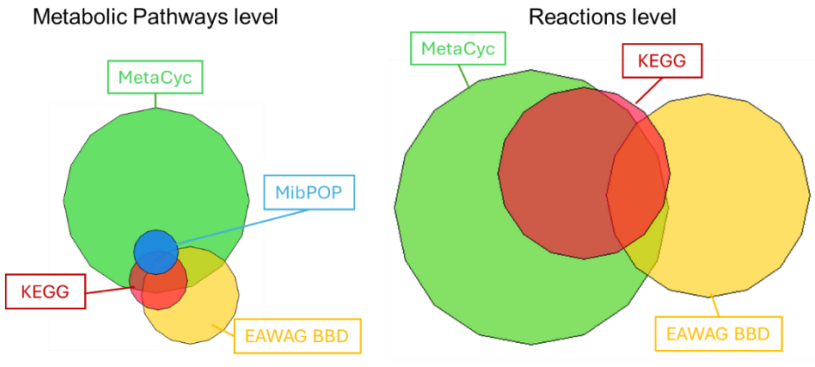
- **Environmental and toxicological databases**: Further compound entries were extracted from the EPA's ECOTOX database[26], EPA'S Structure-Searchable Toxicity (DSSTox) database[27], the CompTox Chemicals Dashboard[28], and PubChem[14], particularly to represent site-relevant pollutants such as pharmaceuticals, pesticides, personal care products, polycyclic aromatic hydrocarbons (PAHs), and heavy metals.
- **BIOSYSMO-specific compilations**: To ensure the database reflects the project's experimental and modelling needs, additional compounds were included based on the pollution profiles of BIOSYSMO target sites, as well as to support descriptor-based clustering and model training.

After curation and de-duplication, the current version of BIOSYSMOdb includes a total of **19,001** unique compound entries, grouped across regulatory, industrial, and structural categories Each compound entry is annotated with available identifiers, including CAS numbers, PubChem Compound Identifiers (CIDs), InChIKey, SMILES, and DTXSID (a unique identifier used in the EPA's DSSTox database). Annotation also includes physicochemical properties (e.g., molecular weight, logP), regulatory classification, and chemical ontology, using hierarchical descriptors generated through the ClassyFire API[15].

Reactions and pathways

The Reactions and Pathways modules in BIOSYSMOdb compile and organize metabolic transformation data relevant to pollutant biodegradation. A total of 3,876 unique reactions and 1,061 unique pathways have been integrated into the database. These were aggregated from multiple sources, including MetaCyc[4], EAWAG-BBD, MibPOP[10], and KEGG[11], which often contain overlapping but complementary information (Figure 4). For example, MetaCyc was the main contributor with 2,685 reactions and 864 pathways, while EAWAG-BBD added 1,477 reactions and 238 pathways, although many of these entities are shared across databases. In the case of MibPOP, most of its entries were covered by other databases.

Reactions are annotated with their respective reactants and products (linked to compound entries), associated enzymes (when available, via EC numbers, see next section), and original data sources. In addition, reactions were categorized by biochemical mechanism. Pathways were also characterized based on their completeness and metabolic context (e.g., aerobic vs. anaerobic degradation).



Database	Total Number
MetaCyc	864
MibPOP	50
EAWAG BBD	238
KEGG	86

Database	Total Number
MetaCyc	2685
MibPOP	No reactions collected
EAWAG BBD	1477
KEGG	1053

Figure 4. Reaction and pathway coverage across sources. Venn diagrams and counting tables illustrate the degree of overlap and complementarity between MetaCyc, EAWAG-BBD, MibPOP, and KEGG in terms of degradation pathways (left) and biochemical reactions (right).

Enzymes

The enzyme dataset in BIOSYSMOdb captures enzymatic activities associated with key degradation reactions, with a total of 3,126 unique enzymes integrated from sources such as MetaCyc, EAWAG-BBD and BRENDA. Enzymes are referenced by their EC numbers, which provide a functional classification, and linked to degradation pathways through their associated reactions. A large proportion of the enzymes belong to oxidoreductase and hydrolase classes, particularly those involved in ring-cleavage, dehalogenation, and metal transformation. To support downstream analyses, each enzyme was annotated based on its level of characterization, ranging from experimentally validated to inferred activities, offering flexibility in the selection of robust candidates for pathway modelling or HMM construction. Where available, enzyme sequences were cross-referenced to UniProt entries, enabling integration with protein-level analysis pipelines.

Organisms

BIOSYSMOdb includes a curated list of microbial and plant organisms linked to pollutant degradation potential, derived from pathway databases and literature sources. Despite being focused on microorganisms, a total of 7,179 different organisms have been incorporated, spanning bacteria, fungi and selected plant species, and annotated with taxonomic information and functional associations. Organisms are linked to the enzymes and pathways they encode or express, and flagged based on their ecological or metabolic traits. In particular, microbial species were classified according to their oxygen requirements (i.e., aerobic, anaerobic, or facultative) based on the pathways they participate in. This allows users to filter and prioritize organisms according to the environmental context of application, and to support the rational design of microbial consortia or plant-microbe systems adapted to specific remediation scenarios.

Data sources

All data sources used in BIOSYSMOdb are properly cited and comply with licensing terms for reuse and redistribution. A detailed list of licenses and database-specific attributions is included in the **README** accompanying the SQL upload package on Zenodo and also in Annex 1.

General statistics

To summarize the scope and coverage of the current version of BIOSYSMOdb, Table 1 provides an overview of the number of curated and non-redundant entries integrated across the main database levels. These figures reflect the outcome of the complete data extraction, cleaning, and harmonization process, and highlight the chemical, metabolic, and biological diversity consolidated within the resource.

It is important to note that while the database currently includes 3,126 unique enzymes/proteins (Table 2), this number refers to non-redundant functional entries, each defined by a distinct enzymatic activity, reaction context, and annotation profile. However, many of these proteins are found across multiple organisms, and in several cases, a single enzyme function is encoded by multiple homologous genes, isoforms, or strain-specific variants. For practical applications (Section 5), which rely on detecting subtle sequence-level conservation patterns, this sequence diversity is essential. Therefore, when retrieving protein sequences for downstream use, all available organism-specific variants and isoforms associated with each unique enzyme were extracted in FASTA format. This approach significantly increases the total number of sequences included in computational pipelines, ensuring that both the taxonomic breadth and intra-functional variability are adequately represented. Such redundancy at the sequence level allows more robust model generation and broader detection capabilities in functional annotation protocols.

Table 1. General statistics of BIOSYSMOdb

Database level	Number of objects
Compounds	19,001
Metabolic Pathways	1,061
Reactions	3,876
Unique Enzymes/Proteins	3,126
Degradative Organisms	7,179

3.2 Database structure

The BIOSYSMOdb database is built upon a modular and relational architecture designed to integrate heterogeneous data relevant to pollutant biodegradation and sequestration. Its structure allows the consolidation of chemical, biological, and biochemical entities, enabling both vertical (from chemical compounds through reactions and pathways to the associated organisms) and horizontal (enabling alignment across source databases, ontologies, and regulatory frameworks) integration.

The database is organized into two main functional blocks (Figure 5):

- **Chemical level section** (left, purple): contains curated information on individual compounds, their structural and physicochemical properties, classification into chemical families, and regulatory metadata.
- **Metabolic level section** (right, green): links compounds to their transformation reactions, the enzymes associated with those reactions, the pathways they belong to, and the organisms in which these elements have been described or predicted.

Each core biological entity (Compounds, Reactions, Enzymes, Pathways, and Organisms) is represented by a dedicated table with clearly defined foreign-key relationships that enable complex querying and ensure data integrity. For example:

i. The Reactions table connects reactants and product compounds via foreign keys to the Compounds table.Each reaction is associated with one or more enzymes (via the Enzymes table) and integrated into one or more metabolic pathways (via the Pathways table).

ii. Pathways are linked to organisms (via the Organisms table), allowing taxonomic filtering and organism-specific retrieval of functional data.

In addition, chemical-level section tables capture:

- **Regulatory provenance** (e.g., via the Agencies table),
- **Chemical groupings** (e.g., Families: PAHs, POPs, etc.),
- **Physicochemical descriptors** (e.g., molecular weight, logP),
- **Ontology terms** (e.g., ClassyFire-based structural categories), and
- **Identifiers** (e.g., CAS numbers, PubChem CIDs, InChIKeys, SMILES).

The entire schema was implemented in PostgreSQL, and deployed using a Dockerized environment to ensure portability, reproducibility, and scalability. The use of Entity Framework Core enabled schema migrations and logic to be synchronized with the application layer, as detailed in Section 3.3.

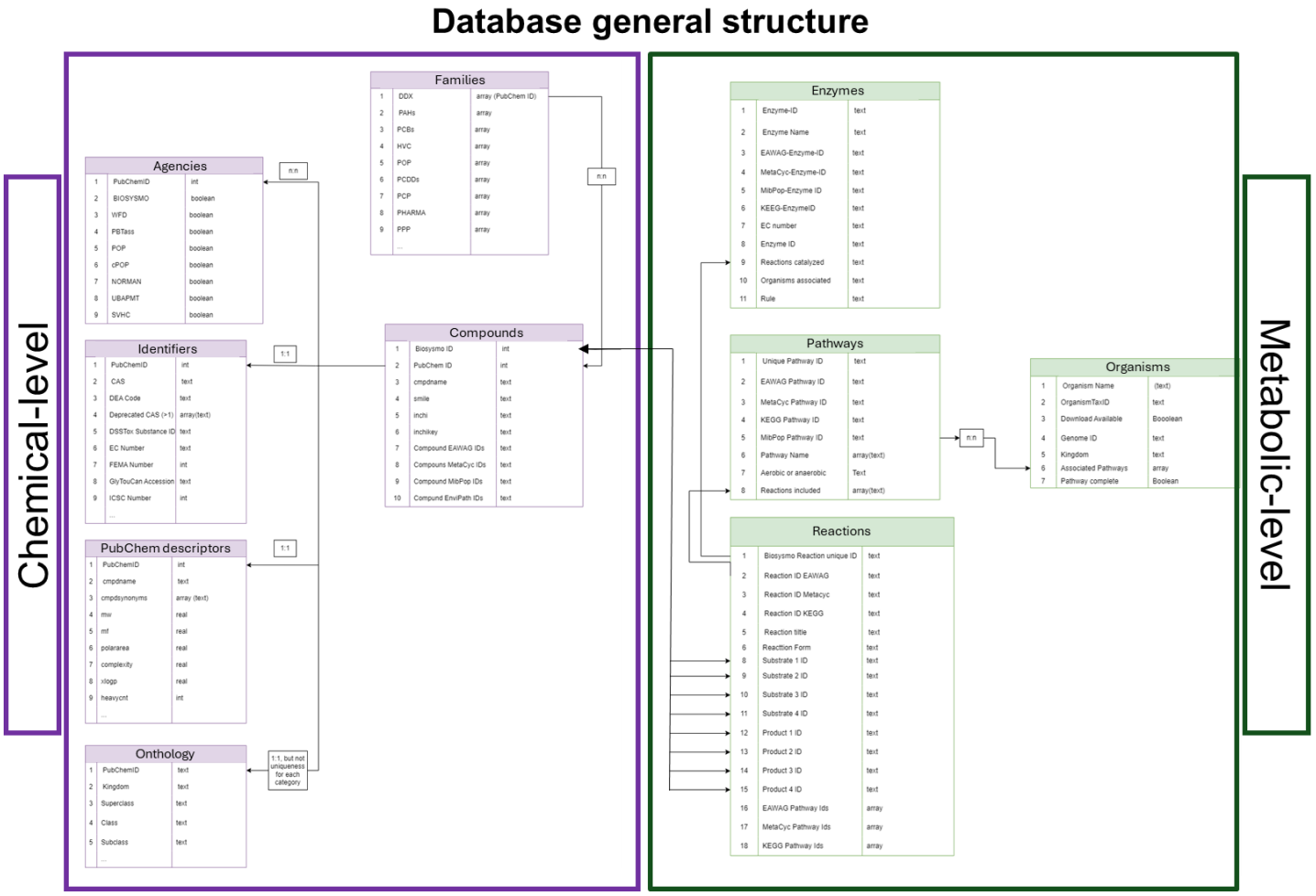


Figure 5 BIOSYSMOdb SQL structure diagram. The database is divided into two main sections: (left, purple) the chemical-level structure; and (right, green) the metabolic-level structure. Arrows indicate primary foreign-key connections between tables.

3.3 Database accessibility

The curated and consolidated SQL implementation of BIOSYSMOdb has been uploaded in Zenodo as an open-access resource for the wider community. In parallel, a dedicated graphical user interface (GUI) has been developed to support consortium partners, providing interactive querying and visualization of degradation pathways together with their associated chemical, enzymatic, and organism-level entities. Access to the GUI is currently restricted to project members via a web client integrated into the BIOSYSMO project platform. Graphical interface

As mentioned above, to facilitate access and exploration of the BIOSYSMOdb content, a graphical web interface was developed and made accessible to project partners. This interface allows users to interactively query the database and visualize the relationships between compounds, reactions, pathways, and organisms involved in pollutant degradation.

The graphical user interface was developed using the .NET Blazor framework[29], providing an interactive environment for querying and visualizing the content of BIOSYSMOdb. This interface enables users to

explore detailed information on chemical compounds, including the reactions they participate in and the metabolic pathways in which they are involved. It also supports side-by-side comparison of compound-specific data to facilitate comparative analysis, as specifically explained in Database structure

For visual representation, two specialized JavaScript libraries were integrated: smilesDrawer[30], used to render compound structures from SMILES notation, and Cytoscape.js[31], employed to dynamically visualize compound-reaction-pathway networks as interactive graphs.

The interface includes three main modules:

- **Compound comparison module**: A dedicated section allows the side-by-side comparison of two chemical compounds. As illustrated in Figure 6, the interface displays key molecular data including structure (rendered from SMILES), formula, molecular weight, canonical SMILES string, and synonyms. This feature is useful for identifying structural similarities and functional analogs across degradation pathways.
- **Interactive network explorer**: Users can investigate the presence and roles of selected compounds within the entire network of degradation reactions. The visualization highlights all reactions in which the selected compounds act as reactants or products, with differentiation of directionality and role (e.g., input, output, or both) (Figure 7). This dynamic view facilitates hypothesis generation regarding alternative pathways, branching points, or convergence between degradation routes.
- **Contextual filtering and metadata display**: Additional filtering options are available to refine results based on reaction context, including aerobicity, compound role, and presence of specific microorganisms. Detailed metadata for reactions and compounds can be accessed directly from the graph (see Figure 8), including reactant/product identity, aerobocity, and the pathways in which the reaction is embedded.

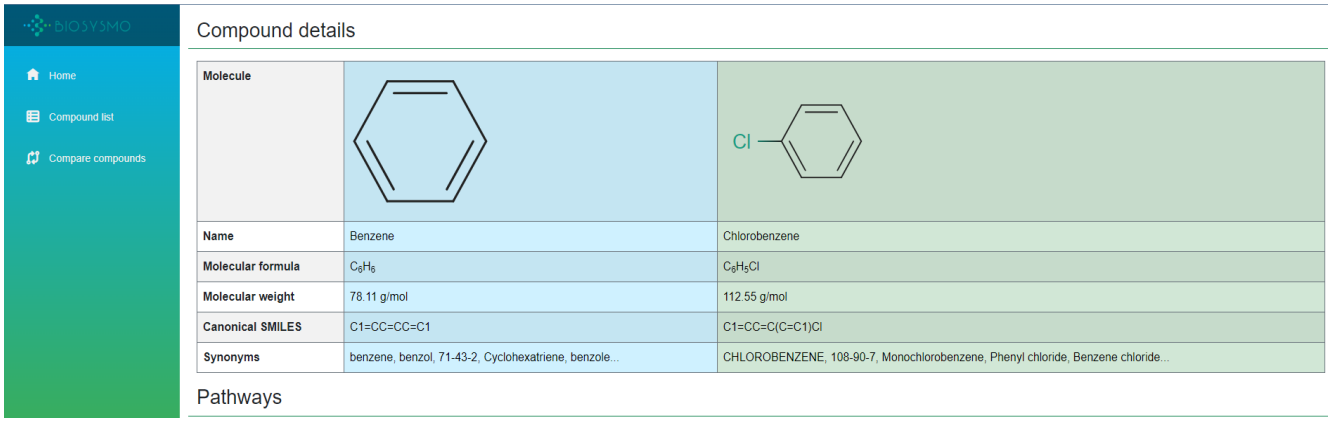


Figure 6 Compound comparison module. Side-by-side comparison of two compounds, displaying molecular structure, formula, weight, SMILES, and synonyms.

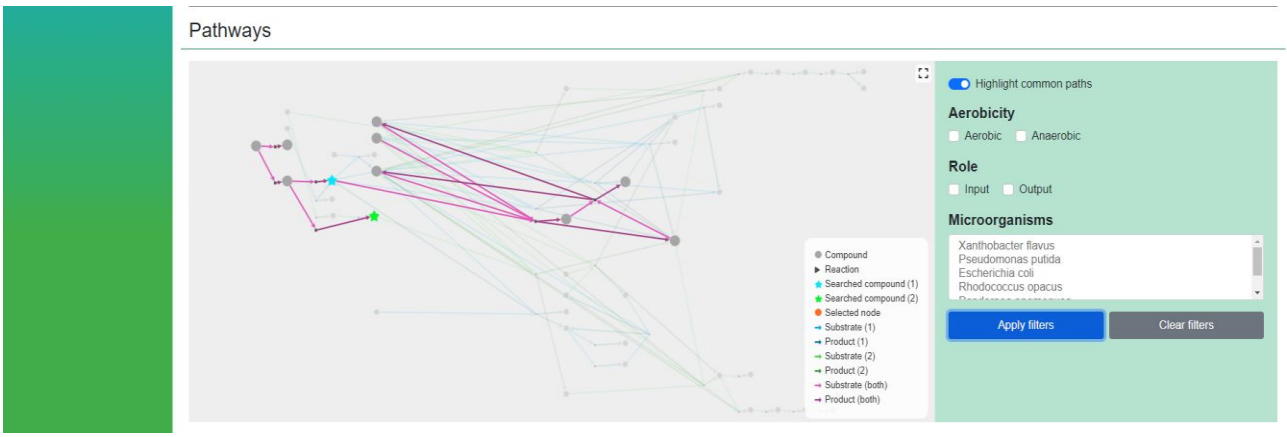


Figure 7 Network explorer (pathway highlighting).Graph highlighting reactions in which the selected compounds participate as substrates or products. Filters allow context-specific exploration.

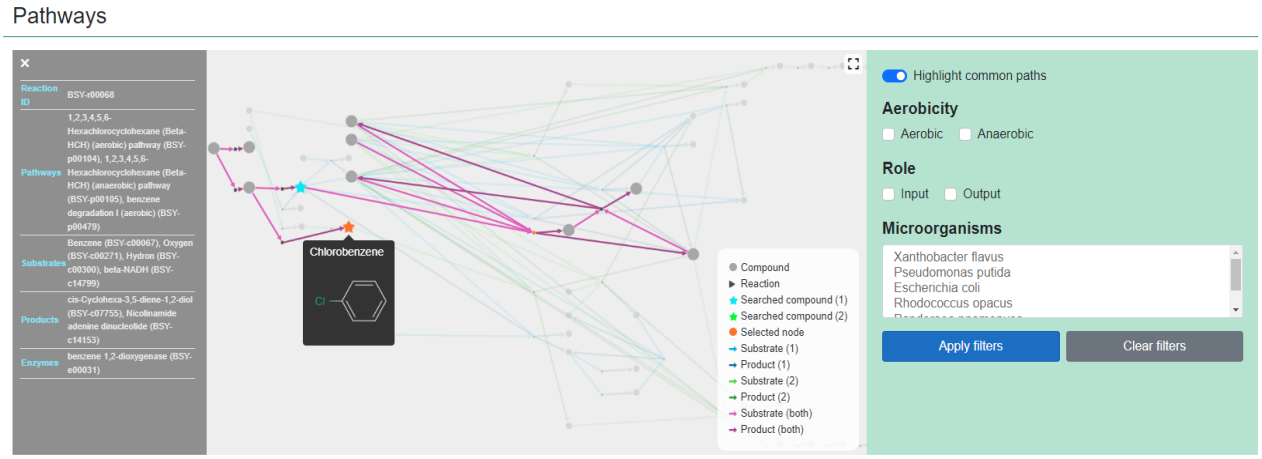


Figure 8 Reaction metadata and filtering panel. Detailed reaction information and filtering tools based on compound role, aerobicity, and associated microorganisms.

4 Collection of complementary datasets

In addition to the core BIOSYSMOdb structure, which focuses on genes and pathways directly involved in the degradation and transformation of organic pollutants, additional protein datasets of strategic importance for the project were collected and curated. Specifically, proteins involved in metal resistance in bacteria, fungi, and plants as the presence of heavy metals often co-occurs at polluted sites and can inhibit microbial growth or disrupt enzymatic activity[32]. Although these datasets are not integrated into the structure of BIOSYSMOdb, they are conceptually aligned with its objectives and were constructed using similar data mining and curation protocols. These include systematic searches in public repositories (e.g., UniProt[1], NCBI[3], PlantPred[33], BacMet[34]) using keyword-driven queries, sequence similarity tools, and literature cross-referencing to ensure the functional relevance of the selected proteins. Each dataset comprises a curated collection of protein sequences, accompanied by a standardized metadata file that provides essential contextual information, including source organism, original database, associated metal(s), predicted or experimentally validated resistance mechanisms, and relevant functional annotations. This modular strategy enabled the inclusion of high-value biological data that, while not embedded within the BIOSYSMOdb core structure, complements and expands the scope for wider applications, particularly those focused on the identification of resistance traits in microbial isolates or engineered consortia exposed to metal stress.

4.1 Bacterial metal resistance proteins

The bacterial metal resistance protein dataset was created to support the identification and functional annotation of microbial traits relevant to metal resistance potential traits. Therefore, including information on resistance mechanisms is essential for selecting resilient bacterial strains. This dataset comprises 542 curated protein sequences linked to bacterial resistance mechanisms against 27 different metals and metalloids, including elements such as arsenic (As), cadmium (Cd), zinc (Zn), mercury (Hg), cobalt (Co), copper (Cu), lead (Pb), and others. A complete list of elements is included in Figure 9

Figure 9 Distribution of bacterial metal resistance-associated proteins across target elements. Pie chart showing the number of protein entries associated with each of the 27 metals and metalloids included in the dataset., which shows the distribution of protein entries by associated metal. These proteins include efflux pumps, metal-binding proteins, oxidative stress response enzymes, and transcriptional regulators.

Protein candidates were identified through targeted queries in BacMet[34], MEGARes[35] and UniProt[1], focusing on sequences with experimentally validated functions or strong functional predictions. In total, these proteins are associated with 139 unique bacterial strain/species, reinforcing their potential value for microbial genome annotation and functional screening.

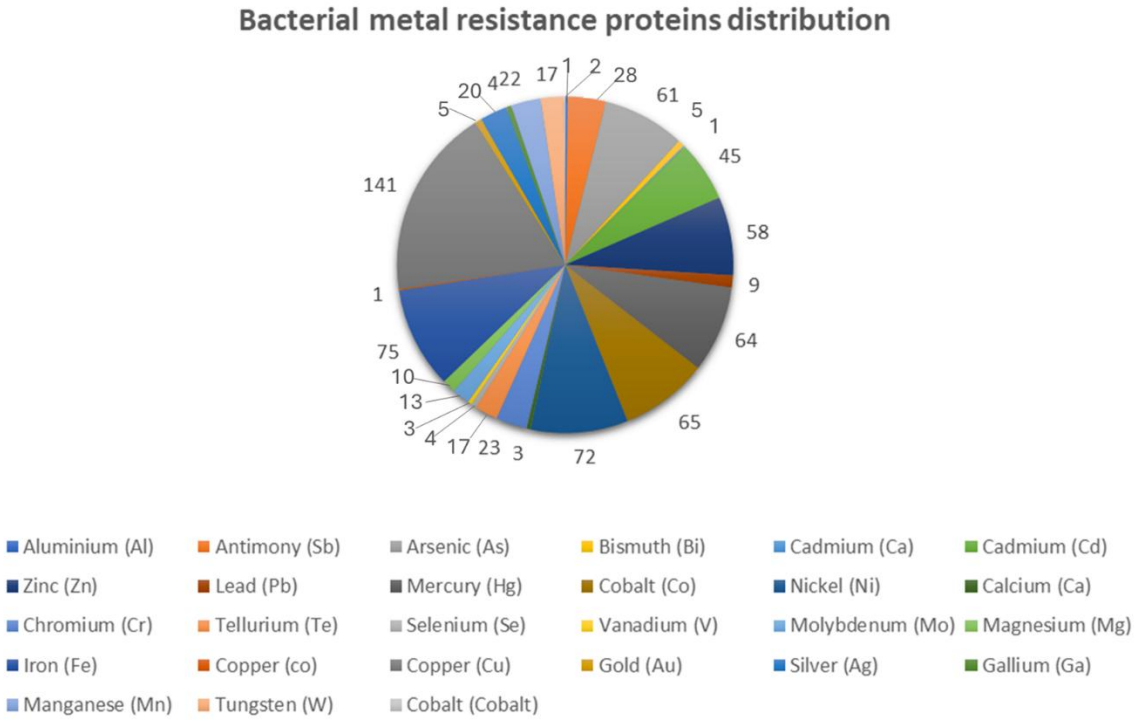


Figure 9 Distribution of bacterial metal resistance-associated proteins across target elements. Pie chart showing the number of protein entries associated with each of the 27 metals and metalloids included in the dataset.

4.2 Fungal metal resistance proteins

The fungal metal resistance protein dataset was developed to support the functional annotation of fungal isolates with potential for inclusion in biosystems operating under metal stress. Fungal species, especially those associated with plant roots or soil environments, are increasingly recognized for their

role in pollutant degradation, yet their metal tolerance mechanisms remain less characterized than in bacteria. To address this gap, a curated set of **730 protein sequences** was compiled, combining annotated entries from UniProt[1] with additional sequences extracted from the scientific literature[36–39]. Literature-derived proteins were selected based on experimental evidence of involvement in resistance to diverse metals. The dataset includes metal-binding proteins, efflux transporters, oxidative stress response enzymes, and regulators associated with metal homeostasis. A large proportion of the proteins in this dataset are associated with arsenic (As) and cadmium (Cd) resistance, reflecting the greater availability of annotated sequences for this metal (Figure 10). Other frequently represented elements include zinc (Zn), copper (Cu) and iron (Fe), which are also known to impose selective pressures on fungal communities in contaminated environments[40,41]. •

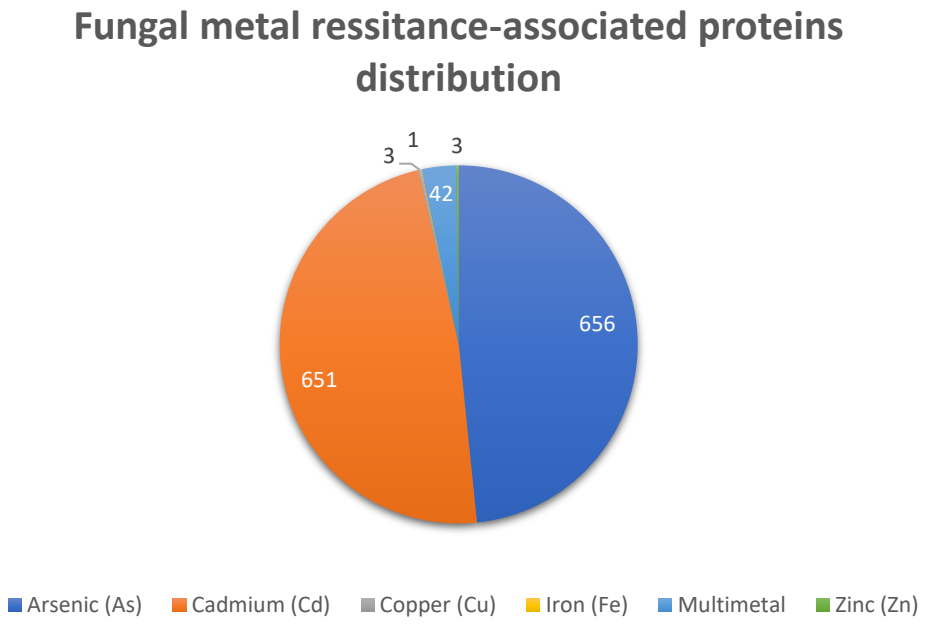


Figure 10 Distribution of fungal metal resistance-associated proteins across target elements. Pie chart showing the number of protein entries associated with each of the 6 metals and metalloids included in the dataset.

4.3 Plant metal resistance proteins

. Plants are key actors in phytoremediation strategies and their ability to tolerate, accumulate, or detoxify metals depends on diverse protein functions spread across multiple pathways and tissues. Given the physiological complexity of plants and their functional specialization across tissues a broader dataset was required to capture the diversity of metal tolerance /resistance mechanisms present across this heterogeneous group. To capture this diversity, a large dataset of 139,259 protein sequences was collected from UniProt and the specialized database PlantPres, which compiles plant proteins with predicted subcellular localization and stress-related roles. The selection focused on proteins involved in metal transport, chelation, oxidative stress response, and regulatory processes related to homeostasis under toxic metal exposure. While not all sequences have been experimentally validated, the dataset was curated to include only those with functional annotations or predictive features relevant to resistance. This collection provides a valuable reference for exploring traits associated with 8 metals collected in Figure 11 in plant genomes and supports the identification of candidate genes for improving phytoremediation performance or facilitating plant-microbe interactions in contaminated environments.

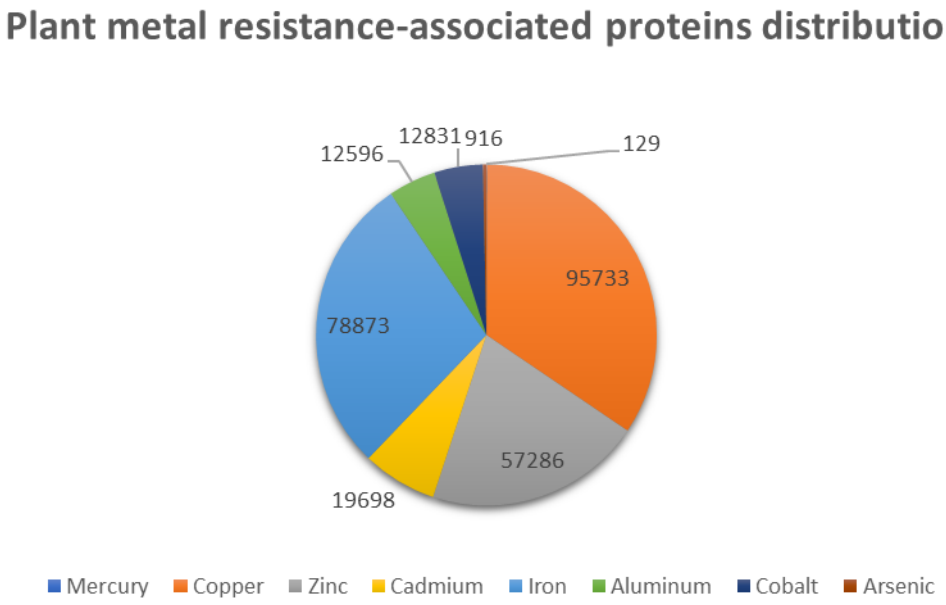


Figure 11. Distribution of Plant metal resistance-associated proteins across target elements. Pie chart showing the number of protein entries associated with each of the 8 metals and metalloids included in the dataset.

5 BIOSYSMOdb and complementary datasets application

The creation of BIOSYSMOdb represents a key step toward organizing and harmonizing biological data relevant to pollutant degradation, providing a structured and standardized resource with broad applicability in environmental genomics and functional annotation. In parallel, the development of complementary datasets (focused on metal resistance proteins in bacteria, fungi, and plants) extends the functional scope of the resource beyond the SQL schema, allowing targeted analyses on traits that are often underrepresented in standard repositories. Within BIOSYSMO, this resource enables the development of targeted bioinformatic tools and supports the design of synergistic biosystems for pollutant degradation. The following sections illustrate practical examples of how BIOSYSMOdb and its complementary datasets have been applied in the project, including the generation of HMM-based functional models and the integration of taxonomic and functional data to predict degradation potential from 16S rRNA sequences.

5.1 Practical example 1 - HMM design and creation

5.1.1 Introduction

By leveraging the modular design of BIOSYSMOdb, both the core protein datasets and the complementary collections specifically curated for metal-related functions have been exploited for the generation of Hidden Markov Models (HMMs). HMMs are probabilistic models that describe a system of hidden states and their associated outputs, and in the context of bioinformatics they are widely used to capture the statistical patterns of conserved regions within protein or DNA sequences. In BIOSYSMOdb, these models are representative of specific biological functions and enable the inference of functional capabilities from metabarcoding data (based on 16S rRNA sequencing), whole genome sequencing (WGS), or metagenome data. HMMs thus constitute a critical asset for downstream functional annotation approaches, allowing the sensitive detection of functionally related sequences and supporting the identification of functional traits in both cultivated isolates and

environmental communities relevant to pollutant degradation and resistance. This strategy ensures that even under-characterized proteins can be assigned to putative functional groups, thereby maximizing the exploitation of available biological data for the BIOSYSMO project.

5.1.2 Methodology

The generation of functional HMMs was carried out through a dedicated computational pipeline designed to extract the maximum functional value from the protein datasets compiled in BIOSYSMO, including both degradation-related proteins from BIOSYSMOdb and resistance-associated sequences from complementary datasets presented in previous section. As shown in Figure 12, the workflow was structured in two main stages: model construction and model verification.

The first stage began with the functional annotation of all protein sequences using EggNOG-mapper[42] and KOfamKOALA[43], enabling the identification of known orthologous groups and KEGG Orthology (KO) terms. Proteins with direct KO matches were assigned to their corresponding reference models. Sequences lacking KO annotations were subjected to unsupervised clustering based on sequence similarity (MMSeq2[44]), to group them into putative functional families. Each resulting cluster was used to generate a custom HMM profile using HMMER tools[45] . In parallel, a consensus amino acid sequence was derived for each cluster, capturing the most representative features of the alignment and supporting further functional inference.

To ensure quality and functional reliability, all resulting HMMs underwent a structured triple verification protocol (Figure 12, right panel), in which the original protein sequences, the consensus sequences, and the constructed HMMs were cross-compared. This step was critical for validating the specificity and internal consistency of each model prior to inclusion in the final collection. The resulting set of validated HMMs and their corresponding consensus sequences are now ready for use in downstream applications such as genome screening, functional trait detection, or the annotation of experimental metagenomic and amplicon data.

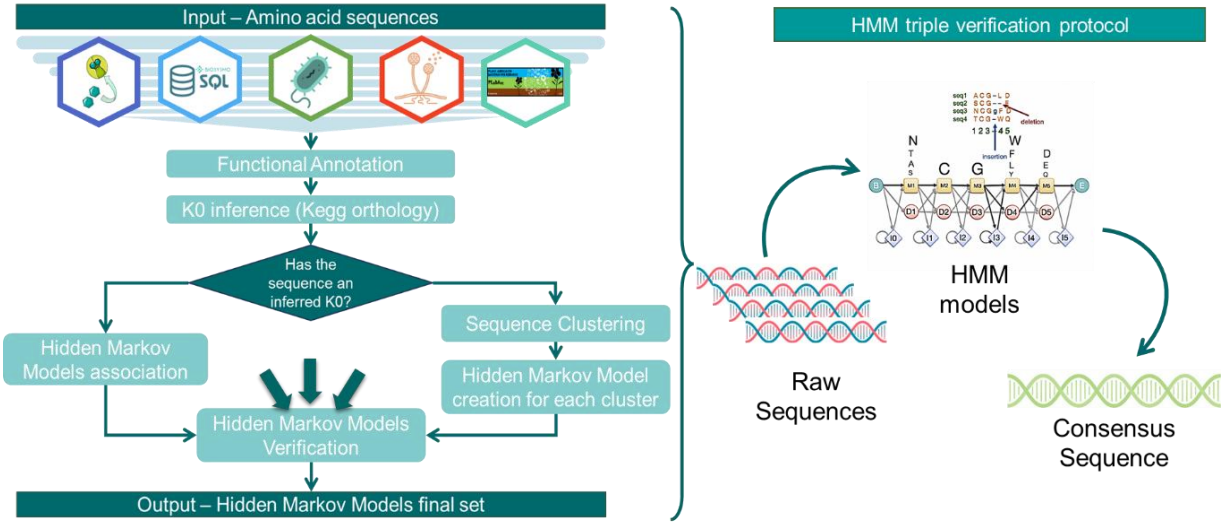


Figure 12 Overview of the HMM construction and validation pipeline. Left: generation of HMMs from direct annotation (using KEEG Orthology) or clustered protein sequences. Right: verification protocol comparing raw input sequences, HMM-derived consensus sequences, and the resulting models to ensure coherence and reliability.

5.1.3 Example of HMM space for BIOSYSMOdb protein dataset construction, evaluation and application - Results

HMM construction statistics

To illustrate how protein sequences were progressively classified through the pipeline, Figure 13 shows the distribution of proteins across each step. On the left, the BIOSYSMOdb protein space is divided into two groups: sequences directly assigned to known KEGG Orthology terms (in blue), and unassigned sequences (in grey) that underwent clustering. Among the latter, a majority were successfully grouped into homologous clusters (dark green), while a subset remained as singletons (light green), reflecting unique or poorly conserved proteins. The right panel summarizes the resulting set of Hidden Markov Models, indicating how many were inherited from reference KOs and how many were generated de novo from clustered sequences.

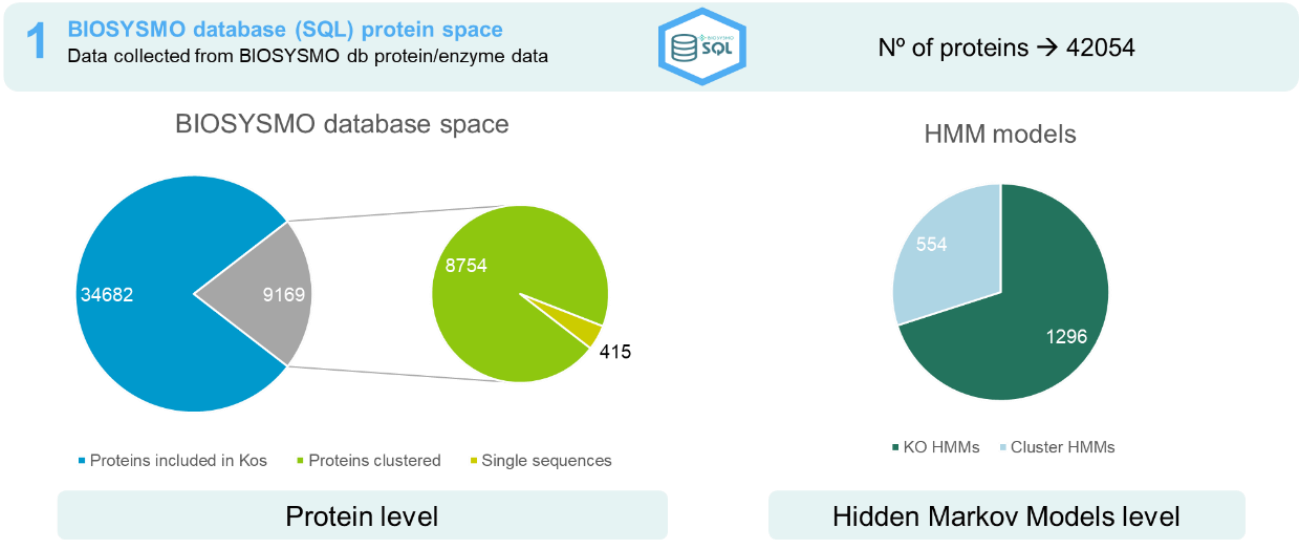


Figure 13 Distribution of protein sequences and resulting HMMs. Left: classification of sequences from BIOSYSMOdb into KO-assigned, clustered, or singletons. Right: total HMMs generated, grouped by origin (KEGG-assigned or clustered).

HMM validation statistics

To ensure the quality and functional specificity of the HMM models generated, a triple validation protocol was applied. This process aims to assess the coherence between the raw sequences, the derived consensus sequences, and the corresponding HMMs. Validation was performed at three levels:

- **Raw sequences vs. assigned HMMs**: Each input protein was aligned back to its assigned HMM profile using hmmer tools. A strong alignment indicates that the HMM accurately captures the conserved features of the sequence cluster and that the classification is robust.
- **Consensus sequences vs. HMMs**: Consensus sequences were aligned to all available HMMs to verify model uniqueness and specificity. Ideally, each consensus should align best to the HMM it was generated from. High-quality models should exhibit high scores for their own consensus and low or negligible matches to others.
- **Raw sequences vs. consensus sequences**: Finally, raw sequences were aligned against their respective consensus sequences using blastp to verify that the consensus faithfully represents the sequence diversity within its cluster. A correct match confirms the internal consistency of the model and supports its biological relevance.

This multi-layered approach allows identifying potential false positives, model overlaps, or overly general profiles, ensuring that only high-confidence HMMs are retained for downstream applications like protein mining, genome screening, or functional annotation from metagenomic data.

Raw sequences vs. assigned HMMs

Once the evaluations had been carried out, the results were analyzed in an appropriate manner for each specific comparison. Figure 14 contrasts the global performance of the HMM models against raw protein sequences with and without singleton clusters. All five metrics are high, and improve slightly when singletons are excluded, indicating that one-member clusters add variability and modestly depress scores. The gap between Top-1 (% of sequences whose best hit is their expected HMM) and Top-3 (% of sequences with your expected HMM correctly associated with the Top-3 hits) shows that, for a subset of sequences, the correct HMM is present but not always ranked first, which is consistent with potential competition among closely related models. The Mean Reciprocal Rank (MRR) summarizes this ranking quality in a single number: it averages 1 divided by the rank of the first correct hit (so it is 1.0 if the correct HMM is always first and approaches 0 when it rarely appears or ranks low). In our results, MRR sits between Top-1 and Top-3, as expected, confirming that most correct assignments are near the top. Overall, the models perform well; further gains are likely from tightening thresholds or disentangling families that are frequently confused.

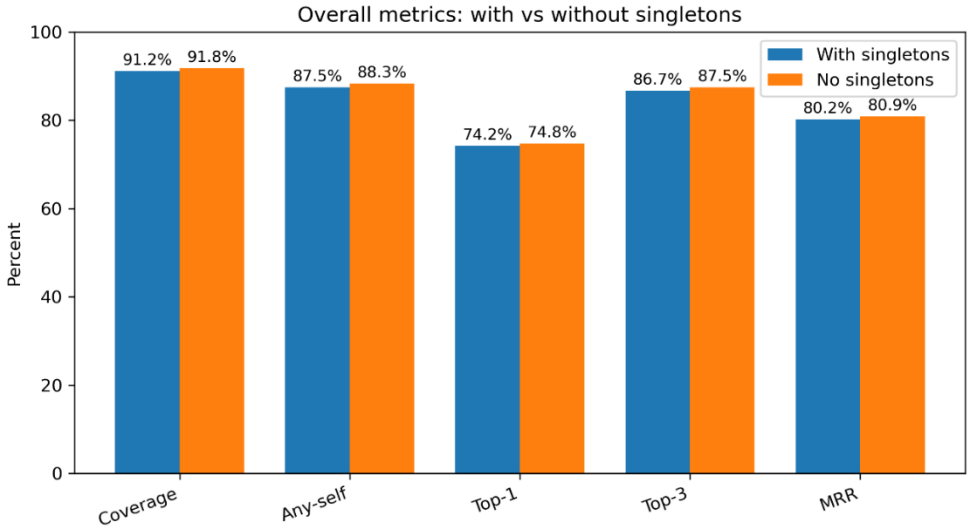


Figure 14 Overall HMM-vs-raw performance with and without singletons. Bars show percentages for five metrics computed from hmmscan results: Coverage = fraction of sequences with any hit; Any-self = fraction with at least one hit to their expected HMM; Top-1 = fraction whose best hit is the expected HMM; Top-3 = fraction where the expected HMM appears among the top three hits; MRR = mean reciprocal rank of the expected HMM (0 if absent). Blue bars include all clusters; orange bars exclude singleton clusters (size = 1). Improvements after removing singletons indicate these groups behave differently (e.g., weaker signal or noisier assignments).

To further interpret this results, the scatterplot shows (Figure 15), for each non-singleton cluster, how often its sequences are ranked first by their own HMM. The x-axis is cluster size (number of sequences in the ascluster). The y-axis is Top-1 self (%): among all sequences in that cluster, the percentage whose best hit is the cluster's expected HMM.

Many clusters sit at or near 100%, meaning their HMM cleanly wins for almost every member, these are well-separated families. A second group appears between ~60–95%, where the correct HMM is often present but sometimes loses the top rank to close competitors; these are candidates for potential

mild threshold tuning. The orange points highlight clusters where the correct HMM does not yet rank first as often as desired. This is common in large-scale profiling and usually reflects one of three benign causes: (i) closely related families that share motifs, (ii) multi-domain proteins where neighboring domains can affect the evaluation , or (iii) very small training sets. These cases are being addressed through routine refinement: reviewing seed alignments and consensus profiles, checking domain composition (e.g., with Pfam), calibrating family-specific Gathering Treshold (GA) thresholds (the curated bitscore threshold used to accept hits from that HMM), and examining the most frequent competing HMM to adjust model boundaries. When two models consistently compete, we assess whether to split or merge families. These steps typically improve rank-1 assignments while preserving sensitivity. Importantly, lower Top-1 values in large clusters point to a modeling boundary we can sharpen (in this case are the less common), whereas low values in very small clusters often reflect limited evidence and are expected to improve as more examples become available.

Figure 15 Cluster size vs. Top-1 self (%). Each point is a non-singleton cluster. X-axis: cluster size (n_members). Y-axis: Top-1 self (%) = fraction of the cluster's sequences whose best hmmscan hit is the cluster's expected HMM. Orange points highlight the lowest-performing decile by Top-1 self.

Consensus sequence vs. assigned HMM

As has been commented at the beginning of the section, each HMM has one consensus sequence derived from its multiple alignment. We scan all consensus sequences against the full correspondent HMM library (e.g., For a consensus named H-consensus, the expected model is H). Performance is assessed per consensus (one per cluster) using similar metrics as in previous case: Any-self (does its expected HMM appear anywhere among hits?), Top-1 (is the expected HMM the best hit?), Top-3, and MRR (0 if absent). Singletons are not part of this analysis because a single sequence cannot yield a consensus. Figure 16 shows the result of this evaluation. The bars show that consensus sequences almost perfectly map back to their own models: Any-self = 100%, Top-1 = 99.94% Top-3 = 100%, and MRR = 99.97%. In other words, for virtually every consensus, its originating HMM is ranked first; in the handful of cases where it is not, the expected model still appears within the top three, indicating very close competition rather than misspecification. This is the expected behaviour if models are well calibrated and the consensuses faithfully capture the family signal.

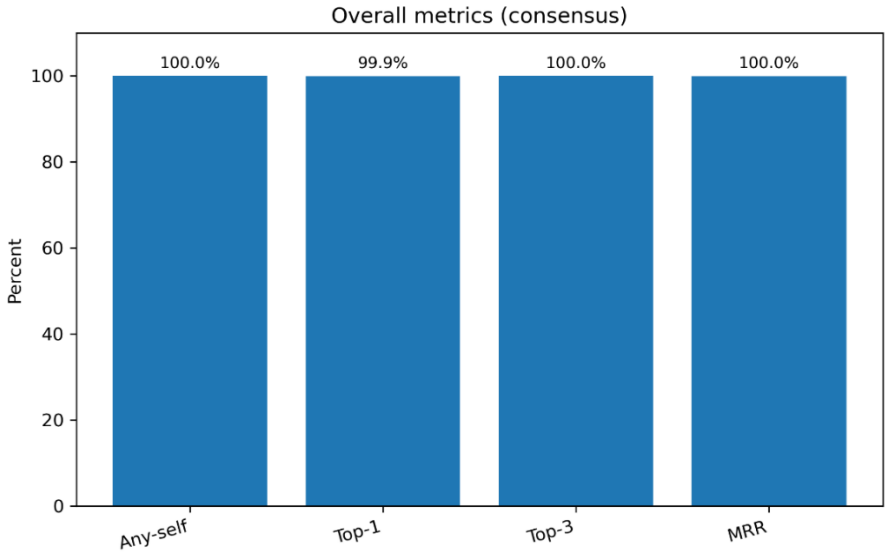


Figure 16 Overall HMM-vs-consensus performance. Bars report percentages across all consensus queries (values rounded to one decimal): Any-self = fraction where the expected HMM appears among hmmscan hits; Top-1 = fraction where the expected HMM is the best hit; Top-3 = fraction where it appears in the top three; MRR = mean reciprocal rank of the expected HMM. Exactly one consensus per cluster is evaluated; singletons are excluded by design.

Raw sequences vs. Consensus sequences

Scanning with HMMs tells us whether each sequence carries the profile signal of its family. Aligning raw proteins to their family consensus with BLAST asks a complementary question: *"Do individual members look most similar to the representative sequence of their own cluster?".* This check validates that the consensus actually represents its cluster in sequence space, and it helps diagnose cases where the HMM recognizes a member (Raw sequences vs HMMs) but the pairwise similarity to the consensus is modest (e.g., broad or multi-domain families, or overly narrow consensus). In short, HMMs test model vs sequence, whereas BLASTp usage tests raw sequence vs representative sequence based on homology alignment.

Figure 17 summarizes performance across non-singleton clusters, the ones that have a consensus sequence generated, based on BLASTp alignment results. Coverage bar (~71%) shows that about seven in ten proteins reach at least one accepted consensus match. Of these, roughly two thirds map to their own consensus at least once (Any-self, ~64%), and in about 60% of cases their own consensus is the top alignment (Top-1). The Top-3 (~64%) and MRR (~62%) indicate that even when the correct consensus does not rank first, it typically appears near the top. These percentages are lower than the ones showed in the HMM plots for two simple reasons. First, BLASTp compares each protein directly to one sequence (the consensus) so it rewards pairwise identity and aligned length. HMMs compare the protein to a profile built from many sequences, which is more sensitive to distant members; a protein can fit the model well yet be too different from the single consensus sequence to pass the identity/coverage cut-offs. Second, BLASTp needs enough aligned length: multi-domain or truncated proteins, or very diverse families, often match only part of the consensus and therefore fall below the coverage threshold. In such cases a neighboring family's consensus may look slightly closer and obtain better alignment statistics. In short, the BLASTp figure reflects a stricter pairwise check that naturally yields lower percentages but complements the stronger, more sensitive HMM results.

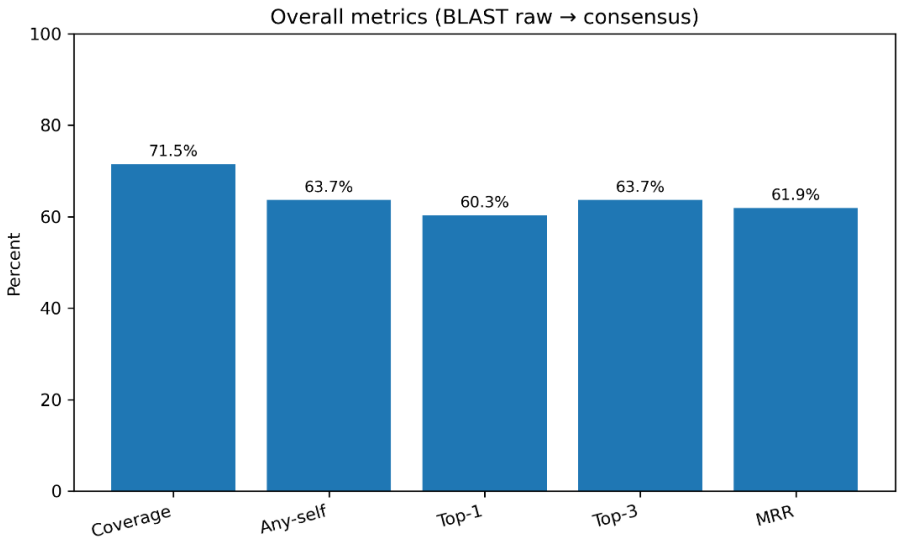


Figure 17 Overall BLAST performance from raw proteins to cluster consensuses (non-singleton families). Bars show Coverage, Any-self, Top-1, Top-3, and MRR; accepted hits require pident ≥ 50% and qcovs ≥ 70%.

As complement to this first insights. Figure 18 asks whether the consensus remains representative as families grow. Each dot is a non-singleton cluster; the x-axis is the number of members in the cluster, and the y-axis is the median identity of its sequences to their own consensus, computed over accepted BLAST self-hits. A dense cloud of small clusters with high medians (many ≳ 90%) can been appreciated, indicating that when families are small the consensus typically captures them very well. As clusters become larger, and therefore more diverse, the medians spread and shift downward (often in the 50–80% range). This is expected: a single representative cannot reflect the full variability of broad families as closely as it does in compact ones. Importantly, some large clusters still keep medians comfortably above the thresholds (see top of y axis), showing that their consensuses generalize well. Clusters with very large sizes and medians near the threshold are priorities for routine refinement.

Figure 18 One dot per non-singleton cluster; X: cluster size (*n_members*). Y: median BLAST identity of members to their own consensus. Higher medians indicate a consensus that represents its family closely in sequence space; lower medians in clusters reflect expected within-family diversity

HMM example of application in project generated data

While the HMM validation protocol described above was designed to ensure model quality using internally generated protein clusters, its practical relevance becomes particularly evident when applied to real project data. One such case study involved the functional screening of candidate proteomes inferred from 16S rRNA sequencing data provided by CIIMAR and isolated from an estuarine site situated in Oporto (Portugal). In these location, microbial isolates were obtained from samples exposed to pharmaceutical contamination, including the antidepressant paroxetine included in this example, and were preliminarily linked to draft genome candidates through taxonomic profiling.

In this context, predicted proteomes were used to infer degradation-related traits by aligning their sequences against the raw protein sequences compiled in BIOSYSMOdb. Although the HMMs constructed from these raw sequences were not directly used in the screening, it was later possible to verify whether the functionalities assigned to the candidate proteins aligned with the functional annotations of the HMM clusters to which those raw sequences belong. This offered an opportunity to retrospectively evaluate the consistency and functional relevance of the HMMs under real environmental conditions.

A specific example is illustrated in Figure 19, corresponding to sample DPS5, a microbial isolate preliminarily identified as *Acinetobacter johnsonii* based on 16S rRNA taxonomic assignment. This isolate was recovered from a site with known paroxetine contamination and selected for trait analysis based on its potential relevance. The figure displays the top 15 KEGG pathways inferred through functional annotation of its candidate proteome using BIOSYSMOdb raw sequences. Among the most represented traits were benzoate degradation, degradation of aromatic compounds, chloroalkane and chloroalkene degradation, and cytochrome P450-associated pathways such as drug metabolism and xenobiotic transformation, all potentially involved in paroxetine degradation[46–48].

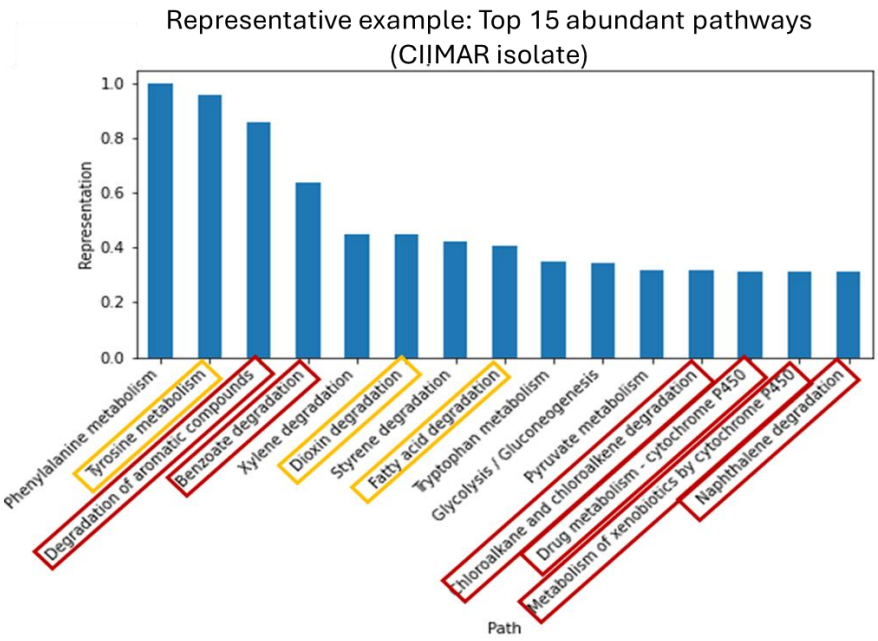


Figure 19 Top 15 KEGG pathways by relative abudance inferred from candidate proteomes in sample DPS5. Bar plot showing the relative representation (normalized per sample) of the most abundant metabolic pathways identified after functional annotation of candidate proteins inferred from 16S rRNA sequencing data. Pathways were assigned based on KEGG-based annotations derived from the BIOSYSMOdb raw sequences. Red-framed labels indicate traits of high interest due to their potential relevance in paroxetine degradation; orange-framed labels denote traits with medium potential relevance.

To further verify the relevance of the constructed HMMs, the annotations obtained from the raw sequences used in this functional screening with the clusters to which those sequences belong were compared. Table 2 summarizes the relationship between candidate protein matches, their matching raw sequences, the traits they were associated with, and the HMM cluster in which they are classified. The comparison revealed a high degree of consistency: proteins that contributed to the identification of relevant traits (e.g., benzoate degradation) were assigned to HMM clusters carrying equivalent annotations (e.g., benzoate/toluate 1,2-dioxygenase subunit alpha).

This convergence reinforces the validity of the HMMs, confirming that the models accurately reflect the functional identity of the sequences from which they were built, and are reliable tools for inferring metabolic capabilities in environmental samples. In summary, this example illustrates how integrating functional screening results with the HMM-based framework provides a robust approach for trait inference and model validation, reinforcing the value of BIOSYSMOdb both as a protein reference space and as a foundation for HMM-assisted functional annotation pipelines.

Table 2 Functional traits detected in example's candidate proteome and associated HMM clusters. The table shows matches between candidate proteins in presented sample and BIOSYSMOdb raw sequences, including the inferred KEGG pathway (Annotated Path/Trait) and the assigned HMM cluster.

Candidate Protein ID	Query (Raw sequence)	Annotated Path/Trait	Path/Trait name	Assigned HMM (cluster ID)	Cluster annotation
WP_004693190.1	P07769	map00362	Benzoate degradation	K05549	benzoate/toluate 1,2-dioxygenase subunit alpha
WP_004979732.1	P07770	map00362	Benzoate degradation	K05550	benzoate/toluate 1,2-dioxygenase subunit beta
WP_004979734.1	P07771	map00362	Benzoate degradation	K05784	benzoate/toluate 1,2-dioxygenase reductase component
WP_004979736.1	P07772	map00362	Benzoate degradation	K05783	dihydroxycyclohexadiene carboxylate dehydrogenase
WP_004979740.1	A0A0C5RP30	map00362	Benzoate degradation	K01031	3-oxoadipate CoA-transferase, alpha subunit
WP_004979749.1	Q59091	map00362	Benzoate degradation	K01032	3-oxoadipate CoA-transferase, beta subunit
WP_004979750.1	Q43935	map00362	Benzoate degradation	K00626	acetyl-CoA C-acetyltransferase
WP_004979750.1	Q43935	map00362	Benzoate degradation	K00626	acetyl-CoA acyltransferase
WP_004979750.1	Q43935	map00362	Benzoate degradation	K00626	3-oxoadipyl-CoA thiolase
WP_004981405.1	Q43935	map00362	Benzoate degradation	K00626	acetyl-CoA C-acetyltransferase
WP_004981405.1	Q43935	map00362	Benzoate degradation	K00626	acetyl-CoA acyltransferase
WP_004981405.1	Q43935	map00362	Benzoate degradation	K00626	3-oxoadipyl-CoA thiolase
WP_004981409.1	A0A0C5RP30	map00362	Benzoate degradation	K01031	3-oxoadipate CoA-transferase, alpha subunit
WP_004693257.1	P25437	map00625	Chloroalkane and chloroalkene degradation	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase

WP_004693190.1	P07769	map01220	Degradation of aromatic compounds	K05549	benzoate/toluate 1,2-dioxygenase subunit alpha
WP_004979732.1	P07770	map01220	Degradation of aromatic compounds	K05550	benzoate/toluate 1,2-dioxygenase subunit beta
WP_004979734.1	P07771	map01220	Degradation of aromatic compounds	K05784	benzoate/toluate 1,2-dioxygenase reductase component
WP_004979736.1	P07772	map01220	Degradation of aromatic compounds	K05783	dihydroxycyclohexadiene carboxylate dehydrogenase
WP_004693257.1	P25437	map01220	Degradation of aromatic compounds	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
WP_004693257.1	P25437	map00982	Drug metabolism - cytochrome P450	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
WP_004979750.1	Q43935	map00071	Fatty acid degradation	K00626	acetyl-CoA C-acetyltransferase
WP_004979750.1	Q43935	map00071	Fatty acid degradation	K00626	acetyl-CoA acyltransferase
WP_004693257.1	P25437	map00071	Fatty acid degradation	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
WP_004981405.1	Q43935	map00071	Fatty acid degradation	K00626	acetyl-CoA C-acetyltransferase
WP_004981405.1	Q43935	map00071	Fatty acid degradation	K00626	acetyl-CoA acyltransferase
WP_004979750.1	Q43935	map01212	Fatty acid metabolism	K00626	acetyl-CoA C-acetyltransferase
WP_004979750.1	Q43935	map01212	Fatty acid metabolism	K00626	acetyl-CoA acyltransferase
WP_004981405.1	Q43935	map01212	Fatty acid metabolism	K00626	acetyl-CoA C-acetyltransferase
WP_004981405.1	Q43935	map01212	Fatty acid metabolism	K00626	acetyl-CoA acyltransferase
WP_004979750.1	Q43935	map00630	Glyoxylate and dicarboxylate metabolism	K00626	acetyl-CoA C-acetyltransferase
WP_004981405.1	Q43935	map00630	Glyoxylate and dicarboxylate metabolism	K00626	acetyl-CoA C-acetyltransferase
WP_171065396.1	A0A485JME7	map00630	Glyoxylate and dicarboxylate metabolism	K01915	glutamine synthetase
WP_004693257.1	P25437	map00980	Metabolism of xenobiotics by cytochrome P450	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
WP_004693257.1	P25437	map00626	Naphthalene degradation	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase
WP_004693257.1	P25437	map00350	Tyrosine metabolism	K00121	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase

5.1.4 Conclusion, utility of the protocol, importance of the HMM strategy.

The pipeline developed for generating and validating Hidden Markov Models (HMMs) from BIOSYSMOdb protein sequences has proven to be a valuable tool for functional annotation. Thanks to this approach, it is possible to detect traits linked to pollutant degradation and resistance, even in proteins that lack prior annotations.

By grouping similar proteins and building HMMs that represent their conserved features, the system allows for the identification of functional traits in data derived from 16S rRNA, whole genomes or metagenomes. The triple validation protocol ensures that each model is specific and reliable, increasing confidence in downstream analyses. Importantly, when these models were applied to real case studies, the traits inferred using raw protein sequences matched the traits associated with their corresponding HMM clusters. This confirms that the HMMs accurately represent the biological functions they were designed to detect.

Overall, this strategy reinforces BIOSYSMOdb not only as a high-quality protein database, but also as a foundation for building sensitive and versatile tools to support functional screening in bioremediation research.

5.2 Practical example 2 - Prediction of metal resistance potential from 16S rRNA data

5.2.1 Introduction

As part of the BIOSYSMO project, UBU conducted a generalized analysis of the microbial communities present at an industrial site ubicated in Sabiñánigo (Aragón, Spain) where pollution is primarily caused by metals and metalloids. These communities, shaped by environmental stressors such as arsenic (As), copper (Cu), nickel (Ni), zinc (Zn), and elevated iron and sulfate levels, represent a valuable model for investigating natural microbial adaptation under toxic conditions. Taxonomic profiling was performed via 16S rRNA amplicon sequencing, providing a foundational overview of community composition and enabling the selection of key Amplicon Sequence Variants (ASVs) for further functional exploration.

In this case study, the downstream functional screening was not performed using HMM-based models but rather by direct alignment of predicted proteomes to the curated complementary datasets of metal resistance proteins developed in Task 2.1. These compact, manually curated protein collections offer an agile and reliable resource for identifying relevant functional traits in environmental isolates. In case of bacterial metal resistance-associated dataset, its reduced size and high annotation quality make it especially suitable for preliminary screenings, guiding the prioritization of candidates for genome sequencing or experimental validation.

This practical example highlights the flexibility and complementarity of BIOSYSMO's data architecture, demonstrating how different components of the resource can be strategically employed to meet specific project needs. By leveraging the bacterial metal resistance-associated dataset, the analysis enabled the inference of potential resistance mechanisms in candidate genomes associated with the dominant ASVs. The results provided critical insights into the adaptive strategies of microbial communities under metal stress and supported the identification of promising isolates for subsequent experimental characterization.

5.2.2 Methodology

This specific analysis began with the taxonomic characterization of microbial communities using 16S rRNA amplicon sequencing. This technique allows researchers to identify the microbial genus/species present in a sample by targeting a highly conserved region of their ribosomal RNA genes. The sequencing results are processed into Amplicon Sequence Variants (ASVs), which are unique DNA sequences representing distinct microbial populations. Unlike traditional methods that group sequences into broad taxonomic bins (Operational Taxonomic Units, or OTUs), ASVs offer higher resolution, enabling finer differentiation between closely related organisms.

Each ASV was compared against the NCBI nucleotide database using a multi-tool taxonomy annotation protocol to infer the most likely genome candidates associated with the detected sequences (Figure 20). From these genome references, the corresponding protein sets (i.e., predicted proteomes) were downloaded using the NCBI Datasets resource.

Once the candidate proteomes were retrieved, each one was screened against the bacterial metal resistance-associated protein dataset described in section 4.1. This alignment was performed using BLASTp, which compares protein sequences to detect similarities and infer potential functional traits. The goal was to identify conserved domains or sequence similarities that may suggest the presence of resistance mechanisms to specific metals.

The resulting matches were processed and formatted into structured tables for interpretation. In parallel, visualizations were generated to support comparative analysis across samples, allowing researchers to evaluate which resistance traits were present and how they varied across different microbial isolates from the same contaminated site.

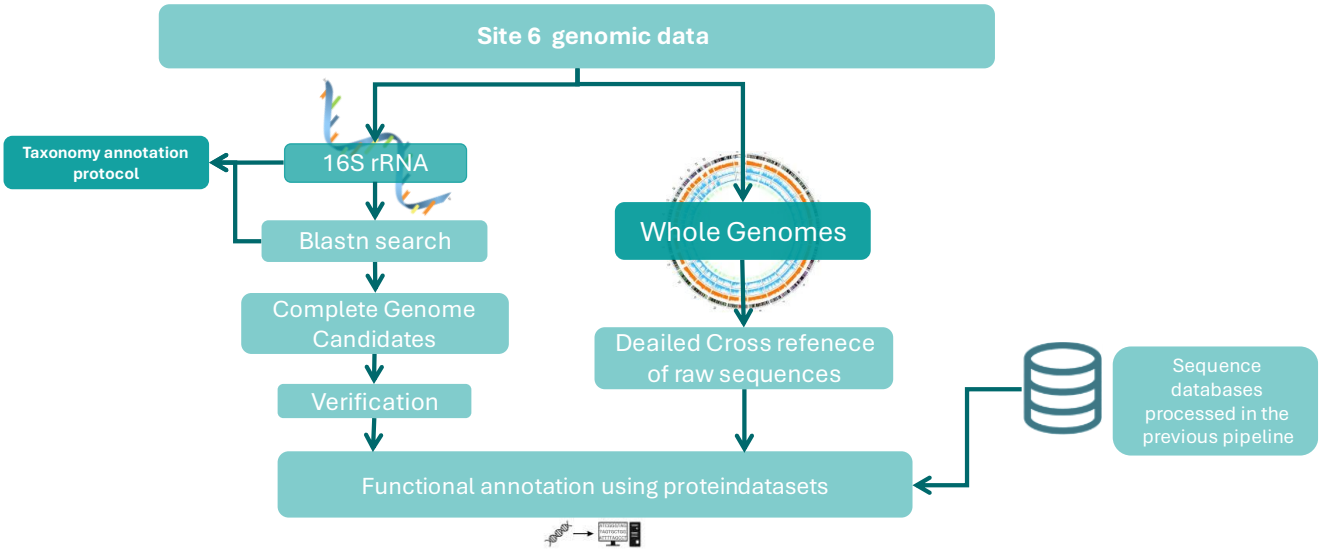


Figure 20. Complete protocol scheme

5.2.3 Functional annotation for metal resistance associated traits - Results

Representative example – UBU sample (metal/metalloid-contaminated site)

To illustrate an example of the pipeline application in this study, the control sample corresponding to *Pseudomonas fluorescens F113* has been selected, which served as a reference in the experimental setup that also included epiphytic and endophytic samples. This control was subjected to the full annotation protocol and is presented here as a representative example. The analysis of its three most abundant ASVs (each exceeding 12% relative abundance) revealed a high level of genomic similarity,

as all of them exhibited an identical profile of potential metal resistance mechanisms (Figure 21). Although these ASVs were taxonomically distinct and linked to different genome candidates, their conserved resistance repertoire suggests they share a closely related taxonomy.

This control sample displayed a specific profile, with copper resistance emerging as the most prominent trait (Figure 21). A total of eight copper-associated proteins were identified across all three dominant ASVs, passing all filtering criteria. These included copper transporter, copper resistance protein C, CopA, CopB, copper resistance protein D, transcriptional activator protein CopR, copper resistance protein C, and copper resistance protein B (Table 3. Unique metal resistance descriptors associated with each ASV candidate proteomes; “pident” represents the highest identity obtained between some candidate protein and a metal resistance protein collected in our dataset.Table 3). Copper resistance mechanisms in bacteria are well-characterized and often involve a combination of efflux transporters, periplasmic detoxification proteins, and transcriptional regulators. The presence of CopA and CopB, which function as P-type ATPase efflux transporters, suggests an active mechanism for exporting excess copper from the cytoplasm[52]. Additionally, the identification of CopR, a transcriptional activator, indicates the potential for regulated expression of copper resistance genes in response to metal exposure[52]. The prevalence of these mechanisms in all dominant ASVs suggests that copper represents a significant environmental pressure in this specific site, possibly due to natural copper bioavailability or plant-driven microbial selection.

Beyond copper, additional metal resistance mechanisms were identified, though at lower diversity. A nickel/cobalt efflux system was detected, which is known to provide resistance to nickel, cobalt, zinc, and cadmium through an energy-dependent efflux process[53]. This finding suggests that these metals, while less dominant than copper in shaping resistance traits, may still exert selective pressure. Additionally, a single DNA starvation/stationary phase protection protein (DpsA) was linked to iron resistance, potentially contributing to oxidative stress protection under iron-rich conditions[54].

The resistance profile of the selected sample contrasts sharply with that of others from the same site, where arsenic resistance was the only detected mechanism. This highlights the potential heterogeneity in metal resistance strategies across different endophytic isolates, suggesting that environmental or host-related factors may play a significant role in shaping microbial adaptation.

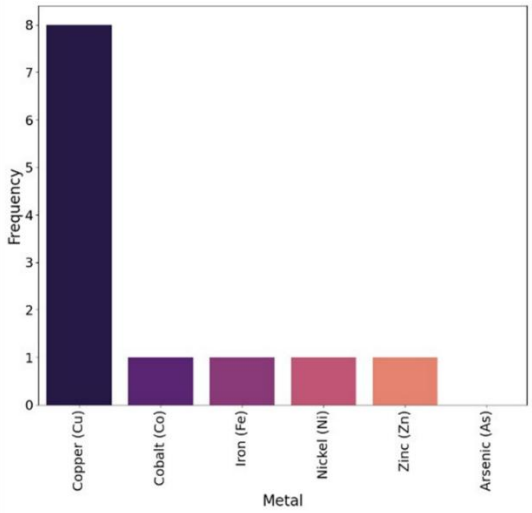


Figure 21. Number of different metal resistance mechanisms segmented by metal which aligns with the top 3 ASVs proteomes associated with control sample. The represented metals are based on study preferences. As the ASVs have an identical metal resistance profile, only one plot has been created for all of them.

Table 3. Unique metal resistance descriptors associated with each ASV candidate proteomes; "pident" represents the highest identity obtained between some candidate protein and a metal resistance protein collected in our dataset.

qseqid	pident	Matched-Protein (Uniprot)	Metal(s) tag	Protein Description
ASV23	98.996	Q8KWW2	Copper (Cu), silver (Ag)	Copper transporter
ASV23	97.581	Q7WYG9	Copper (Cu)	Copper resistance protein C
ASV23	97.522	Q7WYH1	Copper (Cu)	CopA
ASV23	95.088	Q7WYH0	Copper (Cu)	CopB
ASV23	91.409	Q7WYG8	Copper (Cu)	Copper resistance protein D
ASV23	82.301	Q02540	Copper (Cu)	Transcriptional activator protein CopR
ASV23	81.102	Q9KWM9	Copper (Cu)	Copper resistance protein C
ASV23	79.211	Q88IN1	Nickel (Ni), cadmium (Cd), zinc (Zn), cobalt (Co)	Nickel/cobalt efflux system
ASV23	78.351	P12375	Copper (Cu)	Copper resistance protein B
ASV23	78.289	Q8KR86	Iron (Fe), hydrogen peroxide (H_2O_2)	DNA starvation/stationary phase protection protein (DpsA)
ASV23	78.17	P12374	Copper (Cu), silver (Ag)	Copper resistance protein A
ASV23	76	P12376	Copper (Cu)	Copper resistance protein C
ASV34	100	Q8KWW2	Copper (Cu), silver (Ag)	Copper transporter
ASV34	99.194	Q7WYG9	Copper (Cu)	Copper resistance protein C
ASV34	98.053	Q7WYH1	Copper (Cu)	CopA
ASV34	95.439	Q7WYH0	Copper (Cu)	CopB
ASV34	94.845	Q7WYG8	Copper (Cu)	Copper resistance protein D
ASV34	84	Q02540	Copper (Cu)	Transcriptional activator protein CopR
ASV34	81.102	Q9KWM9	Copper (Cu)	Copper resistance protein C
ASV34	79.211	Q88IN1	Nickel (Ni), cadmium (Cd), zinc (Zn), cobalt (Co)	Nickel/cobalt efflux system
ASV34	78.351	P12375	Copper (Cu)	Copper resistance protein B
ASV34	78.289	Q8KR86	Iron (Fe), hydrogen peroxide (H_2O_2)	DNA starvation/stationary phase protection protein (DpsA)
ASV34	78.17	P12374	Copper (Cu), silver (Ag)	Copper resistance protein A
ASV34	76	P12376	Copper (Cu)	Copper resistance protein C
ASV9	98.996	Q8KWW2	Copper (Cu), silver (Ag)	Copper transporter
ASV9	97.581	Q7WYG9	Copper (Cu)	Copper resistance protein C
ASV9	97.522	Q7WYH1	Copper (Cu)	CopA
ASV9	95.088	Q7WYH0	Copper (Cu)	CopB
ASV9	91.409	Q7WYG8	Copper (Cu)	Copper resistance protein D
ASV9	84	Q02540	Copper (Cu)	Transcriptional activator protein CopR
ASV9	81.102	Q9KWM9	Copper (Cu)	Copper resistance protein C
ASV9	79.211	Q88IN1	Nickel (Ni), cadmium (Cd), zinc (Zn), cobalt (Co)	Nickel/cobalt efflux system
ASV9	78.351	P12375	Copper (Cu)	Copper resistance protein B

ASV9	78.289	Q8KR86	Iron (Fe), hydrogen peroxide (H_2O_2)	DNA starvation/stationary phase protection protein (DpsA)
ASV9	78.17	P12374	Copper (Cu), silver (Ag)	Copper resistance protein A
ASV9	76	P12376	Copper (Cu)	Copper resistance protein C

5.2.4 Conclusion, utility of the protocol, importance of the database

This case study illustrates the practical relevance of BIOSYSMO's curated protein datasets for environmental microbiology applications. By applying a streamlined pipeline to link 16S-derived candidate proteomes with metal resistance traits through direct protein alignment, the complementary bacterial protein dataset proved effective in identifying functional traits under environmental stress. The analysis enabled the comparison of isolates, revealing potential resistance profiles and taxon-specific adaptations. This confirms the value of curated complementary datasets as powerful tools for functional inference, especially in early-stage screenings or when working with isolates data. The flexibility of the BIOSYSMO data framework, spanning both a structured SQL database and specialized protein collections, offers an adaptable foundation for addressing diverse analytical challenges across microbial contexts.

6 Conclusions

The development of BIOSYSMOdb and its associated resources represents a major milestone in the consolidation of biodegradation-relevant biological knowledge within the BIOSYSMO project. By integrating heterogeneous data from public repositories, regulatory frameworks, and project-specific compilations into a harmonized and accessible SQL-based structure, BIOSYSMOdb provides a centralized, interoperable platform for querying and analyzing chemical, metabolic, enzymatic, and organismal information related to pollutant degradation and resistance. Beyond its core database, the parallel creation of curated complementary protein datasets, focused on metal resistance-associated traits in bacteria, fungi, and plants, has significantly expanded the functional scope of the resource, enabling detailed screening of traits often underrepresented in conventional repositories. Together, these resources support the identification and prioritization of microbial candidates, the annotation of genomic and metagenomic datasets, and the design of synergistic biosystems tailored to specific bioremediation challenges.

The deliverable also demonstrates how these resources have been effectively applied within the project through two practical case studies: (1) the generation and validation of Hidden Markov Models (HMMs) from BIOSYSMO protein datasets, and (2) the prediction of metal resistance-associated traits in microbial enriched isolates based on 16S-derived proteomes. These examples illustrate both the technical robustness and applied relevance of the BIOSYSMO data architecture.

In sum, BIOSYSMOdb not only provides a stable foundation for current computational tasks within the project but also establishes a flexible and extensible framework that can evolve alongside future experimental results, partner contributions, and external community needs. Its modular structure facilitates the seamless integration and updating of data both during and beyond the project's lifetime, ensuring long-term relevance and adaptability. Designed in accordance with FAIR principles, BIOSYSMOdb will continue to serve as a valuable and reusable asset for the broader scientific community engaged in environmental microbiology, bioremediation, and systems biology.

7 References

1. UniProt: the Universal Protein Knowledgebase in 2023 | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/51/D1/D523/6835362.

2. O'Leary, N. A. *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024).

3. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

4. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).

5. Ellis, L. B. & Wackett, L. P. Use of the University of Minnesota Biocatalysis/Biodegradation Database for study of microbial degradation. *Microb. Inform. Exp.* **2**, 1 (2012).

6. Division, I. A.-B., Gil-Guerrero, S., Otero-Gonzalez, L., Franco de Benito, M. & Hall, A. BIOSYSMOdb: Curated Database for Biodegradation and Bioremediation. Zenodo https://doi.org/10.5281/zenodo.14795254 (2025).

7. Li, C. *et al.* A Review on Heavy Metals Contamination in Soil: Effects, Sources, and Remediation Techniques. *Soil Sediment Contam. Int. J.* **28**, 380–394 (2019).

8. Hung, H. *et al.* Climate change influence on the levels and trends of persistent organic pollutants (POPs) and chemicals of emerging Arctic concern (CEACs) in the Arctic physical environment – a review. *Environ. Sci. Process. Impacts* **24**, 1577–1615 (2022).

9. Verma, A. Bioremediation Techniques for Soil Pollution: An Introduction. in *Biodegradation Technology of Organic and Inorganic Pollutants* (IntechOpen, 2021). doi:10.5772/intechopen.99028.

10. Ngara, T. R., Zeng, P. & Zhang, H. mibPOPdb: An online database for microbial biodegradation of persistent organic pollutants. *iMeta* **1**, e45 (2022).

11. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

12. ECHA website is unavailable. https://echa.europa.eu/information-on-chemicals.

13. CompTox Chemicals Dashboard. https://comptox.epa.gov/dashboard/.

14. PubChem 2025 update | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/53/D1/D1516/7903365?login=false.

15. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* **8**, 61 (2016).

16. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

17. Web Scraper - The #1 web scraping extension. https://webscraper.io/.

18. Beautiful Soup Documentation — Beautiful Soup 4.13.0 documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

19. pandas documentation — pandas 2.3.0 documentation. https://pandas.pydata.org/docs/.

20. Search for chemicals - ECHA. https://echa.europa.eu/information-on-chemicals.

21. The POPs. https://www.pops.int/TheConvention/ThePOPs/tabid/673/Default.aspx.

22. Directive - 2000/60 - EN - Water Framework Directive - EUR-Lex. https://eur-lex.europa.eu/eli/dir/2000/60/oj/eng.

23. Lista de sustancias candidatas extremadamente preocupantes en procedimiento de autorización - ECHA. https://echa.europa.eu/es/candidate-list-table.

24. PBT assessment list - ECHA. https://echa.europa.eu/es/pbt/-/dislist/details/0b0236e1809fea57.

25. NORMAN Database System. https://www.norman-network.com/nds/.

26. Olker, J. H. *et al.* The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment. *Environ. Toxicol. Chem.* **41**, 1520–1539 (2022).

27. Grulke, C. M., Williams, A. J., Thillanadarajah, I. & Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Comput. Toxicol.* **12**, 100096 (2019).

28. Helman, G. *et al.* Generalized Read-Across (GenRA): A workflow implemented into the EPA CompTox Chemicals Dashboard. *ALTEX - Altern. Anim. Exp.* **36**, 462–465 (2019).

29. guardrex. ASP.NET Core Blazor. https://learn.microsoft.com/en-us/aspnet/core/blazor/?view=aspnetcore-9.0.

30. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript | Journal of Chemical Information and Modeling. https://pubs.acs.org/doi/10.1021/acs.jcim.7b00425.

31. Cytoscape.js 2023 update: a graph theory library for visualization and analysis | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/39/1/btad031/6988031.

32. Huang, L., Yao, S.-M., Jin, Y., Xue, W. & Yu, F.-H. Co-contamination by heavy metal and organic pollutant alters impacts of genotypic richness on soil nutrients. *Front. Plant Sci.* **14**, (2023).

33. Mousavi, S. A. *et al.* PlantPReS: A database for plant proteome response to stress. *J. Proteomics* **143**, 69–72 (2016).

34. Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E. & Larsson, D. G. J. BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* **42**, D737–D743 (2014).

35. Bonin, N. *et al.* MEGARes and AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Res.* **51**, D744–D752 (2023).

36. Robinson, J. R., Isikhuemhen, O. S. & Anike, F. N. Fungal–Metal Interactions: A Review of Toxicity and Homeostasis. *J. Fungi* **7**, 225 (2021).

37. Chi, B.-B. *et al.* Sequencing and Comparative Genomic Analysis of a Highly Metal-Tolerant Penicillium janthinellum P1 Provide Insights Into Its Metal Tolerance. *Front. Microbiol.* **12**, (2021).

38. Branco, S., Schauster, A., Liao, H.-L. & Ruytinx, J. Mechanisms of stress tolerance and their effects on the ecology and evolution of mycorrhizal fungi. *New Phytol.* **235**, 2158–2175 (2022).

39. Urquhart, A. S., Chong, N. F., Yang, Y. & Idnurm, A. A large transposable element mediates metal resistance in the fungus *Paecilomyces variotii*. *Curr. Biol.* **32**, 937-950.e5 (2022).

40. Guarino, F., Improta, G., Triassi, M., Cicatelli, A. & Castiglione, S. Effects of Zinc Pollution and Compost Amendment on the Root Microbiome of a Metal Tolerant Poplar Clone. *Front. Microbiol.* **11**, (2020).

41. Oladipo, O. G., Awotoye, O. O., Olayinka, A., Bezuidenhout, C. C. & Maboeta, M. S. Heavy metal tolerance traits of filamentous fungi isolated from gold and gemstone mining sites. *Braz. J. Microbiol.* **49**, 29–37 (2018).

42. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale | Molecular Biology and Evolution | Oxford Academic. https://academic.oup.com/mbe/article/38/12/5825/6379734.

43. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

44. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).

45. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).

Annexes

Annex 1

All datasets integrated into BIOSYSMOdb were carefully reviewed to ensure compliance with reuse and redistribution terms. Table 4Table 4 summarizes the main source databases, their general statistics, and the corresponding licenses under which the data were accessed and incorporated. For complementary datasets (bacterial, fungal, and plant metal resistance proteins), licensing information is also reported. A complete list of license terms, database-specific attributions, and references is provided in this Annex, accompanying the SQL upload package on Zenodo.

Table 4 Original databases and related licenses

Database Name	General statistics of the original database	License
EAWAG	219 pathways 1,503 reactions 1,396 compounds 993 enzymes 543 microorganism entries 249 biotransformation rules	Licensed under **CC BY 4.0.** Open access for the EAWAG-BBD version. All website content and individual documents are protected by copyright and may be downloaded, copied, and printed exclusively for personal, educational, and non-commercial use.
MibPOP	32 target compounds 3,479 organisms 5,736 microbial-associated functional genes	Licensed under **CC BY-SA 4.0**. Database owners were contacted for BIOSYSMOdb integration; permission for data use was granted.
MetaCYC	3,063 pathways 18,285 reactions 18,622 metabolites	Free access to EcoCyc and MetaCyc was available at the time the data gathering process was performed (2023). Access to BioCyc is restricted and requires a subscription.
Brenda	8,331 different enzymes	Licensed under **CC BY 4.0**.

For complementary datsets

Bacterial metal resistance proteins

- **Sources and licensing**:
  - **BacMet**: Creative Commons Attribution 4.0 International CC BY 4.0
  - **UniProt**: Creative Commons Attribution 4.0 International CC BY 4.0
  - **MEGARes:** Creative Commons Attribution 4.0 International CC BY 4.0

Fungal metal resistance proteins

- **Source and licensing**:
    - **UniProt**: Creative Commons Attribution 4.0 International CC BY 4.0
    - **Literature-based entries** – Selected from peer-reviewed publications (no licensing restrictions apply to curated sequence metadata)

Plant metal resistance proteins

- **Sources and licensing**:
    - **UniProt**: Creative Commons Attribution 4.0 International CC BY 4.0
    - **PlantPres**: No licensing associated. (http://www.proteome.ir/ContactUs.aspx)