



**Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.**





<b>Deliverable number:</b>	D6.13
<b>Work package:</b>	WP6 – Terrestrial Biosphere
<b>Intermediate Objective:</b>	IO6
<b>Deliverable type:</b>	<input checked="" type="checkbox"/> Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify .....
<b>Dissemination level:</b>	<input checked="" type="checkbox"/> Public
	<input type="checkbox"/> Restricted
<b>Estimated delivery (bimester):</b>	B13
<b>Actual delivery date:</b>	31/12/2024
<b>Author(s) (Partner-OU):</b>	Nicola Curci, Mauro Di Fenza, Federica De Lise, Angela Capaccio, Beatrice Cobucci-Ponzano.
<b>Reviewed by:</b>	
<b>Note:</b>	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)

Funded by EU - Next Generation EU

Mission 4 “Education and Research” - Component 2: “From research to business” -

Investment 3.1: “Fund for the realization of an integrated system of research and innovation infrastructures”

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

## TABLE OF CONTENTS

<b>1. INTRODUCTION</b>	<b>4</b>
<b>2. STATE OF THE ART: Enzyme Discovery for Sustainable Use of Natural Resources</b>	<b>5</b>
<b>3. AIM OF THE WORK</b>	<b>7</b>
<b>4. HARMONIZATION, AND FARIZATION OF ENVIRONMENTAL METAGENOMIC DATA FROM AN EXTREME ENVIRONMENT</b>	<b>8</b>
4.1 Metagenomic Data Harmonization	8
4.2 Pisciarelli Metagenomic Data Harmonization and Digitalization	9
<b>5. HARMONIZATION OF DATA ON CHARACTERIZED GLYCOSIDE HYDROLASES DERIVED FROM EXTREMOPHILIC SOURCES, WITH POTENTIAL APPLICATIONS IN LIGNOCELLULOSE BIOCONVERSION PROCESSES</b>	<b>14</b>
5.1 Enzyme Data Harmonization	14
5.2 Harmonizing Characterization Data of Thermostable GH	17
<b>6. SCIENTIFIC PRODUCTION</b>	<b>19</b>
6.1 Case Study: Spent Coffee Ground Polysaccharide Valorization	19
6.2 Characterization of New GHs for Biotechnology Identified from Extreme Environment	20
<b>7. CONCLUSION</b>	<b>22</b>
<b>8. REFERENCES</b>	<b>23</b>

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

## 1. INTRODUCTION

The escalating concerns about climate change, dwindling natural resources, and extensive environmental pollution underscore the urgent need for a transition to a Circular Bioeconomy. Bioeconomy is defined as the knowledge-based production and utilization of biological resources, innovative biological processes, and principles to sustainably provide goods and services across all economic sectors. Circular Bioeconomy aims at “closing the loop” to prevent expansive and unfettered extraction of biological resources and define goals such as sustainability and environmental protection. This involves utilizing renewable resources and waste as raw materials for manufacturing bulk and fine chemicals, nutraceuticals, pharmaceuticals, food, feed, and energy (Pal et al. 2024). At the core of the bioeconomy are industrial biotechnology and biomanufacturing. Biotechnology innovations have the potential to transform economies, societies, and lives, addressing existential threats such as climate change and pandemics. By employing biological systems to catalyze reactions and convert feedstocks into products, biotechnology forms the foundation for products and services across various sectors. It holds transformative potential to enhance lives and contribute to achieving Sustainable Development Goals. Biotechnology is an interdisciplinary field, increasingly relying on advanced digital technologies, including big data and artificial intelligence (AI). The bioscience revolution, exemplified by -omics sciences and big data, has profoundly impacted industrial biotechnology, providing essential knowledge for designing and redesigning biological systems. Biomanufacturing involves deploying biotechnology in large-scale processes to create products for industries. In this context, synthetic biology is poised to bring a new generation of catalysts, including enzymes and engineered microorganisms, leading to an unprecedented era of biomanufacturing across various market sectors.

Industrial Biotechnology Innovation and Synthetic Biology Accelerator (IBISBA) is a distributed European research infrastructure for biotechnology, a driver for innovation in biotechnology and biomanufacturing ([www.ibisba.eu](http://www.ibisba.eu)). IBISBA coordinates a network of European biotechnology platforms to deliver end-to-end innovation services in the field. By fostering their integration, IBISBA produces translational Research & Development & Innovation (R&D&I) services for an international community of biotechnology stakeholders. To achieve this, IBISBA actively develops standards and promotes interoperability between facilities. IBISBA-IT ([www.ibisba.it](http://www.ibisba.it)) is the Italian node of IBISBA, coordinated by CNR IBBR-NA, in which it has the specific mission of developing new molecules and processes through enzyme/protein discovery and engineering, and new biotransformation and bioprocesses for the valorization of lignocellulosic biomasses. Natural biodiversity has historically been a rich source of enzymes and microorganisms, and recent advances in genomics and metagenomics, along with high-throughput tools, enable to explore biodiversity at unprecedented levels. Despite the plummeting cost of DNA sequencing, functional validation remains a bottleneck. Current databases, like the Carbohydrate-Active enZymes database (CAZy, [www.cazy.org](http://www.cazy.org) - a repository for sequences encoding putative carbohydrate-acting enzymes and related proteins), reveal that less than 10% of sequences are characterized, emphasizing the gap between DNA sequence determination and experimental attribution of biological function. Moreover, both data are often not harmonized and difficult to find, making them hard to use for further analysis and biotechnological exploitation. Therefore, a strategy that standardizes and catalogs key information on metagenomic dataset and enzymes identification and characterization for their potential use in bioprocesses is essential.

Data is at the core of IBISBA's ambitions, with data generation, management and sharing being central features. To this aim, the IBISBA Knowledge Hub (<https://hub.ibisba.eu/>), a versatile online resource for storing, managing, and sharing a wide range of scientific research data, models, processes, and outcomes has been developed. In the framework of the ITINERIS project, data, and protocols of metagenomic studies and enzymatic identification, characterization, and exploitation for waste biomass valorization will be FAIRified, harmonized, and digitalized. IBBR-NA will manage

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

the collection of these digital assets, ensuring they adhere to data standards, are properly indexed, and accessible through the IBISBAHub and ITINERIS Central Hub, thus supporting harmonized and standardized datasets for integration into process models and accelerating bioprocess development. To this aim, Activities 6.11 and 6.19 worked in strict collaboration (see also Deliverable 6.23, "Digitalized IBISBA-IT distributed platforms, protocols, and validated pipelines for biomolecule/microorganism discovery, characterization, and engineering for bioprocess development and a circular economy transition.").

## 2. STATE OF THE ART: Enzyme Discovery for Sustainable Use of Natural Resources

The “zero-waste” policy of circular bioeconomy has fueled the development of waste biorefineries, microalgal biorefineries, and lignocellulosic biorefineries. The growing emphasis on sustainable processes has heightened the demand for novel biocatalysts. The widespread use of enzymes in the industrial sector has significantly contributed to increased yields and product quality, while simultaneously reducing energy consumption and environmental impact. The global enzymatic market for biocatalysts is continuously expanding. Estimated at 10 billion USD in 2019, it was projected to experience a compound annual growth rate (CAGR) of 7.1% from 2020 to 2027 (Enzymes Market Size & Share Industry Report, 2020-2027). Hydrolases, particularly carbohydrate hydrolases, proteases, and lipolytic enzymes, dominate industrial enzyme usage, catalyzing the breakdown of natural polymeric substrates. The utilization of natural biomass stands out as a promising strategy to mitigate dependence on fossil resources, which not only contributes to pollution but also disrupts the natural balance of greenhouse gasses (Arbige et al. 2019). Agricultural waste and forest land stand out as the two largest potential sources of lignocellulosic biomass, serving as low-cost and sustainable feedstock for energy production. Unlike crops for food, this biomass is abundant and represents Earth's most abundant renewable energy resource. The saccharification of these biomasses paves the way to produce value-added products, aligning with the demand for a circular use of resources (Botha et al. 2018). In this context, (poly)saccharides-degrading enzymes play a pivotal role in the biotransformation of lignocellulosic biomasses, establishing themselves as essential components in lignocellulose-based biorefineries. Despite their increasing significance in bioconversions, the integration of enzymes into industrial and sustainable processes is progressing slowly. This delay is attributed to factors such as the limited commercial availability of certain enzyme classes and the suboptimal performance of many enzymes under the harsh conditions typical of industrial processes. Hence, the imperative arises to uncover novel enzymes exhibiting essential properties such as resistance to organic solvents, thermostability, process robustness, and improved regio- and enantioselectivity compared to existing catalysts (De Lise et al. 2023). The quest for industrial enzymes primarily focuses on microbial sources, and metagenomic approaches provide access to entire microbiomes from virtually any environment, including extreme ecological niches where extremophiles emerge as a rich source of new enzymes to be exploited in biotechnological applications due to their innate resilience to extreme conditions (Curci et al. 2019).

Metagenomics is a culture-independent technique and plays a vital role in enzyme discovery. The choice of sampling location is critical, as environmental conditions determine the characteristics of the enzymes. The identification of new enzymes can be achieved through two primary strategies: sequence-based and function-based screenings.

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

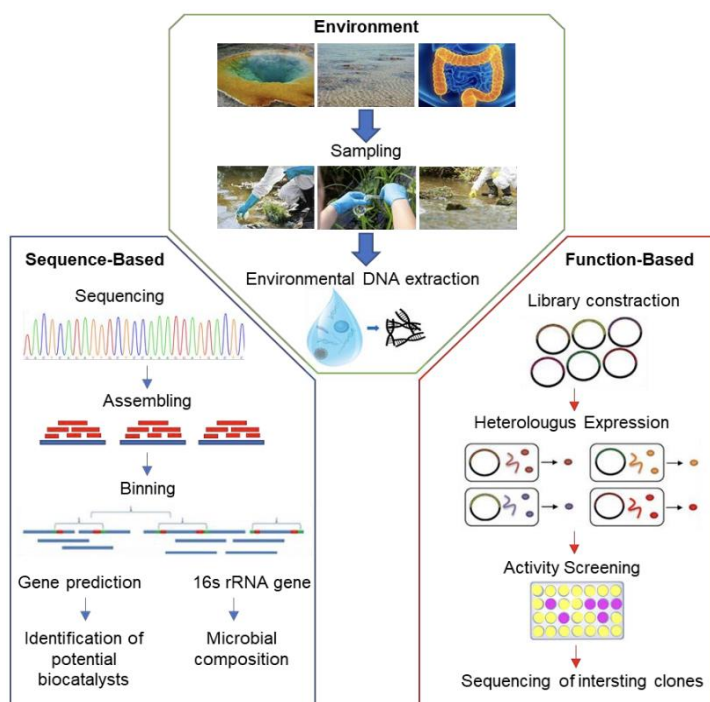


Figure 1: Framework of metagenomic strategies

In the sequence-based metagenomic approach, also known as shotgun metagenomics, environmental DNA (eDNA) is sequenced in-depth, generating numerous short sequences called reads. Bioinformatics plays a crucial role in processing these reads, where they are analyzed and computationally assembled to reconstruct longer contiguous sequences of eDNA (Figure 1). The result of the process is the predictions of coding DNA sequences followed by the functional annotation through homology-based screening. Thus, the sequence-based metagenomic approach mainly identifies variants or distantly related sequences of already known enzyme classes. However, the amount of data generated it's huge and can lead to the identification of novel enzymes with new or better features than those already available. For instance, the metagenomic analysis of the Sargasso Sea and the cow rumen allowed the identification of 70000 and 27000 novel genes for putative enzymes, respectively (Venter et al. 2004; Terper et al. 2006). On the other hand, function-based metagenomics involves cloning mDNA into expression vectors and producing libraries that are screened for activity on selected substrates (Figure 1). This approach is more likely to discover potentially novel classes of enzymes with no homology to known sequences. While the identified enzymes already have assigned activities, the quantity of data generated by this strategy is noticeably less than that of the sequence-based approach (Niu et al. 2018). Shotgun metagenomics has revolutionized enzyme discovery by enabling the direct exploration of genetic material from complex environments without the need to isolate organisms. This approach has accelerated the identification of novel biocatalysts with potential industrial and biotechnological applications from previously inaccessible environments (Robinson et al. 2023). However, metagenomics has also led to a significant discrepancy between annotated sequences encoding enzymes and their functional or structural characterization. Current methods for enzyme characterization are by far less mechanized, and more time-consuming and cost-effective if compared to Next-generation sequencing (NGS) and bioinformatics, thereby, lagging the continuous increment of genes obtained from "Big data". This gap limits the exploitation of new activities for the sustainable use of natural resources. Efforts to characterize novel enzymes are needed (Vanacek et al. 2018). To make more accessible the efforts in enzyme characterization, the enormous number of sequences can be filtered and rationalized through *in silico* analysis for protein function prediction and to preselect attractive targets to support

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

experimental testing. Enzyme function is commonly predicted through sequence similarity using databases that could be complemented by examining highly conserved motifs or domains in the sequences. A valuable (and complementary) alternative for predicting protein function is the structure-based method, as the 3D structure of an enzyme is often more conserved than the sequence. The fold of enzymes is strictly correlated to its function, and proteins with similar fold often have a similar function. Nowadays AlphaFold3 (AF3) has revolutionized the protein fold prediction (Roy et al. 2024). Developed by DeepMind, AlphaFold is a machine-learning model that represents a monumental leap forward in predicting protein structures. In particular, the local-fold similarity, especially around the active site, could help to deduce more accurately the protein function (Jumper et al. 2021). This type of analysis and functional categorization not only provides insights into the possible function of a given sequence but also enables classification into specific protein families and subfamilies, offering a deeper understanding and a more precise indication of potential enzymatic activity. One notable example is the CAZy database. CAZy accurately categorizes the so named CAZymes into families, subfamilies, and clans based on sequence homology, activity, and structural similarities. The classification facilitates preselecting activities in genomic and metagenomic datasets by comparing them with classified enzymes that are functionally and structurally characterized (Drula et al. 2022).

Among the different classes of CAZymes classified in the database, the Glycoside Hydrolases (GHs) are the most abundant enzymes across all living systems and principal actors in the breakdown and processing of carbohydrates. GHs are highly valued across various industries, with those derived from thermophilic microorganisms being particularly well-suited to the harsh conditions often required in industrial bioprocesses (Berlemont et al. 2016). Thermophiles and hyperthermophiles, primarily from the Archaea domain, represent a crucial source of biocatalysts, known as thermozymes, which exhibit remarkable stability under extreme conditions. These enzymes maintain their functionality at high temperatures, extreme pH levels, in the presence of organic solvents, and even in environments with heavy metals or proteolytic threats. The utilization of thermozymes in industrial applications offers several advantages, including reduced contamination risks, enhanced substrate solubility, and accelerated reaction rates. Their ability to perform efficiently under such conditions is a key advantage, ensuring robust and sustainable processes in industrial settings (Cobucci-Ponzano et al. 2015). One of the most promising applications for thermozymes is in the processing of lignocellulosic biomass waste. The utilization of agricultural and food industry waste through enzyme-based treatments aligns perfectly with the principles of circular bioeconomy. This approach not only helps in reducing waste but also in creating value-added products from what would otherwise be discarded. By breaking down complex lignocellulosic materials, these enzymes contribute to the sustainable production of biofuels, biochemicals, and other renewable resources (Chandukishore et al. 2024). A strategy based on the identification of new GHs through the exploration of extreme environments, their characterization, and their use for the valorization of lignocellulosic biomass would align perfectly with the development of sustainable, eco-friendly processes and the reuse of waste to produce high value-added compounds.

### 3. AIM OF THE WORK

In the rich-data era, predictive design and rapid evaluation are fundamental to the successful implementation of any bioprocess. Within this context, the vast amount of data generated by environmental metagenomics represents a valuable reservoir of information with the potential to uncover new enzymes with specific functions. However, much of this data is embedded in scientific publications, making it challenging to analyze. Furthermore, despite significant efforts to identify and characterize novel enzymes of biotechnological interest for bioprocess development the

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

biochemical data remain scattered across publications, making it difficult to access, analyze, and compare them effectively.

To streamline the accessibility of data present in literature and optimize their utilization, it's necessary to improve the **F**indability, **A**ccessibility, **I**nteroperability and **R**euse of digital assets according to the FAIR principles. The first step in using or reusing data is the process of finding them. The principle of findability requires that data have unique and persistent identifiers with distinct labels. It also emphasizes the need for detailed metadata that explicitly includes the data identifier, creating a clear link between descriptive information and actual content. Additionally, it highlights the significance of registering or indexing metadata in a searchable resource for efficient information location and retrieval by users. Once the user finds the necessary data, it is essential to know how to access them, possibly involving authentication and authorization. The (meta)data have to be retrieved by their identifier using a standardized communication protocol. The data need to be integrated with others and interoperate with applications or workflows for analysis, storage, and processing necessitating the use of a formal, accessible, and shareable language for knowledge representation. The final objective of FAIR is to enhance the efficiency of data reuse. To accomplish this task, it is essential to provide comprehensive descriptions of metadata and data, enabling their replication and combination across various contexts. In close connection with the data FAIRization process, the "Data Harmonization" (DH) process also holds a central role in streamlining and ensuring the availability of data. DH consists of integrating and aligning data from various sources to make them compatible and consistent with each other. The goal of data harmonization is to create a unified and standardized structure for the data, enabling better interoperability and analysis. This process is particularly important when working with data from heterogeneous sources, such as those found in scientific publications, and may involve standardizing formats, units of measurement, codes, and other data characteristics to facilitate their seamless integration and interpretation.

To address these challenges, Activity 6.11 focused on the collection, harmonization, and FAIRization of metagenomic dataset and enzyme characterization from extreme environments data to support the development of a bioprocess for the valorization of a lignocellulosic biomass within the framework of the circular bioeconomy. Personnel recruited for Activities 6.11 and 6.19 collaborated closely, enabling coordinated data collection, analysis, harmonization, and FAIRization. This streamlined the processes of digitalization. As a comprehensive example of the bioprocess, Activity 6.11 addressed three key steps of the process:

- I. Harmonization and FAIRization of environmental metagenomic data from an extreme environment.
- II. Harmonization and FAIRization of data on characterized GHs derived from extremophilic sources, with potential applications in lignocellulose bioconversion processes.
- III. Collection, FAIRization, harmonization, and digitalization of scientific data generated during the project.

All collected and processed data will be uploaded to the IBISBA-IT Data Nexus digital infrastructure as part of Deliverable 6.23, in accordance with both WP2, and the guidelines reported in the IBISBA Knowledge Hub.

## 4. HARMONIZATION, AND FARIZATION OF ENVIRONMENTAL METAGENOMIC DATA FROM AN EXTREME ENVIRONMENT

### 4.1 Metagenomic Data Harmonization

Metagenomic data from specific environments are often challenging to retrieve, as these analyses are primarily conducted by independent groups and usually embedded in one or more research articles.

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

Determining if the content is relevant to a particular interest requires a detailed review and extraction process, which is often unavoidable. Additionally, while an environment may appear suitable for a given interest, the data and findings reported in the study can sometimes diverge from what is needed, making often necessary to re-analyses the raw data. Moreover, the comparison of metagenomic data analyses from independent studies is possible only when the analysis workflows are described in a standardized and harmonized manner. Thus, harmonizing and applying FAIR principles to these data should focus on the core elements of metagenomic analysis, particularly for sequence-based approaches.

With the aim of making (meta)genomic data FAIR and enabling its integration, discovery, and comparison through international community-driven standards, the Genomic Standards Consortium (GSC, [www.genc.org](http://www.genc.org)) has defined a core set of information as a standard for genomes and metagenomes in the form of metadata, providing the minimum information required for a given analysis. The core elements identified as minimal information required are compiled in checklists which include a set of minimum descriptors (Field et al. 2011). The Minimum Information about Metagenomic sequences (MIMS), developed by GSC, accurately report information associated with metagenomic sequencing samples. This standard information can be also integrated with various GSC-environmental packages to specify the environmental context of a sequenced microbial community. Therefore, according to the GSC checklist, the MIMS for a given environment are listed in Table 1 (Field et al. 2014).

*Table 1: Standard information according to MIMS model*

<i>SCG - MIMS</i>	<i>Descriptor format</i>
Investigation type	Text
Project name	Fixed value "metagenome"
Geographic location (latitude and longitude)	Decimal degree in WGS84 system
Geographic location (country)	Text
Collection date	ISO8601 date and time
Environment (biome)	ENVO class
Environment (feature)	ENVO class
Environment (material)	ENVO class
Environment GSC package	GSC controlled vocabulary
Sequencing method	Text

This standardization model provides a concise yet detailed overview of information regarding the origin and characteristics of the sequenced sample, as well as the experiment conducted to obtain the data in the form of sequenced reads. However, in a detailed and comprehensive description of a given metagenomic study, important information that should not be overlooked includes the purpose of the study, which is the reason for conducting the metagenomic exploration, the experiment's output, meaning the reads, and the results of the analysis. These results consist of a series of outputs generated from the analysis of the primary sequencing, including sequence assembly and taxonomic annotation. Having information about the purpose of the study would immediately provide, within the metadata, an idea of the type of data generated from the primary sequencing analysis. Regarding the reads, the useful information to share mainly concerns the format and where they have been deposited. More detailed information, generated from the assembly of the reads, such as the number of ORFs obtained and their functional classification, could still add value if standardized and harmonized (Ten Hopen et al. 2017; Cernava et. al 2022).

#### 4.2 Pisciarelli Metagenomic Data Harmonization and Digitalization

In 2019, in collaboration with the IBISBA partner University of Naples Federico II, a metagenomic analysis of the extreme environment Pisciarelli solfatara has been performed (Strazzulli et al. 2019).

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

The Phlegraean Fields is a large volcanic caldera located west of Naples, Italy. It spans an area of about 13 kilometers and is part of a much larger volcanic complex that has been active for thousands of years. Formed over 40,000 years ago, the Phlegraean Fields is one of the most active and dangerous volcanic areas in the world. The landscape features a series of craters, fumaroles, hot springs, mud pools, and sulfur deposits, all of which are visible indicators of the geothermal processes occurring beneath the caldera. The most geothermally active area within the Phlegraean Fields is the Pisciarelli solfatara, rich in fumaroles, hot springs, and mud pools. Unique extremophilic microorganisms have been discovered here, offering significant value for biotechnological research. The enzymes isolated from these organisms exhibit activity and stability under extreme conditions, making them suitable for diverse applications in biotechnology.

In 2019, the environment was characterized by two main pools of geothermal water and mud, so named Pool1 and Pool2, each exhibiting distinct chemical and physical properties. Pool1 had an average temperature of 80 °C with a pH around 5.5, while Pool 2 exhibited even more extreme characteristics, a temperature of 95 °C and a pH of 1.5. Samples from Pool 1, primarily composed of hot muddy sediments, gravel, and water, were collected from the surface of the pool. In contrast, samples from Pool 2 were obtained by scraping the side of the pool, which was submerged in a clear mud-water mixture and consisted predominantly of gravel. The taxonomic identification and a comprehensive view of the microbial composition within the samples revealed a significant portion of the reads obtained from the two samples showed no matches in the NT and NR databases. Specifically, 32% and 45% of the reads from Pool 1 had no match, while the remaining reads were primarily attributed to Archaea, with minor contributions from viruses (0.17%) and bacteria (0.11%). In contrast, Pool 2 exhibited an even higher percentage of unmatched reads, with 62% in the NT database and 68% in the NR database. The remaining reads in Pool 2 were predominantly Archaea (37%) and included a small fraction of archaeal viruses (0.36%). Furthermore, the functional analysis of the metagenomic data revealed that the largest proportion of annotated sequences was involved in carbohydrate metabolism, accounting for nearly 2% of the total ORFs suggesting that the environment offers a unique ecological niche for discovering new thermophilic CAZymes. The CAZymes analysis revealed a total of 278 and 308 putative CAZymes were identified in Pool1 and Pool2 respectively. Among these, glycosyltransferases (GTs) were the most abundant, comprising 60% of the total in Pool 1 and 56% in Pool 2, while glycoside hydrolases (GHs) followed closely with 36% in Pool 1 and 38% in Pool 2 (Strazzulli et al. 2019).

In the Activity 6.11, the Pisciarelli metagenomic data were extracted and harmonized according to the MIMS developed by GSC. Therefore, although originating from the same study, we are dealing with data generated from two different metagenomic analyses from distinct samples, each with its own name, ID, and output. The information of the generated metadata (MIMS\_Pool1 and MIMS\_Pool2) are reported in the Tables 2 and 3.

*Table 2: Standardized information on Pool1 sample*

<i>MIMS_Pool1</i>	
Investigation type	Metagenome
Project name	Pool1Pisciarelli
Geographic location (latitude and longitude)	40°49'45.0768"N, 14°8'49.3512E
Geographic location (country)	Italy
Collection date	2019-08-20
Environment (biome)	ENVO:00003018
Environment (feature)	ENVO:00002120 - ENVO:00002027
Environment (material)	ENVO:00002007 - ENVO:00005793
Environment GSC package	Sediment
Environment temperature	80 °C

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

Environment pH	5.5
Sequencing method	Illumina WGS
Reads deposited	NCBI Sequence Read Archive (SRA)
Reads ID	SRR7545549
Experiment ID	SRX4411842
Functional annotation	KEGG - COG
Study PMID	31595646

Table 3: Standardized information on Pool2 sample

<i>MIMS_Pool2</i>	
Investigation type	Metagenome
Project name	Pool2 Pisciarelli
Geographic location (latitude and longitude)	40°49'45.0768"N, 14°8'49.3512E
Geographic location (country)	Italy
Collection date	2019-08-20
Environment (biome)	ENVO:00003018
Environment (feature)	ENVO:00002120 - ENVO:00002027
Environment (material)	ENVO:00002007 - ENVO:00005793
Environment GSC package	Sediment
Environment temperature	95 °C
Environment pH	1.5
Sequencing method	Illumina HiSeq 2000
Reads deposited	NCBI Sequence Read Archive (SRA)
Reads ID	SRR7545550
Experiment ID	SRX4411841
Functional annotation	KEGG - COG
Study PMID	31595646

In addition to MIMS, the following parameters were added: temperature and pH of the environment, where the reads were deposited along with their respective ID, and the experiment ID. Additionally, information about the strategy used for the functional annotation of the identified ORFs and the PUBMED ID (NCBI) of the study from which the analysis originated were also included.

As previously mentioned, information on the type of output generated from the analysis of the reads could provide added value if made available and standardized. Therefore, we aimed to harmonize the data related to the key elements resulting from the metagenomic analysis, namely the taxonomic classification and functional annotation of the identified ORFs.

The primary and most immediate output of metagenomic analysis is the profile of the microbial population within a given environment. This type of analysis can be performed at different levels of taxonomic precision. At lower taxonomic ranks, information can generally be obtained only on species that have already been isolated and characterized. However, it is also possible to reconstruct genomes of new species from the sequencing data obtained, the so named metagenome assembled genomes (MAGs) (Setubal 2021). In the reported metagenomic exploration of Pisciarelli solfatara the microbial composition was analyzed, covering taxonomic ranks from phylum to species. Therefore, to make these data more accessible and available, the taxonomic data analysis for Pool 1 and Pool 2 were extracted and classified, generating CSV files, providing a percentage-based comparison of the two pools (Figure 2). The generated files were organized into a single folder named "Microbial Community" and labeled according to taxonomic category (e.g., Phylum Comparison, Genus Comparison, etc.). In this case, it was preferable to consolidate the data from both pools, as they originate from the same environment and study. This approach also allows each file to provide

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

dual information: (i) the microbial composition of each pool and (ii) a comparison between the two pools.

#Datasets	Pool1_reads_vs_nt	Pool2_reads_vs_nt
<b>Bacteroidetes</b>	0	0.016052
<b>Alphaproteobacteria</b>	0.014766	0
<b>Enterobacteriaceae</b>	0.081675	0
<b>Cyanobacteria</b>	0	0.010251
<b>Bacillaceae</b>	0	0.011138
<b>Lactobacillales</b>	0	0.015035
<b>Clostridiaceae</b>	0	0.023483
<b>Lachnospiraceae</b>	0	0.010433
<b>Thermoanaerobacterales</b>	0	0.019043
<b>Mollicutes</b>	0	0.010654
<b>Thermoplasmataceae</b>	0.0121	0
<b>environmental samples &lt;archaea,superkingdom Archaea&gt;</b>	0.028422	0.020956
<b>Methanobacteriaceae</b>	0	0.014531
<b>Thermococcaceae</b>	0	0.025105
<b>Acidilobaceae</b>	0	0.022758
<b>Desulfurococcaceae</b>	0.015866	0.061639
<b>Pyrodictiaceae</b>	0	0.023594
<b>environmental samples &lt;crenarchaeotes,order Sulfolobales&gt;</b>	0	0.014743
<b>Sulfolobaceae</b>	62.18251	98.687431
<b>Thermoproteaceae</b>	37.372475	0.035759
<b>Muridae</b>	0	0.013111
<b>Hominidae</b>	0	0.024752
<b>rosids</b>	0.011125	0
<b>Lipothrixviridae</b>	0.036174	0.104719
<b>Ampullaviridae</b>	0	0.118535
<b>Bicaudaviridae</b>	0.21021	0.716277
<b>other entries</b>	0.034662	0

*Figure 2: Example of microbial composition comparison. Image of CSV table for family comparison.*

Additionally, the shared folder has been implemented with an HTML file under the name of "Pool1andPool2\_Krona". This file opens a fully interactive Krona chart (Figure 3), viewable in any modern web browser without requiring additional software or plugins. Krona is a powerful tool for exploring relative abundances and confidence levels in complex metagenomic classifications. Its unique design simplifies metagenomic data interpretation, while the browser-based format ensures portability and easy integration with existing analysis workflows. Each chart functions as an independent document, making it easy to share via email or host on a web server. Krona's compatibility with popular metagenomic analysis tools has been demonstrated across various datasets (Ondov et al. 2011). By applying the Krona tool to analyze the microbial community

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*



generated documents already offers an excellent starting point for determining the relevance of the data to specific areas of interest. A clear example is evident in the case study itself. One of the primary goals of the metagenomic exploration of the Pisciarelli solfatara was to identify new sequences encoding for CAZymes of hyperthermophilic origin. From the KEGG functional annotation results from the two solfataric pools, it was evident that the majority of identified ORFs are associated with carbohydrate metabolism (Strazzulli et al. 2019). The data obtained led to the conclusion that the explored environment could be an ideal site for the identification of new enzymes involved in the carbohydrate modifications, with unique stability properties, which could perfectly match the requirements of various biotechnological and industrial applications. Furthermore, by leveraging the existing classification of these enzymes in the CAZy database, a more targeted analysis was conducted on the identified ORFs to specifically pinpoint CAZymes classes and family within them. Two CSV file (*CAZyme Pool1* and *CAZyme Pool2*) detailing the sequences of the identified enzymes and their classification by class/family has been generated and included in a subfolder (CAZyme ORFs) of the Functional Annotation main folder.

Taking the Pisciarelli metagenomic study as a model for developing standardized criteria, we harmonized the data from both Pools and applied standardized tools. All the highlighted criteria are easily transferable to any other metagenomic study. By applying this standardized approach, researchers can streamline data processing and interpretation, ensuring consistent, accessible outputs across diverse environments. This facilitates not only the description of new sampling sites and experiments but also the identification of specific enzyme types within datasets, enabling targeted exploration for biotechnology and beyond.

## 5. HARMONIZATION OF DATA ON CHARACTERIZED GLYCOSIDE HYDROLASES DERIVED FROM EXTREMOPHILIC SOURCES, WITH POTENTIAL APPLICATIONS IN LIGNOCELLULOSE BIOCONVERSION PROCESSES

### 5.1 Enzyme Data Harmonization

Enzyme usage in biotechnology and industry has expanded significantly due to their ability to catalyze specific biochemical reactions with high efficiency and selectivity, making them valuable for various applications. However, the use of enzymes in biotechnological applications inevitably requires knowledge of optimal activity conditions, substrate specificity, stability, and other characteristics, which can only be determined through thorough enzyme characterization. This detailed analysis is crucial for optimizing enzymatic processes in industrial settings, ensuring that enzymes perform efficiently under the desired conditions while maintaining their stability and activity over time (Cui et al. 2023, Chapman et al. 2018). This type of information is often compiled across various scientific articles, and retrieving it can be challenging, requiring careful analysis and detailed scrutiny. This limitation, coupled with the relatively small number of characterized enzymes compared to the annotated sequences, slows down, or limits the widespread use of new enzymes as biocatalysts. While more general enzyme databases, such as CAZy, focus on broader classifications and properties, such as enzyme families and mechanisms of action, they often lack the in-depth information needed for certain specialized enzymes. Recently, the database has been also updated with the CAZYme activity descriptor (CAZac), which describes CAZymes' mechanisms, glycosidic bond orientations, subsites, and inter-residue connectivity (Lombard et al. 2024). However, despite the significant contribution of the CAZy database, it falls short when it comes to detailing their properties beyond substrate specificity. To obtain crucial information for the exploitation of GHs in bioprocesses, it is still necessary to consult the scientific literature for more comprehensive data.

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

Improving access to these deeper insights would greatly enhance the practical application of these enzymes in industrial and research settings.

Critical characteristics for the biotechnological use of these enzymes include optimal activity conditions, such as temperature and pH, as well as substrate specificity. These data are inherently challenging to collect, interpret, and standardize, as they are widely dispersed across journals from various fields and are often influenced by specific experimental conditions. There have been several attempts to standardize enzymatic characterization data to compare, evaluate, interpret, and reproduce experimental research results published in the literature and databases. Databases such as BRENDA (<http://www.brenda-enzymes.org/>) (Chang et al. 2015) and SABIO-RK (<http://sabio.its.org>) (Wittig et al. 2012) have been developed, serving as the primary collection of enzyme functional data, freely accessible to the scientific community. These databases aim to provide a representative overview of the characteristics and variability of each enzyme. However, for more detailed information, readers must still consult the primary literature.

Another initiative was launched by the STRENDA (Standards for Reporting Enzyme Data) Commission (<http://www.beilstein-institut.de/en/projects/strenda/>). Established in 2003 as part of an initiative by the Beilstein-Institut (<http://www.beilstein-institut.de/>), its goal was to define, through extensive collaboration with the biochemistry community, the minimum information required to accurately describe assay conditions and enzyme activity data. In the web based STRENDA DB (<https://www.beilstein-strenda-db.org/strenda/index.xhtml>), guidelines are provided to assist authors in submitting essential information related to functional enzymology data when preparing manuscripts (<https://www.beilstein-institut.de/en/projects/strenda/guidelines/>). These guidelines help ensure that critical data is accurately reported. The required data include information about the manuscript, the experiments conducted, and details about the enzyme. More specific information on the activity includes pH, temperature, protein concentration, and any cofactors. Additionally, where applicable, kinetic parameters are also provided (Table 4) (Tipton et al. 2014, Swainston et al. 2018).

Table 4: Required data from STRENDA guidelines.

<i>STRENDA guidelines</i>	<i>Format</i>
<i>Manuscript data</i>	
Title	Text
Author Names	Text
Status	Text (Published / Unpublished / Submitted)
PMID	Text (PubMed identifiers)
Creation Date	ISO8601 date and time
Last work Date	ISO8601 date and time
Publication in Journal Date	ISO8601 date and time
Publication Date	ISO8601 date and time
<i>Experiment</i>	
Description	Text (aim of the work)
Methodology	Text
SRN	STRENDA Register Number
DOI	Text (STRENDA DB DOI)
<i>Protein</i>	
Protein name	Text
UniProtKB AC	Text (UniProt identifiers)
EC number	Text (According to the E.C. number format)
Sequence modification	Text (yes/no)
Expression system	Text

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

Organism	Text (Source)
<i>Assay condition</i>	
Substrate	Text (molarity)
Other Compound	Text (molarity)
Product	Text (molarity)
Buffer	Text (molarity)
pH	Text (pH value)
Temperature	Text (°C)
Pressure	Text
Protein concentration	Text (molarity)
<i>Results</i>	
Substrate	Text (molarity)
$K_M$	Text (molarity)
$K_{cat}$	Text (seconds <sup>-1</sup> )
$K_{cat}/K_M$	Text (molarity <sup>-1</sup> seconds <sup>-1</sup> )

STREND A DB has been designed specifically to accept data submission directly from the research community to benefit scientific community from the availability of standardized data.

Databases such as BRENDA, SABIO-RK and STREND A, aim to standardize results from a wide range of enzyme classes, which requires an extensive amount of information and detailed reporting. This is necessary because enzymes can vary significantly in their properties, mechanisms, and functions. However, when focusing on specific enzymes, like GHs, or their particular applications, the amount of required information could naturally be reduced, as the need for generalization across diverse enzyme types is minimized. In these cases, the reporting could be more streamlined and specific to the particular context of the enzyme's use. In a data harmonization process, it is crucial to limit the number of information fields for a specific dataset for several reasons:

- *Simplicity and Efficiency*: Keeping the dataset concise allows easier handling and management of data. When the amount of information is minimized to only the most relevant fields, the complexity of analysis decreases, making it faster to process and compare.
- *Data Consistency*: Limiting the fields ensures that the data collected is consistent across different entries. Too many fields may lead to variability in the way data is recorded, which complicates efforts to harmonize and standardize it across different sources.
- *Avoiding Information Overload*: Including too much unnecessary information can obscure key insights. Focusing on the most critical data points helps to maintain clarity, improving the quality of the interpretation and facilitating decision-making processes.
- *Improved Usability*: In many applications, especially in biotechnology, end-users need quick access to the most relevant data. Reducing the dataset to the essential fields makes it more user-friendly, allowing users to find and apply the information more easily.
- *Error Reduction*: The more data points collected, the higher the chance for errors in data entry, integration, or interpretation. By limiting the scope to only what is essential, the risk of such errors decreases, resulting in higher data quality.

In the case of enzymatic data, for instance, reducing information to key parameters like temperature stability, pH, and kinetic values can streamline how useful these enzymes are for industrial or biotechnological applications. Too much non-essential data could make it harder to identify enzymes with specific desired properties.

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

## 5.2 Harmonizing Characterization Data of Thermostable GH

In Activity 6.11, the harmonization and digitalization of enzymatic characterization data for several thermostable GHs have been carried out, focusing on their potential for the valorization of lignocellulosic biomasses. Some of these enzymes were identified from the metagenomic datasets derived from Pisciarelli solfatara reported above, providing a clear framework of how analysis of an extreme environment, can lead to the discovery and characterization of thermostable enzymes with biotechnological interest. The integration and standardization of these data aim to accelerate the broader use and application of these enzymes in modern biotechnology, fostering their exploitation in industrial processes.

To make information about GH characteristics easily accessible for biotechnological and industrial applications, it is essential to consider a broad range of data and properties. To achieve this, we proposed a set of Minimum Information about Thermostable GHs (MITG). Efforts have been made to establish standards, in line with databases like STRENDA and CAZy mentioned earlier, to harmonize and integrate data, ensuring consistency and comparability across different sources. The goal is to create a unified, standardized dataset that facilitates straightforward analysis and interpretation, while eliminating inconsistencies and redundancies. The data were either collected in Activity 6.11 or curated from relevant scientific publications. Ultimately, the standardized information has been consolidated into a single CSV file (*MITG\_GH\_list*), highlighting the key and essential characteristics of the enzyme within the context of biotechnological relevant data:

- ***Protein information:*** *GenBank* accession link has been included to provide immediate access to the sequence of a given enzyme, along with the *Source*, which may be represented by the organism of origin or metagenomic data. Moreover, *UniProt* accession has been included, as it offers links to specialized databases like BRENDA, STRENDA, and SABIO-RK, helping consolidate information from multiple sources (UniProt consortium 2023). Finally, *GH Family* classification is included.
- ***Temperature and pH:*** these parameters are fundamental for enzymatic activity and their potential use in bioprocesses. These data have also been extracted and cataloged for each enzyme, considering not only the optimal values of these two parameters but also the range within which the enzyme remains active. In fact, relying on a single value for temperature or pH might be limiting. An enzyme could exhibit significant activity even close to or within a certain range of these parameters, which could be useful for a specific application. These details have thus been cataloged in the file accordingly.
- ***Substrate specificity:*** It is always the critical factor, and often the first consideration when selecting an enzyme for a specific biotechnological application. This characteristic is essential since it determines how effectively the enzyme will catalyze reactions on substrates, ensuring optimal performance in the desired process. The collected information from literature are reported in the document under the points: *Activity*, *Substrate*, *Products*, and *Kinetic constants*. A single enzyme could exhibit different activities also depending on the type of substrate used. Therefore, each activity has been listed and linked to the corresponding substrate used to determine that specific activity. Another important characteristic to highlight for this type of enzyme is their modality of action that can be divided into two categories: *exo*-glycosidases and *endo*-glycosidases. *Exo*-glycosidases hydrolyze glycosidic bonds by progressively cleaving sugar units from the ends of oligo- or polysaccharides, either at the reducing or non-reducing terminus. In contrast, *endo*-glycosidases break glycosidic bonds within the middle of the oligo- or polysaccharide chains, leading to the fragmentation of the larger sugar molecules into smaller pieces. This classification based on their site of action provides valuable insight into their distinct roles in carbohydrate breakdown. Since this type of information is very important both for the enzyme's role and its potential action on a given biomass, to make it immediately available and easily accessible, it has been included in the activity section, with the prefix *endo*- or *exo*-

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

preceding the name of a given activity. The same has been done for their anomeric specificity, distinguishing between  $\alpha$ - or  $\beta$ -glycosidic linkages (Figure 4). Next to each activity-substrate pair, the products (both experimental or inferred) were listed. Additionally, if the study from which the information was sourced calculated kinetic constants, the values for  $K_M$ ,  $k_{cat}$ , and  $k_{cat}/K_M$  were included for that specific activity and substrate.

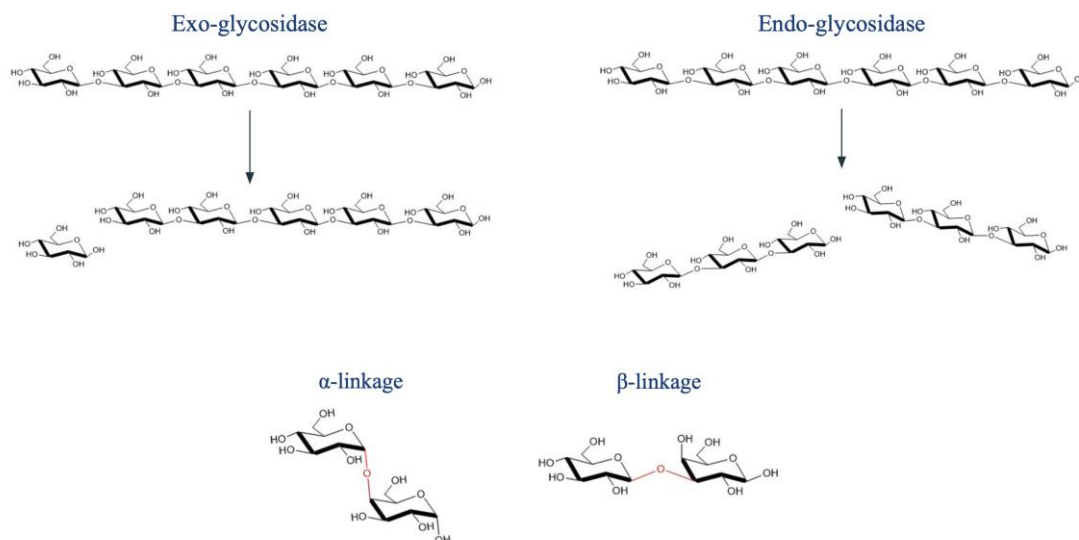


Figure 4: Example of (a) exo-glycosidase, (b) endo-glycosidase activities and (c) alpha and beta glycosidic linkages.

All this information, made easily accessible in the catalog, is crucial to understand what type of lignocellulosic biomass a given enzyme can act on and, consequently, the polysaccharides contained within it. In this regard, it is important to emphasize that when we talk about lignocellulosic biomass, we are referring to complex structures that may contain various types of polysaccharides. Each polysaccharide, in turn, can have more or less complex structures. Therefore, considering their complexity along with enzymatic activities and their potential application in the valorization of lignocellulosic biomass, a detailed list of known polysaccharides from land plants was prepared. In the CSV file named "*Polysaccharide\_description*", a thorough description and classification of polysaccharides are provided, including their monosaccharides composition and structural characteristics.

The MITG required for each enzyme is listed in Table 5 and the CSV file (*MITG\_GH\_list*) was prepared accordingly.

Table 5: Data standardization for MITG

MITG	Format
Name	Text
GenBank	Text (GenBank accession number)
UniProtKB AC	Text (UniProt accession number)
Source	Text (microorganism or metagenomic origins)
GH family	Text (According to CAZy classification)
Temperature range	Text ( $^{\circ}\text{C}$ - $^{\circ}\text{C}$ )
Temperature optima	Text ( $^{\circ}\text{C}$ )
Temperature stability (half life)	Text ( $^{\circ}\text{C}$ )
pH range	Text (pH value - pH value)
pH optimum	Text (pH value)

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

Activity	Text (exo/endo- $\alpha/\beta$ -activity according to CAZy description)
E.C. number	Text (According to the E.C. number format)
Substrate	Text
Products	Text
$K_M$	Text (molarity)
$k_{cat}$	Text (seconds <sup>-1</sup> )
$k_{cat}/K_M$	Text (molarity <sup>-1</sup> seconds <sup>-1</sup> )
Reference	Text (PMID)

The criteria identified, standardized, and listed in the table take into account the essential information needed to determine whether a given enzyme from a hyperthermophilic source could be useful for a specific application. At the same time, with the idea of minimizing the information to be reported for striking a balance between comprehensiveness and focus is essential for effective data harmonization. Regarding the enzyme information, in addition to being collectively stored in the file *MITG\_GH\_List*, dedicated folders have been created for each enzyme. These folders are named using the following format: *GH family\_Enzyme Name\_Microorganism*. Each folder contains the information already present in the list, along with a subfolder that includes files related to AlphaFold3 structural predictions performed during Activity 6.11. This organization helps centralize data for each enzyme while also providing 3D structural models essential for further functional and structural analyses. Moreover, could serve as a valuable resource for guiding the discovery and functional annotation of homologous sequences. Structural models and detailed characterization data support comparative studies, allowing researchers to identify and analyze new homologous enzymes with similar functions or novel applications. This system enhances both the efficiency and accuracy of enzyme annotation in future research efforts.

## 6. SCIENTIFIC PRODUCTION

### 6.1 Case Study: Spent Coffee Ground Polysaccharide Valorization

During the ITINERIS project, within the framework of Activities 6.11 and 6.19, a complete bioprocess was developed at lab scale for the valorization of a waste lignocellulosic biomass. The aim of the bioprocess was to utilize thermostable GHs to selectively degrade the polysaccharides in spent coffee grounds (SCG), producing high-value compounds and thereby converting waste lignocellulosic biomass into valuable products. This approach not only adds value to the waste material, but also demonstrates the potential for sustainable utilization of lignocellulosic resources. SCG are the most abundant waste byproducts generated from coffee beverage production worldwide. Typically, these grounds are seen as waste and end up in landfills. However, SCG contain valuable compounds that can be valorized and used in different applications. Notably, they are rich in carbohydrates, primarily galactomannan, arabinogalactan type II, and cellulose.

The work involved testing different mild pretreatment methods on the biomass to enhance its accessibility for enzymatic hydrolysis. It also included the selection of GH activities by utilizing enzymes already available in our lab, collaborating with companies, exploring exploitable activities from commercial enzyme suppliers, and selecting enzymes based on findings from scientific literature. Identified enzymes were characterized, and the most effective activities were selected with the goal of developing a new thermostable enzymatic cocktail specifically designed for the hydrolysis of the polysaccharides present in SCG. Only the enzymes with best performances on the pretreated biomass were validated and included in the enzymatic cocktail (Table 6).

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

Table 6: The enzymes selected for inclusion in the final cocktail

Target Polysaccharide	Enzyme	Activity	GH family	Source
Galctomannan	<b>TmMan5B</b>	Endo $\beta$ -1,4-mannanase	5	Nzytech
	<b>TmManA</b>	$\beta$ -mannosidase	2	In-house*
	<b>TmGalA</b>	$\alpha$ -1,6-galactosidase	36	In-house*
Arabinogalactan II	<b>Gal3D</b>	Endo $\beta$ -1,3 galactanase	16	Novonesis
	<b>Gal6D</b>	Endo $\beta$ -1,6 galactanase	5	Novonesis
	<b>CelB</b>	$\beta$ -1,3-galactosidase	1	In-house*
	<b>XarS</b>	$\alpha$ -arabinosidase	3	In-house*
Cellulose	<b>12A</b>	Cellulase	12	Nzytech
	<b>CBHI</b>	Cellobiohydrolase	7	Megazyme
	<b>CelB</b>	$\beta$ -1,4-glucosidase	1	In-house*

The developed enzymatic cocktail allowed a conversion yield of 52% of the polysaccharides present in the pretreated biomass into oligosaccharides and monosaccharides. Additionally, microwave pretreatment of SCG, followed by the application of thermostable endo- $\beta$ -mannanase only, allowed the production of high-value mannoooligosaccharides. These compounds were tested for their prebiotic activity on probiotic microbial strains, yielding very promising results by promoting the growth and biofilm formation of five different strains. The work carried out highlights the importance of MITG to have a comprehensive dataset detailing the functional characteristics of characterized thermostable GHs. Furthermore, to build an enzymatic cocktail is crucial to have information on the polysaccharide composition of the biomasses of interest, including their monosaccharide composition and overall structural features. All data resulting from the characterization of the enzymes were standardized, harmonized, and included in document "*MITG\_GH\_list*".

This work has been carried out by the staff working on the ITINERIS project (TD and TI) and in the IBISBA CNR-IBBR lab, by using the new equipment acquired for activities 6.11 and 6.19 of the IBBR-NA. The results have been published in a peer-reviewed international journal (Shaikh-Ibrahim et al 2025) and have been presented at international conferences. All the pipeline and main protocols used were standardized digitalized and will be integrated in the Deliverable 6.23.

## 6.2 Characterization of New GHs for Biotechnology Identified from Extreme Environment

The metagenomic exploration of Pisciarelli solfatara has led to the identification of a significant number of ORFs for enzymes from hyperthermophilic microorganisms. Functional analysis, followed by a search in the CAZy database, identified several thermostable GHs with potential applicability in biotechnology, particularly in the hydrolysis of lignocellulosic biomass. Some of these enzymes had been previously characterized, confirming the expected properties, and have been included in the "*MITG\_list*" file. In line with this approach, as part of the ITINERIS project and inherently related to activities 6.11 and 6.19, novel GH sequences were selected from the metagenomic analysis to investigate their properties and assess their potential for biotechnological applications. For this purpose, 10 enzymes from the GH116 family, one enzyme from the GH77 family, and one enzyme from the GH78 family have been selected. All enzymatic sequences selected were successfully cloned into vectors for heterologous expression in *E. coli* and produced recombinantly. The soluble proteins obtained were then thoroughly characterized biochemically. All information related to their biochemical characterization has been properly standardized and included

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*



in the "MITG\_list" document. The work was carried out and/or supervised by the personnel (TD and TI) involved in activities 6.11 and 6.19, using the equipment purchased for ITINERIS in relation to these two activities. A brief overview of the activities carried out and the results obtained, which will lead to the production and submission of three new scientific articles to peer-reviewed journals, is provided below:

- **GH116 enzymes:** The metagenomic analysis of Pisciarelli solfatara led to the identification of 10 novel sequences encoding enzymes from the GH116, which is an example of a family formed around the characterization of an archaeal enzyme. This family, indeed, was initially discovered through the study of a bifunctional  $\beta$ -glucosidase/ $\beta$ -xylosidase enzyme (SSO1353) from the (hyper)thermophilic archaeon *Saccharolobus solfataricus*. Despite the proposed subdivision of GH116 into subfamilies based on substrate specificity and inhibitor sensitivity, few members have been characterized, with most subfamilies having only one known representative. The discovery of these new sequences provides an opportunity to further explore the functional diversity within this family, especially given its origins in hyperthermophilic archaea and the limited number of characterized members. To address this, expression tests were performed on the various genes cloned into appropriate vectors for *E. coli* expression, yielding the soluble production of only two enzymes, A55 and B08. This outcome underscores the ongoing challenge of achieving successful heterologous expression of archaeal sequences. A55 and B08, which subsequently became the focus of this study, belong to subfamily 2 of GH116. Prior to this, the subfamily had only one characterized member, SSO3039, also originating from a hyperthermophilic archaeon. Phylogenetic studies were conducted on the enzyme family, leading to the classification of the newly identified sequences into their respective subfamilies. AlphaFold models of the enzymes were generated and compared with already resolved structures within the family, revealing differences in the presence or absence of domains across the different subfamilies. Additionally, the amino acids within the active site of the characterized enzymes from subfamily 2 were mapped, highlighting divergences between A55 and B08, which were subsequently characterized in detail along with three mutants of B08. This work has produced new enzymatic characterization data and has significantly expanded our knowledge of enzymes belonging to the GH116, particularly subfamily 2 (*manuscript in preparation*).
- **GH77 enzyme:** GH77 includes amylomaltases (AMs), enzymes with a monospecific 4- $\alpha$ -glucanotransferase activity (EC 2.4.1.25). AMs catalyze the cleavage of  $\alpha$ -1,4 glucosidic bonds in  $\alpha$ -1,4-D-glucans and subsequently transfer the resulting glucan chain to the O-4 position of an  $\alpha$ -1,4-D-glucan acceptor. Their unique transferase activity has made AMs highly valuable for various biotechnological applications, including the synthesis of sugar substitutes, the production of cycloamyloses, and the modification of starch for industrial purposes. Despite the annotation of over 19,000 GH77 sequences in CAZy, only 26 amylomaltases have been characterized to date (December 2024), with just 9 derived from thermophilic sources. In the metagenomic dataset of Pool1 in the Pisciarelli solfatara, an ORF encoding a novel GH77 enzyme from the hyperthermophilic archaeon *Pyrobaculum arsenaticum* was identified. Preliminary characterization revealed optimal activity at 95°C and pH 5.5, with a specific activity of 1500 U/mg on maltotriose. The enzyme also demonstrated disproportionation activity, producing maltooligosaccharides with various degrees of polymerization. Additionally, it was tested on high-amylose starch granules to assess its activity on granular starch, and its cyclization activity was evaluated using amylose as a substrate. The enzyme's robust thermostability, high activity at extreme temperatures, and ability to modify starch make it particularly promising for biotechnological applications, especially in the agri-food sector. Its potential to produce modified starches and oligosaccharides could be highly valuable for developing new food products and enhancing the functional properties of starch in food processing. The characterization is still in progress.

- ***GH78 enzyme***: GH78 enzymes specifically hydrolyze  $\alpha$ -L-rhamnosidic bonds, playing a key role in the degradation of rhamnose-containing polysaccharides and glycoconjugates found in various plant and microbial sources. Their substrates include a wide range of natural glycosides, such as naringin, rutin, quercitrin, hesperidin, dioscin, and terphenyl glycosides, all of which contain terminal  $\alpha$ -L-rhamnose residues. These enzymes, known as  $\alpha$ -L-rhamnosidases, hold significant industrial importance, particularly in the food and pharmaceutical industries, where they are used for the biotransformation of natural products. By hydrolyzing  $\alpha$ -L-rhamnosyl linkages in flavonoids,  $\alpha$ -L-rhamnosidases help produce compounds with enhanced bioavailability and bioactivity, such as prunin, hesperetin 7-*O*-glucoside, and isoquercitrin. In the food industry, these enzymes are employed to improve the sensory properties of juices while increasing their nutritional content by boosting antioxidant levels and enhancing flavonoid bioavailability. In this study, a novel thermostable  $\alpha$ -L-rhamnosidase was successfully cloned, expressed, and characterized for its ability to hydrolyze naringin, a key flavonoid. This marks the first report of a hyperthermophilic  $\alpha$ -L-rhamnosidase, representing the first archaeal GH78 enzyme identified and biochemically characterized. The ability of this enzyme to hydrolyze  $\alpha$ -L-rhamnosidic bonds under high-temperature conditions suggests its potential use in industrial processes that require elevated temperatures, such as the production of high-value compounds in the food industry. Additionally, this discovery underscores the importance of exploring archaeal diversity through metagenomics. As more archaeal genomes are sequenced and analyzed, additional thermostable enzymes with unique catalytic properties are likely to be identified, expanding the toolkit for industrial biocatalysis. The identification of this first archaeal thermostable GH78  $\alpha$ -L-rhamnosidase from metagenomics marks a significant milestone in enzyme discovery from extreme environments (*manuscript in preparation*).

All the data have been standardized according to the MITG guidelines and included into the *MITG\_GH\_list* file. Additionally, individual folders were created for each enzyme. These folders contain the associated metadata following MITG standards, as well as the directory with the AlphaFold models that were obtained for each enzyme with available sequences. The research achievements have significantly expanded our understanding of enzymes from hyperthermophilic microorganisms, particularly those from the GH116, GH77, and GH78 families. The successful identification, cloning, and detailed biochemical characterization of these novel enzymes, combined with the creation of standardized metadata and AlphaFold models, provides valuable insights into their potential industrial applications. The thermostability and unique catalytic properties of these enzymes make them promising candidates for various biotechnological fields, particularly in the agri-food and pharmaceutical industries. This work highlights the importance of metagenomic exploration in extreme environments and sets the stage for future discoveries of robust, industrially relevant biocatalysts.

## 7. CONCLUSION

The work carried out in Activity 6.11 and reported in Deliverable 6.13 led to the identification of the minimum information required for the standardization and harmonization of data related to metagenomic analyses of extreme environments (MMI) and the enzymatic characterization of thermostable GHs with potentially useful activities for the valorization of lignocellulosic biomass (MITG). It is crucial to emphasize that, although metagenomics and enzyme characterization are treated separately, they are closely interconnected in developing bioprocesses aimed at valorizing lignocellulosic biomass. Establishing standardized models for collecting both metagenomic data and enzymatic properties is especially important, considering the vast amounts of data generated daily. These models help streamline the interpretation and application of metagenomic insights, enabling the efficient identification of relevant enzymes for industrial biotechnological processes, particularly

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

in biomass conversion. This improved approach ensures consistency across datasets and facilitates the comparison and integration of findings, ultimately advancing the use of metagenomic and enzymatic data in biotechnology applications.

All files generated and outputs are stored in the Nexus, a network-attached storage (NAS) system that underpins IBISBA-IT's digital infrastructure, ensuring secure and centralized access for collaborative analysis and long-term data management. The Nexus storage system has been developed in collaboration with the activity 6.19 and will be described in the Deliverable D6.23 "Digitalized IBISBA-IT distributed platforms, protocols, and validated pipelines for biomolecule/microorganism discovery, characterization, and engineering for bioprocess development and a circular economy transition".

Pending the FAIRization criteria for the upload into the ITINERIS HUB, the generated and stored data are accessible only to authorized users via VPN at this link <http://172.16.74.3:8080/share.cgi?ssid=f574124c112e44c994e37830f2b1077b>. For authorization, please contact:

Nicola Curci ([nicola.curci@cnr.it](mailto:nicola.curci@cnr.it))

Mauro Di Fenza ([mauro.difenza@cnr.it](mailto:mauro.difenza@cnr.it))

Federica De Lise ([federica.delise@cnr.it](mailto:federica.delise@cnr.it))

Beatrice Cobucci Ponzano ([beatrice.cobucciponzano@cnr.it](mailto:beatrice.cobucciponzano@cnr.it))

## 8. REFERENCES

- Arbige M V, Shetty J K and Chotani G K. Industrial Enzymology: The Next Chapter. *Trends in Biotechnology*. 2019, 37:12.
- Berlemont R, Martiny AC. Glycoside Hydrolases across Environmental Microbial Communities. *PLoS Comput Biol*. 2016, 12:e1005300.
- Botha J, Mizrahi E, Myburg A A and Cowan D A. Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. *Extremophiles*. 2018, 22: 1.
- Bozkurt EU, Ørsted EC, Volke DC, Nickel PI. Accelerating enzyme discovery and engineering with high-throughput screening. *Nat Prod Rep*. 2024 ,15.
- Cernava T, Rybakova D, Buscot F, Clavel T, McHardy AC, Meyer F, Meyer F, Overmann J, Stecher B, Sessitsch A, Schlöter M, Berg G; MicrobiomeSupport Team. Metadata harmonization-Standards are the key for a better usage of omics data for integrative microbiome analysis. *Environ Microbiome*. 2022, 17:33.
- Chandukishore T, Satwika D, Prabir D, Venkata D V, Ashish A. Re-routing the hemicellulosic fraction of lignocellulosic biomass toward value added products: A pragmatic bio refinery approach. *Journal of Environmental Chemical Engineering*, 2024, 12:2.
- Chang A, Schomburg I, Placzek S, Jeske L, Ulbrich M, Xiao M, Sensen CW, Schomburg D. BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res*. 2015, 43:D439-46.
- Chapman J, Ismail AE, Dinu CZ. Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks. *Catalysts*. 2018, 8:238.
- Cobucci-Ponzano B, Strazzulli A, Iacono R, Masturzo G, Giglio R, Rossi M, Moracci M. Novel thermophilic hemicellulases for the conversion of lignocellulose for second generation biorefineries. *Enzyme Microb Technol*. 2015, 78:63.
- Cui J, Ocoy I, Mahmoud MA, Du Y. Editorial: Enzyme immobilization technologies and their biomanufacturing applications. *Front Bioeng Biotechnol*. 2023, 31:1256181.

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

- Curci N, Strazzulli A, Iacono R, De Lise F, Maurelli L, Di Fenza M, Cobucci-Ponzano B, Moracci M. Xyloglucan Oligosaccharides Hydrolysis by Exo-Acting Glycoside Hydrolases from Hyperthermophilic Microorganism *Saccharolobus solfataricus*. *Int J Mol Sci*. 2021, 24;22:3325.
- De Lise F, Iacono R, Moracci M, Strazzulli A, Cobucci-Ponzano B. Archaea as a Model System for Molecular Biology and Biotechnology. *Biomolecules* **2023**, 13.
- Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res*. 2022, 7;50:D571-D577.
- Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I, Klenk HP, Knight R, Kottmann R, Kyrpides N, Meyer F, San Gil I, Sansone SA, Schriml LM, Sterk P, Tatusova T, Ussery DW, White O, Wooley J. The Genomic Standards Consortium. *PLoS Biol*. 2011,9(6):e1001088.
- Field D, Sterk P, Kottmann R, De Smet JW, Amaral-Zettler L, Cochrane G, Cole JR, Davies N, Dawyndt P, Garrity GM, Gilbert JA, Glöckner FO, Hirschman L, Klenk HP, Knight R, Kyrpides N, Meyer F, Karsch-Mizrachi I, Morrison N, Robbins R, San Gil I, Sansone S, Schriml L, Tatusova T, Ussery D, Yilmaz P, White O, Wooley J, Caporaso G. Genomic standards consortium projects. *Stand Genomic Sci*. 2014, 15;9:599.
- Jumper, J., Evans, R., Pritzel, A. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021, 596: 583.
- Krüger A, Schäfers C, Busch P, Antranikian G. Digitalization in microbiology - Paving the path to sustainable circular bioeconomy. *N Biotechnol*. 2020, 25;59:88.
- Iacono R, Cobucci-Ponzano B, De Lise F, Curci N, Maurelli L, Moracci M, Strazzulli A. Spatial Metagenomics of Three Geothermal Sites in Pisciarelli Hot Spring Focusing on the Biochemical Resources of the Microbial Consortia. *Molecules*. 2020, 3;25(17):4023.
- Lombard V, Henrissat B, Garron ML. CAZac: an activity descriptor for carbohydrate-active enzymes. *Nucleic Acids Res*. 2024,11:gkae1045.
- Niu S, Yang J, Mcdermaid A, Zhao J, Kang Y. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief Bioinform*. 2018, 19:1415.
- Ondov BD, Bergman NH and Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*. 2018, 12:385.
- Pal P, Singh AK, Srivastava RK, Rathore SS, Sahoo UK, Subudhi S, Sarangi PK, Prus P. Circular Bioeconomy in Action: Transforming Food Wastes into Renewable Food Resources. *Foods*. 2024, 23;13:3007.
- Robinson SL, Piel J, Sunagawa S. A roadmap for metagenomic enzyme discovery. *Nat Prod Rep*. 2021, 17;38:1994-2023.
- Sabbarese C., Ambrosino F., Chiodini G. *et al*. Continuous radon monitoring during seven years of volcanic unrest at Campi Flegrei caldera (Italy). 2020, *Sci Rep* 10, 9551.
- Setubal JC. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys Rev*. 2021, 4;13:905-909.
- Strazzulli A, Cobucci-Ponzano B, Iacono R, Giglio R, Maurelli L, Curci N, Schiano-di-Cola C, Santangelo A, Contursi P, Lombard V, Henrissat B, Lauro F M, Fontes C M G A and Moracci M. Discovery of hyperstable carbohydrate-active enzymes through metagenomics of extreme environments. *The FEBS Journal* 2019, 287:1116–1137.
- Ten Hoopen P, Finn RD, Bongo LA, Corre E, Fosso B, Meyer F, Mitchell A, Pelletier E, Pesole G, Santamaria M, Willassen NP, Cochrane G. The metagenomic data life-cycle: standards and best practices. *Gigascience*. 2017, 1;6:1-11.
- Shaikh-Ibrahim A, Curci N, De Lise F, Sacco O, Di Fenza M, Castaldi S, Isticato R, Oliveira A, Aniceto JPS, Silva CM, Serafim LS, M Krogh KBR, Moracci M, Cobucci-Ponzano B. Carbohydrate conversion in spent coffee grounds: pretreatment strategies and novel enzymatic

*Harmonized environmental metagenomic and enzyme characterization datasets to build standardized and interoperable data files to be stored, shared, and used for data integration into process models.*

- cocktail to produce value-added saccharides and prebiotic mannoooligosaccharides. *Biotechnol Biofuels Bioprod.* 2025, 7;18:2.
- Swainston N, Baici A, Bakker BM, Cornish-Bowden A, Fitzpatrick PF, Halling P, Leyh TS, O'Donovan C, Raushel FM, Reschel U, Rohwer JM, Schnell S, Schomburg D, Tipton KF, Tsai MD, Westerhoff HV, Wittig U, Wohlgemuth R, Kettner C. STRENDA DB: enabling the validation and sharing of enzyme kinetics data. *FEBS J.* 2018, 285:2193-2204.
  - Terpe K. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol.* 2006, 72:211–222.
  - Tipton KF, Armstrong RN, Bakker BM, Bairoch A, Cornish-Bowden A, Halling PJ, Hofmeyr J-H, Leyh TS, Kettner C, Raushel FM, Rohwer J, Schomburg D, Steinbeck C. Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. *Perspectives in Science.* 2014,1:131–137.
  - UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 2023, 6;51:D523-D531.
  - Vanacek P, Sebestova E, Babkova P, Bidmanova S, Daniel L, Dvorak P, Stepankova V, Chaloupkova R, Brezovsky J, Prokop Z and Damborsky J. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catal.* 2018, 8:2402-2412.
  - Venter J C, Remington K, Heidelberg J F, Halpern A L, Rusch D, Eisen J A et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 2004, 304:66–74.
  - Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algae E, Weidemann A, Sauer-Danzwith H, Mir S, Krebs O, Bittkowski M, Wetsch E, Rojas I, Müller W. SABIO-RK--database for biochemical reaction kinetics. *Nucleic Acids Res.* 2012, 40:D790-6.