DISCOVERY AND FUNCTIONAL CHARACTERIZATION OF NOVEL GLYCOSIDE HYDROLASES: FROM *IN-SILICO* IDENTIFICATION TO BIOTECHNOLOGICAL APPLICATIONS

Francesca Maria Pia Paragliola

Dottorato in Biotecnologie – 37° ciclo

Università di Napoli Federico II

Dottorato in Biotecnologie – 37° ciclo

Università di Napoli Federico II

DISCOVERY AND FUNCTIONAL CHARACTERIZATION OF NOVEL GLYCOSIDE HYDROLASES: FROM *IN-SILICO* IDENTIFICATION TO BIOTECHNOLOGICAL APPLICATIONS

Dottorando: Francesca Maria Pia Paragliola

Relatore: Prof. Marco Moracci

Co-tutor: Prof. Andrea Strazzulli

Coordinatore: Prof. Marco Moracci

Settore Scientifico Disciplinare BIOS07

A chi ha creduto in me
quando io non riuscivo a farlo.
A chi ha visto possibilità
dove io vedevo solo limiti.
A te, che sei stato forza nella mia fragilità,
ogni mio passo avanti è anche il tuo.

INDEX

ABBREVIATIONS

1-DNJ	1-Deoxynojirimycin
4MU	4-Methylumbelliferyl
5F-Xyl	5-fluoro-Xylopyranosyl
AA	Auxiliary Activity
AxC1-OR	OR group axially attached to the anomeric carbon C1 in a sugar ring
CAZyme	Carbohydrate-Active enZymes
CE	Carbohydrate Esterases
Chf	Chloramphenicol
DMSO	Dimethyl Sulfoxide
DSF	*Differential Scanning Fluorimetry*
EC numer	Enzyme Commission number
eDNA	environmental DNA
EqC1-OR	OR group equatorially attached to the anomeric carbon C1 in a sugar ring
ERT	Enzyme Replacement Therapy
FCE	Free Cellular Extract
FPLC	Fast-Performance Liquid Chromatography
GAA	Acid α-glucosidase A
GH	Glycoside Hydrolase
GT	GlycosideTransferase
HMM	Hidden Markov Model
HMNG	High-mannose N-glycans
HPLC	High-Performance Liquid Chromatography
IC50	Half maximal inhibitory concentration
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kan	Kanamicin
LB	Lysogeny Broth
LPMO	Lytic polysaccharide monooxygenase
mDNA	metagenomic DNA
MSA	Multiple Sequence Alignment
NAC	*N*-acetylcysteine
NB-DNJ	*N*-Butyldeoxynojirimycin
NGS	Next-generation sequencing technologies
NHE-DNJ	*N*-hydroxyethyl-DNJ

PCR	Polymerase chain reaction
PCT	Pharmacological chaperon therapy
PET	Poly(ethylene terephthalate)
PL	Polysaccharide Lyase
pNP	para-Nitrophenol
PUF	Protein with unkwon function
PUL	Polysaccharide utilisation loci
Sus	Starch utilisation system
Tm	Melting temperature
WT	Wilde type
β-O-Glc	β-octyl-glucopyranoside

RIASSUNTO

La domanda di approcci più sostenibili ed economicamente vantaggiosi, che sostituiscano i tradizionali processi chimici, è fortemente in aumento nell'ultimo ventennio. In questo contesto, i catalizzatori enzimatici rappresentano una promettente soluzione ecologica. In particolare, l'utilizzo di biocatalizzatori da microorganismi da ambienti estremi permette in molti casi di lavorare in condizioni generalmente più efficenti ai fini del bioprocesso ma che risultano non applicabili per la loro controparte mesofila. Tra le classi enzimatiche più adoperate ritroviamo enzimi attivi sui carboidrati (CAZymes), che rappresentano il 44% del mercato mondiale degli enzimi. Questi infatti svolgono ruoli essenziali nella degradazione, modifica e sintesi di zuccheri complessi. Tra questi spiccano gli enzimi con capacità idrolitica, glicosil idrolasi (GH), che risultano indispensabili in varie applicazioni biotecnologiche, che vanno dalla produzione di biocarburanti, alla filiera alimentare e alla produzione di biopolimeri.

La scoperta di nuove GH tramite analisi *in silico* rappresenta un approccio indispensabile nelle moderne biotecnologie. Sfruttando le grandi quantità di dati genomici e metagenomici, i ricercatori possono identificare potenziali nuovi enzimi con capacità uniche non presenti nei database noti.

Scoperta di nuovi CAZymes da ambienti estremi tramite *sequences-based metagenomics*

Lo metagenomica si occupa dello studio del DNA ambientale (eDNA). Grazie agli avanzamenti nelle tecnologie di sequenziamento di nuova generazione, la metagenomica è cresciuta rapidamente, consentendo la comprensione dell'origine della vita e la scoperta di nuove specie. L'identificazione di nuovi enzimi ipertermofili è cruciale per le biotecnologie industriali grazie alla loro stabilità e attività a elevate temperature, che migliorano l'efficienza e la sostenibilità dei processi. Tuttavia, l'isolamento, la crescita e lo studio in laboratorio dei microrganismi (iper)termofili presentano sfide significative, a causa delle severe condizioni ambientali in cui questi organismi abitualmente prosperano. L'analisi del metagenoma consente di studiare questi microrganismi, rendendo l'analisi del eDNA da ambienti estremi una fonte inestimabile di nuovi termozimi. Per identificare nuove glicosidasi di origine ipertermofila, sono stati selezionati tre dataset metagenomici disponibili in banche dati pubbliche, provenienti da Octopus Spring (USA), Raoul Island (Nuova Zelanda) e le isole Kerguelen (Oceano Indiano). L'analisi di questi dataset ha portato all'identificazione di circa 8.000 putative GH di potenziale interesse per applicazioni industriali, ora raccolte in un nuovo database denominato *Meta*GH.

Esplorazione della biodiversità funzionale della famiglia GH122: identificazione di un nuovo enzima dalla solfatara Pisciarelli tramite un approccio metagenomico

L'utilizzo di metodiche di analisi *in-silico* di metagenomi da ambienti estremi è uno strumento inestimabile per la scoperta di nuovi enzimi. Questo approccio, tuttavia se non affiancato allo studio biochimico e strutturale di questi putativi enzimi, diviene una risorsa non pienamente sfruttata. In particolar modo, la scoperta di nuove sequenze associate allo studio biochimico, permette di ampliare le nostre conoscenze riguardo la biodiversità di alcune famiglie ad oggi poco studiate. Tra queste, GH122 è una famiglia che presenta un solo un membro caratterizzato con attività α-glucosidasica. Inoltre, questa famiglia è caratterizzata da sole sequenze di origine archaeale. Tramite approcci metagenomici, è stato possibile costruire il *CAZome* del sito di campionamento e identificare un nuovo enzima assegnato alla famiglia GH122, denominata GH122_Pool2. Questo enzima è filogeneticamente distante dall'unico caratterizzato e inoltre lo studio biochimico ha rivelato che GH122_Pool2 è attivo su 4-nitrofenil-α-D-glucuronide e xilano, suggerendo una potenziale attività come α-glucuronidasi. Lo scopo principale di questo lavoro dimostrare come l'analisi metagenomica combinata con la caratterizzazione biochimica possa portare alla scoperta di nuovi estremozimi con attività finora mai osservate. Inoltre, lo studio di questo nuovo enzima getta le basi per l'identificazione delle ragioni molecolari dietro la diversa specificità di substrato e di approfondire la biodiversità funzionale della famiglia.

L'analisi di PUL come strategia per la scoperta di nuove famiglie di glicosil idrolasi

I carboidrati complessi necessitano di una vasta gamma di enzimi specifici per la loro completa scomposizione in zuccheri più semplici. Microrganismi come Bacteroidetes hanno sviluppato sistemi poligenici specializzati che permettono la degradazione dei carboidrati definiti ***P****olisaccharides* ***U****tilization* ***L****oci* (PUL). Ogni PUL è costituito da geni presenti nello stesso operone e funzionalmente correlati, coinvolti nel *management* dei polisaccaridi. La prima prova di un sistema molecolare concertato per la degradazione complessa dei glicani è stato il sistema di utilizzazione dell'amido (Sus) del commensale intestinale umano *Bacteroides thetaiotaomicron*. Ciò ha gettato le basi per la formulazione del paradigma dei PUL: modello cellulare generale per lo studio di queste gruppi genici.
Partendo da questo modello, è stato presentato un approccio bioinformatico per la previsione dei PUL utilizzando genomi disponibili e un database dove sono riportati tutti i PUL identificati sperimentalmente e tramite l'analisi *in-silico*, PULDB all'interno del database CAZy (www.cazy.org). Questo ad oggi include 73.782 previsioni PUL in 2.227 specie di Bacteroidetes. Questo approccio altamente efficiente ha

permesso l'annotazione di numerosi nuovi CAZyme all'interno dei PUL a partire da proteine di funzione ignota definite PUF (*Proteins of Unknown Function*),
Tramite l'annotazione di nuovi PUL e l'analisi di sequenze PUF è stato possibile identificare i capostipiti di nuove putative famiglie di GH. Tramite una *pipeline* computazionale innovativa e l'utilizzo di una piattaforma ad alta produttività è stato possibile, partendo da 48 cloni, identificare 9 nuovi enzimi appartenenti a nuove putative famiglie di GH, di cui è stata effettuata una dettagliata caratterizzazione biochimica.

SUMMARY

There is an urgent need to replace traditional chemical processes in industries which generate toxic by-products with more sustainable and economically efficient solutions. Enzyme-based catalysts offer a promising ecological alternative, particularly those derived from microorganisms in extreme environments. These biocatalysts, such as carbohydrate-active enzymes (CAZymes), play essential roles in breaking down and synthesising complex sugars. Glycosidehydrolases (GHs) are especially important for industrial applications, including biofuel production and biopolymer manufacturing.

Enzyme discovery was the keyword of this PhD thesis. The goal is to identify and characterise new GHs through metagenomic and genomic approaches for biotechnological applications and to deepen our understanding of the GH family. Metagenomic analysis performed in this work allows the identification of approximately 8,000 new putative GH enzymes from three extreme environments. It develops a novel database, *Meta*GH, representing a solid reservoir of novel hypothetical glycosidases with potential utility in industrial biotransformations.

Additionally, by combining in-silico metagenomic analysis of the Pisciarelli hot springs in Naples with detailed biochemical studies, a novel member of the GH122 family was identified. This study allowed for a deeper understanding of the functional biodiversity within this, to date, poorly characterised family and provided the background to identify the molecular determinants of the different activities identified so far.

Beyond metagenomics, genome analysis, explicitly focusing on Polysaccharide Utilisation Loci (PUL), has emerged as a resource for identifying new GH families. PUL consists of physically linked and functionally related genes found in microorganisms such as Bacteroidetes, which enable the degradation of complex carbohydrates. Starting with identifying sequences of Proteins of Unknown Function (PUF) present in the PUL, the team of Professor Nicolas Terrapon discovered several new CAZymes through bioinformatics tools and high-throughput platforms. During the period abroad, 6 novel GHs were identified and biochemical characterised. These novel enzymes will be the seed for determining novel GH families.

CHAPTER 1: INTRODUCTION

With the pressing need to replace the traditional chemical processes used in industry, which generate toxic and polluting by-products, the demand for more sustainable and economically advantageous tools is rising. In this context, the potential of enzyme catalysts is not just a beacon of hope but a promising reality. Enzyme-catalyzed reactions, specifically producing fewer by-products with lower toxicity, offer a bright future [1]. Given the extensive use of enzymes in bioprocesses, the market for these biocatalysts continues to thrive. The global enzymes market size was estimated at USD 60.48 billion in 2023 and is expected to grow at a compound annual growth rate (CAGR) of 4.9% from 2024 to 2030. (https://www.grandviewresearch.com/industry-analysis/enzymes-industry).

1.1. Enzymes in biotechnology

1.1.1. The role and impact of enzymes in biotechnology

Enzymes play a pivotal role in biotechnology, acting as biocatalysts to accelerate chemical reactions under mild conditions, aligning with green chemistry principles. Various enzymes have been harnessed for biotechnological applications, significantly impacting pharmaceuticals, biofuels, and environmental remediation industries. Oxidoreductases, transferases, and lyases are crucial in multiple industrial processes; however, enzymes with hydrolytic activity hold a primary role in numerous applications, especially in managing products based on lipids, proteins, and even plastics [2]. Lipases from *Rhizomucor miehei* and *Candida rugosa* are used in the food industry to produce Cocoa Butter equivalents and Human Milk fat substitutes [1]. These enzymes have the advantage of functioning in organic solvents and can be employed as enhancers of stability in methanol [3]. Another significant class of enzymes with biotechnological applications are proteases. For example, in the detergent industry, alkaline proteases are used to break down protein-based stains, enhancing the cleaning power of laundry detergents [4]. Enzymes also play a critical role in bioremediation, where they contribute to the degradation of environmental pollutants. A novel class of enzymes active on Poly(ethylene terephthalate) (PET) is now a focal point for modern research [5]. These examples illustrate the versatility and importance of enzymes in biotechnology, where their applications extend across a wide range of industrial fields. Enzymes' ability to catalyse reactions under mild and environmentally friendly conditions makes them invaluable tools in advancing green chemistry and sustainable processes. Moreover, the ongoing research into novel enzymes, such as those capable of degrading plastics like PET, highlights the dynamic nature of this field, where discoveries continuously expand the potential for biotechnological innovation.

As the demand for sustainable solutions grows, the role of enzymes becomes even more critical, particularly in the context of carbohydrate-active enzymes (CAZymes). CAZymes are specialised enzymes that act on carbohydrates, playing essential roles in complex sugars' degradation, modification, and synthesis. Their unique capabilities make them indispensable in various biotechnological applications, ranging from biofuel production to the creation of functional foods and biopolymers. In the following sections, we will explore the biotechnological significance of CAZymes, examining how they contribute to current industrial processes and future innovations [6].

1.1.2. Biotechnological applications of extremozymes

Thermophilic microorganisms grow at temperature between 55 °C and 80 °C, while those that grow ≥ 80 °C are hyperthermophiles [7]. These organisms live in extreme temperatures, such as volcanic sites and underwater hydrothermal hot springs [8]. This feature is possible due to the evolution of specific molecular adaptations.

As the growth temperature increases from 45 to 65 °C, the lipid composition of membranes changes: diether lipids (archaeol-based) decrease from 80% to 20%, while standard caldarcheol-based and cyclic archaeol-based lipids increase from 10% to 40% [9].

These enzymes show higher ionic interactions than hydrophobic interactions [10], and a prevalence of positively charged amino acids on the surface increases the formation of salt bridges [9]. Moreover, more disulfide bonds solidify the tertiary and quaternary structures, preventing their alteration and denaturation [11]. All these factors represent some of the critical determinants of higher extremozymes thermal stability.

(Hyper)thermophilic microorganisms represent an exciting source for discovering new enzymes since they have exploitable properties in biotechnological and commercial fields. The interest in these enzymes arises from the need to replace the traditional chemical processes used in industry, which generate toxic and polluting by-products, favouring a more bio-sustainable and economically advantageous choice involving biotechnological approaches [12]. An example comes from Resinase®, a company which has patented a hyperthermophilic lipase used in the quality control of paper in the pulp and paper industry in North America, China, Japan, and other countries, as it can eliminate pitch residues [13]. Thermophilic L-asparaginases (L-ASNases) from *Thermococcus sibiricus* hold significant potential as therapeutic enzymes, offering a promising approach in cancer treatment due to their effectiveness against specific cancer cell lines [14]. Another example is the polymerase chain reaction (PCR), based on the exploitation of thermophilic enzymes such as the thermophilic DNA polymerases of *Thermus aquaticus, Pyrococcus furiosus* and *Thermococcus litoralis*, otherwise known as Taq, Pfu and Vent

[15]. As previously mentioned, the search for biocatalysts capable of degrading PET is becoming increasingly critical in enzyme biotechnology research. Thermophilic PETases are particularly valuable because they can operate at higher temperatures than their mesophilic counterparts, which significantly increases the rate of PET hydrolysis and makes the recycling process economically viable [16].
Using (hyper)thermophilic enzymes has different advantages in biotechnology (medical, industrial, and agricultural). In particular, they can be subject to conditions that rapidly denature mesophilic enzymes, such as high temperature, extreme pH and high concentration of detergents and organic solvents. The use of purified enzymes, stable at denaturing conditions for the mesophilic counterpart, has several advantages such as *i*) the reduced possibility of contamination, *ii*) the decreased viscosity and increased diffusion coefficient of the substrates, *iii*) the increased solubility and availability of organic substrates during biocatalysis as well as *iv*) a more efficient catalytic activity [17].

1.2. Carbohydrate-Active enZymes (*CAZymes*) and CAZy database

Carbohydrates are arguably the most abundant and structurally diverse class of molecules. Despite their basic chemical structure, they can form a vast array of complex molecules due to the stereochemical diversity of their hydroxyl groups, the numerous ways in which monosaccharides can be linked, and the variety of non-carbohydrate groups that can attach to them. These complex carbohydrates are ubiquitous and serve various biological roles, including energy storage, structural support, and facilitating recognition processes within and between organisms [1]. Their diversity is regulated by a specific set of enzymes known as Carbohydrate-Active enZymes (*CAZymes*), which play a crucial role in synthesising and breaking down these molecules, thereby shaping the biological landscape. This primary function of CAZymes led to the creation of the CAZy database in 1998, dedicated to the genomic, structural, and biochemical analysis of *CAZymes*. The classification system of the CAZy database is based on significant amino acid sequence similarity, with each family including at least one biochemically characterised member. The database provides detailed information on the catalytic mechanism if reported, available 3D structures, quick access to amino acid sequences and literature references. Importantly, CAZy does not include any families with unknown functions, ensuring that the information within the database is reliable and verified. This reliability makes the CAZy database a trusted resource for the CAZymes [18][19].
Enzyme families within the CAZy database can encompass enzymes with a wide range of substrate specificities. When a strong correlation exists between sequence and substrate specificity, the enzymes within a family can be further divided into sub-families [20]. This classification helps

understand and predict the specific functions of enzymes based on their sequences. Additionally, enzyme families with the same three-dimensional structure are often grouped into "clans." These clans suggest a common evolutionary origin, indicating that the enzymes likely evolved from an ancestral DNA sequence, even if they now act on different substrates [21], [22]. CAZyme family classification is now a fundamental tool for analysing and understanding how organisms and communities synthesise and degrade complex carbohydrates.

The CAZy database is constantly updated, yet as of December 2009, less than 10% of the proteins listed in CAZy had been enzymatically characterised, and fewer than 1% had a resolved structure. In the eight years following CAZy's inception, the number of annotated sequences increased fourteen-fold, while the number of enzymatic and structural characterisations nearly doubled. This disparity highlights the non-linear relationship between the annotation of sequences as CAZymes and the corresponding enzymatic and structural analyses [23]. This discrepancy persisted over time (Fig. 1.1), highlighting a critical gap in functional validation and emphasising the need for more targeted research to verify the vast number of putative CAZymes experimentally.

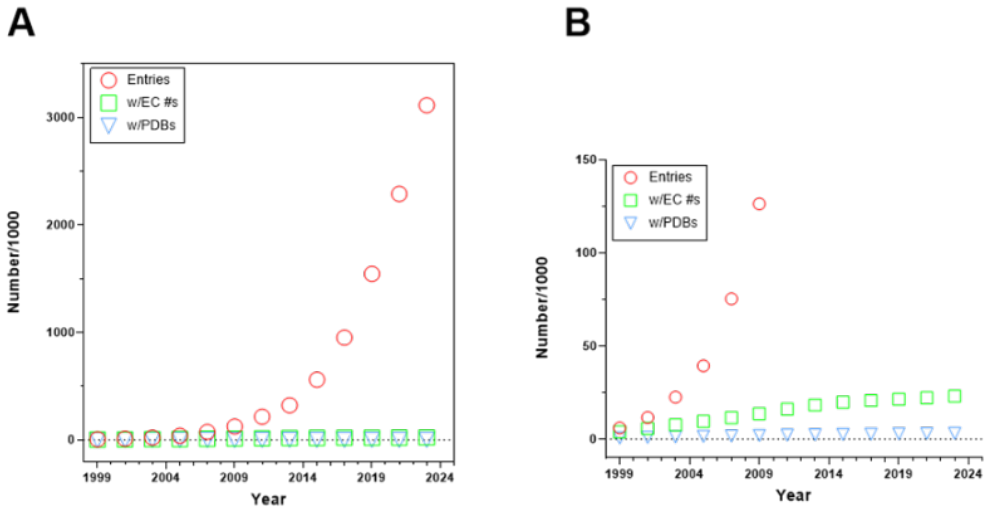

Fig. 1.1: A. the number of the CAZymes noted from 1999 to 2023, compared to the number of CAZymes characterised and had a solved structure; **B**. close-up of the plot in panel A plot to emphasise the growth of the annotated sequences vs the biochemical and structural characterisation. Prof. Pedro M. Coutinho kindly provided the data for this plot from the AFMB - CNRS in Marseille.

Addressing this imbalance is crucial for leveraging the full potential of “omic” data to understand carbohydrate-active enzymes and their applications. To date (August 2024), more than 509 families are known, divided into six enzymatic classes (Fig. 1.2): Glycosyl-transferases (GT), Polysaccharide lyases (PL), Carbohydrate esterases (CE), Carbohydrate-

binding modules (CBM), Auxiliary enzymes for redox reactions (AA), and Glycosyl-hydrolases (GH).

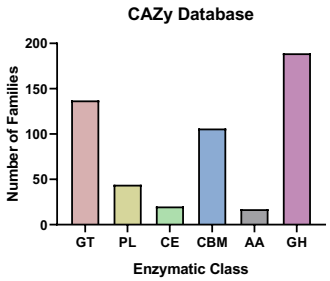

Fig. 1.2: CAZy database enzymatic classes plot, showing the enzymatic class (GT, PL, CE, CBM, AA, and GH) present in the database as a function of the number of families reported to date (October 2024).

1.2.1. Glycoside Hydrolases (GHs)

Glycoside hydrolases (GH) (EC 3.2.1.-) are the class of enzymes capable of cleaving the glycosidic bond, the most stable covalent bond in nature (half-life of $4.7*10^9$ years), accelerating its hydrolysis up to 17 orders of magnitude compared to the spontaneous reaction, making the one implemented by glycoside hydrolases one of the most efficient catalytic mechanisms known. GHs represent a vast and diversified category, present in all organisms and some types of viruses [20], including glycosidases and transglycosidases, responsible for the hydrolysis and transglycosylation of the glycosidic bond. To date (October 2024) this class includes 189 families, 37% of the total families in the CAZy database. Given their relevance in the biotechnological and biomedical fields, they are the enzymatic class best recognised by the CAZy tool [23].

GHs can be classified according to different criteria. Traditionally, glycosidases are distinguished by their substrate specificity, according to the IUBMB Enzyme nomenclature, and in some cases by their reaction mechanism, as is the case of α-amylases EC 3.2.1.1 and β-amylases EC 3.2.1.2 (Enzyme Nomenclature, 1992). This type of classification is however subject to complications, as some glycosidases exhibit a broader substrate specificity, and, in addition, some glycoside hydrolases (EC 3.2.1) also have transglycosidase activity. The substrate's structural characteristics—the configuration of the cleaved glycosidic bond (α and β) and the size of the hemiacetal/ketal ring (pyranosides or furanosides)—further deepen this macro-distinction. Nevertheless, Koshland's 1953 studies allowed glycoside hydrolases to be classified into two main macro-categories: *retaining* and *inverting* ones, based on the configuration of the product's anomeric carbon concerning the original substrate [24].

Retaining glycoside hydrolases are those enzymes that retain the configuration of the anomeric carbon during catalysis. These GHs operate via a two-step S_N2 mechanism involving the formation of a covalent intermediate (Fig. 1.3 A). Key players in this reaction are two residues

within the catalytic site that function as an acid/base and a nucleophile. The first step is glycosylation, where the nucleophile attacks the anomeric carbon, forming a covalent glycosyl-enzyme intermediate with the assistance of the acid group, which donates a proton to the leaving group. In the second step, deglycosylation, the carboxylic acid group, which functions as a base after acting as an acid catalyst, is deprotonated. Upon entering a water molecule into the active site, a proton is abstracted from this molecule, facilitating deglycosylation of the enzyme and the product's release. If a molecule other than water, such as an alcohol or sugar, enters the catalytic site, a transglycosylation reaction may occur [24], [25].
On the other hand, *inverting* glycoside hydrolases are enzymes that reverse the configuration of the anomeric carbon of the substrate in the product. Inverting GHs operate through single-step acid-base catalysis (S_N1) without forming a covalent intermediate. The reaction is driven by a pair of residues designated as acid and base. Catalysis begins with the entry of the substrate and a water molecule into the active site. After being deprotonated by the base residue, the water molecule attacks the anomeric carbon in an orientation opposite side of the O-glycosidic bond, which, assisted by the acid residue donating a proton, is released (Fig. 1.3 B) [24], [25].

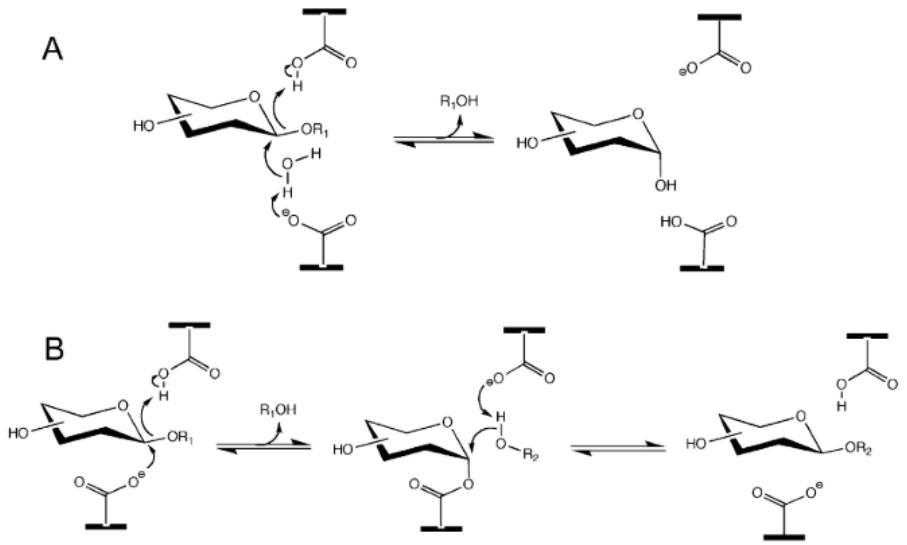

Fig. 1.3: Mechanisms of reaction of (**A**) *inverting* GHs and (**B**) *retaining* β-glycoside hydrolases [25].

Both *retaining* and *inverting* GHs feature acidic amino acids as catalytic residues in their active sites, namely Glutamic Acid and/or Aspartic Acid [26]. Structural studies highlight that *retaining* and *inverting* glycoside hydrolases are characterised by different average distances between their catalytic residues: approximately 9 Å for *inverting* GHs, consistent with the

active site being occupied by both the substrate and water during catalysis, and approximately 4.5 Å for *retaining* GHs [27].
Moreover, concerning the classification of GHs, it is worth highlighting that some GH families do not follow the mechanisms described above. Examples are represented by families GH4 and GH109, which exploit a mechanism that involves a NAD^+ molecule as a cofactor [24], [28]. Moreover, some GH families, such as GH18 and GH20, utilise neighbouring group participation mechanisms, where either a 2-acetamido or 2-hydroxyl group acts as an intramolecular nucleophile, leading to the formation of a reactive intermediate such as an oxazolinium ion or an epoxide, respectively [29].

Finally, in the case of GHs acting on oligo- and polysaccharides, these are subdivided, according to the position in which the hydrolysis occurs, into exo- and endo- refers to the ability of a glycoside hydrolase to cleave a substrate at the end (most frequently, but not always, the non-reducing end) or within the middle of a chain (Fig. 1.4) [19].

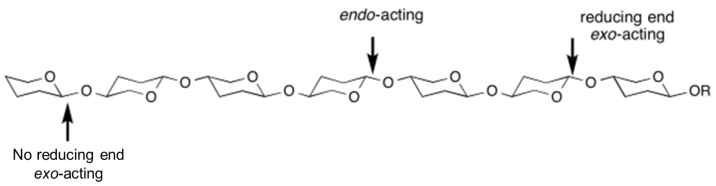

Fig. 1.4: Schematic representation of *exo*- and *endo*-activities (https://www.cazypedia.org/index.php/Glycoside_hydrolases).

1.2.2. The Pivotal Role of GHs in Industrial Biotechnology

According to an analysis by Fact.MR (https://fact.mr/), *CAZymes* represented approximately 44% of the global industrial enzyme market in 2022. This highlights the significant role that this class of biocatalysts plays in various industrial applications, particularly those involving (hyper)thermophilic enzymes such as glycosidases (GHs) of bacterial and archaeal origin. These enzymes are crucial for biomass management, as they facilitate the production of fermentable sugars from lignocellulosic materials and essential bioethanol biorefineries [17] to replace fossil fuels. Biomass management is a vital component of the circular economy, but despite biomasses' vast and heterogeneous availability, it remains underutilised [30]. With global energy demand projected to increase by 28% by 2040, there is an urgent need to develop alternative energy sources [31]. Thermophilic (hemi)cellulases are particularly promising for the degradation of lignocellulosic biomass, which can then be converted into fermentable sugars for biohydrogen production [10] and for second (produced by agricultural waste and forest biomasses) and third-generation (algal biomasses) bioethanol [32] more ethically acceptable. In addition, biomass can be processed to produce oligosaccharides with high

biotechnological value, which is relevant in pharmaceutical (anti-tumour activity [33]), nutraceutical (prebiotics [34]), and agricultural (elicitors of plant immunity [35]). However, the structural complexity of biomass polysaccharides poses challenges for their hydrolysis, requiring the use of tailored enzymatic cocktails containing CAZymes like Glycoside Hydrolases (GHs) and Polysaccharide Lyases (PLs) under extreme conditions [36]. For instance, some GHs, such as cellulases and α-amylases, are key biodegradative agents in modern laundry detergents—for example, the α-amylases from *Bacillus sp.* C37PLCA demonstrate high compatibility with detergents, maintaining optimal activity at alkaline pH and stability in formulations like Ariel® and OMO Matik®, thus enhancing detergent performance [37]. Others are active in the food and beverage industry. For example, β-mannanases from *Bacillus pumilus* (M27) exhibit high stability across a wide range of temperatures (30–80°C) and pH levels (3–11). These enzymes can be used in the fruit juice industry to lower the viscosity caused by the breakdown of the mannan fraction of hemicellulose in fruit juices [38]. GHs are also central to pharmacological studies for discovering and developing new drugs, owing to their defined substrate-binding pockets, which pharmaceutical enzyme inhibitors can target. Integrating classical enzymology with modern experimental and data analysis techniques opens new avenues for drug discovery [39].

1.3 Uncovering New Biocatalysts: the impact of *in-silico* analysis and metagenomics approaches

A significant amount of time, effort and cost are currently associated with discovering enzymes for specific applications. The discovery of new carbohydrate-active enzymes through *in-silico* analysis represents a transformative approach in modern biotechnology. By leveraging the vast amounts of (meta)genomic data, it is possible to identify potential novel enzymes with unique capabilities not yet characterised [40].

In-silico approaches enable the rapid screening of large datasets to predict enzyme function based on sequence homology and the analysis of structural motifs and domains. These computational methods significantly accelerate the discovery pipeline, allowing for the identification of enzymes with desirable properties, such as enhanced stability or specificity, that might be challenging to find through traditional experimental methods alone.

1.3.1. Shotgun metagenomics and sequence-based enzyme discovery

Metagenomics, first defined in 1998 [41], refers to the study of environmental DNA (eDNA) and has since evolved into a crucial field within genomics. This approach is not restricted to traditional natural environments but encompasses various sampling sites, ranging from termite hindguts [42] and cheese rinds [43] to the International Space

Station [44]. Next-generation sequencing technologies (NGS) enabled metagenomics, which has rapidly grown, significantly expanding our understanding of the Tree of Life and identifying novel, deeply rooted microbial lineages, such as the DPANN (superphylum of Archaea including Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota and Nanoarchaeota, the first five groups discovered) archaea and candidate phyla radiation [45].
Due to the extreme conditions in which they thrive, (hyper)thermophilic microorganisms are often challenging to isolate, cultivate, and study in the laboratory, as approximately 99% of bacteria cannot be cultured in vitro in a laboratory environment [40]. Metagenomic analysis is the only way to access the genetic material coding for hyperthermostable enzymes when isolating microorganisms requiring such conditions, which proves unsuccessful [46]. In this context, metagenomic DNA (mDNA) from extreme environments represents a valuable source of novel thermozymes, expanding our metabolic and enzymatic knowledge of the present species [47].
Metagenomics necessitates analysing a substantial number of mDNA sequences to elucidate the genetic diversity within environmental samples. The examination of the metagenome, particularly in the context of enzymatic discovery, encompasses the sequencing of collected environmental samples and subsequent data analysis. This analytical process [48] can be organised as follows: (Fig. 1.5):

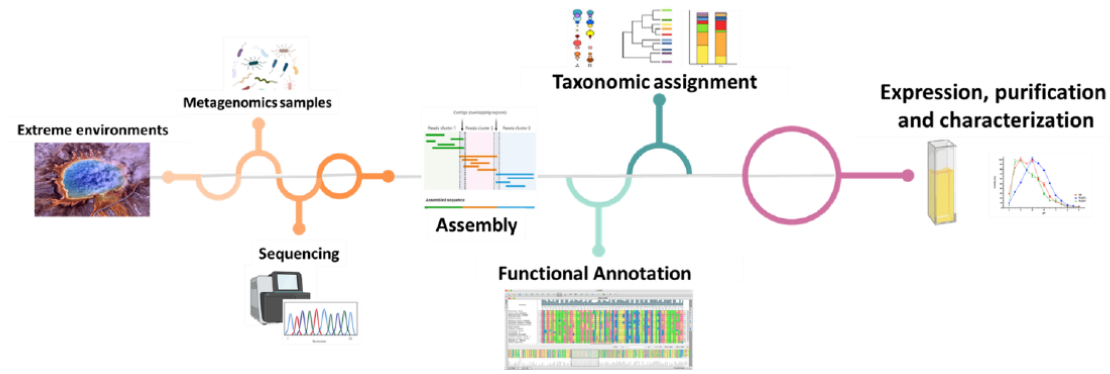

Fig. 1.5: Graphical representation of enzyme discovery by sequence-based metagenomics

a)Extreme environments, Metagenomics samples and sequencing. The first important step is the selection of environments based on geophysical and climatic parameters. However, extracting metagenomic DNA (mDNA) from samples collected in extreme environments is critical to the entire metagenomic analysis workflow. To ensure that the genetic information obtained is representative of the whole microbiome, the preparation of mDNA requires specific protocols that preserve the quality and quantity of the nucleic acids. This is essential to ensure the creation of the best

metagenomic library and the subsequent sequencing approach. Moreover, choosing the best NGS technologies for the proposed project is necessary to perform an assembly with longer and more accurate contigs, which will be required for good metagenomic studies.

b) *Assembly*. It is a process through which the individual sequences read by the NGS technologies are put together in contiguous chains called *contigs*. This step uses complete genomic sequences in the databases to align the contigs obtained from sequencing. It exploits the genomes already known as a guide to find overlaps between contigs and map them.

c) *Functional annotation*. It consists of predicting gene function and identifying the genes encoded by the mDNA sequences and, consequently, their functions. In particular, annotation consists of searching for the homologies between the sequences of the metagenomes and those already present in the databases.

d) *Taxonomic assignment*. By definition, the metagenome is composed of DNA from many different species in terms of "microorganism consortium." Therefore, the first objective of the analysis must be to identify and quantify the most significant possible number of species present in a sample to evaluate the different relative abundances and biodiversity of a specific site.

Activity-guided screening and PCR-based screening represent an alternative for discovering novel enzymes, but they have several limitations. Activity-guided screening is advantageous in detecting new enzymes [45]. This approach centres on the construction and functional screening of libraries of suitable vectors containing metagenomic DNA inserts to identify enzymes of interest [49]. This method effectively identifies industrially relevant enzymes like lipases and cellulases and guarantees that the enzymes discovered are functional in a specific host like *E. coli*. Its downside lies in being limited to genes and gene clusters that can be expressed in the expression host, which might narrow the scope of discovery to only those enzymes that fit this criterion [45].

As the name suggests, PCR-based functional screening uses degenerate primers to amplify genes from eDNA coding for protein domains of interest [50]. PCR-Based Screening excels in sensitivity for detecting low-abundance sequences and effectively identifies variations within specific gene families [51]. It is relatively inexpensive and can pinpoint changes at the single nucleotide level. However, it has several limitations, such as a requirement for conserved DNA motifs and inefficiency in detecting novel sequences. It also provides minimal taxonomic information and is biased against GC-rich sequences, which may skew the discovery results [45].

Shotgun metagenomics is widely regarded as a superior approach for many applications due to its comprehensive, untargeted sequencing of environmental DNA (eDNA), without biases as amplicon PCR [52]. Unlike targeted sequencing methods, shotgun metagenomics enables the

detection of known and unknown genes, providing a holistic view of microorganism communities. This method captures a broad range of genetic material, including from rare or low-abundance organisms, allowing for a more nuanced understanding of community composition and diversity. This method allows for the simultaneous examination of all genetic material within a sample, providing a putative functional and taxonomic profile of the environmental community [45].
Hence, the *in silico* metagenomics approach is a holistic and unbiased method; it represents a suitable approach to identifying potentially novel CAZymes from non-cultured organisms and detecting evolutionary profiles of community function and structure. Thus, this approach provides an immense database for searching for new thermozymes beneficial in the biotechnological field. In Chapter 2, a clear example will be presented to showcase the potential of metagenomic analysis for discovering new biocatalysts. Furthermore, in Chapter 3, the characterisation of an enzyme found through this approach will be detailed, demonstrating how annotation and characterisation must go hand in hand to enable both the advancement of knowledge and the development of practical applications.
In addition, *in-silico* analysis can be complemented with various methodologies to achieve a broader spectrum of data and reduce the risk of missing critical information. For instance, enrichment techniques applied to the biomass of interest from environmental samples can enhance the detection of low-abundance organisms or functional genes, providing a more comprehensive insight into the microbial community and its potential applications. This combined approach allows for a deeper exploration of target enzymes while capturing a more comprehensive array of relevant biological information [53].

1.3.2. Polysaccharide Utilisation Locis (PUL) analysis

Complex carbohydrates, such as oligosaccharides and polysaccharides, comprise various monosaccharide subunits and glycosidic linkages. As a result, a vast array of specific enzymes is necessary for their complete breakdown into simpler sugars, which can then enter primary metabolic pathways. Microorganisms such as Bacteroidetes have developed specialised operon systems focused on polysaccharide metabolism to efficiently break down these complex carbohydrates. [54].
This compartmentalisation allowed these bacteria to develop a specialised strategy for carbohydrate degradation organised within polysaccharide utilisation loci (PULs). Each PUL consists of physically linked and functionally related genes that encode co-regulated proteins responsible for (*i*) binding polysaccharides to the bacterial surface, (*ii*) initiating partial cleavage of these polysaccharides at the outer membrane, (*iii*) transporting the resulting oligosaccharides into the periplasm, (*iv*) completing their breakdown into monosaccharides within the periplasmic space, and (*v*)

sensing the degradation products to regulate PUL expression accordingly [55].
The first evidence of a concerted molecular system for complex glycan degradation in Bacteroidetes bacteria, the starch utilisation system (Sus) of the human gut commensal *Bacteroides thetaiotaomicron*, consists in the PUL paradigm that outlines a general cellular model for the study of other PULs. This PUL consists of eight genes, SusA-G and SusR, including three genes that encode CAZymes. On the outer cell surface, SusDEF binds components of starch, α-amylase SusG (GH13) binds and cleaves the polysaccharide to oligosaccharides, which SusC sequestrates into the periplasmic space of the bacterium. Imported oligosaccharides, sensed by SusR leading to upregulation of the PUL, are depolymerised to glucose by neopullanase SusA (GH13) and α-glucosidase SusB (GH97) (Fig. 1.6) [56].

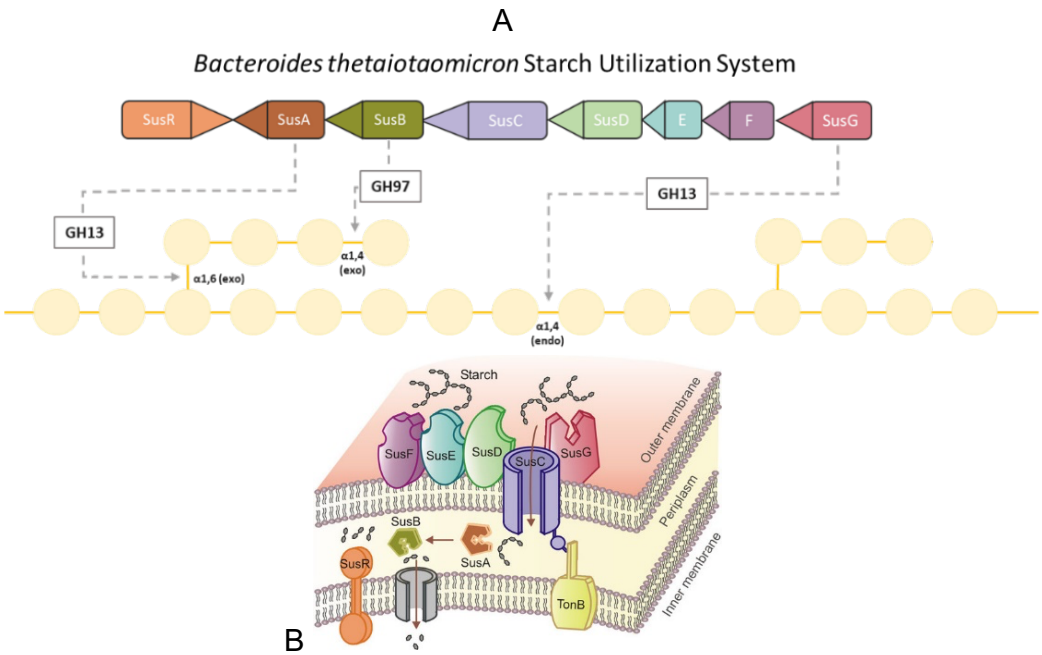

Fig. 1.6: The *B. thetaiotaomicron* Sus PUL, **the classic PUL paradigm** , conting with the genes encoding all components necessary for starch metabolism (**A**) and model for starch catabolism (**B**) [57]

Starting from the PUL paradigm in 2015, the CAZy team presented a bioinformatics approach for PUL prediction using genomes available in public databases [55]. The principle of the PUL prediction is to start from every SusCD-like gene pair and then extend PUL boundaries to operonic genes (based on intergenetic distances between genes on the same strand) and to more distant regulators and CAZyme-coding genes that catalyse polysaccharide breakdown. The PULs, predictions and literature-derived PUL data are reported in a novel database, PULDB [58]. As of

today (October 2024), 73,782 PUL predictions across 2,227 Bacteroidetes species are currently catalogued. The PULDB stands out for its collection of experimentally characterised PULs, which are integrated with the CAZy database (http://www.cazy.org/PULDB/). PULDB is a versatile resource that supports various applications, including in-depth exploration of microbial metabolic pathways; through the efficient CAZy system, these PULs have facilitated the annotation of numerous novel CAZymes. It also provides annotated gene clusters that are valuable for predicting substrate interactions.

By analysing the PUL present in *Bifidobacterium longum*, Cordeiro and co-workers have uncovered the intricate mechanisms behind the breakdown and metabolism of high-mannose *N*-glycans in this microorganism. *Bifidobacterium longum* employs a sophisticated system known as *N*-glycan PULs (N-GULs) to utilise *N*-glycans. In this context, the team investigated the molecular mechanisms driving the demannosylation of high-mannose *N*-glycans (HMNGs). Notably, *Bifidobacterium longum* possesses a diverse repertoire of α-mannosidases, including a GH38 generalist α-mannosidase (Bl_Man38B), two complementary GH38 α-mannosidases (Bl_Man38A and Bl_Man38C), and a GH125-specific α-1,6-mannosidase. This array of enzymes provides a versatile biochemical toolkit for N-glycan processing. Additionally, the specialised cleavage of Man1GlcNAc by a GH5_18 β-mannosidase and the unique metabolic pathway of mannose underscore the complex strategies used by bifidobacteria to exploit host *N*-glycans as sources of carbon and energy [59].

PUL analysis allowed not only to advance our knowledge of GHs but also to identify new putative GH families as demonstrated by the discovery of a novel putative GH129-like family from the identification of a PUL for the breakdown of carrageenan. Carrageenan, a sulfur-containing galactan in red algae, is difficult to degrade due to its protective sulfation patterns. Marine bacteria such as *Zobellia galactanivorans* have evolved PULs that encode specialized enzyme systems to metabolize carrageenan. PULs, including three sulfatases, remove sulfate groups in this bacterium, exposing the polysaccharide backbone to enzymatic degradation. A novel κ-carrageenase, GH16, and three ι-carrageenases, GH82, have been identified to break down carrageenan into smaller oligosaccharides. Furthermore, three GH127 enzymes were discovered, which degrade α-1,3-(3,6-anhydro)-D-galactose bonds and the previously mentioned GH129-like (DagB) with distinct substrate specificity [60], demonstrated the vital role of PUL analysis in advancing our knowledge of GHs.

PULs have proven to be powerful tools in discovering novel CAZymes but also understanding the metabolic pathways responsible for the breakdown of complex polysaccharides and the ecological role of the microorganism. By examining the genetic and functional organisation of PULs, researchers

have identified critical enzymes involved in glycan degradation, expanding our knowledge of microbial carbohydrate metabolism.
Several PULs have been identified in marine microorganisms capable of degrading major algal polysaccharides characterized by high structural complexity [61]. Ulvan is a major polysaccharide in green algae; its degradation plays a crucial role. So, identifying a PUL responsible for the degradation of ulvan can improve the research in this sector. A PUL including enzymes involved in ulvan breakdown by the marine bacterium *Formosa agariphila* was characterised, including P29_PDnc (a novel dehydratase), ulvan lyases, and GH78 and GH3 enzymes. This PUL is conserved in other marine bacteroids, so this study provides insights into how marine bacteria utilise complex algal polysaccharides, contributing to our understanding of marine carbon cycles [62].
In conclusion, this approach enhances our understanding of microbial ecology and uncovers enzymes with significant potential for industrial and biotechnological applications. As research in PUL analysis advances, it allows us to reveal more about enzyme functions and metabolic pathways. Combining the study of PULs with a sophisticated computational system and high-throughput enzyme analysis could allow us to fully harness this invaluable resource to its maximum potential, as reported in Chapter 4.

1.4. Aims of this PhD thesis

This PhD thesis aims to broaden our understanding of carbohydrate-active enzymes by leveraging cutting-edge metagenomic analysis and Polysaccharide Utilization Loci exploration to uncover novel enzymes with significant biotechnological potential. This research will first delve into metagenomic datasets from extreme environments, where the unique conditions offer a rich and promising source of previously undiscovered enzymes (Chapter 2). Inoltre, questa tesi presenterà (Capitolo 3) la caratterizzazione biochimica di un enzima glicoside idrolasi appartenente alla famiglia GH122, precedentemente annotato da un ambiente estremo, utilizzando un approccio metagenomico simile a quello descritto nel Capitolo 2. This case study will illustrate the potential of these enzymes, thus emphasising the relevance of metagenomics in enzyme discovery. Further, by analysing PULs in a high-throughput manner, this research aims to identify and classify new GH families, expanding the repertoire of known CAZymes (Chapter 4). Through these explorations, this thesis seeks to expand the catalogue of available enzymes and demonstrate their potential in diverse and impactful biotechnological applications.

CHAPTER 2: UNCOVERING NEW CAZYmes, METAGENOMICS FROM EXTREME ENVIRONMENTS WORLDWIDE

2.1. Aim of the Chapter

Interest in extremophiles, as they are ideal tools for industrial applications, has given impetus to metagenomics studies of extreme environments. Organisms living in these environments are often not cultivable in the laboratory. Therefore, in silico metagenomics represents a valid method for discovering and studying new CAZymes, and the interest in sequence-based screening has grown increasingly [47]. Metagenome analysis requires the sequencing of environmental samples and the functional and taxonomical annotation of the identified sequences [48].

Three metagenomic datasets from the NCBI SRA Database (https://www.ncbi.nlm.nih.gov/sra) were chosen to populate the *Meta*GH (Metagenomic Glycosidase Hub) with new putative GH sequences for future biochemical studies for the hydrolysis of recalcitrant biomasses (Fig. 2.1).

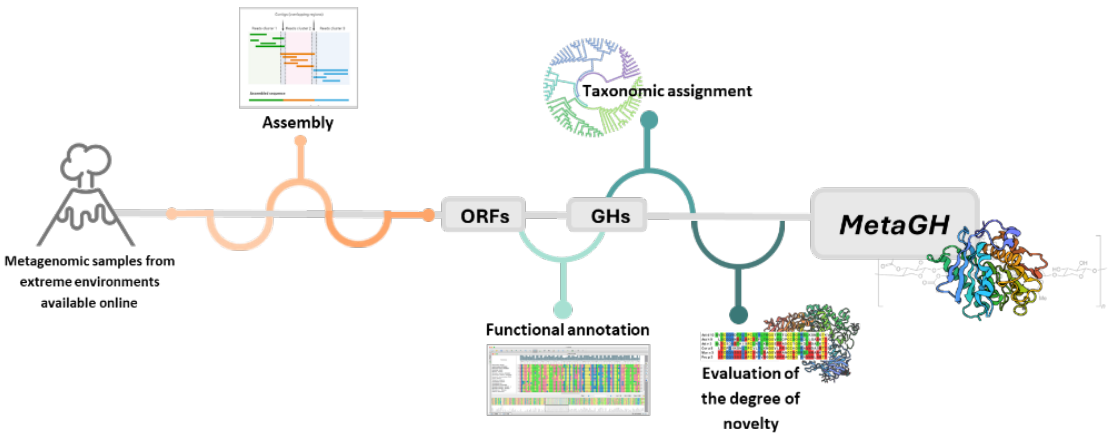

Fig. 2.1: Graphical abstract of the pipeline followed in this work to produce *Metagenomic* Glycosidase Hub (*Meta*GH).

2.2. Results and Discussion

2.2.1 Metagenomes under study

The selection of samples from Octopus Spring, Raoul Island, and the Kerguelen Archipelago was driven by their unique and extreme environmental conditions, which provide a rich source of novel and diverse microorganisms with potential biotechnological applications (Tab. 2.1).

Sample Site	Sample name	SRA Accession	Broad-scale enviromental context	Localization	Alkalinity ($mEqL^{-1}$)	pH	Temperature (°C)	Reference
Octopus spring	GCUZT	SRR10848475	Thermal springs: phototrophic mat microbial communities	44°32'02.7"N 110°47'52.2"W	Not reported		42-90	https://www.ncbi.nlm.nih.gov/sra/?term=SRR10848475
	CYNWO	SRR9650329						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9650329
	CYONH	SRR9650328						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9650328
	CYONG	SRR9650326						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9650326
	CYNWN	SRR9650299						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9650299
	GBTAH	SRR9650298						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9650298
	CYONB	SRR9649890						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9649890
	CYOHT	SRR9649884						https://www.ncbi.nlm.nih.gov/sra/?term=SRR9649883
	CYOHU	SRR9649883						https://www.ncbi.nlm.nih.gov/sra/?term=SRR949883
	CYOHS	SRR9649882						https://www.ncbi.nlm.nih.gov/sra/?term=SRR949882
	CYOHO	SRR9649876						https://www.ncbi.nlm.nih.gov/sra/?term=SRR949876
	CYONC	SRR8863765						https://www.ncbi.nlm.nih.gov/sra/?term=SRR8863765
	CYOHP	SRR8863601						https://www.ncbi.nlm.nih.gov/sra/?term=SRR8863601
	CYOHH	SRR8863542						https://www.ncbi.nlm.nih.gov/sra/?term=SRR8863542
	CYOGY	SRR8863384						https://www.ncbi.nlm.nih.gov/sra/?term=SRR8863384
	CYOGX	SRR8863379						https://www.ncbi.nlm.nih.gov/sra/?term=SRR8863379
Raoul Island	Eastern spring	SRR10484996	Neutral hot spring: depth -0.05m	29°26'25.20"S 177°92'06.24"W	0.69	7.2	91.3	https://www.ncbi.nlm.nih.gov/sra/?term=SRR10484996
Kerguelen Island	RB10	ERR6412193	Spring located on a plateau: water	49°62'53.5"S 68°79'86.9"E	25	8.7	78.4	https://www.ncbi.nlm.nih.gov/sra/ERX6040246[accn]
	RB13	ERS7179887		49°62'56.6"S 68°79'86.5"E	29	7.6	93.0	https://www.ncbi.nlm.nih.gov/sra/ERX6040247[accn]
	RB32	ERS7179888		49°62'57.9"S 68°79'89.5"E	3.2	5.8	97.0	https://www.ncbi.nlm.nih.gov/sra/ERX6040248[accn]
	RB108	ERR6412196		49°62'53.5"S 68°78'61.6"E	83	9.6	101.3	https://www.ncbi.nlm.nih.gov/sra/ERX6040249[accn]

Tab. 2.1: Metadata overview of Octopus Spring (in green), **Raoul Island** (in orange) and **Kerguelen Island** (in light blue). In order: the sample site; the sample name code available on the SRA database; The SRA accessions; Broad-scale environmental context; Localization of the samples; chemical and physical parameters of the sample site (Alkalinity, pH and Temperature); and the last line is the SRA reference.

Octopus Spring is an alkaline (pH 8.8–8.3) hot spring located several miles north of Old Faithful Geyser in Yellowstone National Park (44°32'02.7"N 110°47'52.2" W). The water flows from 95 °C at the source to 83 °C downstream. Pulses of water every 4–5 minutes raise the temperature to 88 °C (Fig. 2.2). *Thermocrinis ruber* thrives in this environment, and photosynthetic organisms, including *Synechococcus* and *Chloroflexus*, grow where temperatures range from 75 to 65 °C [63]. In 2018, Professor Devaki Bhaya's team conducted the project entitled "Hot spring phototrophic mat microbial communities from Octopus Spring, Yellowstone National Park, Wyoming, United States" (https://gold.jgi.doe.gov/project?id=Gp0323971) to explore the genetic diversity in phototrophic cyanobacteria and their interactions with viral predators. They produced 16 metagenomes available on SRA from samples.

Fig. 2.2: **Octopus Spring**: **A**. the position of Octopus spring in Yellowstone National Park, Wyoming, United States; **B**. Octopus spring; **C**. temperature overview of the pool [63]

Raoul Island, or Rangitāhua, is located 980 km northeast of New Zealand and is a prominent volcanic complex summit on the Kermadec Ridge (29°16′S 177°55′W). The island has two central calderas, one in Denham Bay to the south and the central caldera on the island, both of which are active volcanically (Fig. 2.3 A, B). In the study of 2018 by Stewart et al., using a combination of molecular and culture-dependent techniques, characterised for the first time the diversity of thermophilic microorganisms at five hydrothermal spring close to the central Raoul caldera,to investigate the relationship between geochemical makeup and microbial diversity (Fig. 2.3 C, D) [64].

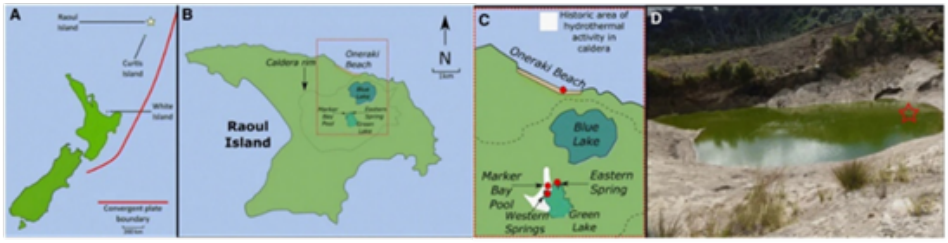

Fig. 2.3: Raoul Island: **A**, its position relative to New Zealand and other geothermally active islands. **B**. Map of the island with the main caldera, Oneraki Beach, and caldera lakes marked. **C**, Map of sampling sites inside and outside caldera and historic area of hydrothermal activity. **D**, photograph of Eastern Spring with the sample site. [64]

The **volcanic Kerguelen Archipelago** is located in the southern part of the Indian Ocean (49°20′S 69°20′E), over 3000 km from the closest inhabited area. It represents the world's third largest volcanic island complex, after Iceland and Hawaii. The current volcanic activity is evidenced by fumaroles, mud pots, hydrothermal discharges and small thermal springs that arise at sea level from at least 300 m above sea level, with temperatures ranging from 35 °C to over 100 °C [65]. This area has not been subjected to any comprehensive microbiological functional study (metabolism, physiology, adaptations) until the recent work of Allioux et al. [66] surveyed Kerguelen's microbial diversity. They performed a metagenomic analysis of the microorganisms present in Kerguelen hot springs, in particular from four different hot springs (Fig. 2.4). From this study, it is possible to obtain a first overview of the microbial diversity at Kerguelen hot springs, allow the identification of putative new taxa, namely

11 new genomic species, seven new putative genera, one new putative family and two new putative orders.

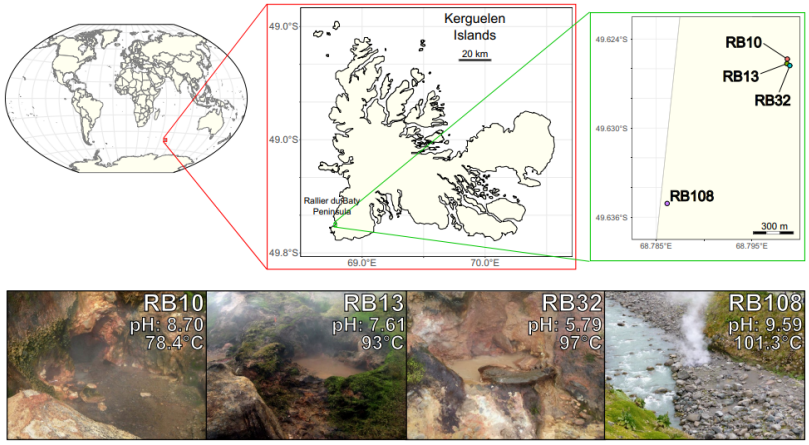

Fig. 2.4: Kerguelen Islands: In order, Location of the Islands in the Southern Ocean; Map of Kerguelen Islands, French Southern and Antarctic Territories; an overview of the Rallier du Baty peninsula with the sampling sites: RB108; RB10, RB13 and RB32.[66]

These sites were chosen for their unique environmental conditions, previous research efforts, and metagenomic data availability, all of which contribute to the potential discovery of novel enzymes and microorganisms with significant biotechnological relevance.

2.2.2. Functional annotation of novel GHs

To identify coding sequences for the enzymes of interest, each dataset was reassembled. Open Reading Frames (ORFs) were identified, and redundancies were reduced through sequence clustering with a 97% identity filter. The functional annotation of the CAZyme GHs was performed using dbCAN [7]. Differently, Kerguelen's annotation was performed in collaboration with Professor Nicolas Terrapon and the CAZy team at the AFMB of Marseille (France). Starting from 1,230,714 ORFs, we identified 7,657 putative GHs (Tab. 2.2). While the number of GHs in Octopus Spring is surprisingly high, as this dataset is the largest in terms of total sequences among the three considered, the Raoul sample, in contrast, exhibits a surprisingly low number of sequences, likely due to the quality and type of reads available in the SRA, which may have affected the assembly. Despite the low number of possible GHs, the Raoul dataset is fascinating, given the presence of two sequences from the GH122 family, which to date has only one characterised member. Hence, these putative GHs and the 9 identified in Kerguelen samples could significantly boost efforts to study and understand this enzyme family.

	ORF	GHs
Octopus Spring	954,973	6,894
Raoul Island	129,844	28
Kerguelen Island	145,897	735

Tab. 2.2: ORFs and GHs certificated in the samples.

Several GH families were identified in the samples. Regarding the enzymes active on equatorial C1-OR bonds (Fig. 2.5), many putative β-glucosidases, β-galactosidase, and β-xylosidases (826 sequences) were annotated belonging to families GH1, GH2, GH3 and GH39. They are essential in bioethanol production from lignocellulosic biomass, as they degrade oligosaccharides and polysaccharides into fermentable sugars [67]. Moreover, in Octopus Spring and Kergulend samples, several sequences belonging to GH5 were annotated. This family, in particular, groups cellulase, endoglucanase and xylanase used in bioethanol production, paper processing and the textile industry [68].
GH23s were annotated in the datasets (280 putative enzymes), a family rich in chitinase and lysozymes. These activities are involved in the degradation of chitin and peptidoglycan, respectively. The main difference between the datasets is the number of GH29 and GH51 in Octopus Spring (232 GH29s and 243 GH51s). GH29 mainly includes α-fucosides, which are crucial for the degradation of substrate with fucose, a sugar found in glycoproteins and glycolipids [69]. Regarding the GH51, the sequences annotated could be putative endo-β-1,4-glucanase, involved in the catalysis of lignocellulose, or α-L-arabinofuranosidases, used for the biotransformation of polysaccharides in food and animal feed, improving digestibility and nutritional value [70].
Regarding the enzymes active on axial C1-OR bonds (Fig. 2.6), several GH13, GH57 and GH77 were annotated. All these families are involved in starch processing as α-amylases, pullulanase, cyclodextrin glycosyltransferases, α-glucosidases, amylomaltases and 4-α-glucanotransferases [71]. In particular, GH57 enzymes represent 50% of the sequences identified in the Raoul Island dataset. This family is recognised for its thermostability (the characterised sequences reported in the CAZy database are almost all from (hyper)thermophilic microorganisms, http://www.cazy.org/GH57_characterized.html), making them suitable for industrial processes that require high temperatures, giving them great biotechnological value.
The analysis conducted on the Kerguelen dataset, in collaboration with the CAZy group, also allowed us to annotate some sequences belonging to families recently added to the database. Different sequences are annotated as GH176 (15), GH177 (11) and GH179 (2), whose families suggest hypothetical isoamylase, exo-α-sialidase and β-*N*-acetylhexosaminidase, respectively. These glycosidase families are each

represented by only one characterised enzyme in the current database. The discovery of new members offers a promising opportunity to expand their functional diversity and potential applications. Interestingly, except for the GH177 enzyme, both families GH176 (https://www.cazy.org/GH176_characterized.html) and GH179 (https://www.cazy.org/GH179_characterized.html) presented just one characterized enzyme belonging to a thermophilic source, as reported in CAZy database. This underscores the immense potential of exploring extreme environments to deepen our understanding of enzymatic biodiversity and highlights the power of metagenomics in discovering new enzymes and generating entirely new CAZyme families.

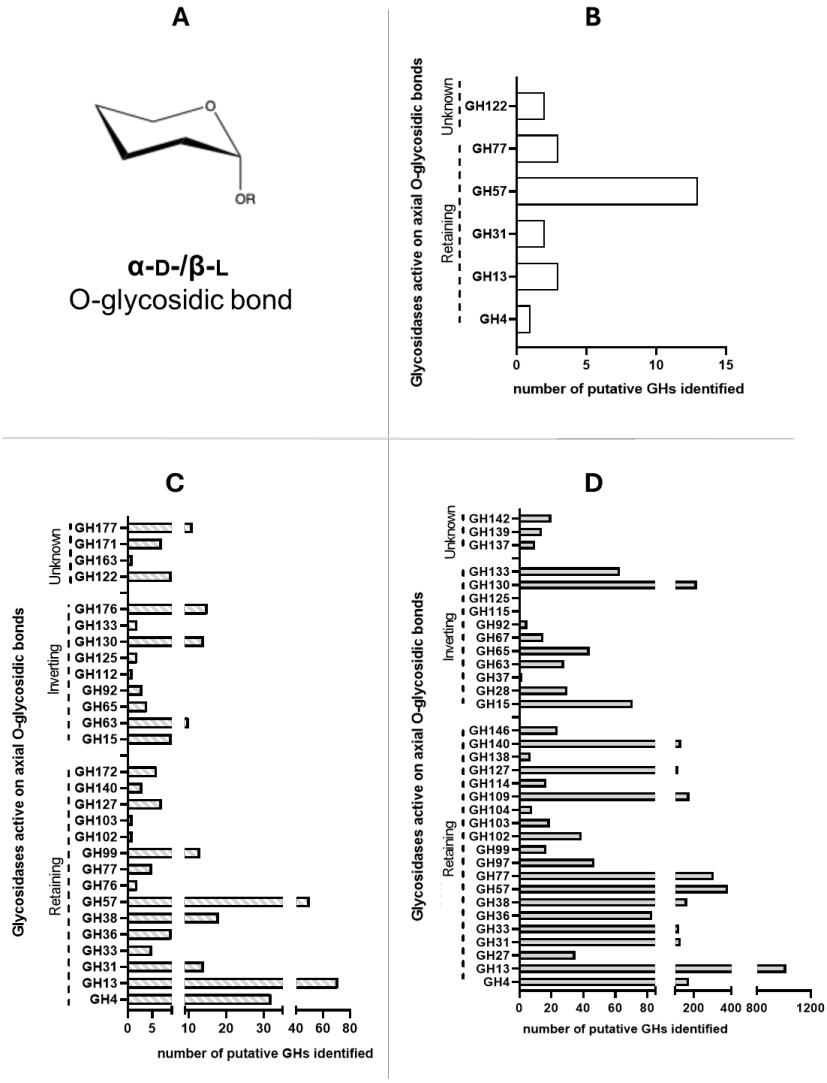

Fig. 2.5: GH identified in the samples analysed by dbCAN with the spatial orientation of the carbon one **bond in axial** (AxC1-OR). In order: (**A**) Structure of axial bond (α-D- and β-L- O-glycosidic bond); AxC1-OR GHs annotated in (**B**) Raoul Island, (**C**) Octopus Spring and (**D**) Kergulen Island samples.

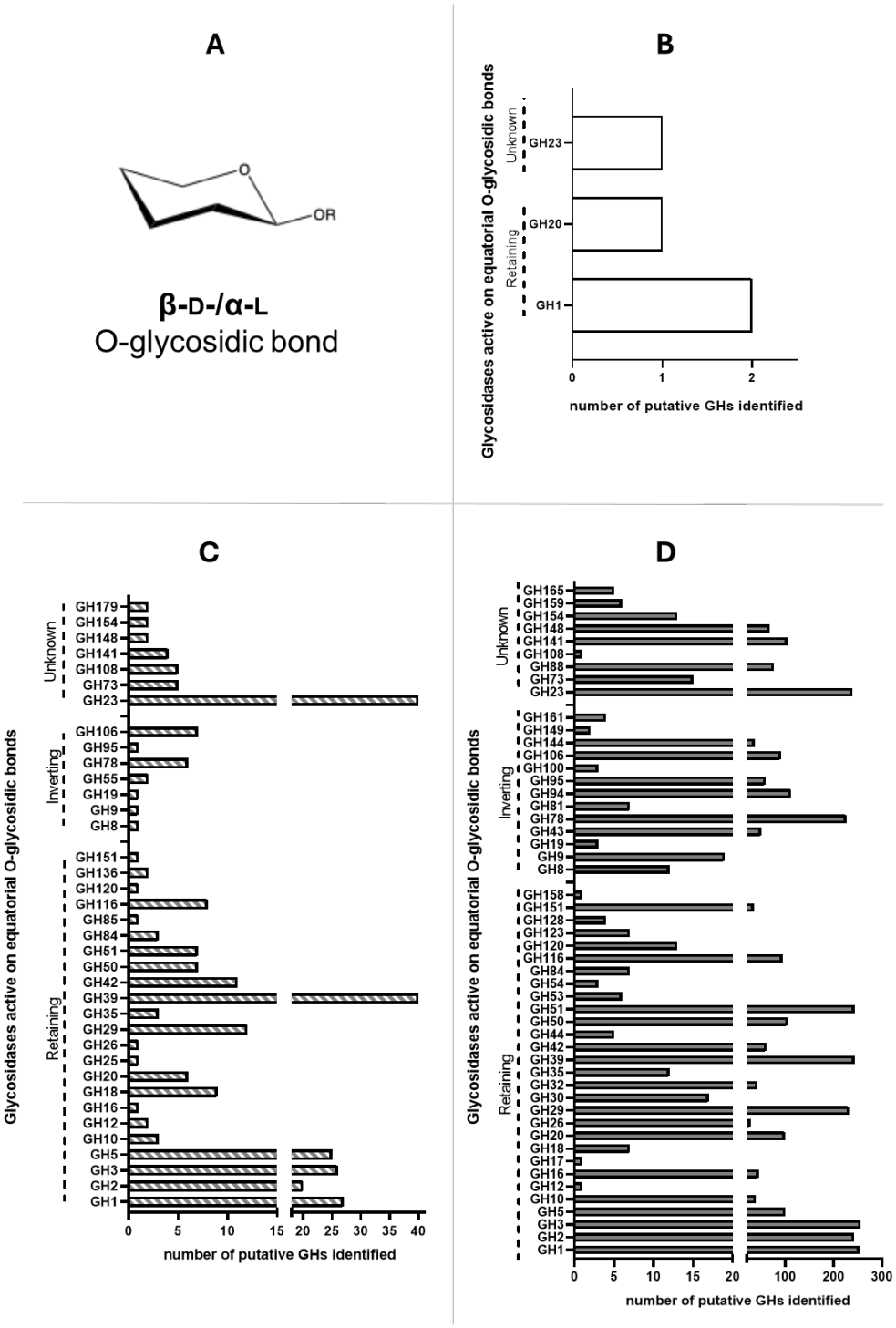

Fig. 2.6: GH identified in the samples analysed by dbCAN with the spatial orientation of the carbon one **bond in equatorial** (EqC1-OR). In order: (**A**) Structure of axial bond (β-D- and α-L- O-glycosidic bond); (**B**) EqC1-OR GHs annotated in Raoul Island, (**C**) Octopus Spring and (**D**) Kergulen Island samples.

Through the annotation of EC numbers present in the identified glycoside hydrolases (GHs), distinct patterns were observed across the different datasets, shedding light on the ecological and environmental contexts of the sampling sites. Approximately 8% of the GHs in each dataset were annotated with an EC number, allowing for an analysis of GHs concerning some of the major polysaccharides. This approach enabled a functional understanding of the GHs distribution, highlighting potential correlations with specific environmental factors.

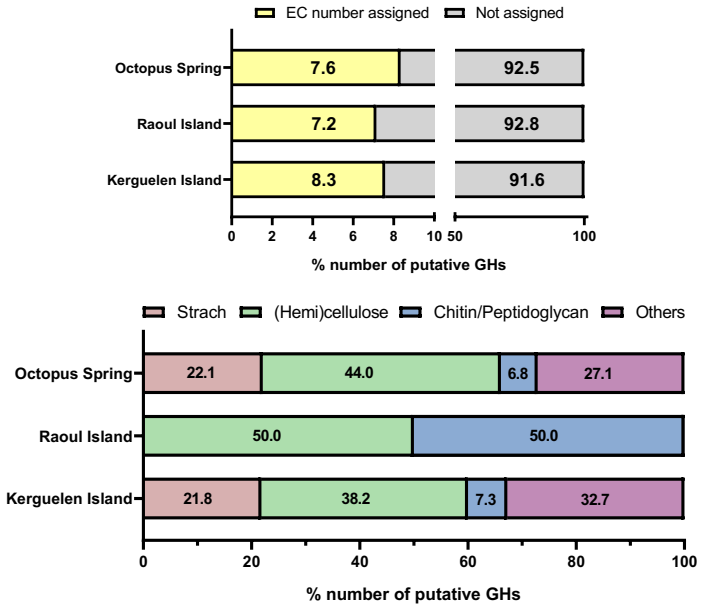

Fig. 2.7: abundance of GH according to the EC numbers vs the main polysaccharides. The number of assigned GHs at the level of EC number (A). The annotated GHs for each dataset are according to the main polysaccharides.

GHs active on starch, a crucial energy source for many microorganisms, are prominently represented in the Octopus and Kerguelen datasets, constituting approximately 22% of the total glycoside hydrolases annotated at the EC number level. This abundance underscores the critical role of starch in microbial metabolism and highlights the evolutionary pressure exerted on microorganisms to develop efficient starch-degrading mechanisms. In these datasets, the predominance of α-amylases from GH13 and GH57 families reflects an adaptive strategy to exploit starch as a primary nutrient source, facilitating microbial growth and survival in starch-rich environments [72].

Moreover, cellulose-degrading enzymes are predominant in all three datasets, representing around 44% of the GHs found in Octopus Spring, 38% from Kerguelen Island and 50% of the annotated GHs from Raoul's sample. This suggests an adaptation to environments where cellulose-rich organic matter is prevalent. These sites, likely abundant in plant material

or other cellulose sources, drive the evolution of cellulolytic enzymes essential for breaking down cellulose. This enzymatic adaptation is crucial for the nutrient cycling and energy dynamics within these ecosystems, where plant debris provides a significant cellulose source for microbial communities [67].
Additionally, around 7% of the annotated sequences of the Kerguelen and Octopus datasets and 50% of Raoul's sample are chitin-degrading enzymes, highlighting another layer of ecological adaptation. Chitin is a significant component of the exoskeletons of arthropods and fungi. The presence of these enzymes reflects an evolutionary response to utilise chitin as a nutrient source, facilitating the recycling of essential nutrients and sustaining the broader ecosystem [35].
Among the other enzymatic activities identified, α-L-fucosidase (EC 3.2.1.51) emerged as the predominant activity across all samples, ~10% of the GHs annotated in Kerguelen and Octopus samples. Enzymes belonging to the GH29 and GH95 families have been demonstrated to be involved in the metabolism of fucoidans [73], complex polysaccharides commonly found in brown algae [74], and their oligosaccharides. Multiple putative enzymes from these two families were identified in this analysis, associated with α-L-fucosidase activity. These findings suggest a potentially significant role for α-L-fucosidases in environments.
Overall, these findings illustrate microbial communities' remarkable metabolic versatility and adaptive strategies in response to the availability of different polysaccharides in their respective environments. The observed patterns of GH distribution provide valuable insights into the ecological roles of these microbes and their evolutionary adaptations to their specific habitats.
The analysis clearly shows the enzymatic diversity in the three studied environments, highlighting significant differences in the composition of GHs and suggesting further studies to understand the possible metabolism involved in the degradation of recalcitrant polysaccharides.

2.2.3. Taxonomic annotation of the novel GHs

A comprehensive taxonomic analysis was performed to elucidate the microbial origins of the putative GHs identified in the samples.
Regarding the first and most massive datasets, in Octopus Spring, 71.2% of the sequences belong to the Bacteria kingdom and only 0.33% to Archaea. However, 28.4% of the sequences could not be assigned, likely due to their significant phylogenetic divergence from the reference sequences in the annotation database used (RefSeq Protein). These environmental samples, which are in the form of microbial mats, are dominated by unicellular cyanobacteria (*Synechococcus* spp.) and filamentous anoxygenic phototrophic bacteria (FAPs) (e.g., *Chloroflexus spp.* and *Roseiflexus spp*.); they also contain newly described anoxygenic

phototrophic bacteria and many organisms involved in microbial mat decomposition [75]. A total of 32 phyla were identified from the dataset. The leading bacterial phyla are Chloroflexi (16.95%) and Acidobacteria (10.08%), often associated with thermophilic ecosystems, followed by Planctomycetes (7%), a widely distributed bacteria occurring in both aquatic and terrestrial habitats. Regarding the Archaea dominant phylum, Euryarchaeota represents the main phylum, with 18 of the Archaea sequences. Of the 4908 sequences annotated as bacteria, Bryobacterales, Verrucomicrobiales, and Chloroflexales correspond to 18.4% of the GHs annotated at the order level. The analysis at the genus level and species allows us to highlight several thermophilic microorganisms: several sequences belonging to *Thermus*, *Thermogutta*, *Pyrinomonas* and *Caldanaerobius* (522 sequences), but the main thermophilic microorganism annotated is genus *Roseiflexus* with 286 GHs, in particular, *Roseiflexus sp, RS-1* isolated for the first time in this hot spring, adapted to slightly higher temperatures (45 to 60°C; optimum) [75] (Fig. 2.8).

In the Raoul Island samples, 64% of GHs were attributed to Archaea and 28% to Bacteria, with the unique GH20 and one GH57 sequence remaining unassigned. This trend of unassigned sequences increased significantly at lower taxonomic levels, with only 9 GHs assigned at the species level. The Crenarchaeota phylum was predominant among the Archaea, with a single sequence belonging to GH1 assigned to Euryarchaeota. Within the Bacterial domain, all sequences annotated at the phylum level belonged to Aquificae, except for one GH77 sequence, assigned to Deinococcus-Thermus, and sequences not annotated at the phylum level were exclusively bacterial. Further examination revealed that 46% of the GH sequences were annotated at the genus level and 32% at the species level. Pyrobaculum is the dominant genus within the Crenarchaeota. Among the bacterial sequences, three were assigned to Hydrogenobacter and only one to Thermus (Fig. 2.9).

In the Kerguelen samples, the analysis revealed that 82% of GHs belonged to Bacteria and 10% to Archaea, with 6% of sequences unassigned. Interestingly, the percentage of taxonomically unassigned GHs rises to 56% at the genus level. The bacterial composition of all samples showed a substantial presence of sequences from the Chloroflexi and Deinococcus-Thermus phyla, with Chloroflexi being the dominant bacterial phylum (182 sequences). The Crenarchaeota phylum was most abundant among the Archaea, with 67 putative GHs. Detailed investigation at the genus and species levels revealed a predominance of sequences related to the Thermus genus, specifically *T. filiformis* and *T. thermophilus*, as well as a higher number of hypothetical GHs associated with the Thermoflexus genus, particularly *Thermoflexus hugenholtzii* (Fig. 2.10). The high number of not annotated sequences in this dataset may be related to the high

number of new putative taxa reported during the first metagenomic overview of the microbial diversity of Kerguelen hot springs: namely 13 new putative genomic species and 6 new putative genera affiliated to Bacteria and Archaea. Many of these MAGs are likely to be derived from populations of thermophilic/hyperthermophilic bacteria and archaea [66].

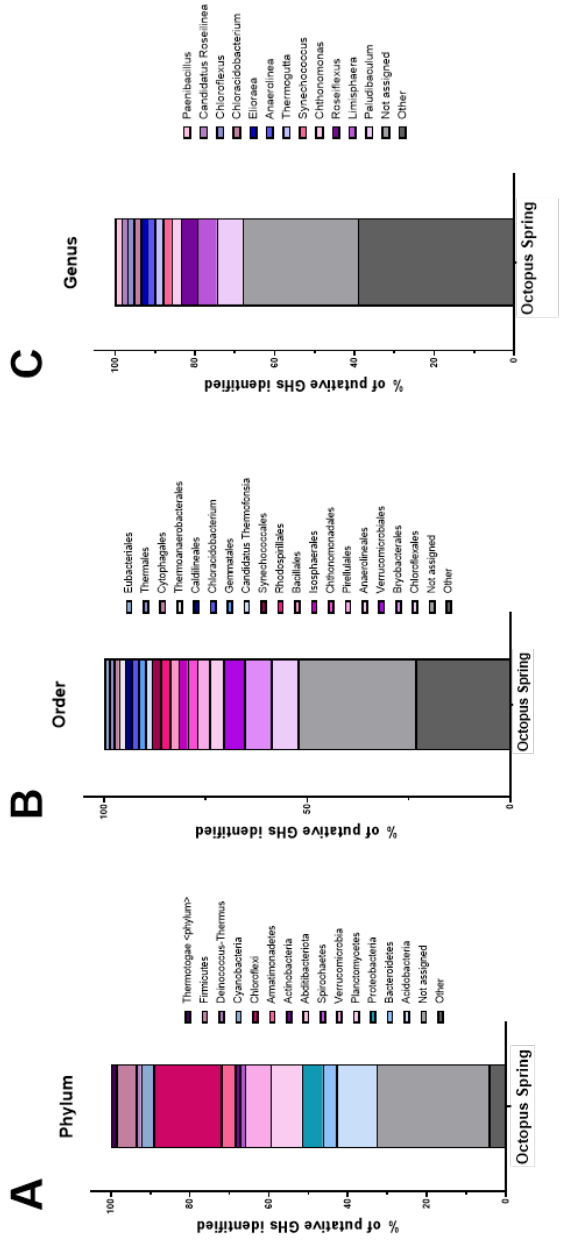

Fig.2.8: **Taxonomic assignment** of the GHs identified in Octopus Spring samples at (**A)** Phylum, Order (**B**) and genus (**C**) level; Others, % of sequences less than or equal to 1%; not assigned, the GHs not taxonomically annotated at the taxa level to date.

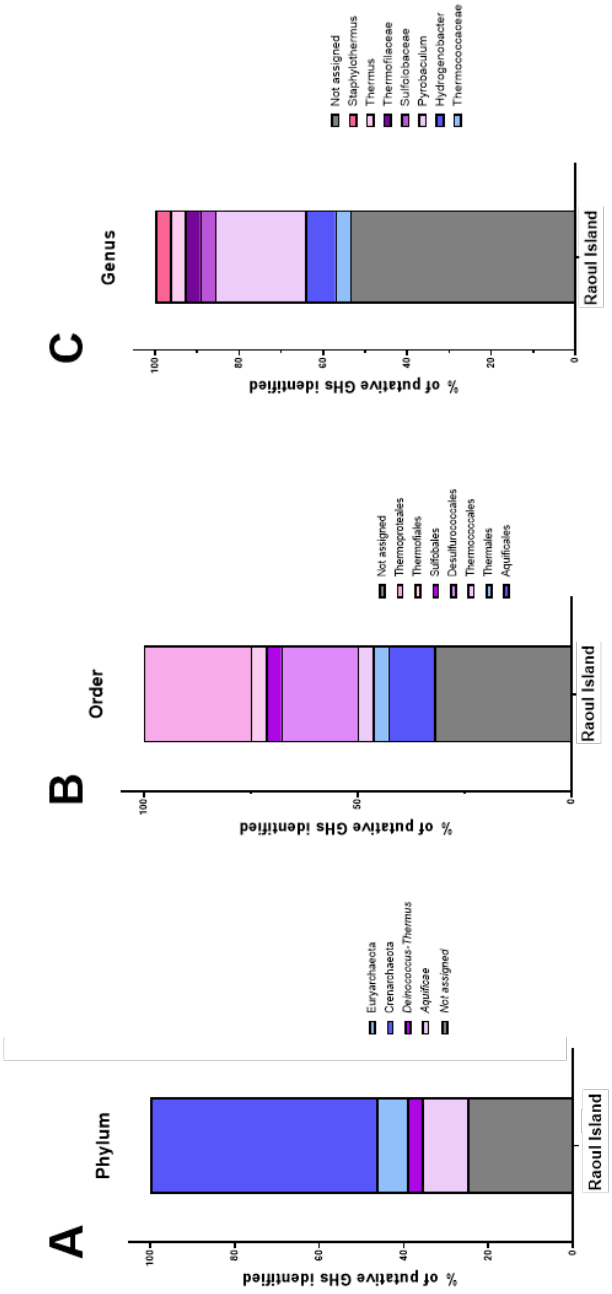

Fig.2.9: **Taxonomic assignment** of the GHs identified in Raoul Island samples at (**A)** Phylum, Order (**B**) and genus (**C**) level; Others, % of sequences less than or equal to 1%; not assigned, the GHs not taxonomically annotated at the taxa level to date.

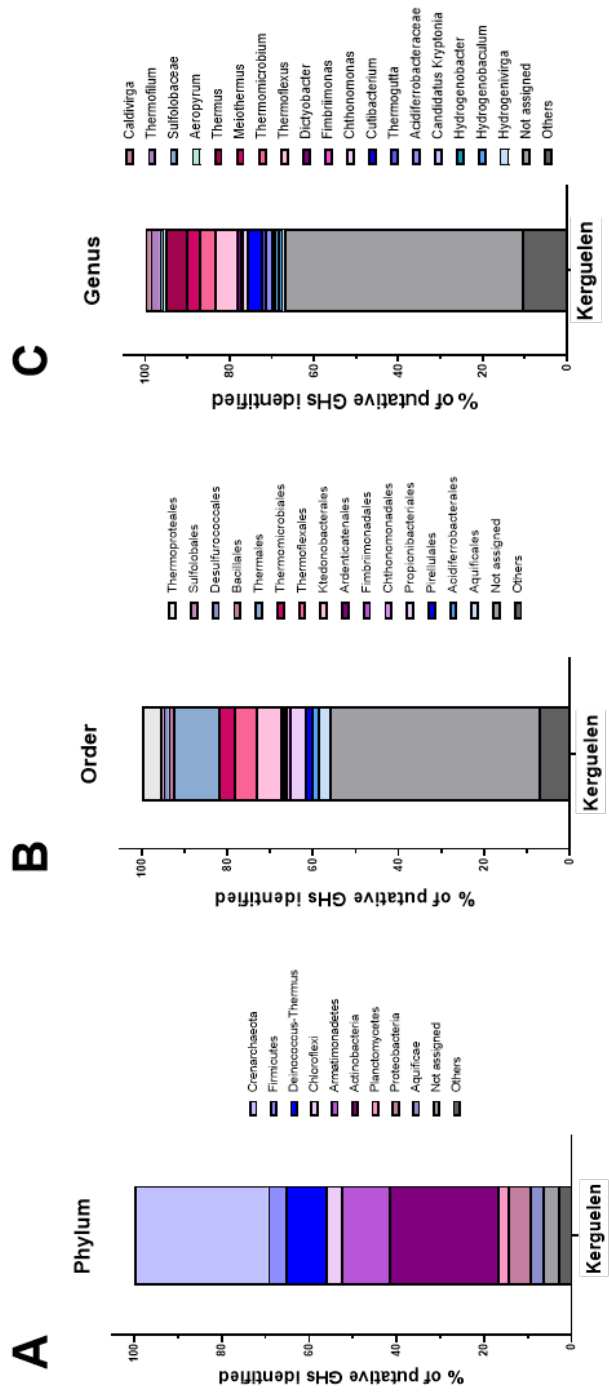

Fig.2.10: **Taxonomic assignment** of the GHs identified in Kerguelen Island samples at (**A)** Phylum, Order (**B**) and genus (**C**) level; Others, % of sequences less than or equal to 1%; not assigned, the GHs not taxonomically annotated at the taxa level to date.

2.2.4. Evaluating the Novelty of MetaGH database

To assess our *Meta*GH's novelty, we evaluate the percentage of the identity of the putative GHs annotated in this study, again the characterised GHs reported to date in the CAZy database. Most GHs annotated in the Octopus Spring samples exhibited a percentage identity ranging from 40% to 60% relative to the CAZy database. In contrast, the Raoul Island sample contained 17 annotated GHs with a percentage identity between 30% and 40%, predominantly comprising GH57, GH20, and GH23. Similarly, the Kerguelen GHs demonstrated a percentage identity primarily between 40% and 60% when aligned with the CAZy database.

Notably, 752 sequences from Octopus Spring (11% of the putative GH sequences annotated) and over 300 GHs from Kerguelen (40% GHs) did not match any characterised entries in the CAZy database. Additionally, only one GH57 annotated in Raoul Island by dbCAN needed a match. These unmatched sequences may represent entirely novel GHs, belong to new families, or be relatively short or fragmented, possibly leading to misannotation. Indeed, only 26% of Kerguelen GHs annotated have a match with characterised GH families (Fig. 2.11)

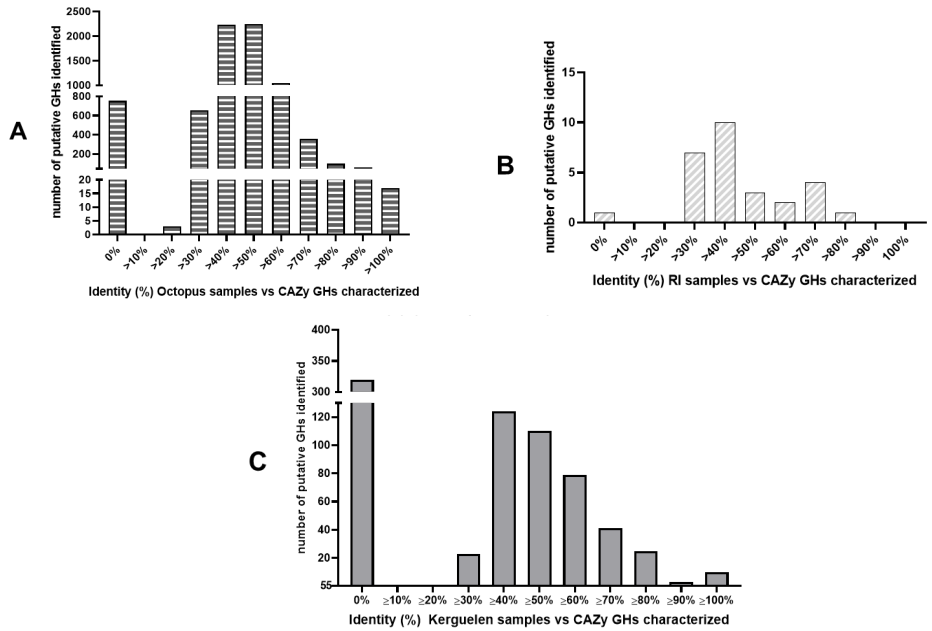

Fig. 2.11: Identity percentages of GHs annotated from (**A)** Octopus Spring, (**B)** Raoul Island, and (**C)** Kerguelen Island against GHs characterised reported in the CAZy database.

2.3. Conclusion

This study highlights the extensive diversity of glycoside hydrolases (GHs) identified through metagenomic analysis of extreme environments, emphasising their potential for novel biotechnological applications. By

examining metagenomic datasets from Octopus Spring, Raoul Island, and Kerguelen Island, we have uncovered significant variations in GH profiles that reflect distinct environmental conditions and adaptive strategies of microbial communities.

The dominance of starch-degrading enzymes in the Raoul Island dataset, particularly α-amylases from the GH13 and GH57 families, underscores the critical role of starch as an energy source in this environment. Similarly, the abundance of cellulose-degrading enzymes in Octopus Spring and Kerguelen Island samples highlights the adaptation of microorganisms to cellulose-rich substrates, which is essential for nutrient cycling in these ecosystems.

Identifying GHs with low similarity to existing CAZy database entries suggests the presence of novel GHs with unique enzymatic activities or belonging to new families. This underscores the need for further characterisation and functional studies to fully explore their potential applications in industrial processes, such as biofuel production, food processing, and biotransformation.

Taxonomic analysis of the identified GHs indicates a substantial presence of (hyper)thermophilic microorganisms, particularly from the phyla Crenarchaeota and Chloroflexi, reflecting the ecological niches occupied by these extremophiles. The predominance of specific genera, such as Pyrobaculum and Thermus, highlights their importance in biogeochemical processes and their potential utility in biotechnological applications. These geothermal ecosystems could be reservoirs of biological and functional novelty.

Overall, this comprehensive metagenomic analysis provides valuable insights into the metabolic capabilities of microorganisms from extreme environments. The discovery of unique GHs and their potential applications underscores the importance of continuing to explore and characterise extremophiles to advance our understanding and utilisation of their remarkable enzymatic properties.

A specialised database, *Meta*GH (*Metagenomic* Glycosidase Hub), containing approximately 8,000 putative GHs active on major polysaccharides, has been generated.

This database is designed to serve as a crucial resource for the targeted selection of GH sequences for further characterisation, enabling the discovery and development of novel enzymes primarily, though not exclusively, with potential activity on recalcitrant biomasses resistant to hydrolysis.

It's worth noting that the ability to update this database in the future, by adding members of newly discovered or novel GH families, will further expand its potential. Indeed, as new GH families are characterised, the database will evolve, broadening its utility for metagenomic enzyme discovery.

2.4. Materials and Methods

2.4.1. Metageenome datasets

All the datasets analysed are available in Read Archive (SRA) data by NCBI (https://www.ncbi.nlm.nih.gov/sra/). Table 1 reports all the metadata of this study.

2.4.2. In silico pipeline used for the discovery of novel GHs

For each sample, the following pipeline was carried out:

1. The quality of the metagenome sequences was tested by FastQC [76].
2. The metagenome was assembled using Megahit assembler, preset with a meta-sensitive mode (k-mers between 21 and 99, with increments of 2) for Kerguelen and Octopus samples [23] and Spades assemblers for Raoul Island sample [77].
3. Identification of ORFs by contigs using Prodigal Gene Prediction [78].
4. dbCAN analysed the identified ORFs [47] to annotate hypothetical CAZymes, considering good only the GHs with three state-of-the-art tools/databases for automated CAZyme annotation: HMMER search for CAZyme family annotation vs dbCAN CAZyme domain HMM database; DIAMOND search for BLAST hits in the CAZy database; and HMMER search for CAZyme subfamily annotation vs dbCAN-sub HMM database of CAZyme subfamilies (derived from eCAMI classification of CAZyDB families).

In collaboration with Professor Nicolas Terrapon's group, the Glycogenomics team (CAZy Team) at AFMB (Architecture et Fonction des Macromolécules Biologiques) in Marseilles, France, Kerguelen dataset was annotated using CAZy tools. This collaboration enabled a comprehensive functional annotation of the identified sequences. The expertise of Professor Terrapon’s team in glycogenomics and advanced bioinformatics tools facilitated an in-depth analysis, ensuring accurate classification of the carbohydrate-active enzymes (CAZymes) present in the datasets.

5. Evaluation of the annotated GHs' novelty, conducting an alignment against all the characterised GHs present in the CAZy database using the Diamond tool [63].
6. Using the AnnoTre Database, an alignment was conducted using the Diamond tool to do functional annotation at the EC number's base [79].
7. Using the NCBI RefSeq Protein Database, identify the microbial source of the putative GHs in the samples. Subsequently, they will be analysed using MEGAN [80].

CHAPTER 3: EXPLORING THE FUNCTIONAL BIODIVERSITY OF GH122 FAMILY: IDENTIFICATION OF A NOVEL ENZYME FROM PISCIARELLI HOT SPRINGS BY METAGENOMIC APPROACH

3.1. Aim of the Chapter

Metagenomics is not just a tool, it's a transformative and powerful force in microbial ecology and enzyme discovery, offering insights not only into the composition of microbial communities but also into their genetic content [81]. This approach has been particularly revolutionary in exploring extreme environments, which are valuable sources for identifying extremozymes [2] and potentially revolutionise our understanding of microbial biodiversity. The potential of this research method is truly inspiring.

This chapter presents a specific example of how metagenomic approaches, combined with subsequent biochemical characterisation, have successfully identified and understood extremozymes' functional potential and biodiversity from extreme environments.

Our primary focus was on the annotation of the CAZome from two mud/water pools in the solfataric field of Pisciarelli, Agnano (Naples, Italy)—an area known for pioneering research in the isolation of hyperthermophilic Archaea. In this unique environment, we made a groundbreaking discovery: a novel sequence belonging to the relatively obscure glycoside hydrolase family GH122 [47]. This discovery adds a new and exciting dimension to our understanding of extremozymes, sparking potential implications for enzyme applications.

The biochemical characterisation of this enzyme is reported here, providing insights that enhance our understanding of GH122 and its potential applications. This case study highlights the significance of integrating metagenomic data with detailed experimental analyses and emphasizes the substantial contribution this work has made in redefining the functional biodiversity of the GH122 family. Furthermore, it underscores the promising future of this enzyme in advancing enzyme discovery and its potential applications in industrial contexts, fostering a sense of optimism about the future of enzyme research and the potential it holds for industrial progress.

3.2. Results and Discussion

3.2.1. Identification of a novel GH122

Recently, the metagenomic analysis of the microbial community populating the Pisciarelli hot springs (Naples, Italy) revealed the entire carbohydrate-

active enzymes portfolio (CAZome) [28]. The sampling site Pool 2 is one of the two main mud/water pools identified at Pisciarelli; this site is characterised by a medium temperature of 92 °C and a pH near 1.5, representing an exciting source of (hyper)thermophilic CAZymes. Metagenomic analysis makes it possible to annotate different CAZymes in Pool2. Notably, a contig assigned to an uncultured Sulfolobaceae, *Contig-4890*, which encodes several Open Reading Frames (ORFs) involved in pentose metabolism, was annotated. Among these, a novel sequence was discovered and classified within the GH122 family, designated as GH122_Pool2 (Fig. 3.1).

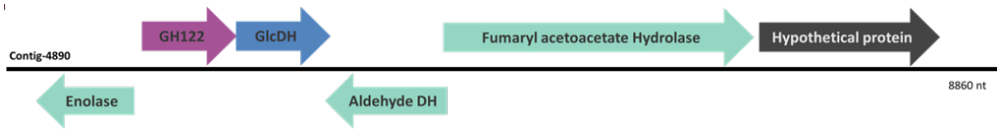

Fig 3.1: Contig 4890 (8860 nt), isolated in Pisciarelli hot spring Pool2, belonging to uncultured Sufulobaces organism. In purple, the ORF-annotated as GH122_Pool2 object of this study.

The GH122 family is a relatively recent addition to the CAZy database, currently comprising only 86 classified Archaeal sequences (68 full lenght sequences), with only one biochemically characterized enzyme, the α-glucosidase (EC 3.2.1.20) from *Pyrococcus furiosus* PF0132 here reported as GH122_P.f [72]. The phylogenetic analysis of these annotated sequences, derived from complete genomes and excluding fragmented sequences, reveals two well-defined significant clusters (Fig. 3.2). One of these clusters predominantly comprises members of the TACK superphylum, while Euryarchaeota dominates the other. Within the Euryarchaeota group, a primary cluster is mainly formed by sequences from *Pyrococcus* species, including GH122_P.f. Another distinct group within the Euryarchaeota, primarily composed of *Thermococcus* species, is more phylogenetically complex and distant from the *Pyrococcus* cluster. The TACK superphylum also demonstrates two distinct subgroups. The first (TACK A) groups mainly sequences from Sulfolobales, including the GH122_Pool2. Sequences from Acidianus and Thermosphaera species dominate the second subgroup (TACK B). Notably, GH122_P.f and GH122_Pool2 are phylogenetically distant, as reflected by their low sequence identity of just 24%. These two enzymes represent distinct evolutionary path within the GH122 family, highlighting the significant structural and functional diversity present among archaeal enzymes in this group.

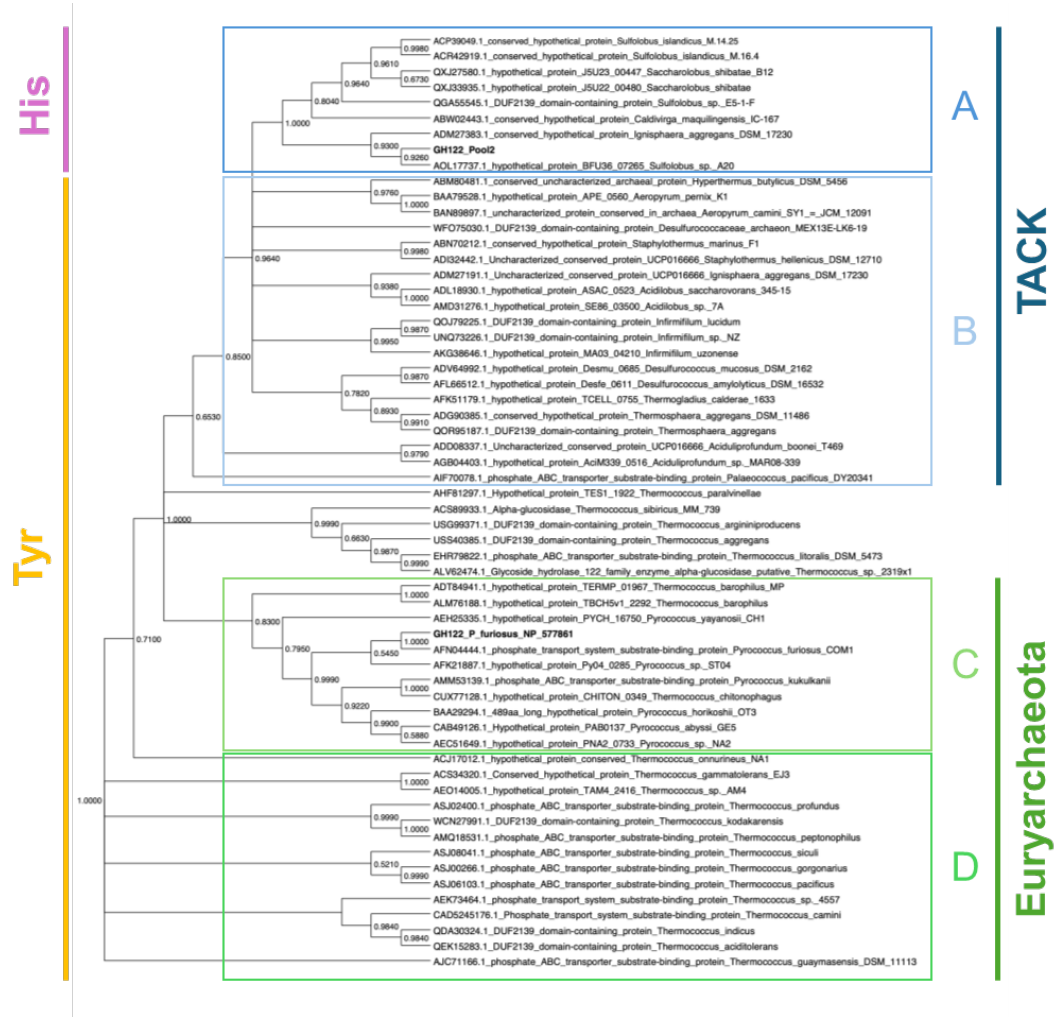

Fig. 3.2: GH122 phylogenetic tree. The blue branches contain only TACK sequences, split into two subgroups: (**A**) Sulfolobacae (**B**) and the other . The green branch is the Euryarchaeota group, divided into sequences of just two genera, (**C**) Thermococcus and Pyrococcus and from the (**D**) genus Thermococcus and the other . **Phylogenetic analysis of GH122 Family according to the substitution of 71-like residues:** The leading group of TACK Desulfurococcaceae and the Thermococcaceae sequences with Tyrosine substitution are in Orange. In purple, the group of Sulfolobaceae was used with the Histidine substitution.

3.2.2. Structural Analysis of GH122_Pool2

The α-glucosidase from *Pyrococcus furiosus* (GH122_P.f) [82] has been purified from the native organism and characterised, displaying activity on pNP-α-glucopyranoside, maltose, isomaltose and pannose [83]. However, its catalytic mechanism and the residues involved in the reaction remain unknown. Using Afold2, a structural prediction model of GH122_1_Pool2 in monomeric form was built and validated with SAVES v6.0, scoring respectively 89.14% and 88.34 % at the VERIFY 3D test [84], [85] while 76.36 and 90.19 as Overall Quality Factor at the ERRAT test [86].

The structural prediction for GH122_Pool2 was compared with the available 3D model of GH122_P.f. Both predicted structures revealed a potential 7-fold β-propeller domain (Fig. 3.3), the same reported in CAZy for that family from AFold prediction. The β-propeller structure is a common all-β protein fold, consisting of 4 to 8 symmetrically arranged β sheets that resemble blades. These β-sheets are organised toroidally around a central axis, forming a funnel-shaped active site. [87].

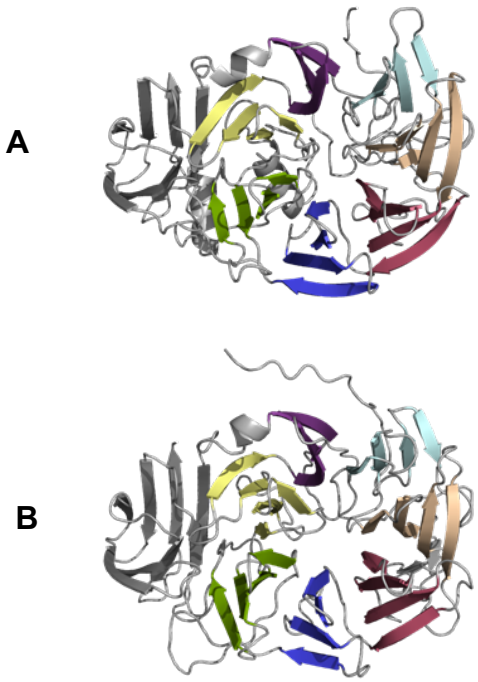

Fig.3.3: 7-fold β-propeller domains highlighted. (**A**) GH122_P.f. possible 7-fold β propeller domain. (**B**) G122_Pool2 possible 7-fold β-propeller domain

The two AlphaFold models for GH122_P.f. and GH122_Pool2 display similar overall architectures, but some differences in secondary structure organisation were highlighted (Fig. 3.4). One primary difference is that the GH122_P.f. model, derived from *Pyrococcus furiosus*, contains two additional α-helices at the surface of the model that are absent in GH122_Pool2. Another significant structural distinction is found at the entrance to the predicted catalytic site (see below): an extended loop region. GH122_Pool2 contains a short β-sheet at the base of this loop, which is absent in the GH122_P.f. structure. This β-sheet within the loop may affect substrate specificity or binding dynamics. Loops at the active site entrance could influence substrate recognition and binding by adjusting the local flexibility and accessibility, and the presence of a β-sheet may help stabilise this loop in a conformation conducive to substrate

interactions. This structural element could thus impact GH122_Pool2's substrate affinity or specificity.

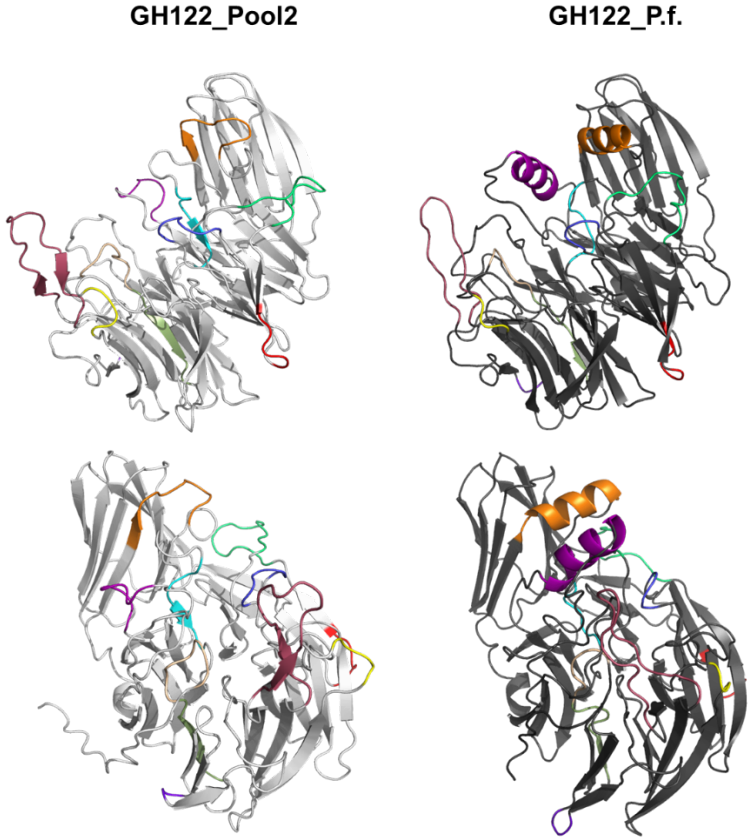

Fig. 3.4: main **structural differences** between **GH122_Pool2** (in white) and **GH122_P.f.** (in grey)

After the alignment of the full-length sequences of all GH122 annotated (including the GH122_P.f.) and GH122_Pool2, it was possible to identify the potential catalytic site, and the candidate residues involved in the hydrolysis reaction (Fig. 3.5). As it is known; generally, the catalytic mechanism of glycoside hydrolases (both *inverting* and *retaining*) involves two negatively charged residues. For this reason, the highly conserved acidic residues within the family were then mapped onto the three-dimensional model of the enzyme. This mapping highlighted three highly conserved acidic residues in the middle of the 7-fold β-propeller domain, namely Asp69, Glu142, and Asp161.
A structural comparison between the two GH122 enzymes was conducted, integrating data from multi-alignment and 3D structure analysis (RMSD of 2.45 Å). This analysis highlighted the putative catalytic sites in both the model, providing insights into their relative distances and the conformation of their catalytic pockets. Although the measured distances are consistent with the catalytic residues typically found in glycosidases in both enzymes,

it is important to note that, accorndig to the models, the putative catalytic pocket of GH122_Pool2 is wider than that of GH122_P.f (see Fig. 3.6 and Fig. 3.7).

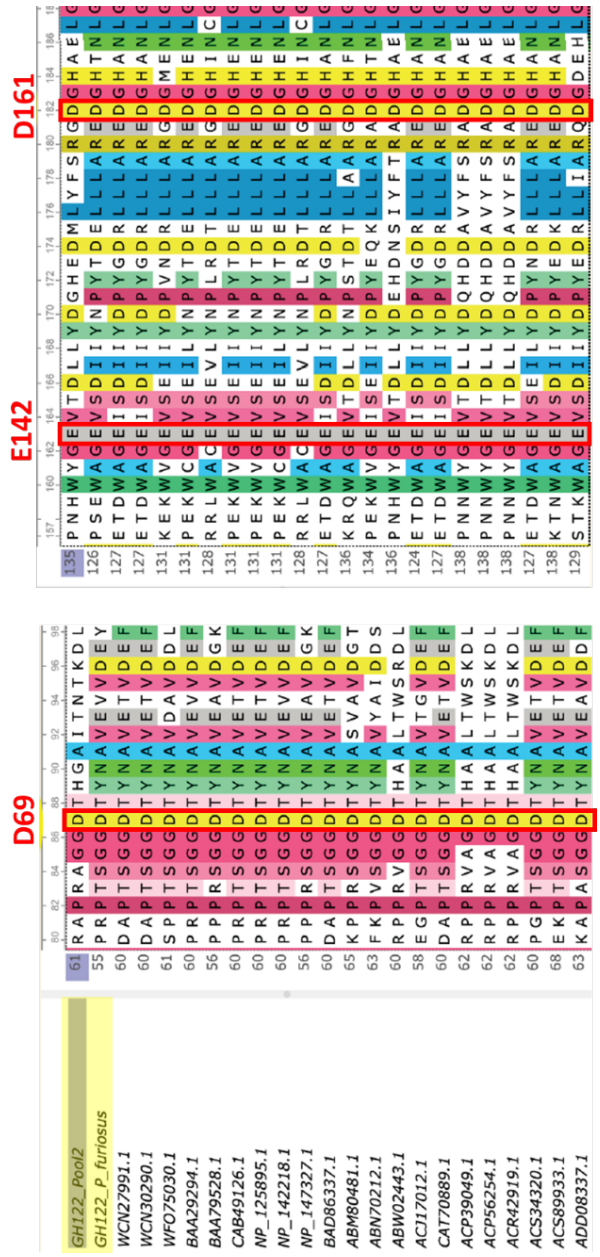

Fig. 3.5: GH122 sequences alignment. The coloured columns represent residues with a conservation percentage higher than 70%. The hypothetical catalytic residues for GH122_P.f and GH122_Pool2 are squared in red, Asp65-Glu135- Asp154 and Asp69-Glu142-Asp161, respectively.

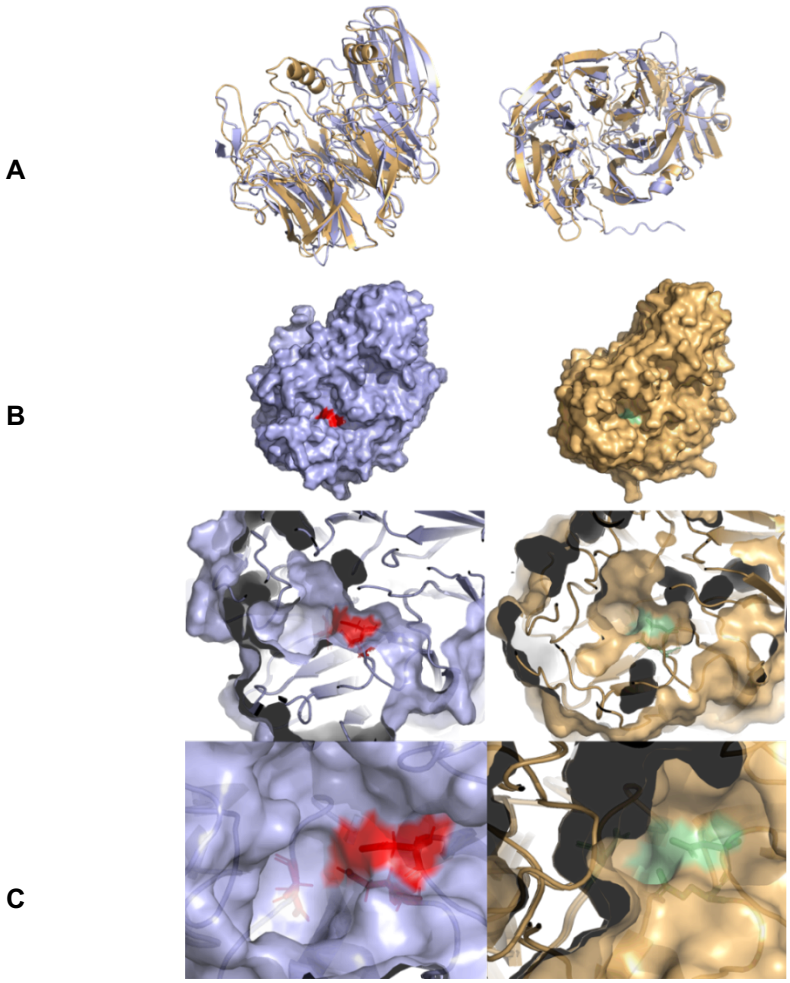

Fig. 3.6: **3D-structure analysis** of the functional model (**A**) and surface model (**B**) of GH122_Pool2 (light blue) and GH122_P.f (light orange) with the respective hypothetical catalytic residues (red and green), with a close-up of the putative catalytic pocket (**C**).

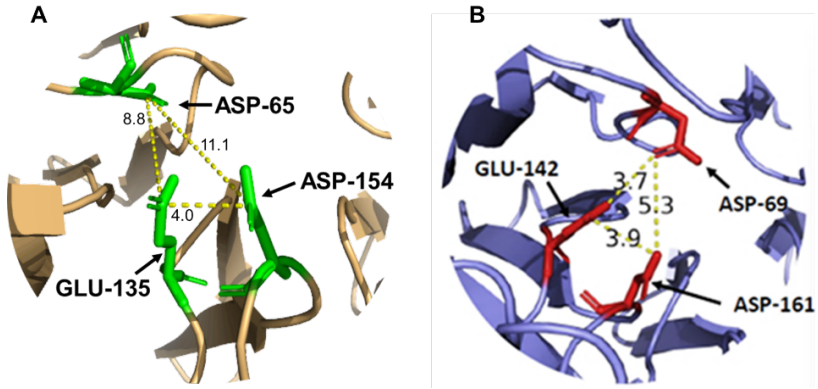

Fig 3.7: Close-up view of the catalytic pocket of GH122_Pool2, in red the **putative catalytic residues** identified by MSA. The table shows the distance between the residues expressed in Ångström (Å).

3.2.3. Biochemical characterisation of GH122_Pool2

GH122_Pool2 was successfully expressed in *an E. coli C43 (DE3)* strain and purified by nickel affinity chromatography (Fig. 3.8). The protein's purification profile was evaluated using SDS-PAGE analysis.

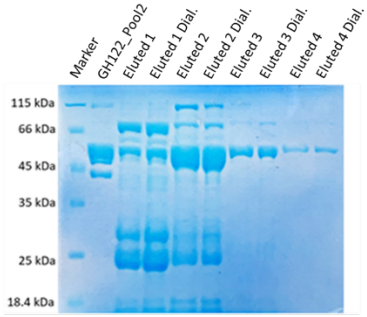

Fig.3.8: SDS-PAGE (8%) GH122_Pool2 (~56 kDa) IMAC Pools and Dialisys Pools, in order: markers; GH122_Pool2, previously purified as reference; Eluted 1, chromatographic run for His-Trap; eluted 1 dialysate in 1X PBS; Eluted 2, chromatographic run for His-Trap; eluted 2 dialysate in 1X PBS; Eluted 3, chromatographic run for His-Trap; Eluted 3 dialysate in 1X PBS; Eluted 4, chromatographic run for His-Trap; eluted 4 dialysate in 1X PBS

The molecular mass of the recombinant GH122_Pool2 in denaturing conditions was 57.39 kDa (including the 6xHis-tag). Size exclusion chromatography in native conditions showed a molecular mass of 256.31 kDa, compatible with a tetrameric quaternary structure (Fig. 3.9).

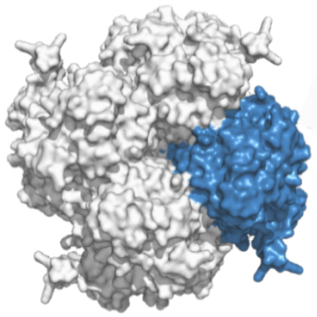

Fig. 3.9: AFold predicted model of the **tetrameric conformation of GH122_Pool2**, and in blues, the monomeric form was highlighted

The purified enzyme was initially tested on pNP-α-glucopyranoside, reported in the literature as a substrate for GH122 from *Pyrococcus furiosus* [83]. However, no enzymatic activity was detected. For this reason, a 16h blanket screening at 65°C was conducted on various substrates α and β, including 4-nitrophenyl-glycosides, malto-oligosaccharides and polysaccharides. The screening showed that GH122_Pool2 is active on 4-nitrophenyl-α-D-glucuronic acid and beechwood Xylan (Fig. 3.10), suggesting that GH122_Pool2 could be active on the α-D-glucuronic acid ramification of xylan. It is worth mentioning that these similar activities, such as Xylan α-1,2-Glucuronidases (EC 3.2.1.131) and α-glucuronidase (EC 3.2.1.139) are

found only in three CAZy families, namely GH4, GH67 and GH115, without archaeal sequences, to date.

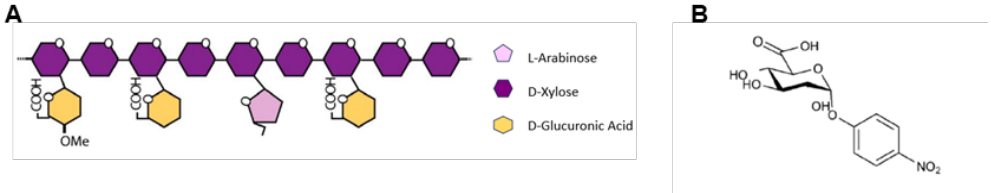

Fig. 3.10: structures of (**A**) beechwood Xylan and (**B**) 4-nitrophenyl-α-d-Glucuronic acid

The preliminary biochemical characterization of GH122_Pool2 on pNP-α-D-Glucuronic Acid (pNP-α-D-GlcA) at pH optimum of 6.0 and a temperature optimum of 85 °C (Fig. 3.11), with a specific activity of 14.8 ± 0.11 U/mg.

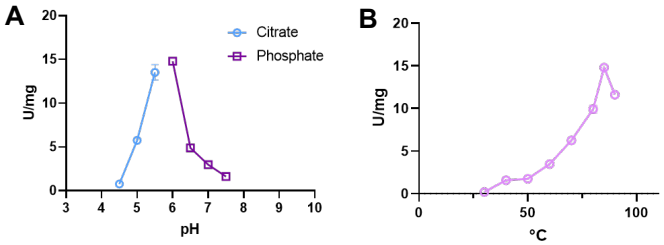

Fig. 3.11: biochemical characterisation of GH122_Pool2 on pNP-α-D-GlcA. (**A**) Specific activity vs pH, the assays were performed in sodium citrate and sodium phosphate buffers in the ranges 4.5–5.5 and 6.0–7.5, respectively; (**B**) Specific activity vs temperature.

The structural stability of GH122_Pool2 was assessed using Differential Scanning Fluorimetry, which revealed a melting temperature (Tm) of 81.5 °C. Moreover, the enzyme exhibited considerable resilience to denaturation, withstanding urea concentrations up to 2.0 M while showing a ΔTm of up to -6 °C. The data obtained from incubation under high concentrations of chemical denaturant and a rigorous denaturation protocol (10 min/°C) demonstrate remarkable stability. This stability aligns with the enzyme's origin from an extreme environment, highlighting its robustness and potential suitability for industrial applications (see Fig. 3.12)

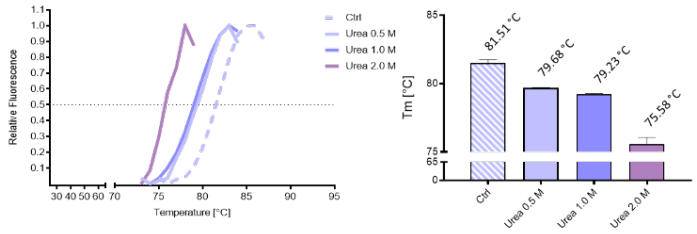

Fig. 3.12: DSF analysis of GH122 to detect the Tm of the enzymes and effect of chemical denaturant agent in different concentrations (from 0.5 to 2.0 M of Urea).

For the assessment of functional thermostability, GH122_Pool2 was incubated under optimal conditions (0.1 M sodium phosphate buffer, pH

6.0, at 85 °C) for up to 24 hours. Enzymatic activity on the substrate pNP-α-D-GlcA was measured at specific intervals throughout the incubation period. After 6 hours, the enzyme retained 95% of its activity, and following 24 hours of incubation at 85 °C, it exhibited 50% residual activity (Fig. 3.13).

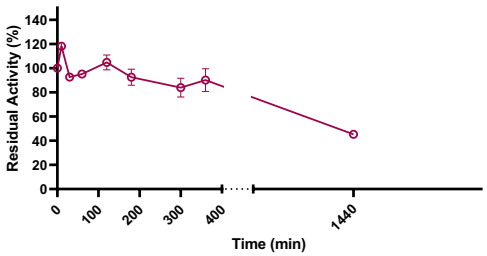

Fig.3.13: Functional thermostability of GH122_Pool2 at 85 °C on pNP-α-D-GlcA under optimal conditions. 100% of activity is equal to 11.0 ± 0.1 U/mg.

Due to the inability to achieve enzyme saturation under the experimental setup, substrate depletion methods were employed to quantify kinetic constants [88]. GH122_Pool2 exhibited a specificity constant (k_{cat}/K_M) of 1.0 ± 0.05 $s^{-1}mM^{-1}$, indicating that pNP-α-D-GlcA is likely not its optimal substrate. In comparison, the thermophilic α-glucuronidase AGU4B from *Thermotoga maritima* MSB8 [89], commonly used in commercial applications for its high catalytic efficiency, has a k_{cat}/K_M of 2.8 $s^{-1}mM^{-1}$ at 80°C [89]. This enzyme's robustness and efficiency make it a valuable model for studying novel thermophilic α-glucuronidases. While GH122_Pool2 demonstrates lower efficiency than AGU4B, it exhibits a higher specificity constant on the same substrate than other thermophilic α-glucuronidases, such as AGU4A (kcat/KM = 0.0018 $s^{-1}mM^{-1}$ at 40°C) [90] and GH67_GC from *Caldicellulosiruptor saccharolyticus* (kcat/KM = 0.05 $s^{-1}mM^{-1}$) [91].

GH122_Pool2 demonstrates not only good catalytic efficiency relative to typical α-glucuronidases but also a potential peculiar ability to hydrolyze the 4-O-methyl-α-D-glucuronic acid branch of xylan, a property not universally present among α-glucuronidases. For example, the GH4 enzyme AGU4A lacks this ability, which limits its effectiveness in possible applications for xylan degradation. This unique trait could provide GH122_Pool2 with a distinct advantage in industrial applications requiring an efficient breakdown of xylan, making it an interesting candidate for biotechnological purposes.

3.2.4 Enzymatic Characterization of GH122_Pool2 on Xylan Substrates

Xylans are complex polysaccharides composed of a xylose backbone linked by β-(1→4) glycosidic bonds, with branches of α-(1→2)-linked glucuronosyl, 4-O-methyl glucuronosyl, and arabinose residues. Due to this specific structure, they are often referred to as

Glucuronoarabinoxylans. This class of hemicellulose represents the most abundant non-cellulosic polysaccharide in the cell walls of dicots and certain monocot species, especially in grasses [92].
The activity of GH122_Pool2 on beechwood xylan was assessed under varying pH levels and temperatures. Results showed optimal activity at pH 5.0, with increasing activity observed up to 85°C (Fig. 3.14).

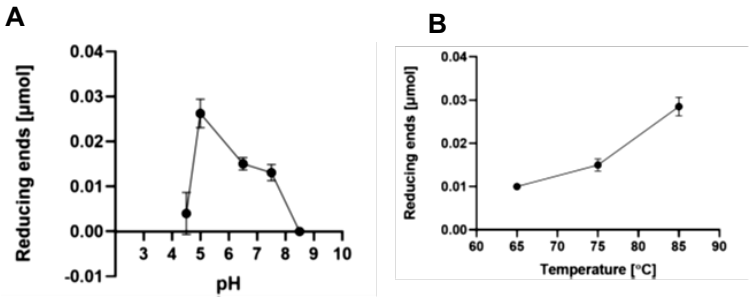

Fig. 3.14: Kinetic studies in the function of pH and temperature. (**A**) activity variations in the function of the pH, (**B**) activity variations in the function of the temperature

The activity of GH122_Pool2 was further characterised by assaying it on Xylan-derived oligosaccharides obtained by incubating beechwood Xylan with the xylanase GH10-XA from *Alicyclobacillus acidocaldarius* [91].
After 1 hour of incubation at 75°C, GH122_Pool2 produced 30.15 ± 2.6 µmol of reducing ends (Fig. 3.15), indicating a higher level of enzymatic activity on pre-treated xylan compared to non-pre-treated xylan. This finding suggests that GH122_Pool2 is more effective on oligosaccharides generated from xylan degradation, potentially due to improved substrate accessibility or specificity. Further studies are needed to investigate this feature in detail, including kinetic analyses on the oligosaccharides to elucidate the enzyme's substrate preference and its potential applications in the hydrolysis of xylan-based materials.

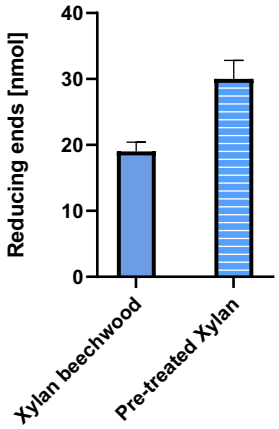

Fig. 3.15: **GH122_Pool2 assays on Xylan and pre-treated xylan**. In light blue the activity on xylan, and from **the assays on pre-treated Xylan**, in blue striped.

3.2.5 Identification of possible molecular determinants underlying substrate specificity

As mentioned above, the GH122 family currently includes only one enzyme characterised with α-glucosidase activity. However, GH122_Pool2 shows no activity on α-Glc– but on α-GlcA–substrates, indicating a new substrate specificity within the family. A detailed examination of the catalytic site was conducted to investigate the molecular basis of this specificity.

A multiple sequence alignment of the available GH122 family sequences identified a highly conserved region. This alignment highlighted a critical residue at position 71 in GH122_Pool2 and located in the hypothetical catalytic pocket (Fig. 3.16), which appears to differentiate the family into two distinct groups (Fig. 3.17).

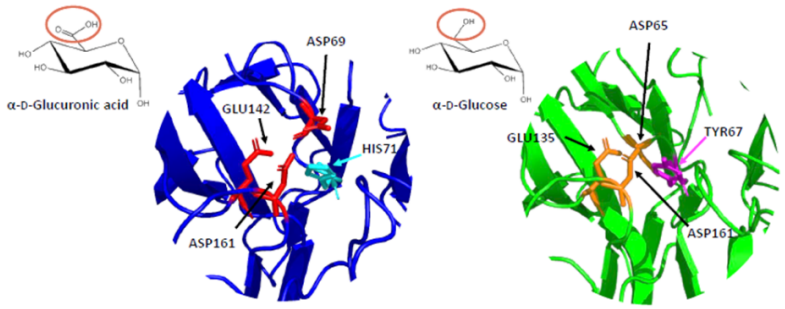

Fig. 3.16: Inspection of the catalytic site of GH122_Pool2 (in blue) **and GH122_P. f** (in green), in particular, highlights the substitution in the catalytic site: in light blue, the His71 in Pisciarelli's enzyme and the Tyr67 in *Pyrococcus's* enzyme in pink. Moreover, the two sugars involved in the respected catalysis and the substitution at the level of C6 of pyranose are highlighted in red.

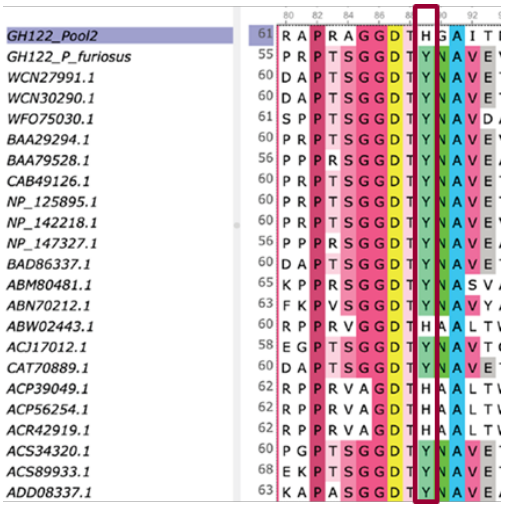

Fig. 3.17: Multiple sequences alignment of GH122 family, full-length sequences, and highlight of residue **H71 in GH122_Pool2** and the homologues in all the family.

In GH122_Pool2, H71 likely forms favourable interactions with the carboxyl group at the C6 position of α-glucuronic acid. In contrast, the Y67 in GH122_P.f could stabilises the hydroxyl group of α-glucose (Fig. 3.17).
Interestingly, the variant residues H/Y are also reflected in the phylogenetic distribution of the GH122 members (Fig. 3.2). The primary group, which includes sequences from the Euryarchaeota and the TACK superphylum within the Desulfurococcaceae family, possess a tyrosine at this critical position. In contrast, the other group, which includes the TACK members of the Sulfolobaceae family—such as the GH122_Pool2—has a histidine at position 71.
These findings suggest that the residue at position 71 may play a critical role in defining the substrate specificity of GH122 members. We hypothesise that this residue functions as a molecular switch, influencing whether the enzyme preferentially interacts with α-Glc or α-GlcA substrates. However, further research is necessary to validate this hypothesis and elucidate the mechanisms underlying this potential specificity.

3.3. Conclusion

This study emphasises the pivotal role of metagenomics in discovering novel enzymes from extreme environments, such as the Pisciarelli hot springs. Identifying a novel GH122 enzyme from geothermal sites with high temperatures demonstrates how a metagenomic approach, combined with biochemical characterisation, could be a successful strategy for discovering new extremozymes with distinct biochemical properties. This study not only expands the GH122 family with the characterisation of a novel member but also reveals a previously unreported α-glucuronidase activity within the GH122 family, representing the first case of an α-glucuronidase activity identified in an archaeal organism.
To fully define the functional biodiversity of the GH122 family, future studies must investigate xylan-derived oligosaccharides to clarify the enzyme's substrate specificity. Essential studies will also include determining the enzyme's catalytic mechanism, identifying the residues involved in catalysis, and resolving its 3D structure. This information will be necessary to bridge the knowledge gap in this family, which will have considerable implications for its potential applications.

3.4. Materials and Methods

3.4.1. Identification, Sequence Analysis, and Structural Prediction studies

The sequence of the enzyme GH122_Pool2 was previously identified and annotated *in silico* and is part of an already reported dataset [47].

The 3D structure was predicted using the ChimeraX v. 1.4 AlphaFold tool [73]. The structure was validated with SAVES v6.0 (https://saves.mbi.ucla.edu/). The structural prediction model of GH122 from *Pyrococcus furiosus* is available online on AlphaFoldDB with the code Q8U4F6 (https://αfold.ebi.ac.uk/entry/Q8U4F6). PyMOL v. 2.4.1 was used to detect the structural features and visualise the superposition of 3D models.

3.4.2. Phylogenetic Analysis

The phylogenetic analysis was performed using MEGA11 [93] with the Maximum Likelihood method and Le Gascuel 2008 model [94]. The percentage of trees in which the associated taxa clustered is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of estimated pairwise distances using the JTT model, and then the topology with superior log likelihood value was selected. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.9246)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 68 amino acid sequences. There were a total of 541 positions in the final dataset.

3.4.3. Heterologous expression and purification

In collaboration with NZYTech (Lisbon, Portugal), the gene encoding GH122_Pool2 was cloned into a pHTP1 vector, and the gene encoding was not subject to optimisation for *E. coli* expression. The GH122_Pool2 gene was expressed in competent C43 (DE3) *E. coli* cells. The cell growth was incubated at 37 °C in LB medium supplemented with 50 ng/μL kanamycin until 0.6 OD_{600} nm was reached, then it was induced with 1mM IPTG (isopropyl-β-D-thiogalactopyranoside) and incubated overnight at 18 °C. Subsequently, cells were withdrawn by centrifugation (5,000 g, 20 min, 4 °C) and resuspended in lysis buffer (150 mM NaCl, 20 mM NaH_2PO_4 pH 7.3, Triton X-100 1%) in a ratio w:v = 1:4. Then, 25 mg/L lysozyme and 5 mg/L DNase I were added, followed by a freeze/thaw lysis. The FCEs were clarified by centrifugation (17,000g, 20 min, 4 °C) and filtered through a 0.45 μm PVDF membrane. FCE was loaded on a His-Trap™ HP (1 mL) exploiting an ÄKTApure chromatography system (Cytiva Life Sciences), equilibrated with 50 mM sodium phosphate buffer, pH 7.5, 150 mM NaCl (Buffer A). After an initial wash-step (20-column volumes) with buffer A, the protein was eluted with a three-step gradient of imidazole in Buffer A (20 mM imidazole, 20-column volumes; 250 mM imidazole, 20-column volumes followed by 500 mM imidazole, 20-column volumes) at a flow rate of 1 $mLmin^{-1}$. The fractions containing the enzyme of interest were pooled, followed by dialysis (cutoff 12 kDa) overnight against 150 mM

NaCl, 20 mM NaH_2PO_4 pH 7.3. SDS-PAGE to evaluate the degree of GH122_Pool2 purification. The concentrations of the purified proteins were determined by Bradford assay, using Bio-Rad Protein Assay Dye Reagent Concentrate by Bio-Rad (California, USA).

3.4.4. Substrates screening

All commercially available substrates were purchased from Sigma–Aldrich (USA), Biosynth (United Kingdom) and Megazyme (Ireland).

The substrate screening was carried out in a final mix volume of 0.2 mL containing the substrate, 20 µL of Na-Phosphate buffer pH 6.5 [1 M], the enzyme and ddH_2O. All the mixes have been incubated overnight at 65°C. The assay parameters were set as previously reported about GH122 from *Pyrococcus furiosus* [83]. Here are the screened substrates and their final concentrations: 4-nitrophenyl-α-D-Glucopyranside [10 mM], 4-nitrophenyl-α-d-Mannopyranoside [10 mM], 4-nitrophenyl-α-D-Galactopyranosie [10 mM], 4-nitrophenyl-α-L-Arabinofuranosyde [10 mM], 4-nitrophenyl-α-D-Xylopyranosyde [10 mM], 4-nitrophenyl-α-*N*-Acetylglucosaminide [10 mM], 4-nitrophenyl-α-L-Fucopyranoside [1 mM], 4-nitrophenyl-α-L-Rhamnopyranside [1 mM], 4-nitrophenyl-α-D-Glucuronic acid [1 mM], 4-nitrophenyl-β-d-Xylopyranoside [10 mM], 4-nitrophenyl-β-D-Mannopyranoside [10 mM], 4-nitrophenyl-β-d-Glucopyranoside [10 mM], 4-nitrophenyl-β-D-*N*-Acetylglucosamiide [10 mM], 4-nitrophenyl-β-L-Fucopyranoside [10 mM], Maltose [10 mM], Isomaltose [10 mM], Tetramaltose [1 mg/mL], Pentamaltose [1 mg/mL], Examaltose [1 mg/mL], Eptamaltose [1 mg/mL], Octamaltose [1 mg/mL], Glycogen from bovine liver type 9 [2.5 mg/mL], Starch Carlo Erba [2.5 mg/mL], Laminarin from Laminaria digitata [2.5 mg/mL], Xylan from beechwood [2.5 mg/mL], Arabinogalactan [2.5 mg/mL], Galactomannan [2.5 mg/mL], Alginate [2.5 mg/mL], λ-Carragenan [2.5 mg/mL], Xyloglucan from Tamarind [2.5 mg/mL] and Carboxymethyl cellulose [2.5 mg/mL].

3.4.5. Standard assays

The standard enzyme assay was conducted as an end-point measurement in a 200 µL reaction mixture at 85 °C in the presence of 100 mM phosphate buffer pH 6.0 and 5 mM pNP-α-D-GlcA, using 1.83 µg of enzyme starting from an enzyme stock of 0.37 µg/µL. After 2 min of incubation, the reaction was blocked in ice and by adding 800 µL of 1 M sodium carbonate buffer (pH 10.2). The absorbance was then measured at 420 nm using a spectrophotometer, with the molar extinction coefficient (ε_{mM}) for p-nitrophenol in sodium carbonate at pH 10.2 is 17.3 $cm^{-1}mM^{-1}$. In all the assays, one unit of enzyme activity was defined as the amount of enzyme catalysing the hydrolysis of 1 µmol of substrate in 1 min at the conditions described. Spontaneous hydrolysis of the substrates was subtracted by using appropriate blank mixtures without enzyme. All kinetic data were calculated as the average of at least two experiments and plotted and

refined with GraphPad Prism 8 (GraphPad Software, San Diego, California, USA).

3.4.6. pH and temperature influence

The pH optimum was determined using a 100 mM buffer, ranging from pH 4.5 to 5.5 citrate buffer and from pH 6.0 to 7.5 in phosphate buffer at 85 °C under standard assay conditions. The GH122_Pool2 was incubated at different temperatures (from 30 °C to 90 °C) at pH 6.0 in phosphate buffer at 85 °C . Functional thermostability was assessed by incubating the enzymes in 100 mM NaH_2PO_4 at pH 6.0 and 85 °C for up to 24 hours. The residual activity was measured in triplicate after 10 min, 30 min, 1 h, 2 h, 3 h, 4 h, 6 h and 24 hours, following the standard assay reported above.

3.4.7. Enzymatic characterisation

Kinetic studies were conducted on pNP-α-D-Glucuronic acid at concentrations ranging from 0.1 to 20 mM under previously established conditions. Substrate depletion methods were employed to quantify kinetic constants.

3.4.8. Activity on Xylan from beechwood and synergic action with GH10-XA

The standard assay mix (200 µL) contained 25 µg of GH122_Pool2, 100 mM Na-Phosphate buffer pH 6.5, and 50 µL of Xylan beechwood [10 mg/mL] at 85 °C. The Somogyi–Nelson assay measured the relative activity, which estimated the amount of reducing sugars released after 1 hour [95]. One unit of activity was defined as the amount of enzyme releasing 1 µmol of reducing equivalents per minute at the conditions described. The tested temperatures have been 65 °C, 75 °C and 85 °C. The pH optimum was determined using a 100 mM buffer: Na-Citrate buffer pH 4.0 and 5.0; Na-Phosphate buffer pH 6.5 and 7.5; and Na-Borate buffer pH 8.5.

For the assay with GH122_Pool2, 100 µL of the pre-treated Xylan oligosaccharides were used in a reaction mixture containing 25 µg of GH122_Pool2 and 10 mM Na-phosphate buffer at pH 6.5. For the assay in the presence of pre-treated xylan, Xylan beechwood has been incubated overnight at 65 °C with the xylanase. The reaction mix (500 µL) has been the following in 200 µL of mix: 5 µg of GH10_XA, 100 mM Na-Phosphate buffer pH 6.5, and 250 µL of Xylan beechwood [10 mg/mL]. Of that reaction mix, 100 µL was used as a substrate for the reaction with GH122_Pool2 in standard conditions. A negative control was also prepared, consisting of the same reaction mixture but with the omission of GH122_Pool2, which was replaced by an equal volume of 1X PBS buffer at pH 7.3. This control was essential to ensure that any increase in reducing ends (R.E.) observed was explicitly due to the activity of

GH122_Pool2 and not residual activity from the GH10-XA used during the pre-treatment.

3.4.9. Molecular mass determination

The molecular mass of the three enzymes was determined by gel filtration on a GE Healthcare Superdex^{TM} 200 10/300GL FPLC column (GE-Healthcare) performed in the presence of 50 mM NaH2PO4 -150 mM NaCl pH 7.5, using AKTA Pure machine at 1 ml min^{-1}. Molecular weight markers were apoferritin (443 kDa), β-amylase (200 kDa), Bovine Serum Albumin (66 kDa) and Carbonic Anhydrase (29 kDa).

3.4.10. Differential Scanning Fluorimetry

The melting temperature (ΔTm) was determined using Differential Scanning Fluorimetry (DSF). A 1.4 µg concentration of purified enzyme was mixed with 0.2 M $Na_2B_4O_7$ pH 8.5 and with 5X SyproTM Orange Protein Gel Stain by Invitrogen (Thermo Fisher Scientific, Massachusetts, USA) to a final volume of 50 µL. The mix was prepared in triplicate and loaded into a transparent sealed PCR 96-well plate in the StepOnePlus Real-Time PCR System by Applied Biosystems (Thermo Fisher Scientific, Massachusetts, USA). The designed program included 10 min at 25 °C, then a ramp of 10 min per °C starting from 25 °C up to 95 °C and lastly 10 min at 95 °C. The data has been recovered and evaluated using GraphPad Prism 8.

CHAPTER 4: DISCOVERY OF NOVEL CAZYMES FAMILIES INVOLVED IN POLYSACCHARIDE UTILIZATION LOCI

4.1. Aim of the Chapter and State of the Art

This chapter reports on the last results of the Operon-based Discovery of Carbohydrate-active Enzymes project (ODE). This project aims to boost the discovery of enzyme families involved in glycan breakdown by analysing Polysaccharides Utilization Loci (PULs). As discussed in Chapter 1, to date, the study of PULs is playing a crucial role in the discovery of new CAZymes. This aim will be achieved by developing a novel and ambitious high-scale computational framework to identify putative CAZyme families and predict their substrates based on the genomic analysis, then select the few candidates for synthesis, production and experimental characterisation. The project involves three principal partners: the Glycogenomics team at AFMB Institute of Marseille, led by Prof. Nicolas Terrapon; the Boost team at BBF Institute of Marseille, led by Dr. Jean-Guy Berrin, and the CERMAV team at CNRS of Grenoble, led by Dr. William Hellbert.

In Bacteroidetes, more than 100 PULs are present, allowing them to become the primary glycan degraders and populate all ecosystems [96]. In 2014, professor Nicolas Terrapon (team leader of the CAZy group) developed an algorithm for PUL prediction based only on CAZy annotation and genomic data [58].

From the analysis of Bacteroidetes PULs, he created a PULDB database containing PULs belonging to thousands of genomes from various environments, including uncultivated species. To date (October 2024), there are 68,500 PULs annotated in the database, and this number is continuously increasing. The synergic job of CAZy/PULDB allowed the discovery of several new enzyme formerly annotated as “Proteins of Unknown Function” (PUFs) [97].

In the ODE project, starting from a list of all PUFs in PULs, the Terrapon’s team performed similarity searches to find the closest homologs and create seeds for new families. Multiple PUFs expected to belong to the same family were detected by overlapping their closest homolog sets and fused into single seeds. Profile Hidden Markov Models (HMMs) were built from each seed, and each PUF family found was analyzed using either PUL/genomic context or intrinsic family features [3][4]. The final aim of this workflow was:

1. Discard protein families that are not statistically over-represented in PULs;
2. Identify position-constrained non-CAZyme families and tag them in PULDB;

3. Rank families with the highest potential to be CAZymes based on several family-wide indicators.

According to the PUL paradigm proposed in 2006 [56], PULs do not harbour random CAZyme associations but instead gather families with similar/complementary activities necessary to orchestrate the breakdown of a given glycan [98]. From known CAZymes, a preliminary analysis was conducted to compute the strength of their associations in PULs and manually identify groups of multi-connected families sharing substrate preferences. This ranking and substrate predictions were used to select 192 sequences for the ODE project.

Under the supervision of Professor Nicolas Terrapon, team leader of the Glycogenomics research group (CAZy group) at the *Laboratoire of Architecture et Fonction des Macromolécules Biologiques* (AFMB https://www.afmb.univ-mrs.fr/) of Marseilles, France and in collaboration with Dr. Nicola Curci of the Institute of Biosciences and Bioresources of the Italian National Research Council (IBBR-CNR), we participated in the ODE project for the expression, purification and activity screening of 48 of the 192 selected clones belonging to new putative 15 GH families (called X Modules), using the high-throughput platform developed during the project at the AFMB.

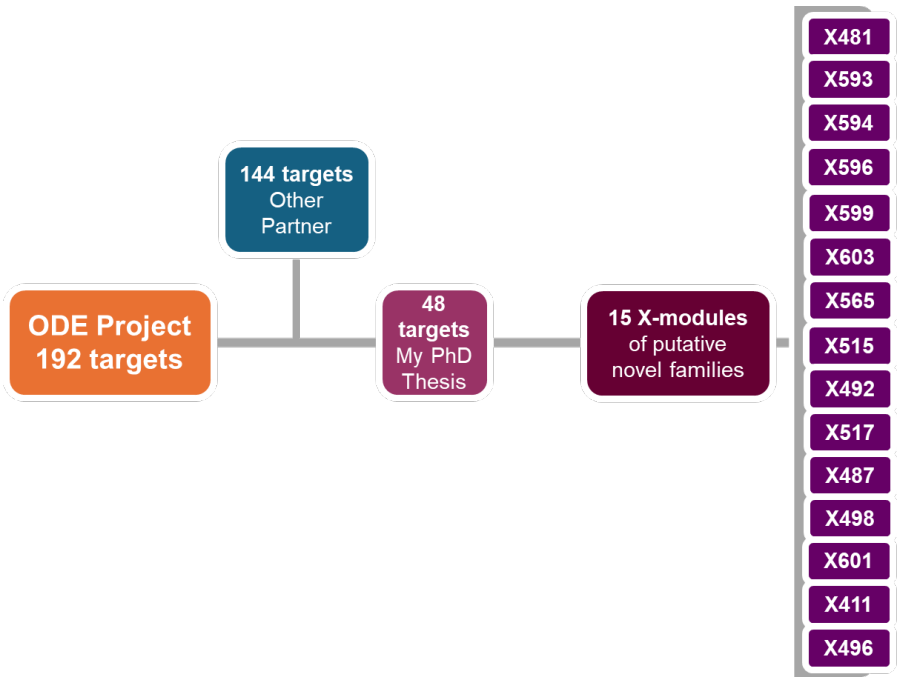

Fig 4.1: diagram of ODE Project according to this PhD project: in orange the total targets selected (192); In blue the 144 sequences object of the studies of the partners of this project; in purple the 48 targets discussed in this section and the belonging 15 X-modules.

4.2. Results and Discussion

4.2.1 High throughput protein production and screening of the 48 ODE targets

Starting from synthetic genes optimized for *E. coli* expression, the protein production and purification were performed at the AFMB laboratory using a high-throughput recombinant protein production platform, thus evaluating production and the purity degree by SDS-PAGE. Out of 48 clones, 37 were successfully expressed and purified (Fig. 4.2).

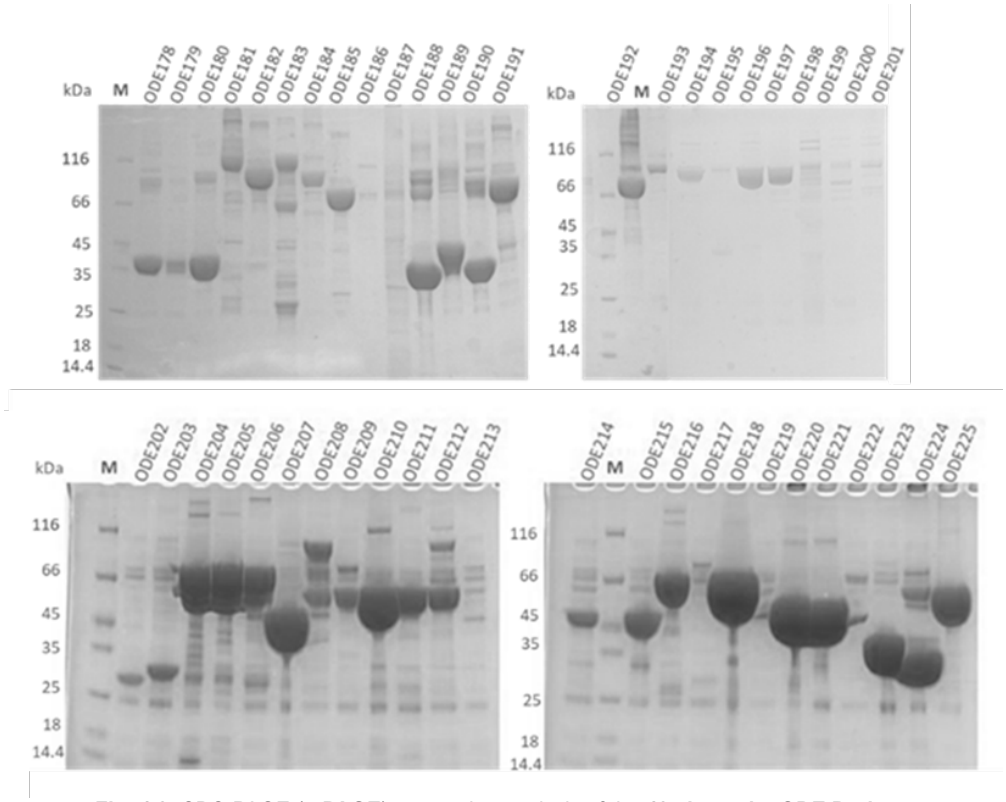

Fig. 4.2: SDS-PAGE (**mPAGE**) expression analysis of the **48 clones by ODE Project**.

All putative GH candidates in the soluble fraction were initially screened against classical colourimetric substrates (pNP-glycosides) to determine their activities. Subsequently, the activity was tested on polysaccharide chains.

Among the 37 clones subjected to functional-qualitative screening, 18 exhibited enzymatic activity (Tab. 4.1).

Target	MW (kDa)	Family	CAZy ID	Organism	Production	pNP Activity	Polysach. Activity
ODE178	38,28	X481	2852871	*Aeoliella mucimassa Pan181*		ND	ND
ODE179	36,41	X481	463719	*Draconibacterium orientale FH5*		ND	ND
ODE180	36,87	X481	258353	*Sunxiuqinia dokdonensis SK*		ND	ND
ODE181	130,57	X593	80318	*Bacteroides thetaiotaomicron VPI-5482*		ND	ND
ODE182	101,54	X593	876097	*Parabacteroides gordonii DSM 23371*		ND	ND
ODE183	125,22	X593	1512523	*Flavobacterium gilvum EM1308*		pNPαMan	ND
ODE184	96,63	X593	1333203	*Paenibacillus glucanolyticus 5162*		pNPαMan	ND
ODE185	75,24	X594	4756644	*Bacteroides caccae ATCC 43185*		pNPαMan	ND
ODE186	73,79	X594	4756705	*Rhizobium leguminosarum ATCC 14479*	ND	ND	ND
ODE187	81,02	X594	4756662	*Lacrimispora saccharolytica FDAARGOS_1340*	ND	ND	ND
ODE188	37,46	X596	3610154	*Bacteroides fragilis 638R*		pNPαMan	ND
ODE189	41,48	X596	4170172	*Lachnospiraceae bacterium CE91-St56*		pNPαMan	ND
ODE190	37,44	X596	3625006	*Candidatus Symbiothrix dinenymphae B4-10h*		ND	ND
ODE191	85,58	X599	4605589	*Bacteroides eggerthii DSM 20697*		ND	ND
ODE192	81,01	X599	2826701	*Crateriforma conspicua Mal65*	ND	ND	ND
ODE193	85,62	X599	1979800	*Polaribacter sp. SA4-12*		ND	ND
ODE194	101,66	X603	4776714	*Siansivirga zeaxanthinifaciens CC-SAMT-1*		ND	ND
ODE195	103,14	X603	4776707	*Spirosoma sp. KCTC 42546*	ND	ND	ND
ODE196	98,21	X603	4776703	*Rhizobium leguminosarum bv. viciae BIHB 1148*		ND	ND
ODE197	97,91	X603	4776711	*Marinovum algicola DG 898*		ND	ND
ODE198	69,75	X565	4398023	*Zobellia galactanivorans DsijT*	ND	ND	ND
ODE199	32,58	X565	4398052	*Wenyingzhuangia fucanilytica CZ1127*	ND	ND	ND
ODE200	31,4	X565	4398072	*Wenyingzhuangia fucanilytica CZ1127*	ND	ND	ND
ODE201	29,74	X565	4398044	*Micromonospora coxensis DSM 45161*	ND	ND	ND
ODE202	28,93	X565	4398046	*Micromonospora coxensis DSM 45161*		ND	ND
ODE203	32,94	X565	4398079	*Alteromonas mediterranea AltCH17*		ND	ND
ODE204	77,14	X515	3873939	*Bacteroides cellulosilyticus WH2*		pNPαGal	ND
ODE205	76,64	X515	3873955	*Zobellia galactanivorans DsijT*		pNPαGal	ND
ODE206	73,97	X515	3874017	*Shewanella cyperi FJAT-53720*		pNPαGal	ND
ODE207	50,8	X492	315839	*Zobellia galactanivorans DsijT*		pNPαGlc	ND
ODE208	105,79	X517	3874243	*Turicibacter sanguinis MOL361*		pNPαGal	ND
ODE209	65,37	X517	3874219	*Bacteroides ovatus ATCC 8483*		pNPβXyl	ND
ODE210	65,41	X517	3874235	*Bacteroides ovatus ATCC 8483*		pNPβGal	Mucin type II
ODE211	62,37	X487	733152	*Bacteroides xylanisolvens NLAE-zl-P727*		ND	ND
ODE212	63,95	X487	683533	*Barnesiella intestinihominis YIT 11860*		pNPαGal	ND
ODE213	48,48	X498	3858976	*Mucilaginibacter ginsenosidivorax KHI28*	ND	ND	ND
ODE214	49,01	X498	3858955	*Pedobacter ginsengisoli T01R-27*		pNPβGlcNAc	ND
ODE215	49,12	X498	3610087	*Flavobacterium gilvum EM1308*		ND	ND
ODE216	73,06	X601	4606052	*Bacteroides eggerthii DSM 20697*		ND	ND
ODE217	83,8	X601	2205386	*Rhizobium leguminosarum ATCC 14479*	ND	ND	ND
ODE218	71,51	X601	1397152	*Flammeovirga sp. MY04*		pNPαXyl	Mucin type II
ODE219	53,58	X411	3610042	*Maribellus comscasis WC007*	ND	ND	ND
ODE220	54,27	X411	2770107	*Mangrovibacterium diazotrophicum DSM 27148*		ND	ND
ODE221	53,56	X411	2770164	*Roseimaritima ulvae UC8*		ND	ND
ODE222	48,59	X411	2783672	*Pseudoalteromonas luteoviolacea S4054*		ND	ND
ODE223	41,08	X496	3858890	*Sphingomonas sp. ABOJV*		pNPβGal	Pectin
ODE224	39,43	X496	3858884	*Niabella soli DSM 19437*		ND	Pectin
ODE225	62,1	X496	3610149	*Flavobacterium johnsoniae UW101*		ND	Pectin

Tab. 4.1: Result of the high-throughput screening of the 48 clones, in order: the clone, the molecular weight of the clone, the X module of belonging, the CAZy Id, the microbial source, the results of the high-throughput expression and purification, the results of the high-throughput activity screening on pNP-glycosides and polysaccharides. ND: activity not detected and in green the enzymes that gave a positive result (soluble/active).

The activity profiles of most putative enzymes aligned with the substrate predictions from the PUL analysis previously performed. However, some

enzymes exhibited activities that did not match the predictions, highlighting the complex nature of enzyme activity and the need for further exploration.

4.2.2 Validation of the positive targets

After the high-throughput screening, the 18 positive putative enzymes (Tab. 4.1) were subjected to activity validation after two chromatographic purification steps (Ni^{2+} IMAC and size-exclusion) to obtain a homogeneous sample. All putative enzymes have been purified to a purity of at least 80% (Fig. 4.3).

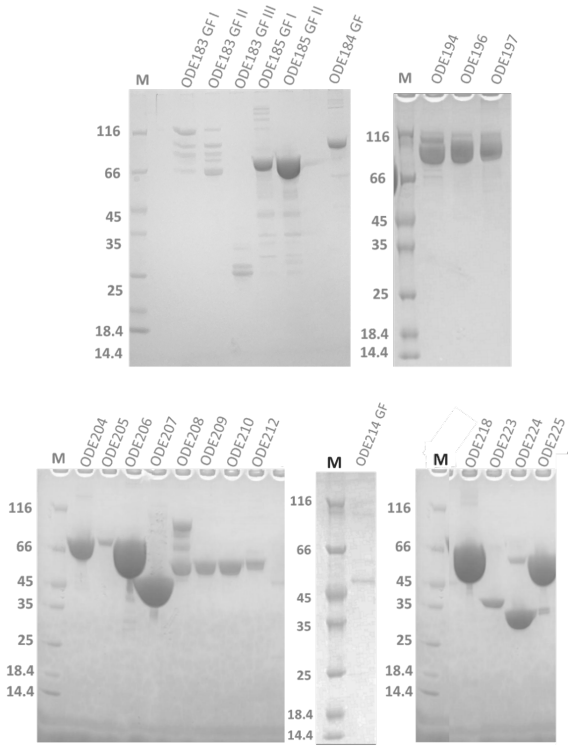

Fig. 4.3: SDS-PAGE (mPAGE) of the previously **active enzymes after two purification steps.**

SDS-PAGE analysis showed that putative enzymes ODE183 and 208 were subjected to a degradation phenomenon during purification; in particular, ODE 183 gave two distinct peaks during gel filtration (ODE183 GF I and II). Clone 205 and clone 214 were affected by high material loss during purification, particularly ODE214, following the removal of imidazole by dialysis precipitates.

All the purified putative enzymes were subject to a final screening activity to validate their activities; 9 enzymatic activities were confirmed (Tab. 4.2).

Target	MW (kDa)	Family	CAZy ID	Organism	pNP Activity
ODE183	125,22	X593	1512523	*Flavobacterium gilvum EM1308*	pNPαMan
ODE184	96,63	X593	1333203	*Paenibacillus glucanolyticus 5162*	pNPαMan
ODE185	75,24	X594	4756644	*Bacteroides caccae ATCC 43185*	pNPαMan
ODE188	37,46	X596	3610154	*Bacteroides fragilis 638R*	ND
ODE189	41,48	X596	4170172	*Lachnospiraceae bacterium CE91-St56*	ND
ODE204	77,14	X515	3873939	*Bacteroides cellulosilyticus WH2*	pNPαGal
ODE205	76,64	X515	3873955	*Zobellia galactanivorans DsijT*	ND
ODE206	73,97	X515	3874017	*Shewanella cyperi FJAT-53720*	pNPαGal
ODE207	50,8	X492	315839	*Zobellia galactanivorans DsijT*	pNPαGlc
ODE208	105,79	X517	3874243	*Turicibacter sanguinis MOL361*	ND
ODE209	65,37	X517	3874219	*Bacteroides ovatus ATCC 8483*	ND
ODE210	65,41	X517	3874235	*Bacteroides ovatus ATCC 8483*	pNPβGal
ODE212	63,95	X487	683533	*Barnesiella intestinihominis YIT 11860*	ND
ODE214	49,01	X498	3858955	*Pedobacter ginsengisoli T01R-27*	pNPαGal
ODE218	71,51	X601	1397152	*Flammeovirga sp. MY04*	pNPαXyl
ODE223	41,08	X496	3858890	*Sphingomonas sp. ABOJV*	ND
ODE224	39,43	X496	3858884	*Niabella soli DSM 19437*	ND
ODE225	62,1	X496	3610149	*Flavobacterium johnsoniae UW101*	ND

Tab. 4.2: results of the activity screening of previously positive clones after two steps of purification, in order: the clone, the molecular weight of the clone, the X module of belonging, the CAZy Id, the microbial source, the results of the activity screening on pNP-glycosides and polysaccharides. ND: activity not detected and in green the enzymes that gave a positive result.

Selected enzymes were subjected to detailed experimental characterisation to understand their biochemical properties and potential applications. This analysis included determining kinetic parameters, substrate specificity, and stability.
The six selected clones were initially biochemically characterised on pNP-substrate, which was detected during the high-throughput screening and validated (Tab. 4.3).

4.2.3. Possible GH38 subfamilies division thanks to the discovery of three novel mannanase

Among the GH families available on the database, the Glycoside Hydrolase 38 (GH38) family is notable for its diverse range of sequences, over 12,000, encompassing approximately 50 well-characterized α-mannosidases. These enzymes are involved in various physiological processes, including the *N*-glycosylation pathway in the Golgi and the lysosomal degradation of glycoproteins. Additionally, some members of the human gut microbiota utilise α-mannan from yeast, highlighting their ecological importance [59].

Sequences within the non-classified Glycoside Hydrolase (GH0) group exhibit remote homology to GH38. This analysis revealed two distinct groups of sequences spanning all kingdoms of life, which prompted a re-evaluation of the GH38 family. The hypothesis is that the evolutionary split within GH38 results in two main groups, with the GH0 sequences lying between these groups, potentially explaining their previous classification difficulties. Intriguingly, members of the Bacteroidetes phylum from these remote groups were found in PULs alongside CAZymes with diverse substrate specificities, suggesting functional roles in carbohydrate metabolism that diverge from well-characterized patterns.

Some of these sequences, such as X-modules X593 and X594, were selected during the ODE Project for further investigation due to their potential as members of a novel enzyme family. Three sequences, designated ODE183, ODE184, and ODE185, were successfully expressed and purified. Their enzymatic activity on pNP-α-D-Man was confirmed, setting the stage for a comprehensive biochemical characterisation to understand these potential subfamilies.

Enzymes belonging to GH38, also named class II α-mannosidases, follow a retaining reaction mechanism [99] and have one large globular domain consisting of an N-terminal α/β domain (approximately 1000-residue), a three-helical bundle, and an all-β C-terminal domain named “mannosidase fold” [10]. The structural analysis was performed by using 3D structure prediction of the three enzymes (Overall quality factor: ODE183 94.7, ODE184 95.02 and ODE185 96.99) and the closest PDB [100] available identified by HHpred [101], the GH38 from *Bifidobacterium longum NCC2705* (Bl_Man38A), a mannosyl-oligosaccharide α-1,6-mannosidase (Fig. 4.4) [59]. Through structural superimposition and multiple sequence alignment (MSA) [102] of the GH38 family with select sequences annotated as X593 and X594 modules, it was possible to reveal that the new enzymes display the characteristic “mannosidase fold” typical of family 38. Additionally, this analysis identified a putative catalytic site (Fig. 4.5A) and three candidate catalytic residues for each enzyme based on their positioning within the distances known for *retaining* enzymes (Fig. 4.5 B, C).

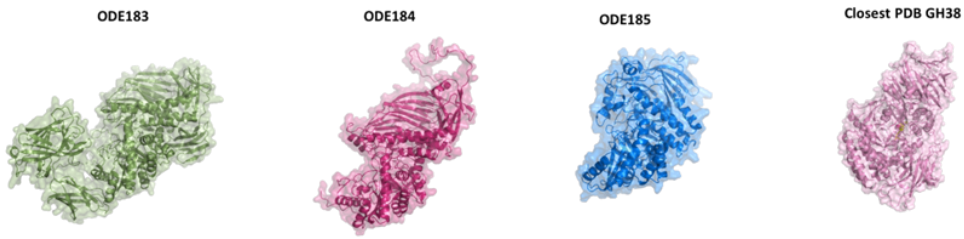

Fig. 4.4: AFold predicted model of ODE183 (green)**, ODE184** (magenta) **and ODE185** (blue) and the closest PDB file available of GH38 from *Bifidobacterium longum* NCC2705 (pink)

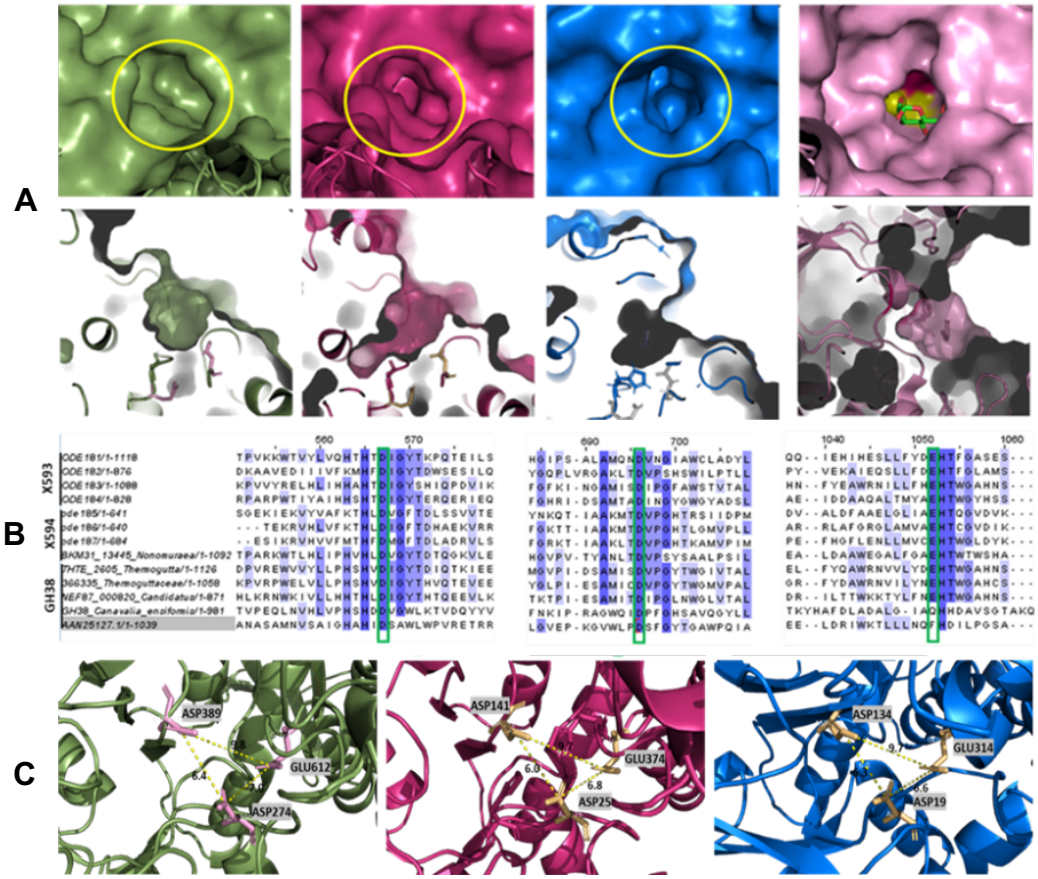

Fig: 4.5: 3D structure analysis of AFold model of ODE183 (green), 184 (magenta) and 185 (light blue) clones, **compared to the closest PDB** (pink), GH38 from *Bifidobacterium longum* NCC2705 (Bl_Man38A). (**A**) close-up of the hypothetical **catalytic pocket** (frontal and lateral view). (**B**) **MSA** of the ODE clones object of this section, some X593 and X594 sequences annotated and family GH38, including the closest PDB (highlighted in grey). (**C**) **The putative catalytic residue** and their distance for each enzyme were detected.

The biochemical characterisation of the three enzymes revealed that ODE184 and ODE185 have the same optimal pH (6.5) and temperature (45 °C). At the same time, ODE183 prefers a more acidic pH (5.0) and lower temperature (37 °C), consistent with its mesophilic nature (Fig 4.6 A, B).

Regarding functional thermal stability (Fig. 4.6 C), the residual activity on pNP-α-D-Man was reported after incubation at optimum temperature. ODE183 displayed a marked decline in activity at 37°C, with less than 40% residual activity after just 30 minutes and less than 10% after four hours. In contrast, ODE184 and ODE185 demonstrated superior thermostability: ODE184 retained approximately 80% of its activity after three hours, while ODE185 showed 60% residual activity under the same conditions. After 24 hours, ODE185 lost all activity, but ODE184 remained notably more stable,

retaining about 40% of its activity. Finally, regarding the melting temperature, all the enzymes are characterised by a value between 58-62 °C (Fig. 4.6 D).

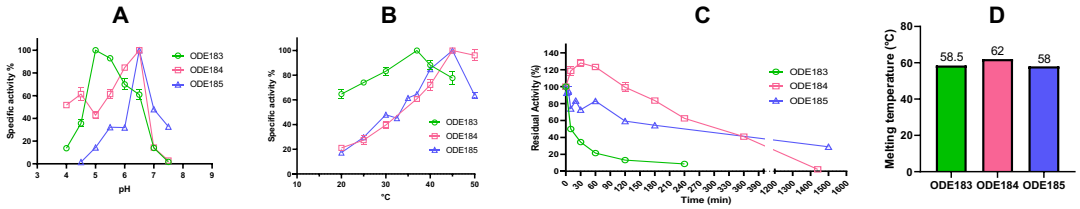

Fig.4.6: Biochemical characterisation of the three mannosidases belonging to X593 and X594: (**A**) activity in the function of pH; (**B**) activity in the function of temperature; (**C**) functional thermostability of the enzymes; (**D**) Structural thermostability by nanoDSF.

The enzymes' kinetic characterization showed a general low affinity for pNP-α-D-Man, with K_M values of ~270 mM for ODE183 and ODE184 and 190 mM for ODE 185. Moreover, the value of k_{cat}/K_M also shows that the enzyme is not erricient in binding this substrate and converting it to a product, suggesting that pNP-α-D-Man is not the enzyme's preferred substrate. Further tests on natural substrates will be necessary to identify the enzymes' preferred substrate (Tab. 4.3).

			K_M	k_{cat} monomer	k_{cat}/K_M
Putative GH family	**Target**	**Substrate**	**mM**	**sec^{-1}**	**$sec^{-1}*mM^{-1}$**
X593	**ODE183**	pNP-α-D-Man	272.8 ± 4.8	3.2 ± 1.3	0.012
	ODE184	pNP-α-D-Man	273 ± 0.66	11 ± 1.2	0.04
X594	**ODE185**	pNP-α-D-Man	190 ± 6.4	21 ± 6.4	0.11

Tab. 4.3: ODE183, 184 and 185 Kinetics constants on pNP-α-D-Man in each optimal condition.

The GH38 family of glycoside hydrolases often features a metal ion in their active site, typically zinc. This metal ion plays a crucial role in the enzyme's catalytic mechanism, forming an integral part of the -1 site. Zinc helps stabilise the enzyme's structure and participates directly in the hydrolysis of glycosidic bonds. It coordinates with key amino acid residues and the substrate, facilitating the proper positioning of the substrate and activating water molecules necessary for the hydrolytic reaction [9]. The presence of the metal ion is essential for the enzyme's catalytic efficiency and specificity, making it a critical component in the function of GH38 mannosidases.

The genetic context of the X593 module allows us to hypothesise the functional role of this enzyme (Fig. 4.8). In this PUL [96], the X593 sequence was found to be associated with a GH43_24, a family characterised by exo-β-1,3-galactanase (EC 3.2.1.145), and two GH92. The first GH92 is the α-mannosidase BT2111 already characterised [103] and a novel GH92 (sharing only 44% of identity with BT2111); X593 is frequently found in the same operon as other α-mannanases, such as GH92, which are present in 71% of X593-containing PULs. This co-

occurrence strongly suggests that these PULs are involved in the degradation of mannose-rich substrates, such as high mannose-type *N*-glycans (HMNG). HMNGs are oligosaccharides with complex mannose branches, including α-(1-2), α-(1-3), and α-(1-6) linkages. These structures are abundant in specific biological contexts, such as apoptosis vesicles and within the human gut. Therefore, this operon, including X593, is likely involved in the degradation of HMNG, contributing to mannose metabolism and the utilisation of these complex glycans. This hypothesis aligns with the observed genetic associations and suggests that X593-containing PULs have a functional role in mannose-rich oligosaccharide processing.

Fig. 4.8: **operon belonging *Bacteroides thetaiotaomicron VPI-5482*.** In order: a X593 annotated (ODE181), GH43_24 with putative exo-β-1,3-galactanase activity; the GH92 BT2111; two sequences with unknown function; SusD, hypothetical protein ; putative outer membrane protein, SusC, probably involved in nutrient binding; MFS, putative transmembrane hexose transporter; ROK family member transcriptional repressor; and GntR, transcriptional regulator [96].

Regarding the X594, the ODE185 was identified in a PUL belonging to *Bacteroides caccae ATCC 43185* (60 PULs predicted by the PULDB database). Given the different hypothetical activities present in the operon, this PUL is less informative for our analysis (Fig. 4.9).

Fig. 4.9: **operon belonging *Bacteroides caccae* ATCC 43185 (ode205).** In order: ECF-σ, RNA polymerase sigma factor; GH92_5, hypothetical 1,4-mannosidase; Anti-σ, Fe^{2+}dicitrate sensor, membrane component; SusD, hypothetical protein; putative outer membrane protein, SusC, probably involved in nutrient binding; sequence of known function protein; GH35, putative β-galactosidase; GH31_4, α-xylosidase; GH43_1, hypothetical arabinoxylan or Arabinofuranohydrolase with CBM91 (binding to xylans); the ODE185 sequences belonging to X594 with α-mannosidase activity; Pept_MA, F5/8 type C domain; GH2_6, putative β-galactosidase/β-glucuronidase; GH120_dist, putative xylan β-1,4-xylosidase; GH95, hypothetical protein linked to the X-modules X424_dist; sequence of known function protein; GH43_10, hypothetical β-xylosidase linked to CBM91; X491_2, sequences belonging to the new module with known function; and GH28, putative Endopolygalacturonase.

In conclusion, the discovery of three novel mannanase enzymes, ODE183, ODE184, and ODE185, has prompted a possible reclassification of the GH38 family into distinct subfamilies.

These findings expand our understanding of GH38 enzymes and suggest a functional role for the X593-containing PULs in mannose metabolism, specifically in degrading complex *N*-glycans. Future research should focus on natural substrates, further structural studies, and investigating metal ion dependencies to complete the family's subdivision and deepen the understanding of key functional differences among subfamilies.

4.2.4. A brand-new family active on galactopyronase substrate

To date, in the CAZy database, there are six main glycoside hydrolase families with α-galactosidase activity (EC 3.2.1.22), capable of hydrolyzing terminal, non-reducing α-D-galactose residues in α-D-galactosides.
These families are GH4, GH27, GH31, GH36, GH57, and GH97. α-Galactosidases are crucial enzymes that play significant roles in various biological processes and have diverse applications in medicine and industry. In plants, α-galactosidases are essential for mobilizing seed storage carbohydrates during germination [104]. They hydrolyze oligosaccharides like raffinose, stachyose, and verbascose into simpler sugars, providing energy for the growing seedlings. In humans and animals, these enzymes aid in the digestion of complex carbohydrates found in legumes and other plant-derived foods, enhancing nutrient absorption and contributing to gut health [105]. In the food industry, α-galactosidases remove anti-nutritional factors from legume-based products, improving their nutritional value and reducing gastrointestinal discomfort associated with their consumption. This application is particularly relevant for producing food products that are easier to digest and more suitable for human consumption [106].
Discovering new α-galactosidases family can significantly advance our knowledge of this class of enzymes. In particular, discovering and characterising novel enzymes with unique properties or enhanced efficiencies can pave the way to novel therapeutic applications.
During the ODE Project, some sequences were annotated as possible enzymes active on the galactopyranose substrate, as X-molude X515. In particular, one of the X515 sequences involved in this study was found in a PUL involved in the degradation of substrates containing galactopyranosides since all the sequences present in the operon belong to families characterised by this activity on this type of substrate (Fig. 4.10).

Fig. 4.10: operon belonging to *Bacteroides cellulosilyticus WH2*, bolded the X515 annotated (ODE204). In order: PL6_1, possible Alginate lyase; GH28_dist, putative exo-polygalacturonase belonging to a potential new subfamily; unk, two proteins with unknow function; A GH28; Two GH97 with putative α-glucosidase or α-galactosidase; GH31 belonging to future new subfamilies with α-galactosidase; GH36_1, α-galactosidase; X515, the ODE204; HTCS, One-component system sensor protein.

During the high throughput screening, two enzymes belonging to the X modules X515, ODE204 and ODE206, showed activity on pNP-α-D-Gal and subsequently validated it.
The biochemical characterisation of the two enzymes in the presence of pNP-α-D-Gal revealed the same pH_{opt} for the two enzymes (pH 5.0) and optimal temperature of 40 °C for ODE204 and 37 °C for the other (Fig 4.11

A, B). As previously shown about the enzymes active on pNP-α-D-Man, the structural thermal stability of the two enzymes was evaluated (Fig. 4.11 C, D). Regarding the functional thermostability, the two enzymes incubated at their optimal temperature showed different profiles. ODE206 after three hours has a residual activity of 20%, while ODE204 retained 100% activity after more than 72 hours. The great stability shown fromODE204 opens new avenues for their application[107]. The kinetic parameters of enzymes ODE204 and ODE206 were determined using pNP-α-D-Gal as substrate under optimal reaction conditions (Table 4.4). ODE204 exhibited a K_M value of 7.8 ± 0.86 mM, indicating a lower substrate affinity than ODE206, which showed a K_M of 5 ± 0.17 mM. However, ODE204 demonstrated a significantly higher k_{cat} value of 7.75 ± 0.46 s^{-1}, in contrast to ODE206's k_{cat} of 0.26 ± 0.046 s^{-1}, indicating a considerably higher turnover. These findings establish ODE204 as the most efficient enzyme, exhibiting a specificity constant 20-fold greater than ODE206.

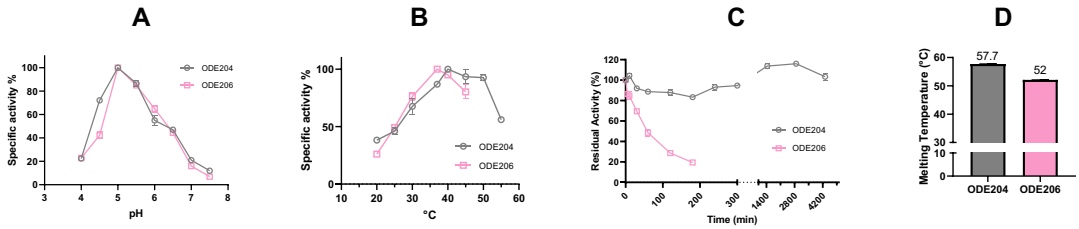

Fig. 4.11: Biochemical characterization of the α-galactosydases belonging to X515 modules: (**A**) activity in the function of pH; (**B**) activity in the function of temperature; (**C**) functional thermal stability of the enzymes at optimum of temperature; (**D**) Structural thermal stability.

Putative GH family	Target	Substrate	K_M mM	k_{cat} monomer sec^{-1}	k_{cat}/K_M $sec^{-1}*mM^{-1}$
X515	**ODE204**	pNP-α-D-Gal	7.8 ± 0.86	7.75 ± 0.46	1
	ODE206	pNP-α-D-Gal	5 ± 0.17	0.26 ± 0.046	0.05

Tab.4.4: Kinetics constants of ODE204 and 206 on pNP-α-D-Gal in each optimal condition.

The identification and characterization of these novel *α*-galactosidases provide a solid foundation for creating a new glycoside hydrolase (GH) family within the CAZy database, significantly expanding the range of enzymes available.

4.2.5. A α-glucosidase as the possible precursor of a novel family

A-glucosidases are a class of enzymes that catalyze, in carbohydrates, the hydrolysis of α-glucosidic linkages, playing a crucial role in the final step of carbohydrate digestion by converting complex sugars into absorbable monosaccharides such as glucose [19]. These enzymes belong to various families, primarily GH13 and GH31, each characterized by distinct structural features and substrate specificities. The biological significance

of α-glucosidases extends beyond digestion, as they are involved in various cellular processes including glycogen metabolism and glycoprotein processing. The discovery of new α-glucosidases is of paramount importance due to their potential therapeutic applications, particularly in the treatment of Pompe disease [108]. Also, α-glucosidases have significant biotechnological applications, such as in the production of biofuels, synthesis of oligosaccharides, and degradation of plant biomasses as a renewable energy source [21]. The ongoing exploration of α-glucosidases, including structural and functional characterisation, thus remains a critical area of research with implications for medicine and industry. During the functional screening of ODE project, novel α-glucosidases annotated as module X492, and closely related to the GH63 family, have been identified. The identified sequence is part of an operon from *Zobellia galactanivorans DsijT*, a Bacteroidota species isolated from red algae (91 PUL belonging to this microorganism was reported in licterature). This Polysaccharide Utilisation Locus (Fig. 4.12) also encodes a GH89 enzyme, which exhibits α-*N*-acetylglucosaminidase activity (EC 3.2.1.50) and comprises eight distinct enzymes. Additionally, the operon includes a GH20 enzyme that shares 27% identity with a β-*N*-acetylglucosaminidase.
Moreover, within the same operon was identified a sequence annotated as ODE205 belonging to the X515 module. However, during high-throughput screening, this module did not exhibit any detectable activity.

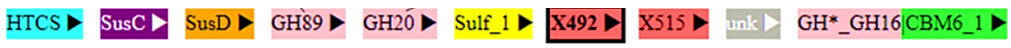

Fig. 4.12: operon belonging to *Zobellia galactanivorans* DsijT, bolded the X492 identified (ODE207). In order: HTCS, One-component system sensor protein; SusD, hypothetical protein; putative outer membrane protein, SusC, probably involved in nutrient binding; GH89, A-N-acetylglucosaminidase; GH20, hypothetical B-N-acetylhexosaminidase; Sulf_1, Sulfatase, family S1-20; X492, the ODE207, object of this study, with α-glucosidase activity; X515, sequence of ODE205 (no activity detected); unk, protein with unknown function; GH*_GH16, sequence with unknown function close to family GH16, linked to a CBM6_1.

Since the identified enzyme differed from every other sequence present in the CAZy database, its biochemical characterization of ODE207 was performed to define its activity and establish a new family. The results revealed an optimal pH of 5.0 and an optimal temperature of 50 °C, which is notably high considering the mesophilic source (Fig. 4.13 A, B). Then, both functional and structural thermal stability of the enzyme were evaluated (Fig. 4.13 C, D). The enzyme exhibited low stability at 50°C (as T_{opt}), with no detectable activity after 80 minutes of incubation. This instability is likely due to the enzyme's structural thermal stability (T_m=51.5 °C).

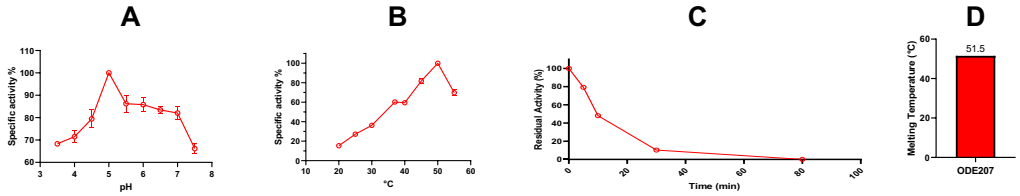

Fig. 4.13: Biochemical characterization of ODE207: (**A**) activity in the function of pH; (**B**) activity in the function of temperature; (**C**) functional thermal stability at optimum of temperature; (**D**) Structural thermal stability by nanoDSF.

The kinetic characterization of ODE207 on pNP-α-D-Glc as the substrate showed a K_M of 0.55 mM, indicating good substrate affinity (Tab. 4.5). However, the k_{cat}/K_M ratio of 0.077 $s^{-1} \cdot mM^{-1}$ reveals that, despite good affinity for the substrate, the enzyme has low catalytic efficiency. These findings suggest that the enzyme may be functionally specialized to act on a different substrate containing α-glucopyranose units. Additional studies will assess substrate specificity through tests on natural oligosaccharides to further investigate the enzyme's substrate specificity and regioselectivity.

			K_M	k_{cat} monomer	k_{cat}/K_M
Putative GH family	**Target**	**Substrate**	**mM**	**sec^{-1}**	**$sec^{-1}*mM^{-1}$**
X492	**ODE207**	pNP-α-D-Glc	0.55 ± 0.04	0.04 ± 0.0006	0.077

Tab. 4.5: **Kinetics constants of ODE207** on pNP-α-D-Glc in optimal condition.

The produced structural prediction of ODE207 (96.65 overall quality factor) was compared with the closest PDB [100], identified by HHpred [101] to highlight the catalytic pocket and the putative residue (Fig. 4.14 A, B). The closest PDB file used for this study is GH37 from *Arabidopsis thaliana* is the crystal structure with trehalase activity (α-D-glucopyranosyl-(1→1)-α-D-glucopyranoside, EC 3.2.1.28). This enzyme belongs to clan GH-G, characterized by $(\alpha/\alpha)_6$ barrel fold similar to other α-toroidal glycosidases [109]. It was possible to identify two candidates as putative catalytic residues, ASP213 and GLU405, at a distance of 8 Å, in accordance with GH distance.

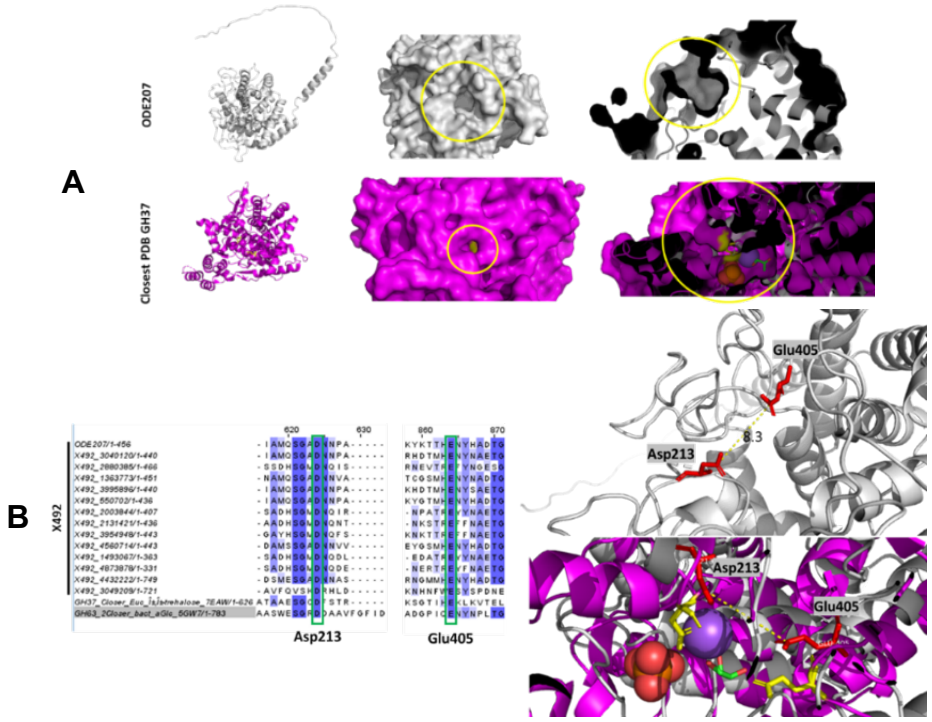

Fig. 4.14: 3D structure analysis of AFold model of ODE207 (grey), (**A**) **compared to the closest PDB** (pink), GH63 (YgjK) from *Escherichia coli* str. K-12 substr. MG1655; and a close-up of the hypothetical **catalytic pocket** (frontal and lateral view). (**B**) **MSA** of the ODE clones object of this section, some X492 sequences annotated, the closest PDB files (GH63 highlighted in grey), the GH37; and the putative catalytic residue detected, and their distance for each enzyme.

In conclusion, the novel α-glucosidase ODE207 from *Zobellia galactanivorans* DsijT, linked to the GH63 family, displays distinct biochemical properties, suggesting a specialized function and establishing a new family based on module X492. The definitive identification of the catalytic residues of enzyme ODE207 and elucidation of its mechanism of action can only be achieved through the generation of mutants, chemical rescue approaches, use of mechanism-specific inhibitors, and structural analysis of the products. [110] these investigations, however, lie beyond the scope of the present study.

4.3. Conclusion

In conclusion, this chapter presents a successful approach to the identification, production, and characterization of novel glycoside hydrolase (GH) families derived from Polysaccharide Utilization Loci (PULs) using high-throughput methods. Our findings led to the discovery of nine new enzymes through a rigorous high-throughput screening and activity validation process, starting from an initial pool of 48 clones. The characterization of three of these novel enzymes allowed for the proposal of a division of family 38 into two distinct subfamilies, as well as the

establishment of two new families. This work significantly contributes to the expanding knowledge of carbohydrate-active enzymes (CAZymes) and underscores the critical importance of integrating computational and experimental approaches in the discovery and characterization of enzymes.

4.4. Materials and Methods

4.4.1 High-throughput recombinant protein production

The strategy of this study, was to produce synthetic genes from a commercial provider (Twist Bioscience, USA). Every clone has its codons optimized for *E. coli* expression and then cloned in a pET29a(+) expression vector containing a His-Tag for protein purification.

Protein production and purification were performed at the AFMB laboratory using a high-throughput recombinant protein production platform. All the clones were expressed in the competent LEMO21 (DE3) *E. coli* strain with 0.5 mM IPTG and induced at 18 °C overnight, using Turbo Broth medium. Proteins were purified by nickel affinity chromatography using the TECAN freedom EVO200 robot. Proteins were purified by nickel affinity chromatography on standardized 96-well plates in the presence of running buffer Tris-HCl 50 mM, NaCl 300 mM, Imidazole 10 mM pH 8; remove the aspecific using Imidazole 50 mM and the elution Imidazole 500 mM. The production and the degree of purity were evaluated preliminarily by SDS-PAGE. Each clone was subject to dialysis to remove imidazole.

4.4.2 High-throughput activity screening

All soluble GH candidates were screened against classical colourimetric substrates from Sigma-Aldrich (USA) and Biosynth (Unite Kingdom) at 5 mM concentrated (pNP-glycosides) to determine their activities: pNP-α-D-Glucopyranoside; pNP-α-D-Galactopyranoside; pNP-α-D-Mannopyranoside; pNP-α-D-Xylopyranoside; pNP-N-acetyl-α-D-Glucosaminide; pNP-*N*-acetyl-α-D-Galactosaminide; pNP-α-D-Glucuronide; pNP-α-D-Maltopyranoside; pNP-β-L-Arabinopyranoside; pNP-α-D-Rhamnopyranoside; pNP-α-L-Fucopyranoside; pNP-β-d-Glucopyranose; pNP-β-D-Galactopyranose; pNP-β-D-Xylopyranose; pNP-N-acetyl-β-D-Glucosaminide; pNP-*N*-acetyl-β-D-Galactosaminide; pNP-α-L-Arabinopyranoside, pNP-β-L-Fucopyranoside; pNP-β-D-Glucuronide; pNP-β-D-Cellobiose.

Moreover, the activity was tested on polysaccharide 0.2% (w/v) from Sigma-Aldrich (USA) and Biosynth (Unite Kingdom) chains chosen according to the PUL of belonging analysis, using colorimetry-based detection of the reducing ends, Cole's ferricyanide method: Beechwood Xylan, Type I Mucin from Bovine Submaxillary Glands, Type II Mucin from Porcine Stomach, β-1,4-Galactan, α-Mannan, Sodium Alginate, Fucoidan, Larch Arabinogalactan, Chitin, Laminarine, Carboxymethyl cellulose

(CMC), Curdlan, Heparin Sodium, L-Rhamnose Monohydrate, Chondroitin Sulfate A, Chondroitin Sulfate C, Citrus Pectin, Pullulan and Starch.
All the reactions were carried out overnight in the presence of 100 mM phosphate buffer pH 6.5 at 25°C.

4.4.3 Production and validation of the activity of the positive clones
Each positive clone was expressed in the same condition as the high-throughput expression but scaled up to 400 ml of Turbo Broth. Proteins were purified by nickel affinity chromatography using HisTrap FF 5ml, followed by a Gel filtration step using HiLoad 16/600 Superdex 200 pg. The production and the degree of purity were evaluated preliminarily by SDS-PAGE, and the protein concentration was determined using Bradford's method.
Each clone was subject to validation in order to confirm the previous activity detected in the High-throughput screening.

4.4.4 Structural analysis
Structural predictions of the six enzymes in their monomeric form were generated using AlphaFold2 [111] through ColabFold [112]. To compare the predicted structures with those of crystal structures, the following closest PDB [100] entries were utilised, identified by HHpred [101], along with their corresponding accession numbers:

- GH38, (BI_Man28A) *Bifidobacterium longum* (strain NCC 2705), PDB code: 7UFR;
- GH94, (ChbP) from *Vibrio proteolyticus,* PDB code: IV7V;
- GH37, (AtTre1) from *Arabidopsis thaliana,* PDB code:7E9U.

The Multi Sequences Alignment (MSA) for the identification of the putative catalytic residue was performed using Clustal Omega online tool [102].

4.4.5 Biochemical characterization
For each enzyme, the standard assay on pNP- substrates was conducted as an end-point measurement in a 200 µL reaction mixture. After incubation, the reaction was blocked in ice, by adding 800 µL of 1 M sodium carbonate buffer (pH 10.2). The absorbance was then measured at 420 nm using a spectrophotometer, with the molar extinction coefficient (ε_{mM}) for p-nitrophenol in sodium carbonate at pH 10.2 17.3 cm^{-1} mM^{-1}.Here reported the detailed conditions for each enzyme:

- ODE183 (~ 6.2 µg of enzymes in Buffer Phosphate 50 mM, NaCl 150 mM pH 7.8) was tested with 10 mM pNP-α-D-Mannopyranoside in the presence of 100 mM citrate-phosphate buffer pH 5.0 at 37 °C in 10 minutes.
- ODE184 (~ 18 µg of enzymes in Buffer Phosphate 50 mM, NaCl 150 mM pH 6.8) was assayed with 10 mM pNP-α-D-Mannopyranoside in the presence of 100 mM citrate-phosphate buffer pH 6.5 at 45 °C in 10 minutes.

- ODE185 (~320 µg of enzymes in Buffer Phosphate 50 mM, NaCl 150 mM pH 7.8) was evaluated with 10 mM pNP-α-D-Mannopyranoside in the presence of 100 mM citrate-phosphate buffer pH 6.5 at 45 °C in 5 minutes.
- ODE204 (~ 8.2 µg of enzymes in Buffer Phosphate 50 mM, NaCl 150 mM pH 6.25) was tested with 10 mM pNP-α-D-Galactopyranoside in the presence of 100 mM citrate-phosphate buffer pH 5.0 at 40 °C in 2 minutes.
- ODE206 (~70 µg of enzymes in Buffer Phosphate 50 mM, NaCl 150 mM pH 6.8) was assayed with 10 mM pNP-α-D-Galactopyranoside in the presence of 100 mM citrate-phosphate buffer pH 5.0 at 37 °C in 5 minutes.

The pH_{op} for each enzyme was determined using a 100 mM citrate-phosphate buffer, ranging from pH 3.0 to 9.0, at 37 °C as reported above. The ODE enzymes were incubated at different temperatures (20 °C to 55 °C) under standard conditions to evaluate thermophilicity. Functional thermostability was assessed by incubating the enzymes in their optimal pH and temperature conditions and measuring residual activity after incubation with 10 mM of the appropriate substrate. The enzymes' melting Tm was determined using nano-Differential Scanning Fluorimetry (nanoDSF) with nanoTEMPER technology (Germany), which monitors protein responses to thermal stress through intrinsic fluorescence.

Kinetic studies for clones X593 and X594 were conducted on pNP-α-D-Mannopyranoside (pNP-α-D-Man) at concentrations ranging from 0.1 to 25 mM under optimal conditions: ODE183 at pH 5.0 - 37 °C, ODE184 at pH 6.5 - 45 °C and ODE185 at pH 6.5 - 45 °C . For α-galactosidase (ODE 183, 184 and 185), kinetic parameters were measured using pNP-α-D-Galactopyranoside (pNP-α-D-Gal) at concentrations from 0.25 to 30 mM at its optimal conditions (pH 5.0 at 40 °C). ODE207's kinetics were assessed with pNP-α-D-Glucopyranoside (pNP-α-D-Glc) over a range of 0.025 to 30 mM at pH 5.0 and at 37 °C.

CHAPTER 5: CLOSING REMARKS

In recent years, the demand for efficient and sustainable catalytic systems has experienced unprecedented growth, driven by the pressing need to minimize environmental impact while maintaining industrial productivity. Biological catalysts, particularly glycoside hydrolases (GHs), have emerged as pivotal players in this landscape due to their exceptional versatility and efficiency. These enzymes, capable of catalyzing the hydrolysis of glycosidic bonds, have become indispensable across a wide array of sectors.

This thesis aims to expand the repertoire of available glycoside hydrolases and deepen our understanding of their potential applications. **Chapter 2** presents a robust pipeline for discovering novel GHs, resulting in the development of *Meta*GH (Metagenomic GH Hub), a comprehensive catalogue comprising approximately 8,000 putative GHs from extreme environments. These enzymes hold significant promise for future studies, particularly in the degradation of recalcitrant polysaccharides. However, the effectiveness of enzyme discovery relies on a combined approach integrating both in-silico and biochemical work. **Chapter 3** exemplifies this interdisciplinary effort by discovering and characterising a novel GH122 enzyme. This enzyme represents a breakthrough, being the first α-glucuronidase identified from an archaeal organism, thereby expanding our understanding of the functional biodiversity within the GH122 family.

Beyond metagenomic analysis, our research explores enzyme discovery by investigating Polysaccharide Utilization Loci (PULs). **Chapter 4**, in collaboration with Professor Nicolas Terrapon, shows the power of PUL analysis in discovering novel CAZymes. This approach led to the identification of nine novel GHs, with the biochemical characterisation of six of these laying the groundwork for establishing new GH families.

In conclusion, this thesis has focused on advancing glycoside hydrolase research on multiple fronts, from *in-silico* identification to biochemical characterisation. Together, these efforts underscore the potential of glycoside hydrolases and the necessity to invest effort in their study.

6. APPENDIX

6.1. Publications

Manuscript in preparation
Title: “A stable GH31 α-glucosidase as a model system for studying mutations leading to human glycogen storage disease type II.”
Authors: Francesca Maria Pia Paragliola, Andrea Strazzulli, Roberta Iacono, and Marco Moracci.

6.2. Communication

6.2.1. Poster Communication: Presenting author

- Meeting Sezione SIB Campania 2022, Complesso di Biotecnologie dell'Università degli Studi di Napoli Federico II - Via Tommaso De Amicis, 95, 80131, Naples, Italy. November 26, 2022
  Presenting author (poster session)
  Title: *Engineering of thermophilic α-glucosidase MalA from the hyperthermophilic crenarchaeota Saccharolobus solfataricus to change the substrate specificity.*
  Authors: Francesca Maria Pia Paragliola, Carmen Ercolano, Andrea Strazzulli, Roberta Iacono and Marco Moracci.
- 3rd edition of the Integrative Structural Biology Meeting (BSI) – Marseille, France, Faculté des sciences médicales et paramédicales Secteur Timone Aix-Marseille Université (AMU), November 13, 2023 -November 17, 2023
  Presenting author (poster session)
  Title: *Exploring the catalytic site of the GH31 MalA α-glucosidase to extend its substrate specificity.*
  Authors: Francesca Maria Pia Paragliola, Andrea Strazzulli, Roberta Iacono and Marco Moracci
- Carbohydrate Bioengineering Meeting XV, Ghent, Belgium - May 5, 2024 – May 8, 2024
  Presenting author (poster session)
  Title: *A stable GH31 α-glucosidase as model system for the study of mutations leading to human glycogen storage disease type II.*
  Authors: Francesca Maria Pia Paragliola, Andrea Strazzulli, Roberta Iacono and Marco Moracci
- 48th FEBS Congress: Mining biochemistry for human health and well-being. Milan, Italy – 29 June to 3 July 2024.
  Presenting author (poster session)
  Title: Inside Carbohydrate-active enzymes: a novel enzyme from Pisciarelli hot spring discovered by a metagenomic approach.
  Authors: Francesca Maria Pia Paragliola, Andrea Strazzulli, Sergio Roberto Scotillo Roberta Iacono and Marco Moracci

6.2.2. Poster Communication: co-author

- 13th International Forum on Industrial Biotechnology and Bioeconomy 2023
  Florence, Italy. September 28, 2024 – September 29, 2024
  Co-author (poster session)
  Title: *Screening and design methods for optimization of lignocellulosic biomass saccharification by new biocatalyst and process strategies.*
  Authors: Adriana Posillipo, Pina Notaro, Francesca Maria Pia Paragliola, Andrea Strazzulli, Marco Moracci, Maria Elena Russo, Piero Salatino and Antonio Marzocchella.
- Carbohydrate Bioengineering Meeting XV, Ghent, Belgium - May 5, 2024 – May 8, 2024
  Co-author (poster session and short talk)
  Title: *Metagenomic identification and enzymatic characterization of a hyperthermophilic glycosidase from Pisciarelli hot springs indicate a novel functional annotation of family GH122.*
  Authors: Andrea Strazzulli, Francesca Maria Pia Paragliola, Sergio Roberto Scotillo Roberta Iacono and Marco Moracci

6.2.2. Oral Communication

- XVI FISV Congress, at the XVI FISV Congress (Federazione Italiana Scienze della Vita) - Reggia di Portici - Via Università, 100, 80055, Portici, Naples, Italy. September 14, 2022 – September 16, 2022
  Presenting author (poster session and short oral presentation)
  Title: *Characterization and engineering of thermophilic α-glucosidase MalA from the hyperthermophilic crenarchaeota Saccharolobus solfataricus.*
  Authors: Francesca Maria Pia Paragliola, Carmen Ercolano, Andrea Strazzulli, Roberta Iacono and Marco Moracci.
- 14th International Congress on Extremophiles, Loutraki, Greece – September 22/26, 2024.
  Presenting author (Oral presentation)
  Title: Some Like It Hot: Uncovering Novel Carbohydrate-Active Enzyme
  from Pisciarelli Hot Spring metagenome
  Authors: Francesca Maria Pia Paragliola, Andrea Strazzulli, Sergio Roberto Scotillo Roberta Iacono and Marco Moracci

6.3. Period spent abroad

From June 2nd, 2023, to December 1st, 2023, under the supervision Professor Nicolas Terrapon, team leader of the research group Glycogenomics group (CAZy group) at AFMB (Architecture et Fonction des Macromolécules Biologiques) of Marseilles, France.

BIBLIOGRAPHY

[1] P. Chandra, Enespa, R. Singh, and P. K. Arora, *Microbial lipases and their industrial applications: A comprehensive review*, vol. 19, no. 1. BioMed Central, 2020.

[2] K.-E. J. A. L. C. Syldatk, *Introduction to enzyme technology in polymer science*, vol. 764. 2024.

[3] P. Chandra, B. Bhimrao, A. Prakash, R. B. Bhimrao, and D. P. Singh, "Isolation of Alkaliphilic Bacterium Citricoccus alkalitolerans CSB1 : An Efficient Biosorbent for Bioremediation of Tannery Waste Water Arsenic adsorption and its mobilization in the soil of arsenic-affected areas View project," *Cell. Mol. Biol.*, no. January, 2016, doi: 10.4172/1165-158X.1000135.

[4] P. Song *et al.*, "Microbial proteases and their applications," *Front. Microbiol.*, vol. 14, no. September, pp. 1–24, 2023, doi: 10.3389/fmicb.2023.1236368.

[5] Y. Kan, L. He, Y. Luo, and R. Bao, "IsPETase Is a Novel Biocatalyst for Poly(ethylene terephthalate) (PET) Hydrolysis," *ChemBioChem*, vol. 22, no. 10, pp. 1706–1716, 2021, doi: 10.1002/cbic.202000767.

[6] R. Iacono, F. De Lise, M. Moracci, B. Cobucci-Ponzano, and A. Strazzulli, "Glycoside hydrolases from (hyper)thermophilic archaea: structure, function, and applications," *Essays Biochem.*, vol. 67, no. 4, pp. 731–751, 2023, doi: 10.1042/EBC20220196.

[7] Z. Zhang *et al.*, "Industrial relevance of thermophilic Archaea," *3 Biotech*, vol. 13, no. 1, pp. 602–618, 2020, doi: 10.1002/bab.1982.

[8] D. I. J. Salas-Bruggink, J. S. S. Martín, G. Leiva, and J. M. Blamey, "Extremozymes: Challenges and opportunities on the road to novel enzymes production," *Process Biochem.*, vol. 143, no. February, pp. 323–336, 2024, doi: 10.1016/j.procbio.2024.04.035.

[9] R. Mehta, P. Singhal, H. Singh, D. Damle, and A. K. Sharma, "Insight into thermophiles and their wide-spectrum applications," *3 Biotech*, vol.

6, no. 1, pp. 1–9, 2016, doi: 10.1007/s13205-016-0368-z.

[10] N. Sources and I. Applications, *Microbial Extremozymes*. 2022.

[11] Naveen Kumar Arora, *Extremozymes and Their Industrial Applications*. 2022.

[12] F. De Lise, R. Iacono, M. Moracci, A. Strazzulli, and B. Cobucci-Ponzano, “Archaea as a Model System for Molecular Biology and Biotechnology,” *Biomolecules*, vol. 13, no. 1, pp. 1–27, 2023, doi: 10.3390/biom13010114.

[13] A. Gutiérrez, J. C. Del Río, and A. T. Martínez, “Microbial and enzymatic control of pitch in the pulp and paper industry,” *Appl. Microbiol. Biotechnol.*, vol. 82, no. 6, pp. 1005–1018, 2009, doi: 10.1007/s00253-009-1905-z.

[14] S. Fusco and B. Siebers, “Extreme Biocatalysts: From Basic Research to Biotechnological Applications,” *Int. J. Mol. Sci.*, vol. 25, no. 12, pp. 10–13, 2024, doi: 10.3390/ijms25126492.

[15] K. Terpe, “Overview of thermostable DNA polymerases for classical PCR applications: From molecular and biochemical fundamentals to commercial systems,” *Appl. Microbiol. Biotechnol.*, vol. 97, no. 24, pp. 10243–10254, 2013, doi: 10.1007/s00253-013-5290-2.

[16] C. Ercolano *et al.*, “Biochemical Characterisation of Sis: A Distinct Thermophilic PETase with Enhanced NanoPET Substrate Hydrolysis and Thermal Stability,” *Int. J. Mol. Sci.*, vol. 25, no. 15, 2024, doi: 10.3390/ijms25158120.

[17] R. Iacono, B. Cobucci-Ponzano, A. Strazzulli, R. Giglio, L. Maurelli, and M. Moracci, “(Hyper)thermophilic biocatalysts for second generation biorefineries,” *Chim. Oggi/Chemistry Today*, vol. 34, no. 5, pp. 34–37, 2016.

[18] D. G. Naumoff *et al.*, “The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics,” *Nucleic Acids Res.*, vol. 14, no. 11, pp. 845–855, 2001, doi: 10.1093/protein/14.11.845.

[19] G. Henrissat, B., & Davies, “Structural and sequence-based

classification of glycoside hydrolases. Current Opinion in Structural Biology," *Curr. Opin. Struct. Biol.*, pp. 297–304, 1997, doi: 10.1109/MCSoC.2016.13.

[20] D. G. Naumoff, "Hierarchical classification of glycoside hydrolases," *Biochem.*, vol. 76, no. 6, pp. 622–635, 2011, doi: 10.1134/S0006297911060022.

[21] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat, "The carbohydrate-active enzymes database (CAZy) in 2013," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 490–495, 2014, doi: 10.1093/nar/gkt1178.

[22] E. Drula, M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat, and N. Terrapon, "The carbohydrate-active enzyme database: functions and literature," *Nucleic Acids Res.*, pp. 1–7, 2021, doi: 10.1093/nar/gkab1045.

[23] B. I. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat, "The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics," *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, pp. 233–238, 2009, doi: 10.1093/nar/gkn663.

[24] D. E. Koshland Jr., "Koshland - original," *Biol. Rev.*, vol. 28, no. February, pp. 416–436, 1953.

[25] B. Cobucci-Ponzano, A. Strazzulli, M. Rossi, and M. Moracci, "Glycosynthases in biocatalysis," *Adv. Synth. Catal.*, vol. 353, no. 13, pp. 2284–2300, 2011, doi: 10.1002/adsc.201100461.

[26] A. White *et al.*, "Recent insights into inhibition, structure, and mechanism of configuration-retaining glycosidases," *Curr. Opin. Struct. Biol.*, vol. 4, no. 5, pp. 750–770, 1997, doi: 10.1021/acscentsci.0c00111.

[27] J. D. Mccarter and S. G. Withers, "McCarter and Withers 1994.pdf," *Curr. Opin. Struct. Biol.*, vol. 4, pp. 885–892, 1994.

[28] N. Curci *et al.*, "Novel GH109 enzymes for bioconversion of group A red blood cells to the universal donor group O," *N. Biotechnol.*, vol. 77,

no. April, pp. 130–138, 2023, doi: 10.1016/j.nbt.2023.08.002.

[29] B. L. Mark, D. J. Vocadlo, S. Knapp, B. L. Triggs-Raine, S. G. Withers, and M. N. G. James, “Crystallographic Evidence for Substrate-assisted Catalysis in a Bacterial β-Hexosaminidase,” *J. Biol. Chem.*, vol. 276, no. 13, pp. 10330–10337, 2001, doi: 10.1074/jbc.M011067200.

[30] James Sherwood, “The significance of biomass in a circular economy,” *Bioresour. Technol.*, 2020.

[31] M. U. Javed, H. Mukhtar, M. T. Hayat, U. Rashid, M. W. Mumtaz, and C. Ngamcharussrivichai, “Sustainable processing of algal biomass for a comprehensive biorefinery,” *J. Biotechnol.*, vol. 352, no. May, pp. 47–58, 2022, doi: 10.1016/j.jbiotec.2022.05.009.

[32] V. P. Santos, N. S. S. Marques, P. C. S. V. Maia, M. A. B. de Lima, L. de O. Franco, and G. M. de Campos-Takaki, “Seafood waste as attractive source of chitin and chitosan production and their applications,” *Int. J. Mol. Sci.*, vol. 21, no. 12, pp. 1–17, 2020, doi: 10.3390/ijms21124290.

[33] J. Ding and Y. Guo, “Recent Advances in Chitosan and its Derivatives in Cancer Treatment,” *Front. Pharmacol.*, vol. 13, no. April, pp. 1–13, 2022, doi: 10.3389/fphar.2022.888740.

[34] M. A. Rahim, F. Saeed, W. Khalid, M. Hussain, and F. M. Anjum, “Functional and nutraceutical properties of fructo-oligosaccharides derivatives: a review,” *Int. J. Food Prop.*, vol. 24, no. 1, pp. 1588–1602, 2021, doi: 10.1080/10942912.2021.1986520.

[35] L. Zhao, *Oligosaccharides of Chitin and Chitosan*. 2019.

[36] M. B. T. D. and U. T. Bornscheuer, “Marine Polysaccharides: Occurrence, Enzymatic Degradation and Utilization.” ChemBioChem, 2021.

[37] E. Ashcroft and J. Munoz-Munoz, “A review of the principles and biotechnological applications of glycoside hydrolases from extreme environments,” *Int. J. Biol. Macromol.*, vol. 259, no. December 2023, 2024, doi: 10.1016/j.ijbiomac.2024.129227.

[38] A. Dawood and K. Ma, “Applications of Microbial β-Mannanases,” *Front. Bioeng. Biotechnol.*, vol. 8, no. December, pp. 1–17, 2020, doi: 10.3389/fbioe.2020.598630.

[39] A. C. Rufer, “Drug discovery for enzymes,” *Drug Discov. Today*, vol. 26, no. 4, pp. 875–886, 2021, doi: 10.1016/j.drudis.2021.01.006.

[40] B. N. Hogg, C. Schnepel, J. D. Finnigan, S. J. Charnock, M. A. Hayes, and N. J. Turner, “The Impact of Metagenomics on Biocatalysis,” *Angew. Chemie - Int. Ed.*, vol. 63, no. 21, 2024, doi: 10.1002/anie.202402316.

[41] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, “Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products,” *Chem. Biol.*, vol. 5, no. 10, 1998, doi: 10.1016/S1074-5521(98)90108-9.

[42] F. Warnecke *et al.*, “Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite,” *Nature*, vol. 450, no. 7169, pp. 560–565, 2007, doi: 10.1038/nature06269.

[43] and R. J. D. 1FAS Benjamin E. Wolfe1,*, Julie E. Button1,*, Marcela Santarelli1,#, “Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity,” *Methods Enzym.*, vol. 538, no. 2, pp. 151–169, 2014, doi: 10.1016/j.cell.2014.05.041.Cheese.

[44] N. A. Be *et al.*, “Whole metagenome profiles of particulates collected from the International Space Station,” *Microbiome*, vol. 5, no. 1, p. 81, 2017, doi: 10.1186/s40168-017-0292-4.

[45] S. L. Robinson, J. Piel, and S. Sunagawa, “A roadmap for metagenomic enzyme discovery,” *Nat. Prod. Rep.*, vol. 38, no. 11, pp. 1994–2023, 2021, doi: 10.1039/d1np00006c.

[46] M. Suleiman, A. Krüger, and G. Antranikian, “Biomass-degrading glycoside hydrolases of archaeal origin,” *Biotechnol. Biofuels*, vol. 13, no. 1, pp. 1–14, 2020, doi: 10.1186/s13068-020-01792-y.

[47] A. Strazzulli *et al.*, “Discovery of hyperstable carbohydrate-active enzymes through metagenomics of extreme environments,” *FEBS J.*,

vol. 287, no. 6, pp. 1116–1137, 2020, doi: 10.1111/febs.15080.

[48] M. Galardini, A. Mengoni, and M. Bazzicalupo, “Produzione e analisi delle sequenze metagenomiche,” *Biologia (Bratisl).*, 2015.

[49] K. N. Lam, J. Cheng, K. Engel, J. D. Neufeld, and T. C. Charles, “Current and future resources for functional metagenomics,” *Front. Microbiol.*, vol. 6, no. OCT, pp. 1–8, 2015, doi: 10.3389/fmicb.2015.01196.

[50] S. Hrvatin and J. Piel, “Rapid isolation of rare clones from highly complex DNA libraries by PCR analysis of liquid gel pools,” *J. Microbiol. Methods*, vol. 68, no. 2, pp. 434–436, 2007, doi: 10.1016/j.mimet.2006.09.009.

[51] J. G. Owen *et al.*, “Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 29, pp. 11797–11802, 2013, doi: 10.1073/pnas.1222159110.

[52] O. G. G. Almeida and E. C. P. De Martinis, “Bioinformatics tools to assess metagenomic data for applied microbiology,” *Appl. Microbiol. Biotechnol.*, vol. 103, no. 1, pp. 69–82, 2019, doi: 10.1007/s00253-018-9464-9.

[53] R. Iacono, A. Strazzulli, R. Giglio, F. Bitetti, B. Cobucci-Ponzano, and M. Moracci, “Valorization of Biomasses from Energy Crops for the Discovery of Novel Thermophilic Glycoside Hydrolases through Metagenomic Analysis,” *Int. J. Mol. Sci.*, vol. 23, no. 18, 2022, doi: 10.3390/ijms231810505.

[54] J. M. Grondin, K. Tamura, G. Déjean, D. W. Abbott, and H. Brumer, “Polysaccharide utilization loci: Fueling microbial communities,” *J. Bacteriol.*, vol. 199, no. 15, pp. 1–15, 2017, doi: 10.1128/JB.00860-16.

[55] N. Terrapon, V. Lombard, H. J. Gilbert, and B. Henrissat, “Automatic prediction of polysaccharide utilization loci in Bacteroidetes species,” *Bioinformatics*, vol. 31, no. 5, pp. 647–655, 2015, doi: 10.1093/bioinformatics/btu716.

[56] G. R. Hemsworth, G. Déjean, G. J. Davies, and H. Brumer, “Learning from microbial strategies for polysaccharide degradation,” *Biochem. Soc. Trans.*, vol. 44, pp. 94–108, 2016, doi: 10.1042/BST20150180.

[57] K. S. Karunatilaka, E. A. Cameron, E. C. Martens, N. M. Koropatkin, and S. Biteen, “Superresolution Imaging Captures Carbohydrate Utilization Dynamics in Human Gut Symbionts,” vol. 5, no. 6, pp. 1–10, 2014, doi: 10.1128/mBio.02172-14.Invited.

[58] N. Terrapon *et al.*, “PULDB: The expanded database of Polysaccharide Utilization Loci,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D677–D683, 2018, doi: 10.1093/nar/gkx1022.

[59] R. L. Cordeiro *et al.*, “Mechanism of high-mannose N-glycan breakdown and metabolism by Bifidobacterium longum,” *Nat. Chem. Biol.*, vol. 19, no. 2, pp. 218–229, 2023, doi: 10.1038/s41589-022-01202-4.

[60] E. Ficko-Blean *et al.*, “Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria,” *Nat. Commun.*, vol. 8, no. 1, 2017, doi: 10.1038/s41467-017-01832-6.

[61] A. Usman, S. Khalid, A. Usman, Z. Hussain, and Y. Wang, *Algal Polysaccharides, Novel Application, and Outlook*. Elsevier Inc., 2017.

[62] M. Bäumgen *et al.*, “A new carbohydrate-active oligosaccharide dehydratase is involved in the degradation of ulvan,” *J. Biol. Chem.*, vol. 279, no. 4, p. 101210, 2021, doi: 10.1016/j.jbc.2021.101210.

[63] L. J. Rothschild and R. L. Mancinelli, “Life in extreme environments,” *Nature*, vol. 409, no. 6823, pp. 1092–1101, 2001, doi: 10.1038/35059215.

[64] L. C. Stewart, V. K. Stucker, M. B. Stott, and C. E. J. de Ronde, “Marine-influenced microbial communities inhabit terrestrial hot springs on a remote island volcano,” *Extremophiles*, vol. 22, no. 4, pp. 687–698, 2018, doi: 10.1007/s00792-018-1029-4.

[65] M. Grégoire, B. N. Moine, S. Y. O’Reilly, J. Y. Cottin, and A. Giret, “Trace element residence and partitioning in mantle xenoliths

metasomatized by highly alkaline, silicate- and carbonate-rich melts (Kerguelen Islands, Indian Ocean)," *J. Petrol.*, vol. 41, no. 4, pp. 477–509, 2000, doi: 10.1093/petrology/41.4.477.

[66] M. Allioux *et al.*, "Metagenome-assembled genomes reveal many novel microbial lineages in the geothermal springs of the subantarctic Kerguelen Islands," pp. 1–15.

[67] L. R. Lynd, P. J. Weimer, W. H. Van Zyl, and S. Isak, "Microbial Cellulose Utilization : Fundamentals and Biotechnology Microbial Cellulose Utilization : Fundamentals and Biotechnology Downloaded from http://mmbr.asm.org/ on February 6 , 2013 by INDIAN INSTITUTE OF TECHNOLOGY MADRAS," *Microbiol. Mol. Biol. Rev.*, vol. 66, no. 3, pp. 506–577, 2002, doi: 10.1128/MMBR.66.3.506.

[68] M. K. Bhat, "Cellulases and related enzymes in biotechnology," *Biotechnol. Adv.*, vol. 18, no. 5, pp. 355–383, 2000, doi: 10.1016/S0734-9750(00)00041-0.

[69] H. Grootaert, L. van Landuyt, P. Hulpiau, and N. Callewaert, "Functional exploration of the GH29 fucosidase family," *Glycobiology*, vol. 30, no. 9, pp. 735–745, 2020, doi: 10.1093/glycob/cwaa023.

[70] D. Shallom and Y. Shoham, "Microbial hemicellulases," *Curr. Opin. Microbiol.*, vol. 6, no. 3, pp. 219–228, 2003, doi: 10.1016/S1369-5274(03)00056-0.

[71] M. J. E. C. Van Der Maarel, B. Van Der Veen, J. C. M. Uitdehaag, H. Leemhuis, and L. Dijkhuizen, "Properties and applications of starch-converting enzymes of the α-amylase family," *J. Biotechnol.*, vol. 94, no. 2, pp. 137–155, 2002, doi: 10.1016/S0168-1656(01)00407-2.

[72] Š. Janeček, "α-Amylase family: Molecular biology and evolution," *Prog. Biophys. Mol. Biol.*, vol. 67, no. 1, pp. 67–97, 1997, doi: 10.1016/S0079-6107(97)00015-1.

[73] A. S. Silchenko, N. K. Rubtsov, A. O. Zueva, M. I. Kusaykin, A. B. Rasin, and S. P. Ermakova, "Fucoidan-active α-L-fucosidases of the GH29 and GH95 families from a fucoidan degrading cluster of the marine

bacterium Wenyingzhuangia fucanilytica," *Arch. Biochem. Biophys.*, vol. 728, no. July, p. 109373, 2022, doi: 10.1016/j.abb.2022.109373.

[74] N. M. A. Ponce and C. A. Stortz, "A Comprehensive and Comparative Analysis of the Fucoidan Compositional Data Across the Phaeophyceae," *Front. Plant Sci.*, vol. 11, no. November, 2020, doi: 10.3389/fpls.2020.556312.

[75] M. T. J. Van Der Meer *et al.*, "Cultivation and genomic, nutritional, and lipid biomarker characterization of Roseiflexus strains closely related to predominant in situ populations inhabiting yellowstone hot spring microbial mats," *J. Bacteriol.*, vol. 192, no. 12, pp. 3033–3042, 2010, doi: 10.1128/JB.01610-09.

[76] S. W. Wingett and S. Andrews, "Fastq screen: A tool for multi-genome mapping and quality control [version 1; referees: 3 approved, 1 approved with reservations]," *F1000Research*, vol. 7, no. 0, pp. 1–13, 2018, doi: 10.12688/f1000research.15931.1.

[77] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "MetaSPAdes: A new versatile metagenomic assembler," *Genome Res.*, vol. 27, no. 5, pp. 824–834, 2017, doi: 10.1101/gr.213959.116.

[78] D. Hyatt, G. L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: Prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, 2010, doi: 10.1186/1471-2105-11-119.

[79] K. Mendler, H. Chen, D. H. Parks, B. Lobb, L. A. Hug, and A. C. Doxey, "Annotree: Visualization and exploration of a functionally annotated microbial tree of life," *Nucleic Acids Res.*, vol. 47, no. 9, pp. 4442–4448, 2019, doi: 10.1093/nar/gkz246.

[80] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Res.*, vol. 17, no. 3, pp. 377–386, 2007, doi: 10.1101/gr.5969107.

[81] T. Thomas, J. Gilbert, and F. Meyer, "Metagenomics - a guide from sampling to data analysis," *Microb. Inform. Exp.*, vol. 2, no. 1, p. 3,

2012, doi: 10.1186/2042-5783-2-3.

[82] S. T. Chang, K. N. Parker, M. W. Bauer, and R. M. Kelly, “A-Glucosidase From Pyrococcus Furiosus,” *Methods Enzymol.*, vol. 330, pp. 260–269, 2001, doi: 10.1016/S0076-6879(01)30381-6.

[83] D. A. Comfort, C. J. Chou, S. B. Conners, A. L. VanFossen, and R. M. Kelly, “Functional-genomics-based identification and characterization of open reading frames encoding α-glucoside-processing enzymes in the hyperthermophilic archaeon Pyrococcus furiosus,” *Appl. Environ. Microbiol.*, vol. 74, no. 4, pp. 1281–1283, 2008, doi: 10.1128/AEM.01920-07.

[84] D. E. JAMES U. BOWIE, ROLAND LtCY and The, “A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Stucture,” vol. 253, no. 58, 1991.

[85] Lüthy et al., “Lüthy R Bowie J U & Eisenberg D (1992) Assessment of protein models with three-dimensional profiles,” *Nat. Publ.*, vol. 356, no. March, pp. 83–85, 1992.

[86] C. Colovos and T. O. Yeates, “Verification of protein structures: Patterns of nonbonded atomic interactions,” *Protein Sci.*, vol. 2, no. 9, pp. 1511–1519, 1993, doi: 10.1002/pro.5560020916.

[87] N. V. Valeyev, A. K. Downing, J. Sondek, and C. Deane, “Electrostatic and functional analysis of the seven-bladed WD β-propellers,” *Evol. Bioinforma.*, vol. 2008, no. 4, pp. 203–216, 2008.

[88] A. Nath and W. M. Atkins, “A theoretical validation of the substrate depletion approach to determining kinetic parameters,” *Drug Metab. Dispos.*, vol. 34, no. 9, pp. 1433–1435, 2006, doi: 10.1124/dmd.106.010777.

[89] C. Suresh, M. Kitaoka, and K. Hayashi, “A thermostable non-xylanolytic α-glucuronidase of Thermotoga maritima MSB8,” *Biosci. Biotechnol. Biochem.*, vol. 67, no. 11, pp. 2359–2364, 2003, doi: 10.1271/bbb.67.2359.

[90] P. Ruile, C. Winterhalter, and W. Liebl, “Isolation and analysis of a gene

encoding α-glucuronidase, an enzyme with a novel primary structure involved in the breakdown of xylan," *Mol. Microbiol.*, vol. 23, no. 2, pp. 267–279, 1997, doi: 10.1046/j.1365-2958.1997.2011568.x.

[91] B. Cobucci-Ponzano *et al.*, "Novel thermophilic hemicellulases for the conversion of lignocellulose for second generation biorefineries," *Enzyme Microb. Technol.*, vol. 78, pp. 63–73, 2015, doi: 10.1016/j.enzmictec.2015.06.014.

[92] H. V. Scheller and P. Ulvskov, "Hemicelluloses," *Annu. Rev. Plant Biol.*, vol. 61, pp. 263–289, 2010, doi: 10.1146/annurev-arplant-042809-112315.

[93] K. Tamura, G. Stecher, and S. Kumar, "MEGA11: Molecular Evolutionary Genetics Analysis Version 11," *Mol. Biol. Evol.*, vol. 38, no. 7, pp. 3022–3027, 2021, doi: 10.1093/molbev/msab120.

[94] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix," *Mol. Biol. Evol.*, vol. 25, no. 7, pp. 1307–1320, 2008, doi: 10.1093/molbev/msn067.

[95] M. SMOGYI, "Notes on sugar determination.," *J. Biol. Chem.*, vol. 195, no. 1, pp. 19–23, 1952, doi: 10.1016/s0021-9258(19)50870-5.

[96] and J. I. G. Eric C. Martens, Herbert C. Chiang, "Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont," *Bone*, vol. 23, no. 1, pp. 1–7, 2008, doi: 10.1016/j.chom.2008.09.007.Mucosal.

[97] W. Helbert *et al.*, "Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 13, pp. 6063–6068, 2019, doi: 10.1073/pnas.1815791116.

[98] R. Campos-Vega, G. Loarca-Piña, H. A. Vergara-Castañeda, and B. Dave Oomah, "Spent coffee grounds: A review on current research and future prospects," *Trends Food Sci. Technol.*, vol. 45, no. 1, pp. 24–36, 2015, doi: 10.1016/j.tifs.2015.04.012.

[99] Y. F. Liao, A. Lal, and K. W. Moremen, "Cloning, expression,

purification, and characterization of the human broad specificity lysosomal acid α-mannosidase," *J. Biol. Chem.*, vol. 271, no. 45, pp. 28348–28358, 1996, doi: 10.1074/jbc.271.45.28348.

[100] C. Zardecki, S. Dutta, D. S. Goodsell, R. Lowe, M. Voigt, and S. K. Burley, "PDB-101: Educational resources supporting molecular explorations through biology and medicine," *Protein Sci.*, vol. 31, no. 1, pp. 129–140, 2022, doi: 10.1002/pro.4200.

[101] F. Gabler *et al.*, "Protein Sequence Analysis Using the MPI Bioinformatics Toolkit," *Curr. Protoc. Bioinforma.*, vol. 72, no. 1, pp. 1–30, 2020, doi: 10.1002/cpbi.108.

[102] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Sci.*, vol. 27, no. 1, pp. 135–145, 2018, doi: 10.1002/pro.3290.

[103] Y. Zhu *et al.*, "Mechanistic insights into a Ca2+-dependent family of αmannosidases in a human gut symbiont," vol. 6, no. 2, pp. 125–132, 2010, doi: 10.1038/nchembio.278.Mechanistic.

[104] G. S. Anisha, "Molecular advances in microbial α-galactosidases: challenges and prospects," *World J. Microbiol. Biotechnol.*, vol. 38, no. 9, pp. 1–16, 2022, doi: 10.1007/s11274-022-03340-2.

[105] M. E. Álvarez-Cao, A. Rico-Díaz, M. E. Cerdán, M. Becerra, and M. I. González-Siso, "Valuation of agro-industrial wastes as substrates for heterologous production of α-galactosidase," *Microb. Cell Fact.*, vol. 17, no. 1, pp. 1–13, 2018, doi: 10.1186/s12934-018-0988-6.

[106] G. S. Anisha, "Microbial α-galactosidases: Efficient biocatalysts for bioprocess technology," *Bioresour. Technol.*, vol. 344, no. PB, p. 126293, 2022, doi: 10.1016/j.biortech.2021.126293.

[107] Y. Fei, W. J. Jiao, Y. Wang, J. Liang, G. Liu, and L. Li, "Cloning and expression of a novel α-galactosidase from Lactobacillus amylolyticus L6 with hydrolytic and transgalactosyl properties," *PLoS One*, vol. 15, no. 7 July, pp. 1–12, 2020, doi: 10.1371/journal.pone.0235687.

[108] V. Roig-Zamboni *et al.*, "Structure of human lysosomal acid α-

glucosidase-A guide for the treatment of Pompe disease," *Nat. Commun.*, vol. 8, no. 1, 2017, doi: 10.1038/s41467-017-01263-3.

[109] M. Frison *et al.*, "The Arabidopsis thaliana trehalase is a plasma membrane-bound enzyme with extracellular activity," *FEBS Lett.*, vol. 581, no. 21, pp. 4010–4016, 2007, doi: 10.1016/j.febslet.2007.07.036.

[110] B. Cobucci-Ponzano *et al.*, "A new archaeal β-glycosidase from Sulfolobus solfataricus: Seeding a novel retaining β-glycan-specific glycoside hydrolase family along with the human non-lysosomal glucosylceramidase GBA," *J. Biol. Chem.*, vol. 285, no. 27, pp. 20691–20703, 2010, doi: 10.1074/jbc.M109.086470.

[111] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.

[112] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," *Nat. Methods*, vol. 19, no. 6, pp. 679–682, 2022, doi: 10.1038/s41592-022-01488-1.