Universidade do Minho
Escola de Engenharia

José Pedro Silva Freitas

Mining metagenomics datasets for novel plastic-degrading enzymes

outubro de 2022

Mining metagenomics datasets for novel plastic-degrading enzymes

José Pedro Silva Freitas

UMinho | 2022

Universidade do Minho
Escola de Engenharia

José Pedro Silva Freitas

Mining metagenomics datasets for novel plastic-degrading enzymes

Dissertação de mestrado

Mestrado em Bioinformática

Trabalho efetuado sob a orientação do(a)
Professor Doutor Miguel Rocha
Doutora Andreia Salvador

outubro de 2022

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

AGRADECIMENTOS

Gostaria de deixar os meus mais sentidos agradecimentos a todos aqueles que, durante o meu percurso académico me ajudaram, seja de que forma for, a ultrapassar os momentos mais difíceis, e que também estiveram comigo nos melhores.

Assim, gostava de começar por agradecer diretamente à Dr. Andreia Salvador, pela experiência e conhecimentos que me transmitiu durante a execução desta tese, por dizer aquilo que tem de ser dito para o meu melhor, ajudando-me no presente, mas também me preparando para o futuro e por permitir uma relação mais informal que sem dúvida ajudou a tornar este trabalho mais divertido.

Agradecer ao Professor Miguel, pelo suporte prestado durante estes dois anos a as linhas a seguir para concluir esta tese.

Agradecer ao João Sequeira, que me aturou com a sua boa disposição e me ajudou a progredir no desenvolvimento desta tese.

Ao Diogo e ao André, que me acompanharam desde o primeiro dia de entrada no ensino superior, e depois de muitas batalhas superadas, continuam ao meu lado.

Aos meus melhores amigos, Nuno, Inês, Tiago e Daniela, são únicas pessoas das quais posso dizer que confio a 100%, e que sei que sempre estarão para mim da mesma forma que eu para eles. As experiências partilhadas, ideias trocadas, os conselhos, as palavras de força, entre muitas outras coisas, são impagáveis. A vós vos agradeço por terem aparecido naquele momento que mais precisei, por me terem dado o prazer de ser vosso amigo e me terem acompanhado até agora. Obrigado.

Ao meu pai e mãe, que estiveram comigo (literalmente) todos os dias, me suportaram, me apoiaram naquilo que podiam e não podiam, me obrigavam a levantar quando tinha sono, fazer o TPC quando não tinha vontade, ralharam quando fazia asneiras, obrigavam a sair e aproveitar a vida enquanto era tempo e sempre aceitaram as minhas decisões. Obrigado por me terem ajudado a crescer e me tornar na pessoa que sou hoje.

À memória da minha madrinha e meu avô, serão para sempre lembrados.

"Empty your mind, be formless. Shapeless, like water. If you put water into a cup, it becomes the cup. You put water into a bottle and it becomes the bottle. You put it in a teapot, it becomes the teapot. Now, water can flow or it can crash. Be water, my friend." – Bruce Lee

RESUMO

A crescente quantidade de dados depositados em bases de dados públicas sem anotação pode ocultar uma série de genes e proteínas cuja função ainda é desconhecida. Com base no conhecimento de algumas enzimas capazes de catalisar reações com interesse ambiental ou biotecnológico, será possível encontrar em bases de dados de proteínas ou em conjuntos de dados ómicos, outras com atividade semelhante, que eventualmente poderão ser mais eficientes. No entanto, não existem ferramentas bioinformáticas projetadas para encontrar proteínas de interesse em grandes conjuntos de dados.

Neste trabalho, uma ferramenta de bioinformática foi desenvolvida e denominada Mining Protein dAtasets foR Targeted enzYmes (M-PARTY) para minerar enzimas alvo em grandes conjuntos de dados. M-PARTY recebe um ficheiro FASTA contendo as enzimas alvo e automaticamente produz bases de dados de Hidden Markov Model, valida e filtra os modelos não validados. M-PARTY procura sequências homólogas em determinados conjuntos de dados e identifica as proteínas mais semelhantes, que apresentam potencialmente as mesmas atividades das enzimas alvo. A M-PARTY é uma Interface de Linha de Comando de uso gratuito, corre no sistema operacional Linux com apenas um comando, é de código aberto e foi desenvolvida em Python.

Esta ferramenta foi testada para encontrar enzimas envolvidas na biodegradação do polietileno em metagenomas hidrotermais e marinhos. A partir de 5 sequências proteicas iniciais, 329 HMMs foram gerados pelo M-PARTY e 103 foram descartados após a etapa de validação. Um total de 19 proteínas apresentaram homologia significativa com as 5 enzimas alvo, sendo enzimas potencialmente degradadoras de polietileno.

Esta ferramenta será muito útil para realizar uma primeira triagem de enzimas de interesse em diferentes ambientes, antecedendo uma posterior confirmação da atividade enzimática e eventual implementação.

Palavras-chave: biodegradação de plásticos; ferramenta bioinformática; mineração de dados ómicos; construção de Hidden Markov Models.

ABSTRACT

There is an increasing amount of data deposited in public databases that is poorly annotated and may hide a number of genes and proteins whose function is yet unknown. By knowing some enzymes that are capable to catalyze reactions with environmental or biotechnological interest, it would be possible to find other enzymes in databases or in omics datasets with similar activity, and which could be even more efficient. However, there are no bioinformatics tools designed to find proteins of interest in large datasets, such as those from metagenomics experiments.

In this work, a bioinformatics tool was developed, named Mining Protein dAtasets foR Target enzYmes (M-PARTY), for mining target enzymes in big datasets. M-PARTY receives a FASTA file containing the target enzymes, and automatically produces Hidden Markov Model databases, validating, and filtering the non-validated models. M-PARTY searches for homolog sequences in given datasets and identifies the most similar proteins, which present potentially the same activities of the target enzymes. M-PARTY is a free-to-use Command-Line Interface, runs on Linux operating system with only a command, is open-source, and was developed in Python.

This tool was tested to find enzymes involved in polyethylene biodegradation in hydrothermal and marine metagenomes. From 5 initial protein sequences, 329 HMMs were generated by M-PARTY, and 103 were discarded after the validation step. A total of 19 proteins showed significant homology to the 5 target enzymes, being potentially polyethylene-degrading enzymes.

This tool will be especially useful for performing a first screening of enzymes of interest in different environments, preceding further enzymatic activity confirmation and eventual implementation on biotechnological processes.

Keywords: plastic biodegradation; bioinformatics tool; omics data mining; Hidden Markov Models construction.

Index

List of Tables

List of Figures

ACRONYMS / ABBREVIATIONS

CLI – Command-Line Interface

CI – Continuous Integration

HMM – Hidden Markov Model

GHG – Greenhouse Gas

PET – Polyethylene Terephthalate

PS – Polystyrene

PP – Polypropylene

PE – Polyethylene

PUR – Polyurethane

PHBV – Poly(3-hydroxybutyrate-co-3-hydroxyvalerate)

PCL – Polycaprolactone

PLA – Polylactic acid

PGA – Poly(glycolic acid)

PHB – Polyhydroxy butyrate

PBAT – Polybutylene adipate terephthalate

MHET – Mono(2-hydroxyethyl) terephthalate

TPA – Terephthalic acid

PCA – Protocatechuic acid

PDB – Protein Data Bank

CDD – Conserved Domain Database

BLAST – Basic Local Alignment Search Tool

MNP – Multilayer Protein Networks

GNN – Genome Neighbourhood Networks

SSN – Sequence Similarity Networks

NR – Non-redundant protein sequences

UPIMAPI – UniProt Id Mapping through API

KEGG – Kyoto Encyclopedia of Genes and Genomes

CD-HIT – Cluster Database at High Identity with Tolerance

T-COFFEE – Tree-based Consistency Objective Function for alignment Evaluation

M-PARTY – Mining Protein dAtasets foR Target enzYmes

Chapter 1.

Introduction

1.1. Context/Motivation

Pollution is a widespread human-based problem, which has been increasing exponentially in the last few years, following economic growth [1], local development [2] and national gross income [3]. This trend is expected to continue as long as distinct and drastic measures are not taken by the richest nations in a global effort to slow and revert an obvious outcome. Directly linked to pollution, we are also consuming dramatically more resources than what is actually needed [4] and therefore affecting, not only humans but all living beings.

Plastic production increased significantly in the last decades, and its discovery represented a great breakthrough for the world economy and trade market around 1950 [3], with great advantages and improvements in general quality of life, gaining popularity in sectors like packing, food, and water conservation, health, transportation, textiles [3], and energy-saving, as plastic is lighter and easier to manufacture than other polymers used at the time [5], [6]. Nevertheless, economic growth rates demanded higher production of these compounds [1], and a culture shift from reusable to single-use packing, was noticeable [7], [8]. Despite the existence of biodegradable plastics, a bigger problem arises when most utilized types are not biodegradable [3], [8] and countries cannot find solutions to settle with ever-growing quantities of generated plastic [7]. The most efficient way to radically erase discarded plastic compounds pass through thermal incineration or recycling, but predictions show insufficient efforts to reduce the global quantities of plastic debris in nature [9], regardless of an also continuous increase of thermal and recycling treatments [3].

Bioinformatics development in the past years has helped the scientific community to fasten large time-consuming processes with tools especially capable of handling large datasets, from sequence alignments and structure prediction to motifs identification, among others.

Taking this into account, a bioinformatics tool able to predict plastic degrading genes or enzymes within metagenomic samples could help to exponentially accelerate the discovery of novel enzymes, and consequently, help to develop new methods to fight this problem through biodegradation. In this work, a novel bioinformatics tool for detecting potential plastics degrading enzymes in metagenomics' derived protein datasets will be developed. This will contribute to discover naturally occurring enzymes similar to known plastic degrading enzymes and possibly more efficient, which can then be tested in the laboratory to confirm their activity against plastic waste. These potentially novel enzymes could then be applied in the treatment of industrial wastewaters, dumping sites groundwater, or even domestic wastes containing plastics [10].

The urgency to find new ways to fight plastic pollution and accumulation, bonded to the tremendous amount of non-reviewed metagenomic data, obtained from different parts of the globe, opens new possibilities to identify novel microorganisms and enzymes as biocatalysts for plastics biodegradation.

This work intends to take a step forward in the combat against plastic pollution, by helping the scientific community targeting a large time consuming, and complex step as the discovery and research of novel and more efficient enzymes produced by microorganisms. It is observable a current lack in available tools for this specific subject, so the prediction of protein function through sequence similarity and homology in metagenomic data is an imperative task. Currently, there are no available tools, online or modules, capable predict plastic degrading enzymes in omics datasets.

1.2. Objectives

The main objective of this thesis is to deploy a fully operational bioinformatics tool for identifying, in protein datasets, homologies between the enzymes in the dataset and groups of target proteins . This will be done by using methods like structural and homology annotations based on Hidden Markov Models (HMMs). To test the tool, the target enzymes were those involved in polyethylene (PE)) degradation, a synthetic plastic highly abundant in plastic waste.

1.3. Thesis structure

The context, motivation and aim of this thesis will be given in Chapter 1. In Chapter 2, the state of the art, will be reviewed by covering relevant knowledge regarding plastics biodegradation, plastic degrading enzymes, protein annotation methods, HMM and their validation and benchmarking procedures, and a review of available tools with similar goals. In Chapter 3 the methodologies utilized in this thesis will be described, and final tools, software, frameworks, pipelines, packages used, and testing methods will be provided and thoroughly explained step by step. In Chapter 4 the results will be presented, including the tool validation and testing. The discussion will be presented in Chapter 5. The main conclusions and future perspectives will be shown in Chapter 6.

Chapter 2.

State-of-the-art

2.1. Plastics characteristics, pollution, and biodegradation considerations

Plastic is the name given to a material that, in some stage of production, can flow, be applied as a coating, extruded, or moulded as liking [5], [11]. Also, distinguished proprieties of the finished material go through a high range of temperature, mechanical, chemical, and light-resistant [5], [6]. Plastics are synthetic hydrocarbon long polymer chains of the same molecule – or monomer – with very high molecular weight, hydrophobic [5]. Most of the everyday used plastics, from domestic to industrial, branch from fossil fuels, such as oil, coal, or natural gas, being so generally called non-biodegradable [12]. On the other hand, bio-based plastics, as the name suggests, are made from renewable sources (parts of plants, animals, algae, etc) [12], and are commonly degradable when in contact with both abiotic and biotic factors [13]. Additionally, not all bio-based plastics are biodegradable, because of the cases of bio-PET, bio-PE, synthetised by living beings from renewable sources, but keeping chemical characteristics from their fossil fuel analogues [13]. Biodegradable plastics are the only type of polymer that poses no risk with environmental disposure since can be fully digested and assimilated by microorganisms [14]. Poly(3-hydroxybutyrate-co-3-hydroxyvalerate) (PHBV), Polycaprolactone (PCL) [15], Polylactic acid (PLA), and Poly(glycolic acid) (PGA) [9], are some examples of proven biodegradable synthetic polymers used in human applications. However, biodegradable plastics do not have the same mechanical and thermal resistance as conventional synthetic plastics, dismantling when interacting with e.g., water and enzymes [12], and so do not experience the same use as the latter, as non-biodegradable plastics are much more durable, permeable, and mechanically resistant [5].

Nowadays, pollution has expanded to all earth ecosystems [1], [16], ranging from soil, air, groundwater, oceans, and others, each one with its challenges, but all in connection with one another. A greater part of ocean pollution is originated from ground dumping [2]. Different objects, materials, and compounds from domestic trash to industrial residues, can be found anywhere across the earth's surface. However, one item can be highlighted: plastic. Plastic pollution is universal [7], and it is present in large concentrations in dumping sites [17], soils, oceans [5], lakes, groundwaters, and even in the atmosphere and animal organs [7]. Other studies have revealed trace elements of microplastics in households' air [18], seafood [2], and has made its way to human placentas and breastmilk [19]. Prior facts arise concerns about public health, as plastic compounds are toxic and not metabolized by human enzymes [20]. Upstream of the disposal/recycling process, plastics also cause environmental distress when in production. Being a petroleum-based material, fossil fuel burn results in high outputs of GHG (greenhouse gas) mostly CO_2 emissions into the atmosphere, increasing even more the existing problems related to the greenhouse effect, and contributing to an increase of the earth's mean temperature and so a greater number of climate change phenomena [21].

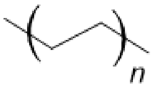
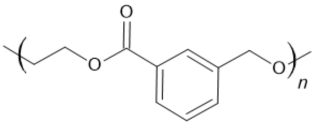
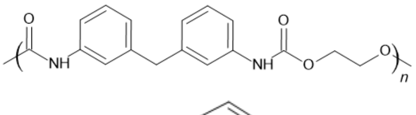
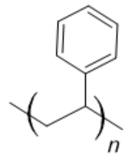
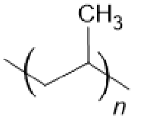
Generating strategies for plastic pollution reduction is complex and depends on a global understanding of this problem and efforts from main pollutant nations [7]. This question can be addressed from distinct perspectives, starting from reducing overall plastic quantities, plastic substitution by biodegradable materials, reducing plastic demands, and focusing on the plastic post-consumption, like upgrading the collecting sector, increase recycling capacities in towns, and alleviating or treating environmental wastes [7]. However, estimates say that even with the best scenarios, until 2040, plastic disposal will still represent a major issue, and environmental footprints noticeable due to plastics' long extensive degradation periods [7].

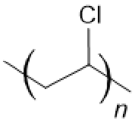
This problem's urgency demand high commitment from all counterparts involved, and so are new methods capable of countering this growth tendency [7], [22], since countries cannot store such huge amounts of plastic, nor dispose of them in a safe matter without their accumulation in undesired places [12].

Diverse strategies can be practised reverting plastic accumulation in the future, yet sustainable solutions are preferable, avoiding the creation of different/unrelated problems. Natural polymers like the ones referenced before, introduce mechanical and cost-related disadvantages, and so, research and development of novel polymers able of controlled biological degradation for replacing traditional fossil fuel-based plastics posts an interesting topic [22], [23], [24]. Experiments have been conducted with the combination of plastics from different monomers, showing promising results in biodegradability levels [25], [26].

First-ever records of plastics synthesis go back to the early 1900s, but only halfway through the century gained severe popularity due to their unique functionalities [2], [21]. Records show an early scientific advance in the synthesis of new monomers in the first half of the latter century [6]. Over the years, the production of plastic exponentially raised to become indispensable and essential to human everyday lifestyle [3]. Despite the existence of countless sustainable and environment-friendly biodegradable plastics, their use deprecated over the years, giving place to petroleum-derived plastics [27], owing to facilitate processes, higher volumes, and cheaper production costs [6]. **Table 1** shows the characteristics and applications of the most commonly used plastics.

Table 1: List of most used non-biodegradable plastics, showing the chemical formula, monomer chemical 2D structure (drawn with ACD/ChemSketch [28]), their main applications, and potential hazards of their uncontrolled disposure on the environment and consequent human exposure [29].

Plastics	Formula	Structure	Field of use	Hazards
Polyethylene (PE)	$(C2H4)_n$		Packing, fuel tanks	Toxic
Polyethylene terephthalate (PET)	$(C10H8O4)_n$		Food and liquids packing	Irritant
Polyurethane (PUR)	$(C17H16N2O4)_n$		Furniture, Electronics,	Irritant, Toxic
Polystyrene (PS)	$(C8H8)_n$		Food packing, auto parts, toys	Toxic, Carcinogenic
Polypropylene (PP)	$(C3H6)_n$		First-aid, machinery	Irritant

Polyvinyl chloride (PVC)	$(C2H3Cl)_n$		Building, health care, packing	Highly toxic, Irritant, Teratogenic

Analysis of chemical structures from **Table 1** tells us a dominant prevalence of carbon atoms as their backbone composition, with carbon-carbon, and carbon-hydrogen bonds in PE, Polystyrene (PS), and Polypropylene (PP), unlike the remaining polymers (Polyethylene Terephthalate (PET)), Polyurethane (PUR), Polyvinyl chloride (PVC)) with heteroatoms O, N-C-O connections, and Cl respectively [24]. Naturally, the structure has a proven impact on both polymer properties and biodegradation [13], as higher ratios of aromatic constituents result in fewer options for enzymatic catalysis [30] whereas plastics with esters or amide bonds are much more likely to suffer from hydrolytic attacks [24], [31]. Adding to this, biotic factors also do not pose a major threat to non-biodegradable plastics' integrity [23], once again because of the above characteristics.

2.2. Known plastics-degrading microorganisms and enzymes

As a consequence of the accumulation of plastic materials in soils, oceans, and landfill sites [14], a natural microbial adaptation of microorganisms living in these environments occurs, which is reflected in their ability to partially biodegrade and catalyze synthetically non-biodegradable and biodegradable plastics as their carbon and energy sources in a bioremediation or biodegradation process [12], [32], [33]. Biodegradation can be defined as the event associated with the mechanisms of degradation and assimilation [12] or their secretion products through the action of enzymes coming from living organisms [9], [23].

Ever-growing plastic quantities build up in soils [34] and waters [5], [35]–[37], changing microorganisms' genotype, to a form passive of digesting these compounds, and so nourishing upon hydrocarbon chains [32]. Different studies performed in distinct parts of the globe, in places highly affected by mismanaged wastes, are already showcasing levels of mass reduction of multiple kinds of plastics by part of naturally altered organisms [38], trying to isolate and characterize the resident microorganisms, as well as documenting the digesting process [39].

Metagenomics refers to the process of sampling all genome sequences available from a community of organisms in an ecosystem, instead of genomics, which looks through hard sequencing of a single isolated organism sequence [40]. Metagenomic research and surveillance reveals a much higher effectiveness when compared to traditional isolation methods, both in concerns of time, associated costs, and scalability, as most bacteria are not reproducible in a laboratory context, due to high cultivation conditions requirements [41]. Research studies mainly focus on traditional laboratory methods, posing a barrier to the still-unknown microbial world and its yet not known potential against plastic pollution. This way, metagenomic analysis enables a fast DNA assembly from a wide range of microorganisms in an environmental sample like soil, water, wastewater, etc [17], [42]. As microorganisms are ubiquitous in all ecosystems [42] metagenome analysis is feasible wherever a sample is collectable and makes possible the reconstruction of a community's whole genome, from prokaryotes, eukaryotes, and even viruses [40]. Finally, metagenomics contributes to a deep understanding of microbial communities' structure (strains predominance) and metabolic abilities [40], by providing next-generation sequencing tools able to identify enzymes and catalytic pathways, without the need for microbial cultivation and isolation [42].

Several studies show similar methodologies for microorganisms/enzymes research and isolation, and parallel notice in plastic degradation [43]. Examples of such include the collection of plastic samples appreciably deteriorated under aerobic conditions and posterior strain isolation from respective cultures [44], plastic contaminated soil and wastewaters samples for Polyethylene Terephthalate film degradation screening [30], local dumpsite soil and compost samples metagenomic analysis and sequence assembly [17], polymer weight loss by specific microbial strains biodegradation [45], among others [43].

Knowledge on the enzymes acting on plastics was very scarce before 2000. For example, regarding PET degrading enzymes and pathways, information became only available during the last years [24], (since the discovery of *Fusarium oxysporum* and *F. solani* [30] activity in 2007-2008), and for the remaining polymers, information is not yet noticeably clear and complete [23].

Because plastics are enormous molecules chaining monomers together, basic cleavage reactions are expected to occur extracellularly by enzymatic attack [44]. **Figure *1*** shows the polymeric chemical structure of PET and PUR, containing hydrocarbon chains and heteroatoms hydrocarbon monomers, respectively.

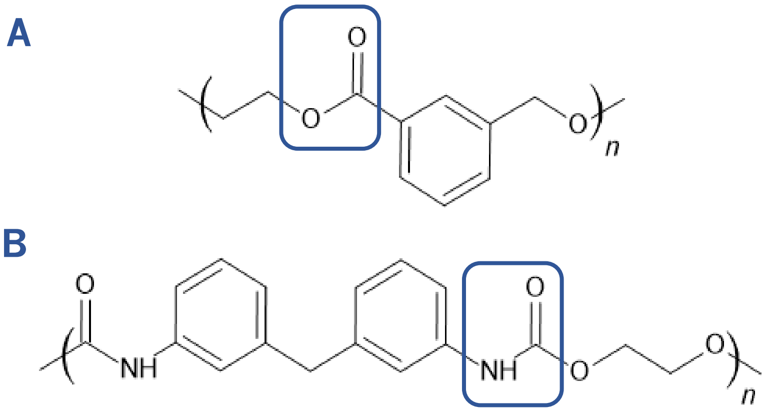


Figure 1: Chemical structure representation of (A) PET and (B) PUR, emphasizing ester and amide bonds, respectively (adapted from [23])

Monomers with atoms like O and N (e.g., PET/PUR in **Figure 1**) usually account for ester or amide bonds, that are susceptible to hydrolytic attacks, either with or without enzymatic mediation [24] in specific conditions, as shown in **Figure 2.** It is also worth noting that, depolymerization of non-biodegradable plastics is a slow process in environmental conditions, due to their inert nature [23], and the availability of groups passive of hydrolysis [12], which can be tricky due to plastics' crystalline structure, where different chains of linked monomers are held together by Van der Waals and H bridges, making these compounds extremely hydrophobic [31].

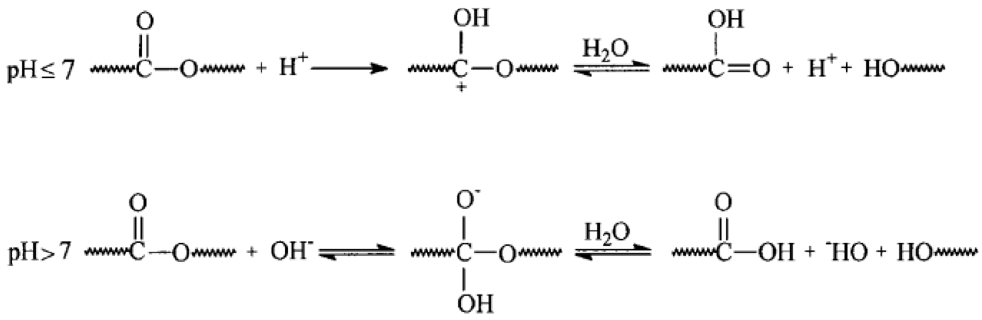


Figure 2: Hydrolysis of an ester bond in acidic and alkaline conditions (adapted from [23]).

It is biochemically impossible for long polymer chains to get assimilated intracellularly by microorganisms, and so, while in this inert state, microbes rely on the production of extracellular enzymes, that are secreted to the environment and function as depolymerizers [12], [13]. As this process occurs, smaller and lighter molecules emerge and are absorbed through microbes' cellular wall/membrane into the cytoplasm, to be further degraded [46].

Degradation mediated by microbes can be divided into two categories: aerobic or anaerobic biodegradation, for the presence and absence of oxygen, respectively (**Figure *3***) [12].

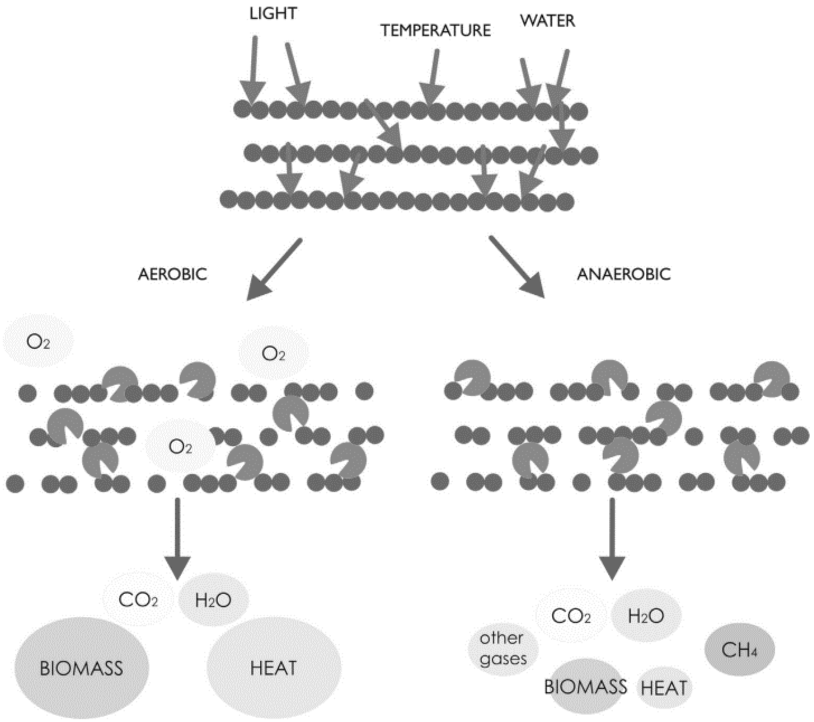


Figure 3: Illustration of both types of biodegradation, aerobic and anaerobic. Adapted from [33].

While in aerobic conditions, microorganisms make use of polymer chains as carbon sources and uptake oxygen as electron acceptors to later form CO_2, H_2O, heat, and biomass [12], [13]. Similarly, without oxygen, anaerobic degradation uses CO_2 as an electron acceptor and produces methane, H_2O, CO_2, heat, and biomass [12], [13]. Other alternative electron acceptors can also be used as for example, sulphate, iron, and nitrate.

Being PET highly popular and consequently, experiencing high indexes of utilization and environmental accumulation [3], knowledge on possible biodegrading pathways in PET-degrading bacterial have been acquired, and PET degradation bacterial metabolism is not a secret anymore. According to S. Yoshida *et al.* (2016) [30], the bacteria *Ideonella sakaiensis* 201-F6 strain can use and survive using PET carbon backbone as its main carbon source for energy conversion. This process is accomplished by the bacterial synthesis of two related and dependent enzymes – PETase and MHETase – both essential for reactions to proceed. In the first step, PETase breaks down PET polymers by hydrolytic cleavage to form Mono(2-hydroxyethyl) terephthalate (MHET) and terephthalic acid (TPA) in different quantities. MHET way outweighs TPA as a result of PETase catalysis. Then, MHETase helps with new hydrolysis from MHET to final TPA (**Figure 4**).

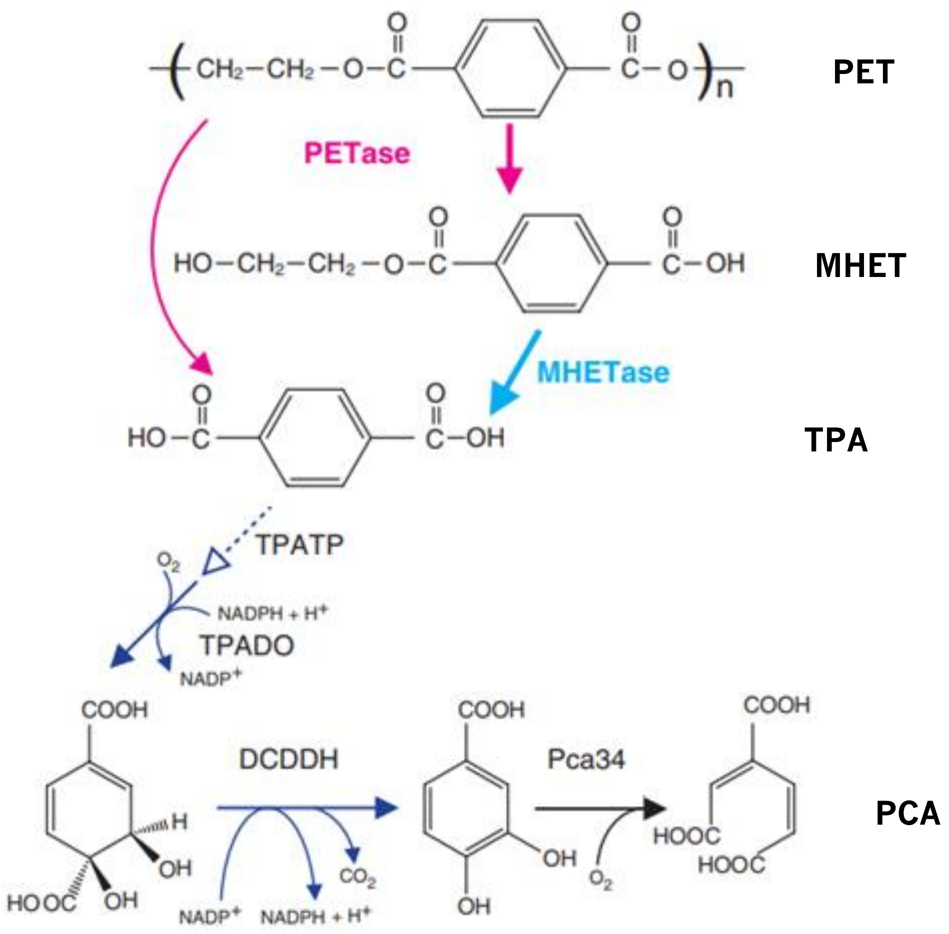


Figure 4: Schematical representation of hydrolytic cleavage of PET polymer, through PETase and MHETase, resulting in MHET and TPA, and TPA, respectively. TPA is transported with TPA transporter (TPATP) and oxygenated by TPA 1,2-dioxygenase (TPADO), which product is decarboxylated by 1,2-dihydroxy-3,5-cyclohexadiene-1,4-dicarboxylate dehydrogenase (DCDDH). Another oxygenation is followed by means of PCA 3,4-dioxygenasedegradation pathway. Adapted from [30].

After this, TPA degradation takes place, facing a series of catabolic reactions of oxygenase and decarboxylation, resulting in the final and low-weight acyclic compound protocatechuic acid (PCA).

Other authors propose a different pathway for the prior steps, this time with an IsPETase enzyme, also depolymerizing PET into MHET, but also referring to an intermediate reaction, first producing Bis(2-hydroxyethyl) terephthalate (BHET) and only after converting to MHET [31]. In this study, additional information is provided concerning the catalysis location, starting in an extracellular environment, followed by absorption and further reactions [31].

Degradation pathways of other synthetic plastics have also been studied, from PUR to other more recalcitrant like PS and PE, with rigid carbon-carbon backbone structures. In the case of PUR, some microorganisms show the ability to degrade this polymer (as shown in **Table *2***). This biodegradation process can be explained by the wide variety of polyurethanes structures, and depending on the monomer linkage type, PUR can be depolymerized through an esterase/protease mechanism but is very dependent on the enzyme-ligand proximity [31], but also by hydrolases [46].

For the case of linear full carbon body PE, biodegradation is a challenge for microorganisms, but it is known to occur by photodegradation with UV light or heat, resulting in smaller molecules, proceeded by biodegradation with a series of oxidation reactions [47].

Reviews on microorganisms that are able to degrade plastic components are abundant, however, this does not expand to the enzymes involved and respective pathways [39], [48], [49]. As a consequence, characterization of enzymes synthesized by these species is a slow process, and information on involved enzymes are often not curated and few information is associated with those enzymes. **Table *2*** shows a set of curated enzymes and others which are assumed to digest plastic. Relevant available information about each enzyme, including the information of the microorganisms associated and the identifiers in distinct protein databases is given. Due to big amount of unreliable data, only entries with more information and better annotation were included.

Table 2: List of enzymes with presumable present plastic-degrading activity towards PE, PET, PUR and PS.

Enzyme name	Plastic	Microorganism	NCBI Protein ID	PM ID	PDB ID	UniProt ID	Reference
Alkane hydroxylases	PE	*Pseudomonas* sp. E4	—	23360778	—	—	[50]
Hydrolase	PE	*Pseudomonas* sp. AKS2	—	23242625	—	—	[45]
Leaf-branch compost cutinase	PE	Unknown prokaryote	—	22194294	6THT	G9BY57	[51]

Leaf-branch compost cutinase	PE	Unknown prokaryote	—	22194295	6THS	G9BY57	[51]
Leaf-branch compost cutinase	PE	Unknown prokaryote	—	22194296	4EB0	G9BY57	[48]
PE-H (DLH domain-containing protein)	PE	*Pseudomonas aestusnigri* VGXO14	—	—	6SBN	A0A1H6AD45	[31], [52]
C208 laccase	PE	*Rhodococcus ruber*	—	—	—	—	[38], [53]
MnP1 (Manganase Peroxidase 1)	PE	*Phanerochaete chrysosporium*	AAA33743	15850380	1YZP	Q02567	[38], [54]
Soybean Peroxidase	PE	Glycine max Soybean	—	11266599	1FHF	O22443	[38], [55]
Rubredoxin-naD(+) reductase	PE	*Pseudomonas aeruginosa*	NP_254036	14574114	2V3A	Q9HTK9	[38], [56]
Alkane 1-monooxygenase 1	PE, PET	*Pseudomonas aeruginosa*	NP_251264	14574114	—	Q9I0R2	[56]
Alkane 1-monooxygenase 1	PE, PET	*Alcanivorax borkumensis*	CAL18155	14871210	—	Q0VKZ3	[38], [57]
Alkane 1-monooxygenase 2	PE, PET	*Pseudomonas aeruginosa*	NP_250216	14574114	—	Q6H941	[56]
Alkane 1-monooxygenase 2	PE, PET	*Alcanivorax borkumensis*	CAL15570	19953301	—	Q0VTH3	[58]
Poly(ethylene terephthalate) hydrolase/PETase	PET	*Ideonella sakaiensis*	GAP38373	26965627	5XFY	A0A0K8P6T7	[30], [38], [59]
BTA-hydrolase 1	PET	*Thermobifida fusca*	—	—	5ZOA	Q6A0I4	[26]
Cut190	PET	*Saccharomonospora viridis* AHK190	—	32882044	7CEH	W0TJ64	[60]
Cutinase	PET	*Thermobifida fusca*	—	—	4CG1	E5BBQ3	[61]
Cutinase	PET	*Thermobifida fusca*	—	—	4CG2	E5BBQ4	[61]
Cutinase	PET	*Thermobifida fusca*	—	—	4CG3	E5BBQ5	[61]
Extracelular esterases	PET	*Rhodococcus rubber*	—	16534612	—	—	[62]

Cutinase	PET	*Fusarium oxysporum*	—	26291558	5AJH	X0BTD8	[63]
Triacylglycerol lipase	PET	*Thermomonospora curvata*	ACY96861	29427431	—	D1A9G5	[64]
LipIAF5-2	PET	uncultured bacterium	—	—	7ECB	C3RYL0	[65]
Lipase	PET	*Oleispira antarctica* RB-8	—	—	—	—	—
Triacylglycerol lipase	PET	[Polyangium] brachysporum	AKJ29164	29427431	—	A0A0G3BI90	[64]
IsPETase	PET	*Ideonella sakaiensis 201-F6*	GAP38373	26965627	5XH3	A0A0K8P6T7	[52], [66]
nsHiCutinase	PET	*Humicola insole*	—	24832484	4OYY	A0A075B5G4	[52], [67]
MHETase	PET	*Ideonella sakaiensis*	GAP38911	32989159	6QZ4	A0A0K8P8E7	[31], [33]
PE-H (DLH domain-containing protein)	PET	*Pseudomonas aestusnigri* VGXO14	—	—	6SBN	A0A1H6AD45	[31], [52]
Serine hydrolase/Thh_Est	PET	*Thermobifida halotolerans*	AFA45122	—	—	H6WX58	[52]
PulA	PUR	*P. protegens* pv. Fluorescens Pf-5	—	—	—	—	[38]
Polyurethanase esterase A	PUR	*Pseudomonas chlororaphis*	WP_011061486.1	10754242	—	Q9X3C0	[68], [69]
Polyurethanase B	PUR	*Pseudomonas chlororaphis*	WP_011061489.1	—	—	Q9R9H2	[69], [70]
pudA	PUR	*Comamonas acidovorans* TB-35	—	—	—	—	[71]
hydroquinone peroxidase	PS	*Azotobacter beijerinckii* HM 121	—	—	—	—	[72]

In this table, information is filled accordingly with what is found in the literature, and only then complemented with the respective data available in different recognized databases like UniProt [73] and Protein Data Bank (PDB) [74], justifying the missing values, since only over 1 % of all UniProt entries are manually curated [75].

Also, in **Table *2***, are referenced distinct families of enzymes, going from hydrolases, laccases, cutinases, and lipases. While some can catalyse the depolymerization (e.g., IsPETase, PETase, LC Cutinase), other are involved in subsequent steps (MHETase). Generally, even belonging to different families, plastic degrading enzymes all evolved and converged to similar 3D structures taking to account their physiological function. They present low molecular weight and volume, to allow dissipation through microbial membranes and to facilitate access on crystalline polymer chains, and beyond that, they have vast and flexible catalytic pockets, to bind to the long chain substrates [31].

Several families of enzymes have been associated with plastic degradation (). For example, hydrolases are involved in depolymerizing the big plastic chains, so features an open catalytic pocket (**Figure 5**). Cutinases are known for their cutin degradation, but these enzymes also showed activity in PET and PE digestion (**Table *2***), due to their ester hydrolysis action. However, the catalytic site (**Figure 6**) is narrower and shallower when compared with the remaining known enzymes, featuring a catalytic triad in its centre [31]. Another family of enzymes, laccases, act as oxidases and feature a binding site with cooper, and are known for their degradation of lignin, a natural polymer very similar to synthetic polymers (e.g., PS, PET), with an aromatic structure [31].

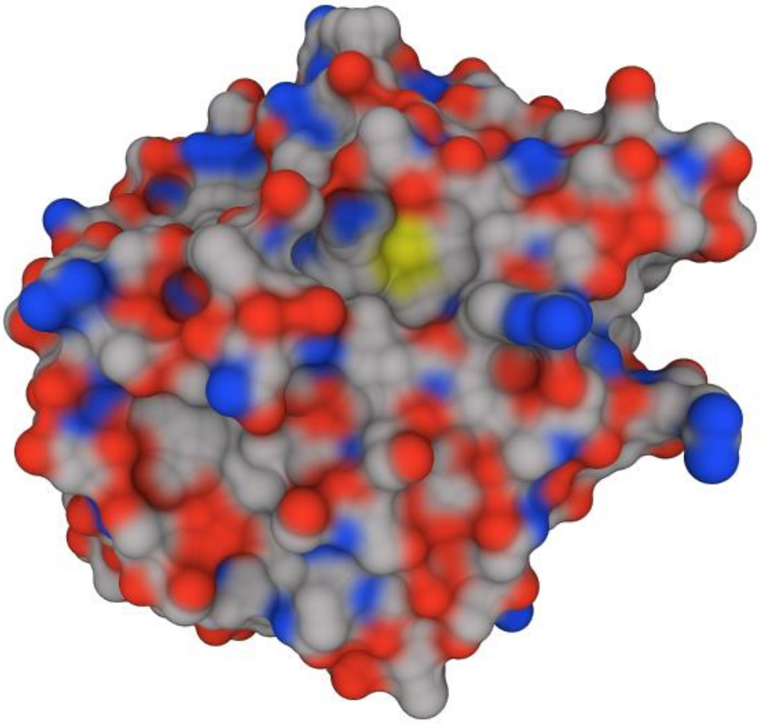


Figure 5: 3D structure of PETase 5XFY retrieved from PDB “3D view”, with Gaussian surface representation and colouring set to element symbol. Blue atoms represent nitrogen, grey atoms carbons, red atoms oxygen and yellow atoms sulphur.

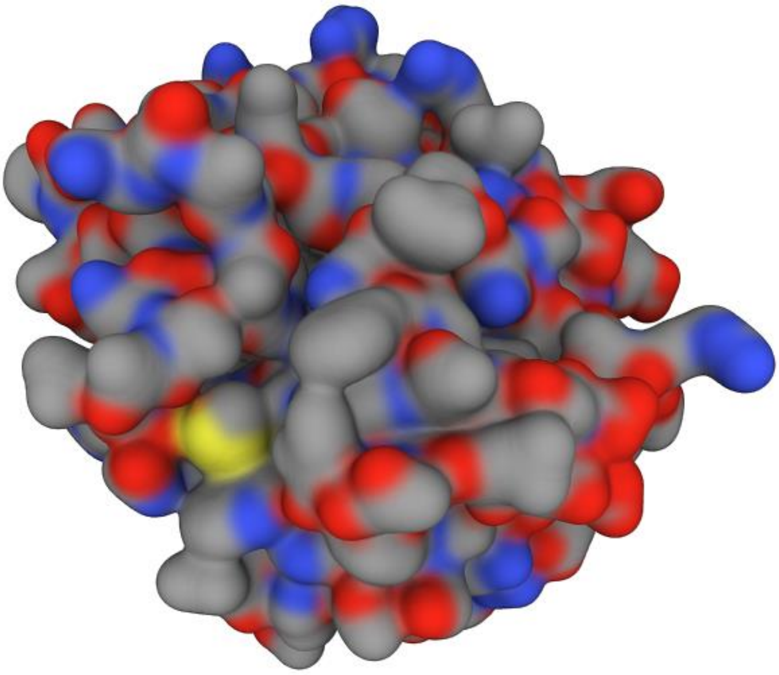


Figure 6: 3D structure of cutinase Cut190 7CEH retrieved from PDB “3D view”, with Gaussian surface representation and colouring set to element symbol. Blue atoms represent nitrogen, grey atoms carbons, red atoms oxygen and yellow atoms sulphur.

Table *2* helps with the following perceptions: on one hand, that PET is the plastic with more known and well characterized enzymes, mostly with laboratory proven activity; on the other, non-biodegradable plastics like PUR and PS have little to no information regarding this matter. PET revealed interest owing to its major percentage of utilization when compared with all other plastics [29].

2.3. Methods for protein annotation

Protein annotation has the potential to extend the knowledge of already well-characterized proteins, through the prediction of functions and metabolic pathways of other unclassified proteins [76]. As more data arrives with metagenomics emergence, new methods are required to analyse such an overwhelming amount of unknown protein sequences [77].

Typical methods for protein annotation are based on the sequence/structure homology (ProtSComp [78]) of protein sequences or domains [75],[79], domain conservation (with the Conserved Domain Database (CDD) from NCBI [80]), and protein-ligand interactions (Ssnet [81], Pupil [82]). Also, metabolic pathways are most of the time associated with gene clusters, and a protein function can sometimes be deduced according to their relative positioning in the genome [76]. More recent annotation methodologies have been developed to consider biological information such as Multilayer Protein Networks (MPN) [83] and Genome Neighbourhood Networks (GNN) [77]. Different studies have been introducing the concept of genomic enzymology to study and predict enzyme function and reactions, based on genome context in enzyme superfamilies [84], [85]. Sequence Similarity Networks (SSNs) are an alternative method to phylogenetic trees produced from multiple sequence alignments of entire protein families [84], which can be too demanding for computers to process for large superfamilies [76]. SNNs were created to handle this problem, performing single sequence alignment between all sequences within the given protein families, utilising *Basic Local Alignment Search Tool* (BLAST). Given a similarity score threshold, SNNs outputs can be compared to a computational graph, where nodes are the sequences and edges makes the connections between sequences with scores higher than the given initially. As this score rises, nodes tend to aggregate into groups, representing distinct functions within the same protein superfamily [85]. This results in a faster process with fewer computer resources required, a user-friendly representation of sequence-function relations, and annotation of enzymes with unknown physiological functions [76]. So far, these methodologies have not been used for annotation of plastic-degrading enzymes. On the contrary, Hidden Markov Models have been applied with that objective, showing satisfactory results.

2.4. Hidden Markov Models and respective validation methods

Hidden Markov Models are statistical/probabilistic models first used in speech recognition problems [86], but with an ever-growing interest for molecular biology sequence analysis, specifically protein structure prediction, gene finding, and homology annotation [87]. The main reason behind the increasing interest for computation biology is the ability of HMM to detect homologies between sequences with low similarity. An HMM is composed by a number of states, corresponding to the number of, for example, characters in a multiple sequence alignment file. Depending on the letters on each position, each state is assigned with an emission probability, and contiguous states are connected by state-transition probabilities [86]. On the other hand, profile HMMs look forward to solving inherent problems of simple HMM, by computing emission and transition probabilities by training HMMs with homologous sequences [87]. This also makes the validation of this models challenging, and some works perform manual curation of the built model [88], or of the results of cross validation with a negative control [89], but also exists dedicated studies to the HMM validation. These are studies from benchmarking distinct methods of validation [90], a method to optimize discrimination thresholds and emissions probabilities by cross validation [91], parameter estimation procedures in place of scoring matrix [92] and the use of logic programming language and machine learning systems [93].

Some articles have already presented methods for the discovery of plastic degrading enzymes in metagenomes collected from distinct ecosystems and in metagenomic databases, in particular for PET enzymes [43], [64], [94]. D. Danso *et al.* (2018) [64] constructed a single HMM based on nine well-characterized PET hydrolase enzymes, which were later used to scan for similar sequences in known databases. Then, after HMM search against UniProtKB, approximately 11,000 enzymes were retrieved, but only the sequences corresponding to matches with a bit score higher than 180 were further aligned using BLAST and the non-redundant protein sequences (NR) database [95]. 13 sequences were very similar to the ones used to train the models and were further used for HMM refinement. It is also worth noting that HMM training opened the possibility to identify motifs and catalytic site composition and enzyme-ligand linkage information. These new PETase sequences were also classified as matters of superfamily lineage and taxonomic phylogeny. Once a protein is identified as a potential plastic degrading enzyme, by using this bioinformatics approach, it can be synthesized and tested for PET enzymatic activity. In case the activity on PET is positive, the sequence can be added to improve the HMM [64]. Other paper from J. Zrimec *et al.* (2021) [96] also applied HMMs to a wider range of plastics, such as PET, Polyhydroxybutyrate (PHB), PLA, Polybutylene adipate terephthalate (PBAT), PUR, PS, PVA, and other, in order to search proteins in assembled metagenomes.

2.5. Available bioinformatic tools to predict enzymes with targeted enzymatic activity

The are no available tools developed with the specific goal of predicting potential plastic-degrading enzymes in biological sequences, derived for example from metagenomics studies. However, there are tools predicting other types of enzymes in this kind of datasets. For example, FeGenie was developed to predict genes with iron oxidation and reduction properties from microbial isolates or metagenomic samples. FeGenie compares queried sequences with pre-characterized and known genes or user-inputted databases for cross-referencing [88]. Furthermore, FeGenie also references default parameters for protein sequence similarity decisions [97] which can serve as a reference. Another identified approach for metagenomic functional profile analysis of communities – SUPER-FOCUS – tries to fight the increased demand for tools able of handling large-scale data samples. This tool makes use of homology methods against reference databases to identify protein families with similar functions as the inputted data, combining BLASTx and DIAMOND, resulting in faster execution speed but lower sensitivity when compared with other methodologies [98].

Chapter 3.

Methodology and Software Development

A bioinformatics tool was developed to predict enzymes with certain activities in protein datasets. This tool was named “M-PARTY – Mining Protein dAtasets for Targeted enzYmes”. M-PARTY is a local tool, free-to-use, open-source, with a friendly Command-Line Interface (CLI) implemented workflow and database, which detects homologue enzymes in protein sequences through homology-based annotation using Hidden Markov Models. These sequences can be originated from genomics/metagenomics samples and have to be inputted as protein FASTA sequence files.

M-PARTY was fully developed with python3 [99], using default python libraries, packages like pandas [100], argparse (python3), NumPy [101], as well as Anaconda packaged tools [102], [103]. M-PARTY has a CI (continuous integration) workflow to help testing the tool while coding and adding new features.

By default, M-PARTY will identify similarities in protein sequence datasets to enzymes existing in M-PARTY databases developed for PE enzymes, in the form of HMMs.

The development of the tool included five general steps (**Figure 7**). In step 1, the enzymes with PE-degrading activity were collected from the literature. In step 2, the enzymes poorly characterized (e.g., enzymes with function predicted or uncertain from UniProtKB) were excluded, and not considered for the database construction. In step 3 coding and development was performed as well as continuous integration and validation. In step 4, the HMM database was built and posteriorly validated in step 5. In step 6 the tool was tested against real datasets, and in step 7 the tool was made available for public utilization.

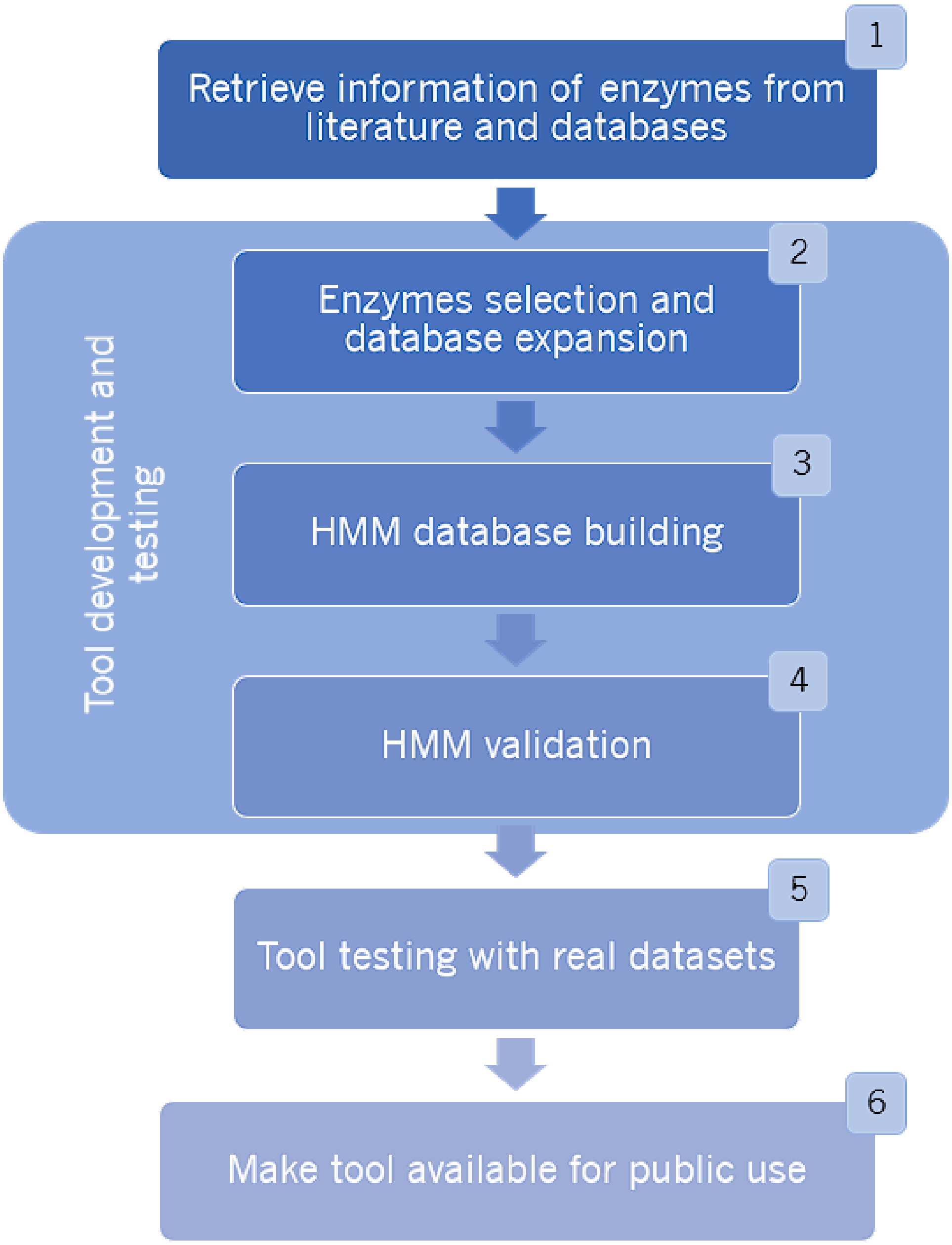

1
Retrieve information of enzymes from literature and databases
Tool development and testing
2
Enzymes selection and database expansion
3
HMM database building
4
HMM validation
5
Tool testing with real datasets
6
Make tool available for public use

Figure 7: Schematic representation of the work performed in this thesis.

For the accomplishment of step 1, deep research work was performed to confirm the state-of-the-art of enzymes involved in PE biodegradation ([31], [38], [48], [52], [54], [56]). Additional functional and structural information on the selected enzymes was retrieved from UniProt, but also from PDB (Protein Data Bank) for the 3D structures and Kyoto Encyclopedia of Genes and Genomes (KEGG) for reports of metabolic pathways where the enzymes participate.

In step 2, the enzymes with more complete information in the databases were the ones selected for further database construction. The number of sequences, curation of the available information, presence or absence of structure and taxonomy were some of the criteria considered. To increase the number of enzymes to be used in the tool training and model construction, the selected enzyme sequences were aligned to other sequences in the UniProtDB by using *UniProt Id Mapping through API* (UPIMAPI) [79]. The enzymes with similarity percentages between 60 % and 90 % were selected, added to the database, and used to develop the models. This range was defined based on the work of D. Danso *et al.* (2018) [64]. The detailed procedure will be described in section 3.2.1.

In step 3 the HMM models were constructed. For that purpose, the protein sequences (from the protein database, step 2) were clustered with CD-HIT (Cluster Database at High Identity with Tolerance) [102] and aligned with T-COFFEE (Tree-based Consistency Objective Function for alignment Evaluation) [103]. This methodology was based on the work of J. Zrimec *et al.* (2021) [96]. The resulted alignments were the input for the HMMs construction. This procedure is detailed in section 3.2.1.

In step 4 the HMMs "leave-one-out" cross-validation procedure was performed as described in section 3.2.2. This validation was done to prevent the appearance of false positive models. The dataset used for the validation, as negative control was retrieved from UniProt with human gut metagenome keyword (60759 sequences, in July 2022), considering that it does not include PE degrading enzymes. This dataset was used as a default by M-PARTY. This strategy was previously used in similar approaches [89]. However, because eventually proteins with PE-degrading activity might still exist in this dataset, another dataset was used for validation, which consisted of polymerase protein sequences downloaded from UniProt (by the search of the keyword "polymerase", which resulted in 27730 entries from Swiss-Prot database). M-PARTY tool can receive a custom negative dataset to perform the validation. The models that returned matches against the negative control datasets were excluded.

The tool was tested by running the HMMs developed in the previous step against 4 different datasets (step 5). The datasets selected were those from marine metagenomes (1,495,503 proteins), hydrothermal metagenomes (218,451 proteins), which were obtained from the UniProt database in September 2022 by searching with the keywords "marine metagenome" and hydrothermal metagenome", respectively. Beyond this, also a negative control dataset, to further check for poorly built models was withdrawn and inputted to M-PARTY. The negative control dataset contained gut metagenome proteins (which were retrieved from the NCBI database by searching in the Taxonomy Browser the term "human gut metagenome", 61074 sequences were obtained). In addition, a positive control dataset containing the enzymes for tool construction (**Table 4**). The tool prediction procedure if fully explained in section 3.2.3.

The final task (step 6) consisted of turning M-PARTY available for the general public over a package manager like anaconda [104]. There is a fully functional tool version available at bioconda https://anaconda.org/bioconda/m-party, and tool coding, modules, and respective scripts can be conferred in a public repository in GitHub with https://github.com/ozefreitas/M-PARTY. All steps needed for local installation are present in Anaconda, GitHub page and in section 3.2.4.

3.1. General architecture

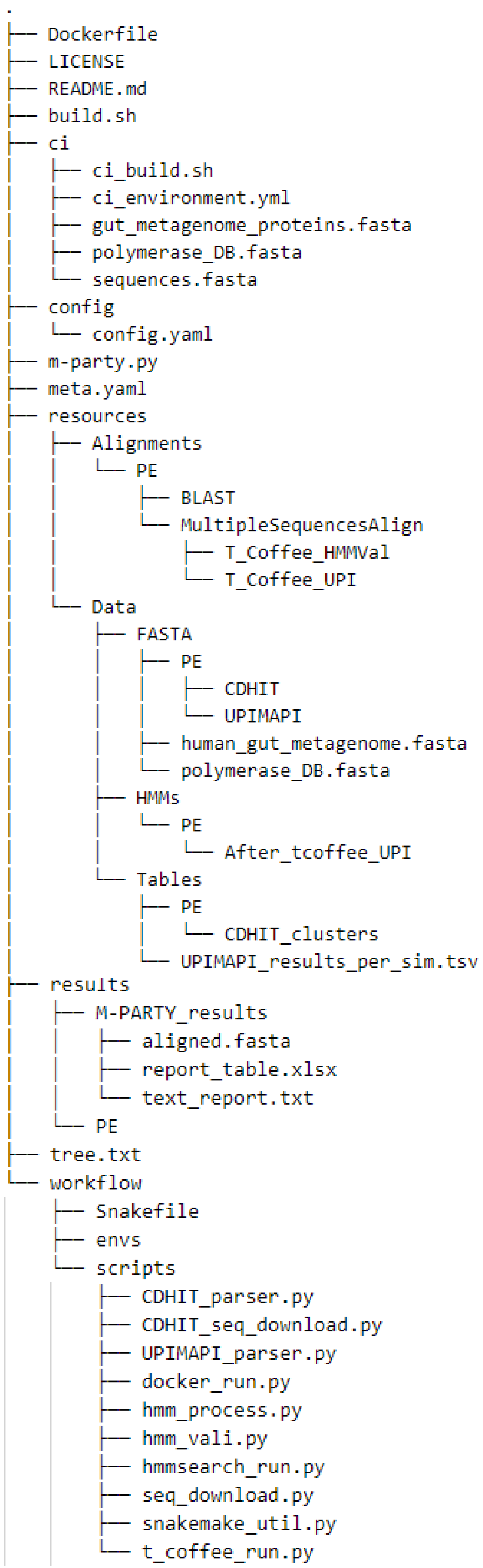
M-PARTY includes both a tool for homology detection and an HMM database. It takes advantage and uses external tools, with extended applications and utilized by many in the most distinct studies. It is the case of CD-HIT [102], a clustering tool that nests sequences by their sequence similarity, that will be of major importance later in the database construction for sequence slicing and selection for each model. Prokaryotic taxonomy [105] and phylogenetic markers [106] tools have been built around CD-HIT, as well as methodologies to work with [107] and condensate large protein databases [108] are using CD-HIT. To align multiple sequences, T-COFFEE tool was picked. T-COFFEE is a versatile multiple sequence alignment tool, with the possibility to align all sequences from DNA to

RNA and proteins [103]. It also offers a variety of alignment methods and output formats, revealing its convenience for posterior steps. Finally, HMMER [109] was used to build the final models and to perform the search of the eventual user inputted against those.

M-PARTY structure follows the standard methodology adopted by several tools implemented with Snakemake [110] workflow manager. Snakemake recommends groping every file involved in workflow execution inside a "workflow" (where a Snakefile is located) directory and "scripts" subdirectory. As a matter of easier readability, M-PARTY is organized by modules instead of all scripts clustered together. Furthermore, all files needed for tool operation (FASTA and HMM files) are within the "resources" folder. Output files from the tool execution are written in a "results" directory.

M-PARTY operability is ensured by its main script "**m-party.py**". Here, it is visible the command line interface from which the user will interact with the tool, all auxiliary functions needed and the main script, where M-PARTY processes the given information and proceeds to take the necessary actions. M-PARTY also offers reduced verbose options for each work section to help and guide the user and states the time of execution for each run.

Erro! A origem da referência não foi encontrada. is displayed as a tree-like r epresentation of the general structure of M-PARTY. *Dockerfile* and *ci* folder are related to continuous integration testing; *meta.yaml* and *build.sh* are the files needed for the bioconda recipe; in the resources folder, there are two main directories – Data and Alignments – with FASTA, HMM, tables and both UPIMAPI and MSA runs, respectively. Inside all these subfolders, the database name given by the user is incorporated, and all data generated for each run is assigned to those folders. The results directory has a sample for all M-PARTY outputs, and the workflow includes the Snakefile, with the task of constructing the databases, and all auxiliary scripts.



```
.
├── Dockerfile
├── LICENSE
├── README.md
├── build.sh
├── ci
│   ├── ci_build.sh
│   ├── ci_environment.yml
│   ├── gut_metagenome_proteins.fasta
│   ├── polymerase_DB.fasta
│   └── sequences.fasta
├── config
│   └── config.yaml
├── m-party.py
├── meta.yaml
├── resources
│   ├── Alignments
│   │   └── PE
│   │       ├── BLAST
│   │       └── MultipleSequencesAlign
│   │           ├── T_Coffee_HMMVal
│   │           └── T_Coffee_UPI
│   └── Data
│       ├── FASTA
│       │   ├── PE
│       │   │   ├── CDHIT
│       │   │   └── UPIMAPI
│       │   ├── human_gut_metagenome.fasta
│       │   └── polymerase_DB.fasta
│       ├── HMMs
│       │   └── PE
│       │       └── After_tcoffee_UPI
│       └── Tables
│           ├── PE
│           │   └── CDHIT_clusters
│           └── UPIMAPI_results_per_sim.tsv
├── results
│   ├── M-PARTY_results
│   │   ├── aligned.fasta
│   │   ├── report_table.xlsx
│   │   └── text_report.txt
│   └── PE
├── tree.txt
└── workflow
│   ├── Snakefile
│   ├── envs
│   └── scripts
│   │   ├── CDHIT_parser.py
│   │   ├── CDHIT_seq_download.py
│   │   ├── UPIMAPI_parser.py
│   │   ├── docker_run.py
│   │   ├── hmm_process.py
│   │   ├── hmm_vali.py
│   │   ├── hmmsearch_run.py
│   │   ├── seq_download.py
│   │   ├── snakemake_util.py
│   │   └── t_coffee_run.py
```

Figure 8: Schematical representation of M-PARTY file structure, with all in-build folders and most important files, from CI, main and auxiliary scripts, resources, and results.

The M-PARTY command line interface can read and save multiple parameters/arguments, making its actions depend on what the user gives or flags. To maintain M-PARTY workflows independent, it does not require any positional arguments, and so only compiles a series of optional arguments. But on the other hand, specific workflows require sets of optional arguments that, if not given, will stop execution. M-PARTY can accept the following arguments:

- *-i, --input,* file, or path to a file of FASTA format containing a list of protein sequences to be analysed;
- *--input_seqs_db_const,* file, or path to FASTA format file from which the user would like to build the model database from scratch. Must be a FASTA file of protein sequences, an exception is raised, otherwise.
- *-db, --database,* file, or path to a FASTA file with a large sequence number to serve as a database for BLAST runs against the prior user inputted sequences from the *--input_seqs_db_const* argument. DIAMOND is the chosen tool to query this task. Defaults to “UniProt” and will download this database.
- *--hmm_db_name*, name to be assigned to the database built from the database_construction workflow. It is recommended to give a name that describes the family or other characteristic of the sequences in–-*input_seqs_db_const.* Mandatory when a new database construction workflow is started.
- *-it, --input_type*, defines the nature of the sequences in the *–input* file between “protein”, “nucleic” or “metagenome”. Defaults to “protein”.
- *-o, --output,* name, or path to the desired output directory. Can be an existent or a non-existent directory, in which case, all the directories and sub-directories will be created as needed. If only a name is given, a folder is made in the current pwd. Defaults to “M-PARTY_results”.
- *--output_type,* chooses the output report table format from “TSV”, “CSV” or “excel”. Defaults to “TSV”
- *-rt, --report_text,* decides whether to produce or not a friendly report in .TXT format with easy-to-read information about the events from M-PARTY execution. Defaults to False. Call flag to set to True.

- *--hmms_output_type*, chose the output type of hmmsearch run from "out", "TSV" or "pfam" format. Defaults to "TSV".
- *--validation*, decides whether to perform models' validation and filtration with the "leave-one-out" cross-validation methods. Defaults to False. Call flag to set to True.
- *-p, --produce_inter_tables,* if user wants to save intermediate tables as parseable .csv files (tables from hmmsearch results processing). Defaults to False. Call flag to set to True.
- *--negative_database*, file or path to a file containing a defined negative control database. The default use of human gut microbiome (already in-built).
- *-t, --threads,* integer number of threads for Snakemake to use. Defaults to 1.
- *-hm, --hmm_models, a* path to a directory containing HMM models previously created by the user. Can be articulated with validation if desired.
- *--concat_hmm_models,* concatenates HMM models into a single file. Defaults to True. Call flag to set to False.
- *--unlock*, could be required after forced workflow termination.
- *-w, --workflow,* defines the workflow to follow between "annotation", "database_construction" and "both". The latter keyword makes the database construction first and posterior annotation. Defaults to "annotation".
- *-c, --config_file,* user-defined config file. Only recommended for advanced users. If given, overrides config file construction from the input.
- *--display_config,* declare to output the written config file together with results. Useful in case of debugging. Defaults to False. Call to set to True.

To summarize, M-PARTY accepts a protein FASTA file or a FASTA with protein sequences from metagenomic samples processing, to search similarities against a pre-built HMM models database, and produces a set of results files. Almost every feature of the tool can be changed or swapped, starting from the own database to the intermediate files that it can generate. Different users may have different objectives and so utilize M-PARTY for different purposes. Additionally, allows running different workflows on their own, if the user has a portion of the work previously done, allowing to skip redundant jobs.

M-PARTY started with a single workflow, *annotation*, which runs the pre-built models against a user inputted FASTA file with a set of sequences using the hmmsearch algorithm. This algorithm performs function prediction in all inputted sequences against each given model, and only the best-suited result is showcased in the output. This output format is composed of several metrics divided into “full sequence” and “best 1 domain”. The most important is the bit score and e-value score. But first, direct output from hmmsearch is not iterable by any mainstream data processing package for python, despite being a “space-delimited” file. Because of this, a script was developed to make the “.OUT” file readable. This script also offers a quick filtering procedure where the bit and E-value scores are overlooked and must obey to predefined thresholds. These values were withdrawn from D. Danso *et al.* (2018) [64] where the authors follow an experimental plan for discovering PET degrading enzymes, by first defining a single Hidden Markov Model. If both metrics are inside these values, both the model number, respective query sequence, metrics and description are shown in the final M-PARTY resulting excel.

3.2. Modules and scripts

M-PARTY is divided into 4 modules, each including scripts to help tool execution. *Database construction*, *annotation* and *validation* modules represent the individual workflows performed by M-PARTY, while *MPARTY_util* has as the scripts and functions to help the latter modules. This distribution is shown in **Table 3**. Inside *database construction,* the Snakefile performs all the work. A Snakefile is a python-based file with distinct “rules” that manage the dependencies between a given input and output and what is needed to make that conversion. It calls other python files, that have the functions to execute each rule, and a final script to transform input into the outputs demanded by snakemake. It reveals its value in wide-scale immutable pipelines with the need for automation. *UPIMAPI_parser.py* receives the output from UPIMAPI execution and returns a ".TSV" comprising the query sequences sliced in the different thresholds. *seq_download.py* will download these same sequences from UniProt, through the UniProt API (https://www.uniprot.org/help/api_retrieve_entries). This step makes new folders with the desired sequences, which are then clustered by CD-HIT. The CD-HIT default output is made of 2 files, a FASTA with the representative sequences from each cluster and a ".TXT" file with the number of clusters found all sequences within each cluster and the percentages of similarity in comparison with the main sequence. To process the latter file, *CDHIT_parser.py* comes to play, creating a ".TSV" with row names as the number of each cluster and a single column with the query ID. In the same way, as done with UPIMAPI, *CDHIT_seq_download.py* proceeds to download the whole same sequences from UniProt once again. Finally, M-PARTY calls T-COFFEE to align all sequences in a format accepted by HMMER.

After the database workflow is complete, the user can decide whether to validate the newly created models. If the validation flag is raised, execution of *hmm_vali.py* is started. This script contains functions to perform all steps for the "leave-one-out" cross-validation method. The result is the filtration of the models not obeying the defined parameters, those being each model possessing a value of strict recall of at least 80 % (explained in section 3.2.2)

Additionally, *annotation* needs to run the hmmsearch algorithm, helped by *hmmsearch_run.py,* which plays with some of the algorithm's options depending on the options inputted by the user. Beyond this, *hmm_process.py* processes the output from hmmsearch (not readable by any data processing package) and produces all the final files so that M-PARTY can display its outputs.

MPARTY_util is a small module with a few useful functions for the snakemake workflow execution and hmmsearch commands run. Some modules are connected, like *validation* and *annotation*, that must process the results from the *hmmsearch* run, and so utilize the same scripts.

Finally, all modules are managed by **m-party.py** and its executions will depend on the arguments given by the user, which are read by the *argparse* python module.

Table 3: M-PARTY modules distribution. As headers, the name is given to each module, and subsequently, below each model the enumeration of the scripts inside each module.

Database Construction

□ Snakefile
□ seq_download
□ CDHIT_seq_download
□ CDHIT_parser
□ UPIMAPI_parser
□ t_coffee_run

Validation

□ hmm_vali
□ hmm_process

Annotation

□ hmm_process
□ hmmsearch_run

MPARTY_util

□ docker_run
□ hmmsearch_run
□ snakemake_util

3.2.1. HMM database construction workflow

M-PARTY can build a complete HMM database from scratch if the *-w/--workflow* option is set to "db_construction" and the *--input_seqs_db_const* option is filled with a protein FASTA file. If only the latter is triggered, M-PARTY will produce the database and cease operation. Options like *--validation* together with *--input* can be added to execute other workflows after the database is concluded. As mentioned in section 3.2, Snakemake performs automatically all steps leading to a desired output. Snakemake individualizes each task into smaller jobs, defined by rules, automatically deducing dependencies between them, simply by comparing input and output filenames. These rules are present in a Snakefile found in the main root or a workflow folder (**Erro! A origem da referência não foi encontrada.**), in this case, of the tool, from where Snakemake can get access. Each M-PARTY run is unique, and produced files vary in both number and name, Snakemake offer of wildcards reveals to be of major importance in some tasks. Wildcards allow Snakemake to generalize a rule and apply it to variable quantities of items, regardless of what is given in each rule iteration.

In M-PARTY's case, a Snakefile is already present with all imperative rules to transform a set of proteins in a vast collection of HMM models. The config file is always written at the beginning of every M-PARTY run, but an external one can be added for Snakemake to use (*-c/--config_file*). This can cause tool interference, making it impossible to run, so is not recommended as the config file is utilized throughout the distinct workflows from M-PARTY.

The input to the database construction workflow is a FASTA protein file (FAA format), and disrespecting this will print an error message.

The methodology used for the creation of an HMM database was based on the work from J. Zrimec *et al.* (2021) [96] and D. Danso *et al.* (2018) [64]. For M-PARTY, the specific steps are described.

Selected proteins are inputted to UPIMAPI, to expand the number of sequences and build a protein database of close related proteins. UPIMAPI runs DIAMOND against the UniProt database with all default parameters and returns a ".TSV" file with similarity percentages, query and target sequence IDs, E-values, bit scores, and more information from annotation. Following the methodology, *UPIMAPI_parser.py* will then divide and group the results by thresholds of similarity. All sequences outside the range of 60 % to 90 % similarity to those inputted are discarded. The sequences left are grouped by increments of 5 %. For this, a file is written with thresholds intervals as row names and the UniProt IDs in front. Every result outside this line is discarded, considering that matches below 60 % identity are too distinct and will introduce error to models, making it more susceptible to identify homologies for proteins with similar domains but distinct functions, but also the decision not to include results above 90 %

similarity, because of redundancy issues, avoiding model “overfitting” and solely detection of training sequences.

The file product from *UPIMAPI_parser.py* is read by *seq_download.py*, which makes the connection with UniProt via API, downloading the sequences through their IDs. Every sequence is downloaded by its threshold, and it is written in a file with the interval as name, like “60-65.fasta”.

Inside each threshold file just mentioned, sequences are then clustered together. CD-HIT default program is executed for these files, with a sequence identity threshold set to 0.9 and word length to 5. CD-HIT produces two files, a FASTA with the representative sequences for each computed cluster and a text file with the number of clusters, the number of sequences inside each cluster and the respective sequence’s entries, accompanied by sequence similarities percentages.

As already mentioned in section 3.2, CD-HIT “*.CLSTR”* file, despite being space-delimited, is not readable as an ordinary “.TSV” or “.TXT” file. *CDHIT_parser.py* makes a similar job as the *UPIMAPI_parser.py* script, processing this file by its characteristic format and generating a “.TSV”, with the cluster number as row names, and the corresponding UniProt IDs in the successive columns.

Latter files sequences are downloaded by *CDHIT_seq_download.py*. The difference now is that *CDHIT_seq_download.py* script will create – for each threshold interval – a FASTA file for every cluster returned by CD-HIT. The number of files exponential raises in this step. Simultaneously, it is also counted the number of sequences on each cluster. This was done to decide whether to include or not, single sequence clusters, like what was done in Peter Skewes-Cox (2014) [89] work. In this article, the authors cluster together a great number of sequences, but only consider clusters with a minimum number of 2 sequences. *CDHIT_seq_download.py* was so instructed to discard all single sequence clusters, if present.

Unfortunately, CD-HIT does not offer a tool to display the alignment performed to do the clustering in any supported format by HMMER. In the article this methodology is being based on, no mentions of the used program are observable, not even mentions that a multiple sequence alignment was performed. This blank in the procedure opens the possibility to use any MSA tool thought appropriate, in this case, T-COFFEE was chosen. T-COFFEE run was performed with all default parameters except the *-output* option set to "clustalw_aln" and *-type* to "protein", to generate a *clustalw* format file. For every FASTA file from the latter step, a new ".CLUSTALW_ALN" is created.

M-PARTY last step is to build the models with HMMER. *hmmbuild* algorithm only accepts alignment format files like .*STOCKHOLM, .PHYLIP, .CLUSTAL*, etc [109].This conditioning was known and influenced the choice of the MSA tool, being one of the few able to output such format.

Summarizing, all steps explained during this section are shown in the following graphic present in **Figure 9**.

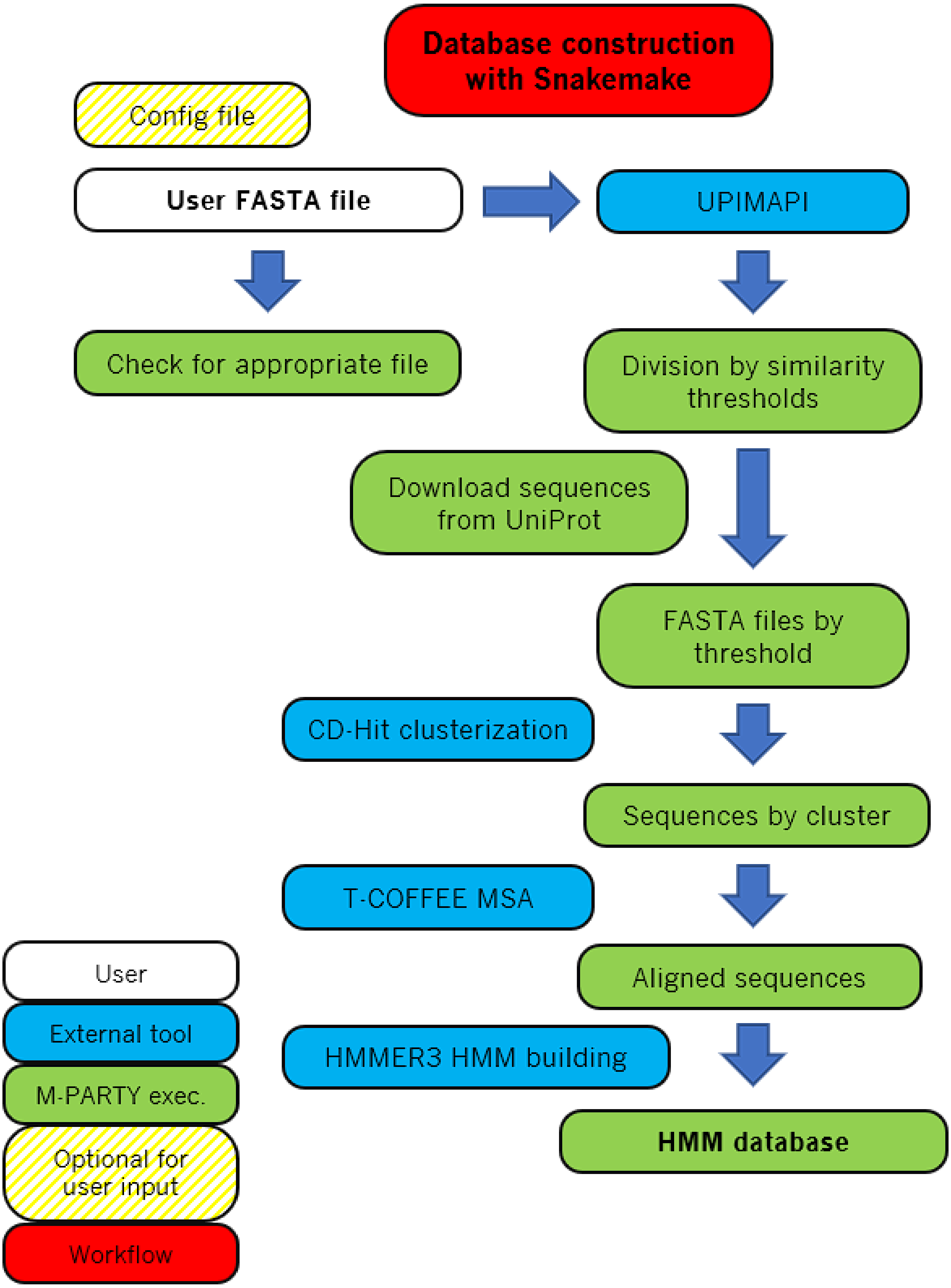


Figure 9: Schematical representation of the steps performed by the database construction workflow.

Not shown in this illustration is a task that helps with file condensation and practicality, to latter fasten and facilitate the next workflows: model concatenation. For each threshold interval, tens or even hundreds of models are generated and saved. This makes further jobs time-consuming, as a new file must be given and opened by the used algorithm, for every existing model. To avoid this, HMMER algorithms allow HMM models to be stacked together, and the subsequent programs that may be used can distinguish among different profiles and instantly give the best results for all present models.

3.2.2. HMM validation workflow

Between the methods to validate HMMs and those described in section 2.4, the "leave-one-out" revealed the most interesting for this case of study. This method was performed with a similar objective and back methodology, as published by Peter Skewes-Cox *et al.* (2014) [89]. The authors tried to scan metagenomic data for virus sequences using Hidden Markov models, by first expanding the number of sequences using BLAST, clustering the resulting database into similar clusters and consequent multiple sequence alignment and models building. Performing this method helped the authors confirm the successfulness of these models training in terms of ambiguity, which means, verifying if the models could distinguish between viral and non-viral sequences and searching for other sequences homologous to the ones inside each model [89]. To accomplish this, the article describes a very detailed procedure: First, having all models built from the HMMER3 hmmbuild function, from each model constituted by N number of sequences, one sequence is removed, and a new model is constructed with N-1 sequences, giving place to a rebuilt model (R). This is repeated for all the sequences making up each model, for all the initial models (M). Worth mentioning that to rebuild these models, each set of N-1 sequences must be aligned again. For every iteration, each deleted sequence is saved and the hmmsearch algorithm is used to try and recall it. Beyond this, the same R models are set against a dataset serving as negative control and against all other sequences not belonging to the model that gave rise to it. A summary of M-PARTY's model validation is represented in the following diagram (**Figure 10**).

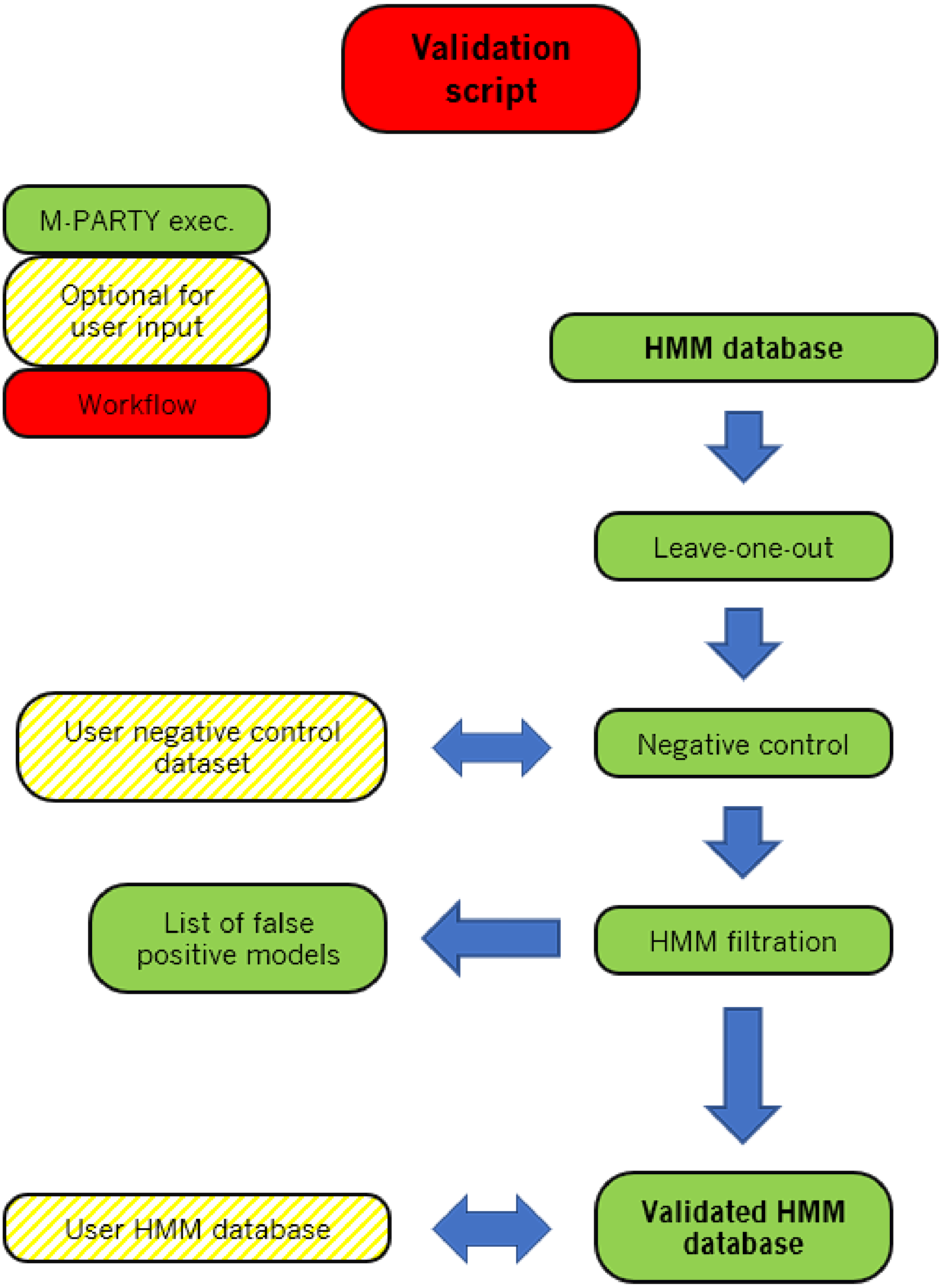


Figure 10: Schematical representation of the validation workflow performed by M-PARTY.

M-PARTY utilizes the human gut metagenome protein sequences from NCBI as a default negative control dataset to be run against each reconstructed model. The goal is to check whether any models match with any sequence from this dataset, concluding for models that would result in false positive matches. In this step, the user can swap this database to one that matches the model database to be validated. Besides this, a well-trained HMM also should not recall sequences used to train other models, and so sequences previously grouped in different clusters [89].

The next step is the filtration of HMMs that did not go through validation. To do this, two metrics are calculated: recall and strict recall. Recall, as the name suggests, estimates the ability of a model to re-collect all the left-out sequences, while strict recall makes a comparison between this measurement and the results given by both negative control procedures, with a literal negative dataset and with sequences from distinct models. To calculate the recall, for all models, **Equation 1** was used:

$$if\ Eval\ \leq 10^{-6}, \qquad then\ R \mathrel{+}= 1, \qquad and\ \frac{R}{N} * 100$$

Equation 1: Representation of the steps taken to get to the recall values for each HMM.

where the e-value was obtained with hmmsearch runs of every reconstructed model with the respective left-out sequence, R is the number of sequences recalled and N is the total number of sequences inside each model. On the other hand, strict recall demands obtaining the evalues for all 3 tasks, direct recall, negative control, and search against the sequences from the other models. When established, these values are compared considering **Equation 2**:

$$if\ Eval_R\ \leq\ \min\left(Eval_{neg}\right) \cap\ \min\left(Eval_{dist.seqs}\right), \qquad then\ SR\ \mathrel{+}=\ 1, \qquad and\ \frac{SR}{N} * 100$$

Equation 2: Representation of the steps taken to get to the strict recall values for each HMM.

were $Eval_R, Eval_{neg}$ and $Eval_{dist.seqs}$ are the evalues of the hmmsearch run of each reconstructed model against the out sequence, negative control database, and all other sequences not belonging to the current model, respectively.

Resulting percentages decide the HMMs' outcome, whether to make part of the final validated HMM database or to be eliminated. Strict Recall is one of M-PARTY parameters and was set to 80 %, just like Peter Skewes-Cox *et al.* (2014) [89] publication. This value means that at least 80 % of the sequences inside a model must respect the condition given in **Equation 2** to proceed.

Finally, the validation workflow *hmm_vali.py* script also allows to re-concatenate the models in single ".HMM" files with interval thresholds as filenames, since it was needed to separate and dissect each individually. Optionally, following M-PARTY philosophy, the user can also input a series of already valid HMM or replace the existing one, and so continue to annotation.

3.2.3. Annotation workflow

The annotation workflow is the simplest and more straightforward pipeline M-PARTY has to offer. Running this pipeline alone without prior jobs aims at users trying to predict PE metabolic activity in a chosen protein sample. The annotation workflow was set up as follows.

An illustrating diagram explaining the steps inside this workflow is displayed in **Figure 11**.

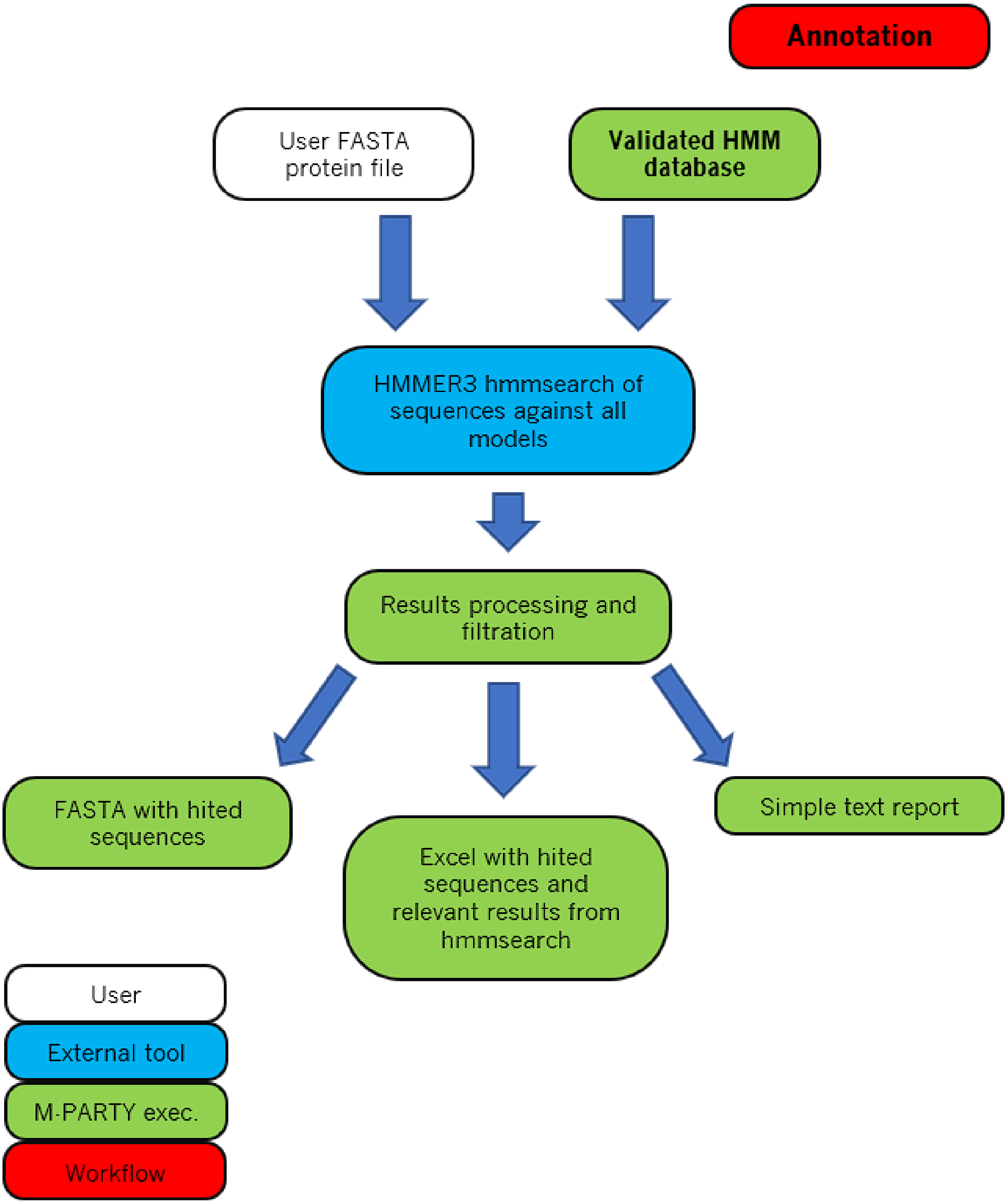

Annotation
User FASTA protein file
Validated HMM database
HMMER3 hmmsearch of sequences against all models
Results processing and filtration
FASTA with hited sequences
Excel with hited sequences and relevant results from hmmsearch
Simple text report
User
External tool
M-PARTY exec.
Workflow

Figure 11: Schematical representation of the annotation workflow performed by M-PARTY.

The only action the user needs to perform is to input the FASTA file with the sequences to be searched against the HMM database and M-PARTY runs the predictions by default against the pre-build PE models. If the database construction workflow is executed beforehand, the annotation will then use the recently built database, if the user so indicates. Beyond this, validation can be performed either for pre-built databases as well as for newly created models Either way, validation execution means that the prior databases get deprecated and M-PARTY instead considers the new validated models, but without deleting the original models from their structure. *hmmsearch* run is set with the "tblout" option, as the *hmms_output_type* option from M-PARTY is defaulted to "tsv", and the space-delimited file is generated. "tblout" is the most similar format that *hmmsearch* program can produce to the standard ".TSV" file. These results are processed and further filtered taking to account their bit scores and e-values. Parameters used for this step were taken from D. Danso *et al.* (2016) [63] and J. Zrimec *et al.* (2021) [96], with bit scores set to greater than 180 and e-values lower than 1e–10, respectively. This filtration step is always fulfilled, whether models have been validated or not. Finally, results range from two mandatory files – a FASTA and a table file – and an optional text file. Both first files contain information about the sequences left after filtration, with the model number accompanied by the plastic prefix and correspondent metrics. FASTA is generated by parsing the input and writing on a new file the pretended sequences. The report file is optional by flagging the *-rt* or *--report_text* option and writes a plain language text file with general information from the jobs performed through the last workflow. Ultimately, the user can call for the *--display_config* flag to get, together with the results, the written config file by the beginning of M-PARTY execution, which can turn out to be useful to trace back possible input errors.

3.2.4. Installation

M-PARTY is available for Linux platforms though GitHub repository cloning, using the following line in a git bash terminal inside the desired (empty) folder:

```
cd path/to/desired/dir
git clone https://github.com/ozefreitas/M-PARTY.git
```

It is highly recommended for users to create an appropriate conda environment with the required dependencies, so M-PARTY executes smoothly, with:

```
cd workflow/envs/

conda env create -n <name of env> -f mparty.yaml

conda activate <name of env>

cd ../..
```

Cloning though GitHub is only recommended in last case scenario, as this as deprecated in detriment of bioconda distribution application.

M-PARTY is available as a conda package from bioconda. Simply open an Anaconda prompt or a command line interface with Anaconda or Miniconda distributions installed and:

```
conda install -c conda-forge -c bioconda m-party
```

If something goes wrong, it is suggested to first create a conda environment with:

```
conda create -n <name of env> -c conda-forge -c bioconda m-party
```

due to possible compatibility issues that may occur.

Chapter 4.

Results

The main result obtained was a bioinformatics tool which predicts certain enzymatic functions in protein datasets. The target enzymes are defined by the user. This tool was named M-PARTY – Mining Protein dAtasets foR Target enzYmes and was tested to predict potential PE-degrading enzymes.

M-PARTY is a python-based tool, runs locally only in LINUX platforms, is free-to-use, its coding open-sourced, and was written with a simple command line interface for easy user interaction.

4.1. Enzyme selection for HMM construction

The enzymes found to be involved in PE degradation were hydrolases, cutinases and peroxidases. The ones selected to construct the database are listed in **Table 4**, where the proteins ID, organisms from which they were assigned, E.C. number and KEGG ID are given if available.

Between the selected enzymes are the cutinase LC cutinase with PDB entry 6THT, Rubredoxin-NAD(+) reductase with peroxidase activity and PDB entry 2V3A, Manganese peroxidase 1 (MnP1) with PDB entry 1YZP, DLH containing protein with the PDB entry 6SBN, and a peroxidase from Glycine max with PDB entry 1FHF. From UniProt, all useful information can be collected, and it is also shown in **Table 4**.

Table 4: List of the enzymes used for the construction of the tool.

UniProt ID	Name	Organism	E.C. Number	KEGG ID	UniProt link	Cellular location	Catalytic activity	Domains	Reference s
G9BY57	LC Cutinase	Unknown thermophilic bacterium	EC:3.1.1.74	K08095 (orthology)	https://www.uniprot .org/uniprot/G9BY5 7	Secreted/ Extracellular	Hydrolysis of cutin	Signal	[48], [51]
Q9HTK9	Rubredoxin-NAD(+) reductase	*Pseudomonas aeruginosa*	EC:1.18.1.1	PA5349	https://www.uniprot .org/uniprot/Q9HTK 9	Cytoplasm	Hydrocarbon hydroxylating system	FAD/NAD binding	[38], [56]
Q02567	Manganese peroxidase 1	*Phanerochaete chrysosporium*	EC 1.11.1.1 3	K20205 (orthology)	https://www.uniprot .org/uniprot/Q0256 7	Secreted/ Extracellular	Oxidation of Mn^{2+} to Mn^{3+}, lignin compounds	Signal, Heme binding site	[38], [54]
A0A1H6A D45	DLH-domain-containing protein	*Halopseudomo nas aestusnigri*	—	—	https://www.uniprot .org/uniprotkb/A0A 1H6AD45	—	—	Signal, hydrolytic catalytic domain	[31], [52]
O22443	Rubredoxin-NAD(+) reductase	*Glycine max* (Soybean) (*Glycine hispida*)	EC:1.11.1.7	—	https://www.uniprot .org/uniprotkb/O22 443	Secreted/ Extracellular	Removal of H_2O_2, lignin degradation	Signal, Heme binding site, peroxidase domain	[38], [55]

The enzymes selected were isolated from a thermophilic bacterium [48], [51], a *Pseudomonas aeruginosa* [38], [56], *Phanerochaete chrysosporium* [38], [54], *Halopseudomonas aestusnigri* [31], [52], and from a soybean plant [38], [55]. Despite the existence of a great number of PE degrading enzymes, the selected proteins are the best characterized and with evidence of its involvement in this pathway.

All enzymes have 3D crystal structure representations in PDB [74], which are shown in **Figure 12**. Beyond this, information like intracellular enzyme location, binding site and catalytic residues, domains and motifs were retrieved from UniProt.

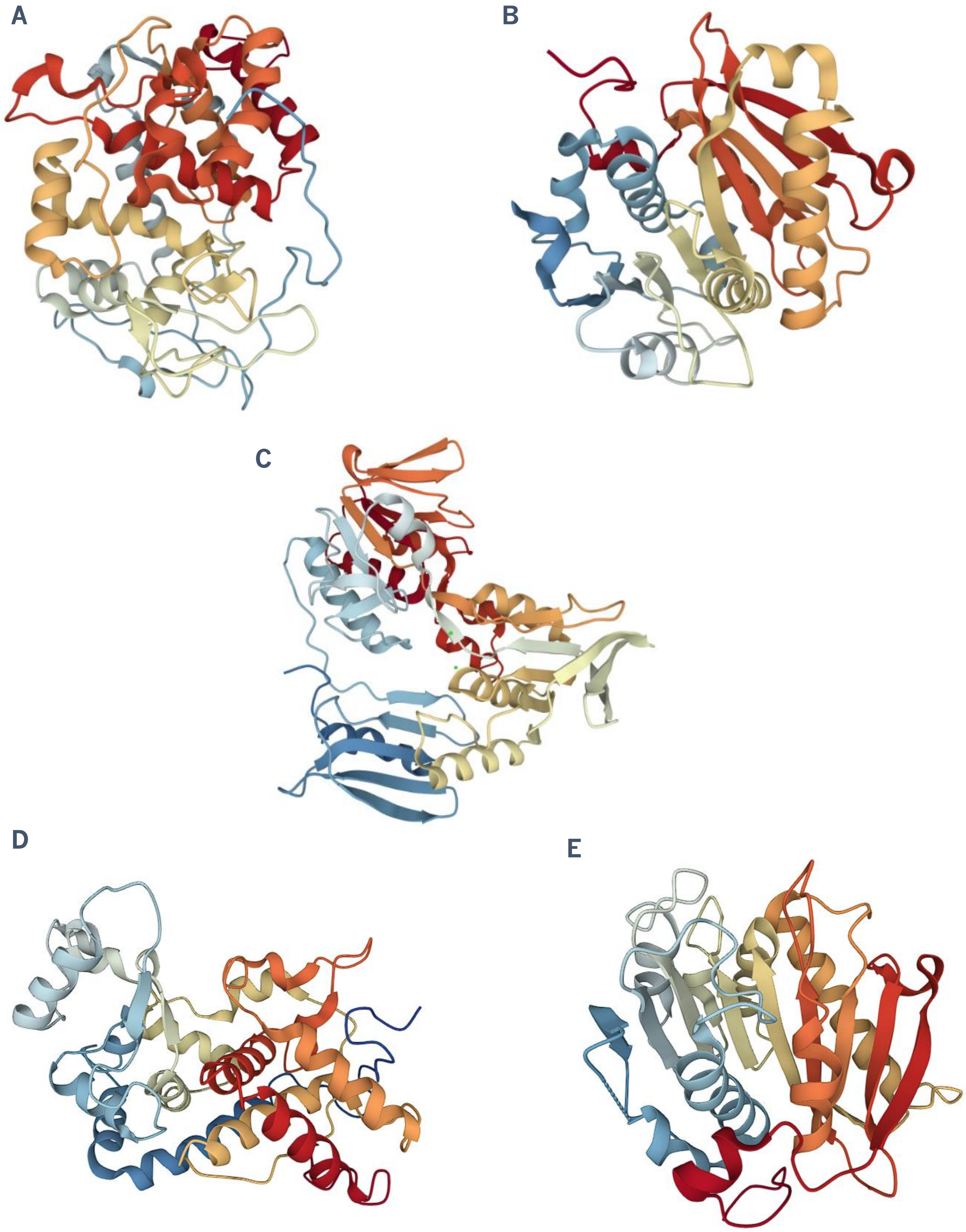


Figure 12: Three-dimensional structure of (A) Manganese peroxidase 1 (PDB ID 1YZP), (B) LC cutinase (PDB ID 6THT), (C) Rubredoxin-NAD(+) reductase (PDB ID 2V3A), (D) peroxidase from Glycine max (PDB ID 1FHF) and (E) DLH containing protein (PDB ID 6SBN), adapted from PDB.

Analyzing these images, is notorious their distinction between the select enzymes' structure, expected from the different family's selection. Differences are observed in the number of α-helixes and β-sheets and the tertiary structure formed by interactions between these structures.

4.2. HMM database construction for PE-degrading enzymes

Table 5 shows a summary of steps taken to build the PE based HMMs, utilized tools and output locations.

Table 5: Summary table of steps leading to database construction, and corresponding used tools, the type of output produced and a link referring to the location of the resulting files in M-PARTY GitHub repository.

Step number	Step description	Tool	Output file type	Results link
1	Sequence database expansion	UPIMAPI	.TSV	https://github.com/ozefreitas/M-PARTY/tree/main/resources/Data/FASTA/PE/UPIMAPI
2	Clustering	CD-HIT	.CLSTR	https://github.com/ozefreitas/M-PARTY/tree/main/resources/Data/FASTA/PE/CDHIT
3	Multiple Sequence Alignment	T-COFFEE	.CLUSTALW_ALN	https://github.com/ozefreitas/M-PARTY/tree/main/resources/Alignments/PE/MultipleSequencesAlign/T_Coffee_UPI
4	HMM building	HMMER	.HMM	https://github.com/ozefreitas/M-PARTY/tree/main/resources/Data/HMMs/PE/After_tcoffee_UPI

The analysis of the 5 sequences with UPIMAPI against the TrEMBL database (approximately 230 million sequences) resulted in almost 50000 sequences, from which only 1304 were selected (showing homology percentages from 60 % to 90 % to the initial 5 sequences) and processed into a single table with the grouped IDs by threshold intervals of 5 %, as introduced in 3.2.1, and can be seen in **Table 6**. Step 1 of database expansion is shown in **Table *5*Erro! A origem da referência não foi encontrada.**, and full results in the respective link.

Table 6: Example of the processed result table from *UPIMAPI_parser.py*.

File number	Homology thresholds	UniProt ID	UniProt ID	UniProt ID	UniProt ID	
1	**60-65**	A0A399NAG9	A0A399NX42	M5B8Q1	A0A8F5W1Z5	+563
2	**65-70**	A0A1H2H9F1	A0A1I6HGD3	A0A4R5UK84	X7F4G7	+216
3	**70-75**	A0A1I0CBD2	A0A1H2H9H3	A0A1H1PFY8	A0A0S4I921	+99
4	**75-80**	A0A1H1RJ19	A0A031MKR8	E9KJL1	A0A031MKR8	+190
5	**80-85**	Q1I2R6	A0A120G870	A0A0F7Y5W5	A0A109KJY2	+194
6	**85-90**	A0A024HPL2	A0A127N1R3	A0A1G8JRQ0	A0A4Z0INC7	+18

Six FASTA files were obtained after analysis with UPIMAPI corresponding to the sequences inside each homology threshold interval. These FASTA files were then submitted to CD-HIT (**Table *5*Erro! A origem da referência não foi encontrada.** step 2) with sequence identity and worth length parameters set to 0.9 and 5 respectively. The cluster file obtained from CD-HIT (".CLSTR") was converted to a ".TSV" file. TSV files containing only one sequence were discarded (**Table 7**). This table shows the first 5 clusters of a total of 88 obtained for the threshold 60-65. For the thresholds 65-70, 70-75, 75-80, 80-85 and 85-90 a total of 41, 20, 19, 6 clusters were obtained, respectively.

Table 7: Example of the processed result table from CDHIT_parser.py for the 60-65 threshold, showing the first 5 cluster from a total of 88 clusters.

Cluster number	UniProt ID	UniProt ID	UniProt ID	UniProt ID	...
1	A0A067FI92	A0A2H5PCA3	A0A2H5PC82	A0A2H5PC95	+13
9	A0A2H5PC80	A0A067FHH9	—	—	—
12	A0A087G5A7	A0A565CIG9	—	—	—
13	A0A3S3PQZ1	A0A3S3N4S4	—	—	—
19	A0A199W3C6	A0A6V7PPX0	A0A6P5FHE8	—	—
+83	...	...	...	...	...

The FASTA sequences of the protein IDs in the TSV files were inputted to T-COFFEE (output parameter set to "clustalw_aln"). This is step 3 from **Erro! A o rigem da referência não foi encontrada.**.

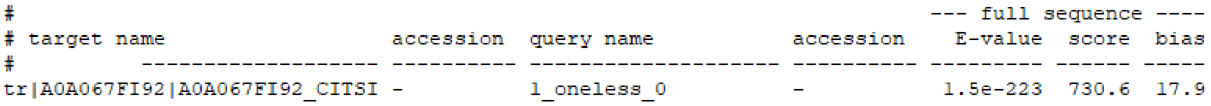
CUSTALW files resulting from T-COFFEE were inputted to HMMER hmmbuild algorithm with all default parameters.

These resulted in a total of 329 HMM composed the PE database organized by threshold, according to **Table 6**. The HMMs can be found on **Table *5*Erro! A origem da referência não foi encontrada.** step 4.

4.3. HMM validation

From the 329 HMM obtained (**Erro! A origem da referência não foi encontrada.** step 4), 103 were excluded, resulting in 226 validated models.

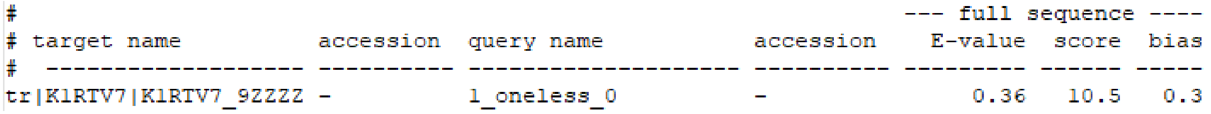
An example of the validation results, with the detail for each step of the validation, is given for one of the 329 HMM models, that is the HMM number 1, corresponding to the 60-65 interval (**Table 7**). **Figure 13** shows the output of the hmmsearch results. In this case, a low E-value was obtained (1.5e–223) meaning that the model was able to successfully recall the excluded sequence.



```
#                                                                --- full sequence ----
# target name                  accession  query name           accession    E-value  score  bias
#         ------------------- ---------- -------------------- ---------- --------- ------ -----
tr|A0A067FI92|A0A067FI92_CITSI -          1_oneless_0          -           1.5e-223  730.6  17.9
```

Figure 13: Extract of a table result from *hmmsearch* run of a reconstructed model without one of its initial sequences against that same sequence.

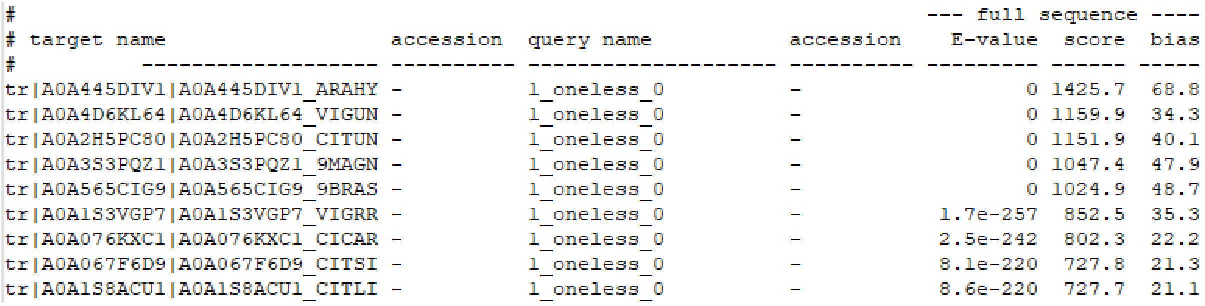
The ".HMM" file without the sequence was queried against the negative control dataset (human gut metagenome proteins), as shown in **Figure 14**. The E-value was very high, which means that the model could not match any of the sequences present in negative control dataset.



```
#                                                             --- full sequence ----
# target name          accession  query name           accession    E-value  score  bias
#  ------------------- ---------- -------------------- ---------- --------- ------ -----
tr|K1RTV7|K1RTV7_9ZZZZ -          1_oneless_0          -               0.36   10.5   0.3
```

Figure 14: Extract of a table result from *hmmsearch* run of a reconstructed model without one of its initial sequences against a negative dataset.

The sequences of the model HMM 1 with one less sequence was aligned to those of the remaining models, as input against all these other sequences. For this, prior job is needed to group together in a single file all those sequences. Results are observable in **Figure 15**.



```
#                                                                         --- full sequence ----
# target name                  accession  query name           accession    E-value  score  bias
#          ------------------- ---------- -------------------- ---------- --------- ------ -----
tr|A0A445DIV1|A0A445DIV1_ARAHY -          1_oneless_0          -                  0 1425.7  68.8
tr|A0A4D6KL64|A0A4D6KL64_VIGUN -          1_oneless_0          -                  0 1159.9  34.3
tr|A0A2H5PC80|A0A2H5PC80_CITUN -          1_oneless_0          -                  0 1151.9  40.1
tr|A0A3S3PQZ1|A0A3S3PQZ1_9MAGN -          1_oneless_0          -                  0 1047.4  47.9
tr|A0A565CIG9|A0A565CIG9_9BRAS -          1_oneless_0          -                  0 1024.9  48.7
tr|A0A1S3VGP7|A0A1S3VGP7_VIGRR -          1_oneless_0          -          1.7e-257  852.5  35.3
tr|A0A076KXC1|A0A076KXC1_CICAR -          1_oneless_0          -          2.5e-242  802.3  22.2
tr|A0A067F6D9|A0A067F6D9_CITSI -          1_oneless_0          -          8.1e-220  727.8  21.3
tr|A0A1S8ACU1|A0A1S8ACU1_CITLI -          1_oneless_0          -          8.6e-220  727.7  21.1
```

Figure 15: Extract of the expect table result from *hmmsearch* run of a reconstructed model without one of its initial sequences against all other sequences not belonging to the current HMM.

The results obtained showed low E-values, meaning that the model was able to recall sequences from the other models. This represents an example of a reconstructed HMM that did not pass the validation step, since the E-value from direct recall is lower than the E-value from negative control, but higher than the minimum E-value from the search of the sequences of other models. In order to drop a model, this condition must be applicable to the results of at least 20 % of the reconstructed models from the initial HMM. In this example, the condition above was confirmed in 100 % of the resulting reconstructed models from HMM 1, discarding this model. All the other HMMs followed the same methodology to validate or exclude the model.

With polymerases as negative control, the same 226 models were maintained after validation and filtering (data not shown).

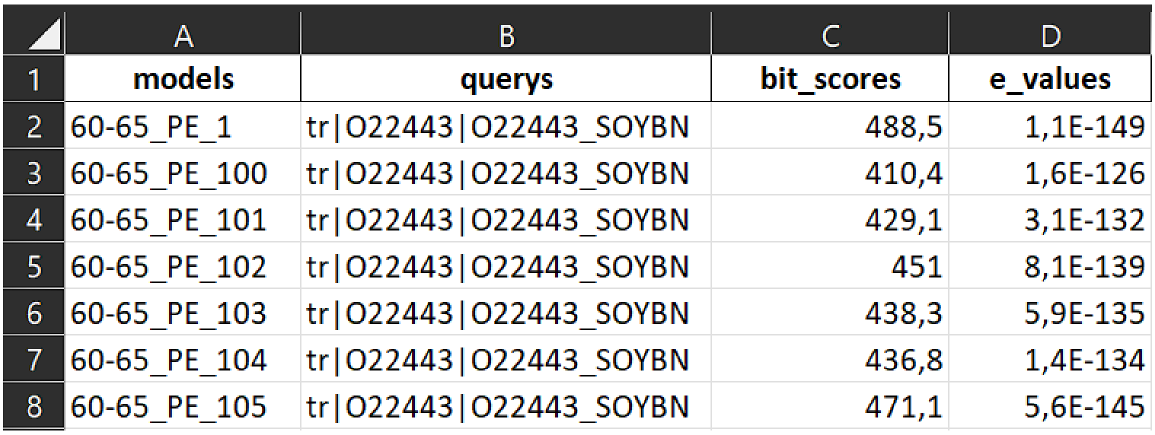
The results suggest that the validation step with the negative control datasets do not influence the final filtered models, what is probably because the filtering occurred during the step of alignment with the sequences from the other models.

4.4. Tool validation

The positive control dataset containing the 5 sequences that were the base to build the HMM models, should be completely identified by M-PARTY and outputted in the results.

The LC Cutinase was identified by 2 HMMs, Rubredoxin-NAD(+) reductase from Pseudomonas aeruginosa by 25, and from Glycine max (Soybean) by 111, Manganese peroxidase 1 by 51 and finally DLH-domain-containing protein by 2 HMMs. These results show that M-PARTY could successfully identify all the 5 proteins.

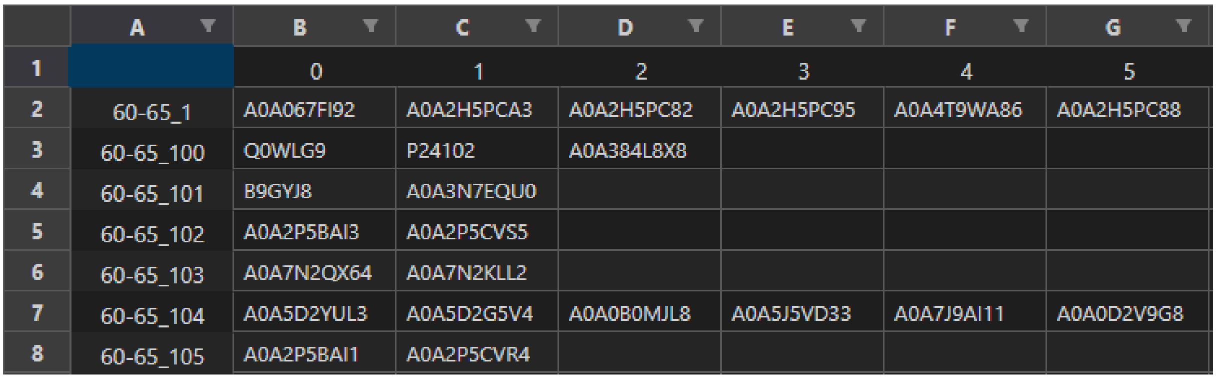
The example of the header of the output excel table report for the protein Rubredoxin-NAD(+) reductase (UniProt ID O22443) is shown in **Figure 16**. The figure shows that this protein was identified at least by 7 different models. It is also shown that the E-values are low meaning that the homology between the queries and the sequence in the model is remarkably high, which was expected since they correspond to the enzymes used for the model's construction.



	A	B	C	D
1	**models**	**querys**	**bit_scores**	**e_values**
2	60-65_PE_1	tr|O22443|O22443_SOYBN	488,5	1,1E-149
3	60-65_PE_100	tr|O22443|O22443_SOYBN	410,4	1,6E-126
4	60-65_PE_101	tr|O22443|O22443_SOYBN	429,1	3,1E-132
5	60-65_PE_102	tr|O22443|O22443_SOYBN	451	8,1E-139
6	60-65_PE_103	tr|O22443|O22443_SOYBN	438,3	5,9E-135
7	60-65_PE_104	tr|O22443|O22443_SOYBN	436,8	1,4E-134
8	60-65_PE_105	tr|O22443|O22443_SOYBN	471,1	5,6E-145

Figure 16: Example of the output report table in ".XLSX" format for the 5 initial sequences. Lines represent the hmmsearch results for each model, with the best matched query sequence. Respective bit scores and E-values are also shown.

M-PARTY also outputs in the excel file the sequences that compose each HMM. Part of this excel sheet is shown in **Figure 17**.



	A	B	C	D	E	F	G
1		0	1	2	3	4	5
2	60-65_1	A0A067FI92	A0A2H5PCA3	A0A2H5PC82	A0A2H5PC95	A0A4T9WA86	A0A2H5PC88
3	60-65_100	Q0WLG9	P24102	A0A384L8X8			
4	60-65_101	B9GYJ8	A0A3N7EQU0				
5	60-65_102	A0A2P5BAI3	A0A2P5CVS5				
6	60-65_103	A0A7N2QX64	A0A7N2KLL2				
7	60-65_104	A0A5D2YUL3	A0A5D2G5V4	A0A0B0MJL8	A0A5J5VD33	A0A7J9AI11	A0A0D2V9G8
8	60-65_105	A0A2P5BAI1	A0A2P5CVR4				

Figure 17: Example of the output report table second sheet with the UniProt IDs of all sequences inside each model, by threshold.

Erro! A origem da referência não foi encontrada. shows the sequences from the 4 datasets tested that could be identified by M-PARTY as potential PE-degrading enzymes.

Table 8: Summary table of the UniProt IDs of the predicted enzyme sequences from each sample dataset.

Positive control (5 initial sequences)	Negative control (Gut microbiome proteins)	Hydrothermal metagenome	Marine metagenome
G9BY57	—	A0A160T8A6	A0A0F9UIZ8
Q9HTK9	—	A0A3B0ZER7	A0A0F9X315
Q02567	—	A0A3B1AKZ9	A0A0F9UNI5
A0A1H6AD45	—	A0A3B0ZJ29	A0A381Z9M3
O22443	—	—	A0A381N483
—	—	—	A0A0F9Q4B9
—	—	—	A0A1Z9EHN2
—	—	—	A0A0F9YHM5
—	—	—	A0A1Z8ZD93
—	—	—	A0A3R7VCY8
—	—	—	A0A424RI56
—	—	—	A0A0F9VWF8
—	—	—	A0A0F9RRP0
—	—	—	A0A0F9YW96
—	—	—	A0A1Z9VMZ9

In the case of negative control dataset, the HMMs did not recall any sequence from the gut microbiome proteins. No results were outputted neither in excel nor in FASTA files, indicating that potential PE-degrading enzymes are not present in this dataset, or at least do not present significant homology with the 5 protein used as reference for the database construction.

M-PARTY returned a total of 4 distinct sequences (**Erro! A origem da referência não foi encontrada.**) from the hydrothermal metagenome dataset. A total of 193 HMMs matched the 4 sequences. All 4 sequences were assigned to the

Rubredoxin-NAD(+) reductase family, the same family of one of the initial sequences used for database building.

The highest number of distinct proteins identified by M-PARTY were obtained when searching against the marine metagenome dataset (**Erro! A origem da referência não foi encontrada.**). A total of 547 HMMs matched 15 different sequences. Similarly, to the results obtained with the hydrothermal dataset, most sequences were assigned to DLH domain-containing and rubredoxin like proteins (10 proteins), only 1 was classified as a hydrolase, and the remaining 4 to FAD-dependent oxidoreductases. The first 10 enzymes identified were most likely homologues to the initial proteins Q9HTK9, O22443 and A0A1H6AD45, and the hydrolase to G9BY57 (**Table 4**). In the tested datasets, no enzyme was matched against the enzyme "Manganese Peroxidase 1". Information about the 19 proteins identified was collected by sequence alignment, by running the NCBI BLASTp against the NR database. **Erro! A origem da referência não foi encontrada.** presents the results from the alignments showing the microorganisms to which the proteins were assigned, and the identity between the identified protein and the closest relative in NCBI database.

Table 9: Summary table with the matched sequences from each dataset with the closest relatives searched by BLAST from NCBI and respective percentage of identity.

Results (from M-PARTY)			Closest relative (in the NCBI database)			
Test dataset	**UniProt ID**	**Enzyme name (UniProt)**	**Enzyme name (NCBI)**	**NCBI ID**	**Microorganism**	**Identity Percentage**
Hydrothermal metagenome	A0A160T8A6	Rubredoxin-NAD(+) reductase	FAD-dependent oxidoreductase	APR65694.1	*Thalassolituus oleivorans*	99.74 %
	A0A3B0ZER7	Rubredoxin-NAD(+) reductase	FAD-dependent oxidoreductase	MBI1423830.1	*Gammaproteobacteria bacterium*	58.73 %
	A0A3B1AKZ9	Rubredoxin-NAD(+) reductase	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	ACL71773.1	*Thioalkalivibrio sulfidiphilus* HL-EbGr7	53.93 %

	A0A3B0ZJ29	Rubredoxin·NAD(+) reductase	rubredoxin·NAD+ reductase	TCV80223.1	*Sulfurirhabdus autotrophica*	49.48 %
Marine metagenome	A0A0F9UIZ8	· DLH domain·containing protein	alpha/beta hydrolase	PKM05449.1	*Gammaproteobacteria bacterium* HGW·Gammaproteobacteria·6	82.69 %
	A0A0F9X315	· DLH domain·containing protein	alpha/beta hydrolase	MAQ49969.1	*Pseudomonas* sp.	81.99 %
	A0A0F9UNI5	· DLH domain·containing protein	alpha/beta hydrolase	MBF78136.1	*Pseudomonadales bacterium*	82.76 %
	A0A381Z9M3	· DLH domain·containing protein	dienelactone hydrolase family protein	MCH2463354.1	*Gemmatimonadetes bacterium*	98.90 %
	A0A381N483	Abhydrolase_5 domain·containing protein	alpha/beta hydrolase	HIF56551.1	*Gemmatimonadetes bacterium*	98.94 %
	A0A0F9Q4B9	FAD·dependent oxidoreductase	FAD·dependent oxidoreductase	HDZ39543.1	*Marinobacter* sp.	100 %
	A0A1Z9EHN2	Rubredoxin reductase	MAG: rubredoxin reductase	OUV68964.1	*Cellvibrionales bacterium* TMED122	100 %
	A0A0F9YHM5	FAD·dependent oxidoreductase	rubredoxin·NAD+ reductase	SDS70520.1	*Halopseudomonas sabulinigri*	86.20 %
	A0A1Z8ZD93	Rubredoxin reductase	rubredoxin reductase	MAJ52495.1	*Halieaceae bacterium*	100 %
	A0A3R7VCY8	Rubredoxin reductase	rubredoxin reductase	RPH12443.1	*Alteromonadaceae bacterium* TMED101	100 %
	A0A424RI56	Rubredoxin··NAD(+) reductase	MAG: rubredoxin··NAD(+) reductase	RPG91173.1	*Cellvibrionales bacterium* TMED148	100 %
	A0A0F9VWF8	FAD·dependent oxidoreductase	FAD·dependent oxidoreductase	HDZ46358.1	*Halomonas* sp.	100 %
	A0A0F9RRP0	Rubredoxin·like domain·containing protein	rubredoxin·NAD(+) reductase	HDY92841.1	*Pseudoalteromonas* sp .	100 %
	A0A0F9YW96	Rubredoxin·like domain·containing protein	FAD·dependent oxidoreductase	MCH4811147.1	*Halomonas neptunia*	91.61 %
	A0A1Z9VMZ9	FAD·dependent oxidoreductase	hypothetical protein	MAH61712.1	*Legionellales bacterium*	100.00 %

BLASTp results show that some predicted sequences are actually existing proteins with assigned functions, and most have been matched to an enzyme with the same function. Hydrothermal results show the worst percentages, not of identity, suggesting that these might correspond to Rubredoxin-NAD(+) reductase enzymes assigned to not yet described microorganisms which inhabit hydrothermal environments. The results are interesting as these proteins might tolerate higher temperatures, which can be beneficial for application in PE-biodegradation strategies.

Due to the low number of enzymes detected in the marine and hydrothermal datasets, the names of the enzyme families were searched directly on the dataset's annotation. It was found that number of enzymes belonging to the families of interest were much higher than the ones detected by M-PARTY (**Table *10***).

Table 10: Percentage of enzymes detected by M-PARTY relatively to the total number of enzymes, with the same name, found in the marine and hydrothermal datasets.

Metagenome dataset	Enzyme family	Enzymes detected by M-PARTY	Total enzymes in datasets matching the enzyme family name	Percentage of enzymes detected by M-PARTY
Marine	Rubredoxin reductase	6	60	10 %
	DLH-domain containing	4	283	1,4 %
	FAD-dependent oxidoreductase	4	514	0,8 %
	Hydrolase	1	8168	≈ 0,01 %
Hydrothermal	Rubredoxin reductase	4	54	7,4 %

For instance, 60 proteins annotated as rubredoxin reductase could be found with this search, but only 6 could be identified by M-PARTY, meaning that M-PARTY retrieved only 10 % of the expected enzymes. For the remaining enzymes the percentages were even lower (**Table *10***). This means that M-PARTY tool is very restrictive and should be reviewed in the future to allow a higher number of identifications.

Regarding the results obtained with M-PARTY, it was found that the microorganisms to which the enzymes were assigned were diverse (**Table 9**), but most of them were not previously related to plastics biodegradation. However, 3 of those were close related to microorganisms described as PE, PET, PHB, PS, PP and PLA biodegradation (**Table *11***). While *Alteromonadaceae bacterium* TMED101 and *Marinobacter* sp. enzymes Rubredoxin reductase and FAD-dependent oxidoreductase, respectively, which were 100 % identical to the enzymes found in the marine metagenomics dataset, were reported to degrade PHB, the DLH domain-containing protein could degrade much more plastics including recalcitrant plastics and showed relatively low percentage of identity to the enzyme of a bacterium belonging to *Gammaproteobacteria*. These results suggest that the marine sediment metagenome contains microorganisms and enzymes capable of degrading a high diversity of plastics.

Table 11: Enzymes previously associated to plastic biodegradation that could be identified by M-PARTY in the metagenomics datasets.

Enzyme name and UniProt ID of the enzymes identified with M-PARTY	Taxonomic assignment in UniProt database	Closest related microorganism obtained by BLAST (percentage of identity)	Degrading Plastic
DLH domain-containing protein (A0A0F9UIZ8)	Marine sediment metagenome	*Gammaproteobacteria bacterium* HGW-Gammaproteobacteria-6 (82.69%)	PE [111], PET [112], PHB [113], PS [114], PP [115], PLA [116]
Rubredoxin reductase (A0A3R7VCY8)	*Alteromonadaceae bacterium*	*Alteromonadaceae bacterium* TMED101 (100%)	PHB [117]
FAD-dependent oxidoreductase (A0A0F9Q4B9)	Marine sediment metagenome	*Marinobacter* sp. (100%)	PHB [113]

Chapter 5.

Discussion

M-PARTY is a functional, easy-to-use, straightforward tool where independent workflows can be executed, depending on the preference of the user. It offers a methodology to build a complete HMM database, starting from protein sequences, and an efficient method to validate the models generated, i.e., avoiding the appearance of false positive results. Then M-PARTY perform function predictions in given protein datasets. M-PARTY main feature is the ability to perform all described steps to any kind of enzyme, and predict its targeted function, just by changing the sequences inputted. Nevertheless, the default version of M-PARTY is applied to PE-degrading enzymes.

The relevance of plastic pollution problem has motivated the development of M-PARTY tool to find enzymes degrading plastics in metagenomes, aiming at discovering novel and highly efficient proteins. There are already several enzymes known to act on PET polymers, but not so many degrading other synthetic plastics, such as, PUR, PE, PS (**Table 2**). PET and PE are the most abundant in plastic waste [118], [119], therefore this tool was first developed to target PE degrading enzymes in order to expand the number of proteins that can potentially degrade this polymer.

By searching against hydrothermal and marine metagenomes 19 proteins matched the criteria and were outputted by M-PARTY (**Erro! A origem da referência não foi encontrada.**). Marine metagenomes were chosen because of the plastic pollution in marine environments, and hydrothermal metagenomes because they contain extremophiles with proteins tolerating high temperatures and other extremophile conditions. The number of proteins obtained could be higher if the HMM were constructed considering a lower similarity cutoff, i.e., lower than 60 %. Indeed, when the same reference proteins were submitted to sequence alignment with BLAST against the UniProt database, maximum sequence similarities obtained were 34 % (for reference enzyme: Q02567), 38 % (O22443), 50 % (Q9HTK9), 62 % (G9BY57) and 63 % (A0A1H6AD45) (data not shown). These results are in agreement with the results obtained with M-PARTY, since few proteins showed similarities higher than 60 %. It is important to note that these percentages cannot be compared directly, as different alignment algorithms were used, for instance BLAST makes whole sequence multiple alignment with position independent substitution matrices based on individual sequences while HMMER considers profiles from the given sequences to check for evolutionary patterns [109], [120].

The backbone of M-PARTY was based on the work from J. Zrimec *et al.* [96]. In that work the authors do no mention which tool was used to perform the multiple sequence alignment. Therefore, several tools were tested including USEARCH, VSEARCH, MUSCLE AND T-COFFEE. HMMER demands a multiple sequence alignment file like stockholm, phylip or clustalw_aln. From those, only MUSCLE and T-COFFEE outputted any of these files. Because TCOFFE is more versatile with the available alignment methods, and was used in D. Danso *et al.* (2018) [64] work, it was chosen to incorporate in M-PARTY.

The validation step is supposed to provide a list of proteins that do not include the proteins of interest. In the case of this work, the chosen datasets with that characteristic were the human gut metagenome. However, because microplastics were already detected in the human body, it would be possible that enzymes with the ability to degrade plastics could also exist in this dataset. So, a second dataset with a set of enzymes to be sure not to have plastic activity, like polymerases, was used to confirm the viability of the human gut metagenome proteins to serve as negative control for this work.

The same HMMs were discarded after validation with both negative datasets (data not shown). Assuming that polymerase sequences does not share any similarities with the model sequences, assumptions can be made that this human gut metagenome proteins sample does not include any enzyme with potential to degrade PE, at least enzymes with some degree of similarity with the M-PARTY database.

However, the last step of validation is very harsh when running against all sequences not belonging to each given model. After a thorough analysis, intermediate results from this step shows extremely low E-values, most even reaching zero. This makes the model immediately discarded, independently of the results from the negative control search. This issue should be improved in the next version of M-PARTY.

Only few enzymes were identified by M-PARTY with potential to biodegrade PE and this may be due to the fact that the tool was quite restrictive. This is concluded because there are much more proteins in the datasets assigned to proteins containing exactly the same name of the initial set of enzymes (the ones associated to PE-biodegradation (**Table 4**)). This issue must be overcome in the future to obtain a tool that correctly identifies enzymes with the same function but that is not too restrictive, excluding positive matches.

Contrarily to what was obtained with the marine metagenome, the enzymes identified by M-PARTY in the hydrothermal metagenome showed relatively low identity percentages to those annotated in the databases. This suggest that these enzymes are distant from those deposited to the NCBI database.

Overall, both metagenomes show a high potential for PE-biodegradation given the resulted obtained in this thesis.

Chapter 6.

Conclusions and Future Work

In this work a tool was developed in python, with a free CLI, open-source code, for Linux platforms – M-PARTY. This tool makes predictions based on sequence homology with HMMs, accepting protein sequences as input and outputting the prediction results in a simple form. M-PARTY offer the full workflow to build and validate the HMMs to be used for prediction.

M-PARTY is operational and could successfully predict and retrieve enzymes with the potential to degrade polyethylene. M-PARTY's models matched a total of 19 protein sequences in metagenomes samples, 4 in hydrothermal and 15 in marine metagenome as possible PE-degrading enzymes.

However, there is some room for improvement in different aspects. For example, the implementation of more advanced and precise methods to further confirm the results, such as multilayer protein networks (MPN) [83] and genome neighbourhood networks (GNN) [77]. Also, to complement the outputs, instead of only returning the IDs for the matched sequences, attention can be given to additionally provide deeper information through automatic sequence mapping, e.g., taxonomy, number of domains, a family of enzymes, EC number. Another feature waiting to be implemented would be the model's refinement after each M-PARTY run. Every highly positive result could be inserted in the corresponding HMM, further maturing each model, which was a strategy already employed by other authors [64]. Adding the step to process metagenomic samples (gene FASTA sequences) to this tool and integrating it into the present workflows would extensively help reduce the user spent time to fully complete a job starting from metagenomes, as M-PARTY currently only receives protein FASTA files.

More work on coding, structure and runtime optimization can always be done, and new ways of displaying the outputs, providing stats about *hmmsearch* results filtration, as well as ways to communicate with the user regarding the steps performed during M-PARTY execution can be explored to make it a more pleasing experience.

References

[1] G. M. Grossman, "Pollution and growth: what do we know?," in *The Economics of Sustainable Development*, Cambridge University Press, 2010, pp. 19–46. doi: 10.1017/cbo9780511751905.003.

[2] M. Smith, D. C. Love, C. M. Rochman, and R. A. Neff, "Microplastics in Seafood and the Implications for Human Health," *Current environmental health reports*, vol. 5, no. 3. Springer, pp. 375–386, Sep. 01, 2018. doi: 10.1007/s40572-018-0206-z.

[3] R. Geyer, J. R. Jambeck, and K. L. Law, "Production, use, and fate of all plastics ever made," 2017. [Online]. Available: https://www.science.org

[4] J. J. Fried, *Groundwater pollution*. Elsevier, 1975.

[5] W. Y. Chia, D. Y. Ying Tang, K. S. Khoo, A. N. Kay Lup, and K. W. Chew, "Nature's fight against plastic pollution: Algae for plastic biodegradation and bioplastics production," *Environmental Science and Ecotechnology*, vol. 4. Elsevier B.V., Oct. 01, 2020. doi: 10.1016/j.ese.2020.100065.

[6] A. L. Andrady and M. A. Neal, "Applications and societal benefits of plastics," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1526, pp. 1977–1984, Jul. 2009, doi: 10.1098/rstb.2008.0304.

[7] W. W. Y. Lau *et al.*, "Evaluating scenarios toward zero plastic pollution," *Science (80-.).*, vol. 369, no. 6509, Sep. 2020, doi: 10.1126/SCIENCE.ABA9475.

[8] D. K. A. Barnes, F. Galgani, R. C. Thompson, and M. Barlaz, "Accumulation and fragmentation of plastic debris in global environments," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1526, pp. 1985–1998, Jul. 2009, doi: 10.1098/rstb.2008.0205.

[9] W. Amass, A. Amass, and B. Tighe, "A review of biodegradable polymers: uses, current developments in the synthesis and characterization of biodegradable polyesters, blends of biodegradable polymers and recent

advances in biodegradation studies," *Soc. Chem. Ind. Polym. Int.*, vol. 47, pp. 89–144, 1998.

[10] S. Facchin, P. D. D. Alves, F. de F. Siqueira, T. M. Barroca, J. M. N. Victória, and E. Kalapothakis, "Biodiversity and secretion of enzymes with potential utility in wastewater treatment," *Open J. Ecol.*, vol. 03, no. 01, pp. 34–37, 2013, doi: 10.4236/oje.2013.31005.

[11] R. C. Thompson, S. H. Swan, C. J. Moore, and F. S. vom Saal, "Our plastic age," *Phil. Trans. R. Soc. B*, 2009.

[12] A. Shah and F. Alshehrei, "Biodegradation of Synthetic and Natural Plastic by Microorganisms," *J. Appl. Environ. Microbiol.*, vol. 5, no. 1, pp. 8–19, 2017, doi: 10.12691/jaem-5-1-2.

[13] V. Bátori, D. Åkesson, A. Zamani, M. J. Taherzadeh, and I. Sárvári Horváth, "Anaerobic degradation of bioplastics: A review," *Waste Manag.*, vol. 80, pp. 406–413, Oct. 2018, doi: 10.1016/j.wasman.2018.09.040.

[14] D. Adamcová and M. Vaverková, "Degradation of biodegradable/degradable plastics in municipal solid-waste landfill," *Polish J. Environ. Stud.*, vol. 23, no. 4, pp. 1071–1078, 2014.

[15] T. Ishigaki, W. Sugano, A. Nakanishi, M. Tateda, M. Ike, and M. Fujita, "The degradability of biodegradable plastics in aerobic and anaerobic waste landfill model reactors," *Chemosphere*, vol. 54, no. 3, pp. 225–233, 2004, doi: 10.1016/S0045-6535(03)00750-1.

[16] K. A. V. Zubris and B. K. Richards, "Synthetic fibers as an indicator of land application of sludge," *Environ. Pollut.*, vol. 138, no. 2, pp. 201–211, Nov. 2005, doi: 10.1016/j.envpol.2005.04.013.

[17] R. Kumar *et al.*, "Landfill microbiome harbour plastic degrading genes: A metagenomic study of solid waste dumping site of Gujarat, India," *Sci. Total Environ.*, vol. 779, Jul. 2021, doi: 10.1016/j.scitotenv.2021.146184.

[18] A. I. Catarino, V. Macchia, W. G. Sanderson, R. C. Thompson, and T. B. Henry, "Low levels of microplastics (MP) in wild mussels indicate that MP

ingestion by humans is minimal compared to exposure via household fibres fallout during a meal," *Environ. Pollut.*, vol. 237, pp. 675–684, Jun. 2018, doi: 10.1016/j.envpol.2018.02.069.

[19] A. Ragusa *et al.*, "Raman Microspectroscopy Detection and Characterisation of Microplastics in Human Breastmilk," *Polymers (Basel).*, vol. 14, no. 13, pp. 1–14, 2022, doi: 10.3390/polym14132700.

[20] Holger M. Koch and Antonia M. Calafat, "Human body burdens of chemicals used in plastic manufacture," *Phil. Trans. R. Soc. B*, 2009.

[21] N. I. S. Abdul-Latif, M. Y. Ong, S. Nomanbhay, B. Salman, and P. L. Show, "Estimation of carbon dioxide (CO2) reduction by utilization of algal biomass bioplastic in Malaysia using carbon emission pinch analysis (CEPA)," *Bioengineered*, vol. 11, no. 1, pp. 154–164, Jan. 2020, doi: 10.1080/21655979.2020.1718471.

[22] B. Imre and B. Pukánszky, "Compatibilization in bio-based and biodegradable polymer blends," in *European Polymer Journal*, Jun. 2013, vol. 49, no. 6, pp. 1215–1233. doi: 10.1016/j.eurpolymj.2013.01.019.

[23] D. Kint and S. Muñoz-Guerra, "A review on the potential biodegradability of poly(ethylene terephthalate)," *Soc. Chem. Ind.*, vol. 48, pp. 346–352, 1999.

[24] R.-J. Mü, I. Kleeberg, and W.-D. Deckwer, "Biodegradation of polyesters containing aromatic constituents," 2001. [Online]. Available: www.elsevier.com/locate/jbiotec

[25] U. Witt, R. J. Müller, and W.-D. Deckwer, "Biodegradation of Polyester Copolymers Containing Aromatic Compounds," *J. Macromol. Sci. Part A*, vol. 32, no. 4, pp. 851–856, Apr. 1995, doi: 10.1080/10601329508010296.

[26] I. Kleeberg, K. Welzel, J. VandenHeuvel, R. J. Müller, and W. D. Deckwer, "Characterization of a new extracellular hydrolase from Thermobifida fusca degrading aliphatic-aromatic copolyesters," *Biomacromolecules*, vol. 6, no. 1, pp. 262–270, Jan. 2005, doi: 10.1021/bm049582t.

[27] S. M. Emadian, T. T. Onay, and B. Demirel, “Biodegradation of bioplastics in natural environments,” *Waste Management*, vol. 59. Elsevier Ltd, pp. 526–536, Jan. 01, 2017. doi: 10.1016/j.wasman.2016.10.006.

[28] Advanced Chemistry Development Inc., “ACD/ChemSketch.” www.acdlabs.com, Toronto, ON, Canada, 2022.

[29] A. Ali Shah, “Role of Microoganisms in biodegradation of plastics,” 2007.

[30] S. Yoshida *et al.*, “A bacterium that degrades and assimilates poly(ethylene terephthalate),” 2016. [Online]. Available: http://science.sciencemag.org/

[31] C. C. Chen, L. Dai, L. Ma, and R. T. Guo, “Enzymatic degradation of plant biomass and synthetic polymers,” *Nature Reviews Chemistry*, vol. 4, no. 3. Nature Research, pp. 114–126, Mar. 01, 2020. doi: 10.1038/s41570-020-0163-6.

[32] Y. Shinozaki *et al.*, “Biodegradable plastic-degrading enzyme from Pseudozyma antarctica: Cloning, sequencing, and characterization,” *Appl. Microbiol. Biotechnol.*, vol. 97, no. 7, pp. 2951–2959, Apr. 2013, doi: 10.1007/s00253-012-4188-8.

[33] H. P. Austin *et al.*, “Characterization and engineering of a plastic-degrading aromatic polyesterase,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 19, pp. E4350–E4357, May 2018, doi: 10.1073/pnas.1718804115.

[34] Z. Montazer, M. B. Habibi-Najafi, M. Mohebbi, and A. Oromiehei, “Microbial Degradation of UV-Pretreated Low-Density Polyethylene Films by Novel Polyethylene-Degrading Bacteria Isolated from Plastic-Dump Soil,” *J. Polym. Environ.*, vol. 26, no. 9, pp. 3613–3625, Sep. 2018, doi: 10.1007/s10924-018-1245-0.

[35] G. K. A., A. K., H. M., S. K., and D. G., “Review on plastic wastes in marine environment – Biodegradation and biotechnological solutions,” *Marine Pollution Bulletin*, vol. 150. Elsevier Ltd, Jan. 01, 2020. doi: 10.1016/j.marpolbul.2019.110733.

[36] A. K. Urbanek, W. Rymowicz, and A. M. Mirończuk, “Degradation of plastics

and plastic-degrading bacteria in cold marine habitats," *Applied Microbiology and Biotechnology*, vol. 102, no. 18. Springer Verlag, pp. 7669–7678, Sep. 01, 2018. doi: 10.1007/s00253-018-9195-y.

[37] E. Schmaltz *et al.*, "Plastic pollution solutions: emerging technologies to prevent and collect marine plastic pollution," *Environment International*, vol. 144. Elsevier Ltd, Nov. 01, 2020. doi: 10.1016/j.envint.2020.106067.

[38] J. Ru, Y. Huo, and Y. Yang, "Microbial Degradation and Valorization of Plastic Wastes," *Frontiers in Microbiology*, vol. 11. Frontiers Media S.A., Apr. 21, 2020. doi: 10.3389/fmicb.2020.00442.

[39] G. J. Palm *et al.*, "Structure of the plastic-degrading Ideonella sakaiensis MHETase bound to a substrate," *Nat. Commun.*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-09326-3.

[40] P. Hugenholtz and G. W. Tyson, "Metagenomics," *Nature*, 2008, [Online]. Available: www.genomesonline.org.

[41] R. I. Amann, B. J. Binder, R. J. Olson, S. W. Chisholm, R. Devereux, and D. A. Stahl', "Combination of 16S rRNA-Targeted Oligonucleotide Probes with Flow Cytometry for Analyzing Mixed Microbial Populations," 1990.

[42] A. Madhavan, R. Sindhu, B. Parameswaran, R. K. Sukumaran, and A. Pandey, "Metagenome Analysis: a Powerful Tool for Enzyme Bioprospecting," *Appl. Biochem. Biotechnol.*, vol. 183, no. 2, pp. 636–651, Oct. 2017, doi: 10.1007/s12010-017-2568-3.

[43] V. R. Viljakainen and L. A. Hug, "New approaches for the characterization of plastic-associated microbial communities and the discovery of plastic-degrading microorganisms and enzymes," *Computational and Structural Biotechnology Journal*, vol. 19. Elsevier B.V., pp. 6191–6200, Jan. 01, 2021. doi: 10.1016/j.csbj.2021.11.023.

[44] G. H. Booth, A. W. Cooper, and J. A. Robe, "Bacterial Degradation of Plasticized PVC," 1968.

[45] P. Tribedi and A. K. Sil, "Low-density polyethylene degradation by

Pseudomonas sp. AKS2 biofilm," *Environ. Sci. Pollut. Res.*, vol. 20, no. 6, pp. 4146–4153, Jun. 2013, doi: 10.1007/s11356-012-1378-y.

[46] A. Loredo-Treviño, G. Gutiérrez-Sánchez, R. Rodríguez-Herrera, and C. N. Aguilar, "Microbial Enzymes Involved in Polyurethane Biodegradation: A Review," *Journal of Polymers and the Environment*, vol. 20, no. 1. pp. 258–265, Mar. 2012. doi: 10.1007/s10924-011-0390-5.

[47] Vasile C., "Degradation and decomposition," *Handb. polyolefins Synth. Prop.*, 1993.

[48] S. Sulaiman, D. J. You, E. Kanaya, Y. Koga, and S. Kanaya, "Crystal structure and thermodynamic and kinetic stability of metagenome-derived LC-cutinase," *Biochemistry*, vol. 53, no. 11, pp. 1858–1869, Mar. 2014, doi: 10.1021/bi401561p.

[49] I. Karunatillaka, L. Jaroszewski, and A. Godzik, "Novel putative polyethylene terephthalate (PET) plastic degrading enzymes from the environmental metagenome," *Proteins Struct. Funct. Bioinforma.*, 2021, doi: 10.1002/prot.26245.

[50] M. Gyung Yoon, H. Jeong Jeon, and M. Nam Kim, "Biodegradation of Polyethylene by a Soil Bacterium and AlkB Cloned Recombinant Cell," *J. Bioremediation Biodegrad.*, vol. 03, no. 04, 2012, doi: 10.4172/2155-6199.1000145.

[51] S. Sulaiman *et al.*, "Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach," *Appl. Environ. Microbiol.*, vol. 78, no. 5, pp. 1556–1562, Mar. 2012, doi: 10.1128/AEM.06725-11.

[52] 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. John Wiley and Sons Inc, pp. 1706–1716, May 14, 2021. doi: 10.1002/cbic.202000767.

[53] M. Santo, R. Weitsman, and A. Sivan, "The role of the copper-binding enzyme - laccase - in the biodegradation of polyethylene by the

actinomycete Rhodococcus ruber," *Int. Biodeterior. Biodegrad.*, vol. 84, pp. 204–210, Oct. 2013, doi: 10.1016/j.ibiod.2012.03.001.

[54] Y. Iiyoshi, Y. Tsutsumi, and T. Nishida, "Polyethylene degradation by lignin-degrading fungi and manganese peroxidase*," 1998.

[55] A. Henriksen, "Structure of soybean seed coat peroxidase: A plant peroxidase with unusual stability and haem-apoprotein interactions," *Protein Sci.*, vol. 10, no. 1, pp. 108–115, Jan. 2001, doi: 10.1110/ps.37301.

[56] Smits TH, Witholt B, and van Beilen JB, "Functional characterization of genes involved in alkane oxidation by Pseudomonas aeruginosa," *Antonie Van Leeuwenhoek*, vol. 84, no. 3, pp. 193–200, 2003, doi: 10.1023/a:1026000622765.

[57] J. B. Van Beilen *et al.*, "Characterization of two alkane hydroxylase genes from the marine hydrocarbonoclastic bacterium Alcanivorax borkumensis," *Environ. Microbiol.*, vol. 6, no. 3, pp. 264–273, Mar. 2004, doi: 10.1111/j.1462-2920.2004.00567.x.

[58] M. Miri, B. Bambai, F. Tabandeh, M. Sadeghizadeh, and N. Kamali, "Production of a recombinant alkane hydroxylase (AlkB2) from Alcanivorax borkumensis," *Biotechnol. Lett.*, vol. 32, no. 4, pp. 497–502, 2010, doi: 10.1007/s10529-009-0177-0.

[59] Y. Cui *et al.*, "Computational Redesign of a PETase for Plastic Biodegradation under Ambient Condition by the GRAPE Strategy," *ACS Catal.*, vol. 11, no. 3, pp. 1340–1350, Feb. 2021, doi: 10.1021/acscatal.0c05126.

[60] A. Senga *et al.*, "Multiple structural states of Ca 2+ regulated PET hydrolase, Cut190, and its correlation with activity and stability", doi: 10.1093/jb/mvaa102/5901052.

[61] C. Roth *et al.*, "Structural and functional studies on a thermostable polyethylene terephthalate degrading hydrolase from Thermobifida fusca," *Appl. Microbiol. Biotechnol.*, vol. 98, no. 18, pp. 7815–7823, Sep. 2014,

doi: 10.1007/s00253-014-5672-0.

[62] I. Y. Gilan Hadar A Sivan, “APPLIED MICROBIAL AND CELL PHYSIOLOGY Colonization, biofilm formation and biodegradation of polyethylene by a strain of Rhodococcus ruber,” *Appl Microbiol Biotechnol*, vol. 65, pp. 97–104, 2004, doi: 10.1007/s00253-004-1584-8.

[63] M. Dimarogona, E. Nikolaivits, M. Kanelli, P. Christakopoulos, M. Sandgren, and E. Topakas, “Structural and functional studies of a Fusarium oxysporum cutinase with polyethylene terephthalate modification potential,” *Biochim. Biophys. Acta - Gen. Subj.*, vol. 1850, no. 11, pp. 2308–2317, Nov. 2015, doi: 10.1016/j.bbagen.2015.08.009.

[64] D. Danso *et al.*, “New Insights into the Function and Global Distribution of Polyethylene Terephthalate (PET)-Degrading Bacteria and Enzymes in Marine and Terrestrial Metagenomes,” 2018, doi: 10.1128/AEM.

[65] A. Nakamura, N. Kobayashi, N. Koga, and R. Iino, “Positive charge introduction on the surface of thermostabilized PET hydrolase facilitates PET binding and degradation,” *ACS Catal.*, vol. 11, pp. 8550–8564, 2021, doi: 10.1021/acscatal.1c01204.

[66] X. Han *et al.*, “Structural insight into catalytic mechanism of PET hydrolase,” *Nat. Commun.*, vol. 8, no. 1, Dec. 2017, doi: 10.1038/s41467-017-02255-z.

[67] D. Kold *et al.*, “Thermodynamic and structural investigation of the specific SDS binding of humicola insolens cutinase,” *Protein Sci.*, vol. 23, no. 8, pp. 1023–1035, 2014, doi: 10.1002/pro.2489.

[68] R. V Stern and G. T. Howard, “The polyester polyurethanase gene (pueA) from Pseudomonas chlororaphis encodes a lipase,” *FEMS Microbiol. Lett.*, vol. 185, pp. 163–168, 2000, [Online]. Available: www.fems-microbiology.org

[69] C. S. Hung *et al.*, “Carbon catabolite repression and Impranil polyurethane degradation in Pseudomonas protegens strain Pf-5,” *Appl. Environ. Microbiol.*, vol. 82, no. 20, pp. 6080–6090, 2016, doi:

10.1128/AEM.01448-16.

[70] G. T. Howard, B. Crother, and J. Vicknair, “Cloning, nucleotide sequencing and characterization of a polyurethanase gene (pueB) from Pseudomonas chlororaphis,” 2001. [Online]. Available: www.elsevier.com/locate/ibiod

[71] Y. Akutsu, T. Nakajima-Kambe, and N. Nomura, “Purification and Properties of a Polyester Polyurethane-Degrading Enzyme from Comamonas acidovorans TB-35,” 1998.

[72] K. Nakamiya and S. Kinoshita, “Non-Heme Hydroquinone Peroxidase from Azotobacter beijerinckii HM 12 1,” 1997.

[73] A. Bateman, “UniProt: A worldwide hub of protein knowledge,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[74] S. K. Burley *et al.*, “RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D,” *Protein Sci.*, vol. 31, no. 1, pp. 187–208, Jan. 2022, doi: https://doi.org/10.1002/pro.4213.

[75] S. Das and C. A. Orengo, “Protein function annotation using protein domain family resources,” *Methods*, vol. 93. Academic Press Inc., pp. 24–34, Jan. 15, 2016. doi: 10.1016/j.ymeth.2015.09.029.

[76] J. A. Gerlt *et al.*, “Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks,” *Biochimica et Biophysica Acta - Proteins and Proteomics*, vol. 1854, no. 8. Elsevier B.V., pp. 1019–1037, Aug. 01, 2015. doi: 10.1016/j.bbapap.2015.04.015.

[77] S. Zhao *et al.*, “Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks,” *Elife*, vol. 3, Jun. 2014, doi: 10.7554/elife.03275.

[78] S. Srivastava *et al.*, “An efficient algorithm for protein structure comparison using elastic shape analysis,” *Algorithms Mol. Biol.*, vol. 11,

no. 1, Sep. 2016, doi: 10.1186/s13015-016-0089-1.

[79] J. C. Sequeira, M. Rocha, M. M. Alves, and A. F. Salvador, "UPIMAPI, reCOGnizer and KEGGCharter: Bioinformatics tools for functional annotation and visualization of (meta)-omics datasets," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1798–1810, 2022, doi: 10.1016/j.csbj.2022.03.042.

[80] S. Lu *et al.*, "CDD/SPARCLE: The conserved domain database in 2020," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D265–D268, Jan. 2020, doi: 10.1093/nar/gkz991.

[81] N. Karki, N. Verma, F. Trozzi, P. Tao, E. Kraka, and B. Zoltowski, "Predicting potential sars-cov-2 drugs-in depth drug database screening using deep neural network framework ssnet, classical virtual screening and docking," *Int. J. Mol. Sci.*, vol. 22, no. 3, pp. 1–16, Feb. 2021, doi: 10.3390/ijms22031392.

[82] M. Radifar, N. Yuniarti, and E. P. Istyastono, "Software PyPLIF: Python-based Protein-Ligand Interaction Fingerprinting," 2013, [Online]. Available: http://code.google.com/p/pyplif.

[83] B. Zhao, S. Hu, X. Li, F. Zhang, Q. Tian, and W. Ni, "An efficient method for protein function annotation based on multilayer protein networks," *Hum. Genomics*, vol. 10, no. 1, pp. 1–15, 2016, doi: 10.1186/s40246-016-0087-x.

[84] J. A. Gerlt, "Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence-Function Space and Genome Context to Discover Novel Functions," *Biochemistry*, vol. 56, no. 33, pp. 4293–4308, Aug. 2017, doi: 10.1021/acs.biochem.7b00614.

[85] R. Zallot, N. Oberg, and J. A. Gerlt, "The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways," *Biochemistry*, vol. 58, no. 41, pp. 4169–4182, Oct. 2019, doi: 10.1021/acs.biochem.9b00735.

[86] S. R. Eddy, "Hidden Markov models," *Seq. Topol.*, pp. 361–365, 1996, doi: https://doi.org/10.1016/S0959-440X(96)80056-X.

[87] J. S. Bernardes, A. M. R. Dávila, V. S. Costa, and G. Zaverucha, "Improving model construction of profile HMMs for remote homology detection through structural alignment," *BMC Bioinformatics*, vol. 8, pp. 1–12, 2007, doi: 10.1186/1471-2105-8-435.

[88] A. I. Garber *et al.*, "FeGenie: A Comprehensive Tool for the Identification of Iron Genes and Iron Gene Neighborhoods in Genome and Metagenome Assemblies," *Front. Microbiol.*, vol. 11, Jan. 2020, doi: 10.3389/fmicb.2020.00037.

[89] P. Skewes-Cox, T. J. Sharpton, K. S. Pollard, and J. L. DeRisi, "Profile hidden Markov models for the detection of viruses within metagenomic sequence data," *PLoS One*, vol. 9, no. 8, 2014, doi: 10.1371/journal.pone.0105067.

[90] S. Sinha and A. M. Lynn, "HMM-ModE: Implementation, benchmarking and validation with HMMER3," *BMC Res. Notes*, vol. 7, no. 1, pp. 1–11, 2014, doi: 10.1186/1756-0500-7-483.

[91] P. K. Srivastava, D. K. Desai, S. Nandi, and A. M. Lynn, "HMM-ModE - Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences," *BMC Bioinformatics*, vol. 8, pp. 1–17, 2007, doi: 10.1186/1471-2105-8-104.

[92] C. Barrett *et al.*, "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods," *J. Mol. Biol.*, vol. 284, no. 4, pp. 1201–10, 1998, [Online]. Available: http://www.ncbi.nlm.nih.gov/sites/entrez

[93] S. Mørk and I. Holmes, "Evaluating bacterial gene-finding hmm structures as probabilistic logic programs," *Bioinformatics*, vol. 28, no. 5, pp. 636–642, 2012, doi: 10.1093/bioinformatics/btr698.

[94] P. Pérez-García, D. Danso, H. Zhang, J. Chow, and W. R. Streit, "Exploring

the global metagenome for plastic-degrading enzymes," in *Methods in Enzymology*, vol. 648, Academic Press Inc., 2021, pp. 137–157. doi: 10.1016/bs.mie.2020.12.022.

[95] N. A. O'Leary *et al.*, "Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, 2016, doi: 10.1093/nar/gkv1189.

[96] J. Zrimec, M. Kokina, S. Jonasson, F. Zorrilla, and A. Zelezniak, "Plastic-Degrading Potential across the Global Microbiome Correlates with Recent Pollution Trends," *MBio*, vol. 12, no. 5, 2021, doi: 10.1128/mBio.02155-21.

[97] B. Rost, "Twilight zone of protein sequence alignments," 1999. [Online]. Available: https://academic.oup.com/peds/article/12/2/85/1550637

[98] G. G. Z. Silva, K. T. Green, B. E. Dutilh, and R. A. Edwards, "SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data," *Bioinformatics*, vol. 32, no. 3, pp. 354–361, Feb. 2016, doi: 10.1093/bioinformatics/btv584.

[99] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[100]W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, vol. 1, no. Scipy, pp. 56–61, 2010, doi: 10.25080/majora-92bf1922-00a.

[101]C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[102]W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006, doi: 10.1093/bioinformatics/btl158.

[103]C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: A novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.*, vol. 302, no.

1, pp. 205–217, 2000, doi: 10.1006/jmbi.2000.4042.

[104]Anaconda Documentation, “Anaconda Software Distribution.” Anaconda Inc., 2020. Accessed: Feb. 09, 2022. [Online]. Available: https://docs.anaconda.com/

[105]J. H. Lee, H. Yi, Y. S. Jeon, S. Won, and J. Chun, “TBC: A clustering algorithm based on prokaryotic taxonomy,” *J. Microbiol.*, vol. 50, no. 2, pp. 181–185, 2012, doi: 10.1007/s12275-012-1214-6.

[106]B. Niu, L. Fu, S. Sun, and W. Li, “Artificial and natural duplicates in pyrosequencing reads of metagenomic data,” *BMC Bioinformatics*, vol. 11, 2010, doi: 10.1186/1471-2105-11-187.

[107]W. Li, L. Jaroszewski, and A. Godzik, “Tolerating some redundancy significantly speeds up clustering of large protein databases,” *Bioinformatics*, vol. 18, no. 1, pp. 77–82, 2002, doi: 10.1093/bioinformatics/18.1.77.

[108]L. Weizhong, J. Lukasz, and G. Adam, “Clustering of highly homologous sequences to reduce the size of large protein databases,” *Bioinforma. Appl. Note*, vol. 17, no. 8, pp. 282–283, 2000.

[109]S. Eddy, “HMMER user’s guide: biological sequence analysis using prole hidden Markov models,” 1998, [Online]. Available: citeseer.ist.psu.edu/eddy98hmmer.html

[110]F. Mölder *et al.*, “Sustainable data analysis with Snakemake [version 1 ; peer review : 1 approved , 1 approved with reservations],” *F1000Research*, no. May, pp. 1–25, 2021.

[111]A. Delacuvellerie, V. Cyriaque, S. Gobert, S. Benali, and R. Wattiez, “The plastisphere in marine ecosystem hosts potential specific microbial degraders including Alcanivorax borkumensis as a key player for the low-density polyethylene degradation,” *J. Hazard. Mater.*, vol. 380, no. November 2018, p. 120899, 2019, doi: 10.1016/j.jhazmat.2019.120899.

[112]R. Sarkhel, S. Sengupta, P. Das, and A. Bhowal, “Comparative

biodegradation study of polymer from plastic bottle waste using novel isolated bacteria and fungi from marine source," *J. Polym. Res.*, vol. 27, no. 1, 2020, doi: 10.1007/s10965-019-1973-4.

[113]K. I. Kasuya, T. Takano, Y. Tezuka, W. C. Hsieh, H. Mitomo, and Y. Doi, "Cloning, expression and characterization of a poly(3-hydroxybutyrate) depolymerase from Marinobacter sp. NK-1," *Int. J. Biol. Macromol.*, vol. 33, no. 4–5, pp. 221–226, 2003, doi: 10.1016/j.ijbiomac.2003.08.006.

[114]J. Savoldelli, D. Tomback, and H. Savoldelli, "Breaking down polystyrene through the application of a two-step thermal degradation and bacterial method to produce usable byproducts," *Waste Manag.*, vol. 60, pp. 123–126, 2017, doi: 10.1016/j.wasman.2016.04.017.

[115]H. J. Jeon and M. N. Kim, "Isolation of mesophilic bacterium for biodegradation of polypropylene," *Int. Biodeterior. Biodegrad.*, vol. 115, pp. 244–249, 2016, doi: 10.1016/j.ibiod.2016.08.025.

[116]T. Bubpachat, N. Sombatsompop, and B. Prapagdee, "Isolation and role of polylactic acid-degrading bacteria on degrading enzymes productions and PLA biodegradability at mesophilic conditions," *Polym. Degrad. Stab.*, vol. 152, pp. 75–85, 2018, doi: 10.1016/j.polymdegradstab.2018.03.023.

[117]C. Kato, A. Honma, S. Sato, T. Okura, R. Fukuda, and Y. Nogi, "Poly 3-hydroxybutyrate-co-3-hydroxyhexanoate films can be degraded by the deep-sea microbes at high pressure and low temperature conditions," *High Press. Res.*, vol. 39, no. 2, pp. 248–257, 2019, doi: 10.1080/08957959.2019.1584196.

[118]A. Maurya, A. Bhattacharya, and S. K. Khare, "Enzymatic Remediation of Polyethylene Terephthalate (PET)–Based Polymers for Effective Management of Plastic Wastes: An Overview," *Front. Bioeng. Biotechnol.*, vol. 8, no. November, pp. 1–13, 2020, doi: 10.3389/fbioe.2020.602325.

[119]A. Sivan, "New perspectives in plastic biodegradation," *Curr. Opin. Biotechnol.*, vol. 22, no. 3, pp. 422–426, 2011, doi: 10.1016/j.copbio.2011.01.013.

[120] M. Madera and J. Gough, “A comparison of profile hidden Markov model procedures for remote homology detection,” *Nucleic Acids Res.*, vol. 30, no. 19, pp. 4321–4328, 2002, doi: 10.1093/nar/gkf544.