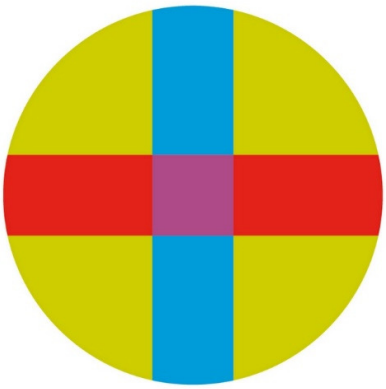
UNIVERSITY CEU - SAN PABLO

POLYTECHNIC SCHOOL

BIOMEDICAL ENGINEERING DEGREE



BACHELOR THESIS

Computational workflow for the identification of antimicrobial targets

Author: Noelia Aubá Arribas
Supervisors: Carlos Óscar Sorzano Sánchez and Javier Tejedor Noguerales

June 2026

UNIVERSIDAD SAN PABLO-CEU
ESCUELA POLITÉCNICA SUPERIOR
División de Ingeniería

Datos del alumno

NOMBRE: Noelia Aubá Arribas

Datos del Trabajo

TÍTULO DEL PROYECTO: Computational workflow for the identification of antimicrobial targets

Tribunal calificador

PRESIDENTE:	FDO.:
SECRETARIO:	FDO.:
VOCAL:	FDO.:

Reunido este tribunal el _____/_____/_____, acuerda otorgar al Trabajo Fin de Grado presentado por **Doña Noelia Aubá Arribas** la calificación de _____.

ACKNOWLEDGMENTS

May be written in Spanish

ABSTRACT

The widespread multidrug-resistant bacteria are considered a global public health threat, increasing the demand for innovative, precise and automated workflow approaches for the identification of antimicrobial targets. Regarding this, computational subtractive genomics provides an efficient strategy based on the reduction of complete bacterial proteomes into a smaller and biologically relevant subset of potential candidate therapeutic targets.

This thesis presents the implementation and optimization of a bioinformatics workflow based on the integration of multiple protein filtering stages, including plasmid-encoded filtering, the retention of transmembrane proteins based on DeepTMHMM topology prediction, BLASTp-based filtering for homology against the human proteome and protein essentiality, and finally, clustering through CD-HIT for the retention of those proteins that are similar and showed conservation across organisms. The combination of these criteria aims to reduce the initial multidrug-resistant bacterial dataset for the subsequent prioritization of candidate proteins based on eggNOG functional annotations for the identification of potential targets that are essential for bacterial growth and survival while minimizing potential off-target effects.

The workflow analysis handled thousands of protein sequences across multiple proteomes using Bash scripting for the orchestration and execution of the sequential steps provided in Python-based bioinformatics scripts, resulting in a prioritization of membrane-associated and transport proteins by the ranking of functional and biological features.

RESUMEN

El aumento de la propagación de bacterias multirresistentes se trata hoy en día como un peligro global de salud pública, lo que provoca un aumento de la demanda de flujos de trabajo innovadores y automatizables para la identificación de dianas antimicrobianas. Como resultado, se emplea la genómica sustractiva computacional para reducir proteomas bacterianos completos a un subconjunto más pequeño y biológicamente relevante como posibles dianas terapéuticas.

Este trabajo se centra en la implementación y optimización de un flujo de trabajo bioinformático basado en la integración de múltiples filtros de proteínas de forma secuencial, que incluyen el filtrado de secuencias interpretadas como plásmidos, la retención de proteínas transmembranas basadas en la predicción de topología de DeepTMHMM, el filtrado a partir de BLASTp para la homología de estas proteínas frente al proteoma humano y su esencialidad y por último la agrupación mediante CD-HIT para retener aquellas proteínas que son similares y están conservadas entre organismos. La combinación de todos estos criterios tiene como objetivo principal reducir el conjunto inicial de proteínas asociadas a bacterias multirresistentes para la posterior priorización de estas proteínas candidatas, a partir de las anotaciones funcionales aportadas por eggNOG para la obtención de potenciales dianas terapéuticas que se muestran como esenciales para el crecimiento y la supervivencia bacteriana, minimizando al mismo tiempo los posibles efectos fuera de diana (off-target).

El análisis del flujo de trabajo procesó miles de secuencias bacterianas a través de múltiples proteomas utilizando scripts en Bash para la orquestación y ejecución de los pasos secuenciales proporcionados en scripts bioinformáticos de Python, obteniendo potenciales dianas proteicas asociadas al transporte y a la membrana mediante la priorización y clasificación de sus características funcionales y biológicas.

INDEX

FIGURE INDEX

TABLE INDEX

LIST OF ABREVIATIONS

Abbreviation	Terminology	Meaning
OM	Outer Membrane	External lipid membrane present in Gram-negative bacteria, which provides protection and selective permeability.
PBP	Penicillin-Binding Proteins	Enzymes involved in bacterial cell-wall synthesis that serve as therapeutic targets.
DNA	Deoxyribonucleic Acid	Molecules that store genetic instructions for the development of organisms.
mRNA	Messenger Ribonucleic Acid	Temporary molecules that carry genetic instructions from DNA to the ribosome for protein synthesis.
AMR	Antimicrobial Resistance	Ability to survive despite the exposure to antibiotics or antimicrobial treatments.
WHO	World Health Organization	International agency responsible for global public health and disease control.
NCBI	National Center for Biotechnology Information	American biomedical research database repositor that acts as a central global hub for molecular biology, genetics and healthcare data.
AA	Amino Acid	Fundamental building blocks of proteins.
GO	Gene Ontology	Standardized system to describe biological functions, processes, and cellular locations of genes and proteins.
KEGG	Kyoto Encyclopedia of Genes and Genomes	Bioinformatics database that integrates genomic, chemical and systemic functional information.
KO	KEGG Orthology	Functional annotation system that assigns genes to metabolic pathways through orthologous groups.

COG	Clusters of Orthologous Groups	Classification that groups proteins from different species based on shared evolutionary history and function.
HGT	Horizontal Gene Transfer	Non-reproductive transmission of genetic material between different organisms.
RAM	Resistance-Associated Mutation	Genetic mutations that lead to antimicrobial resistance.
SP	Signal Peptide	A short amino acid terminal signal that directs proteins to their specific cellular destination.
TM	Transmembrane Region	Protein sequence characterized by spanning the membrane and anchoring the protein within the lipid bilayer.
DEG	Database of Essential Genes	Repository of genes experimentally essential for organism survival.
ATP	Adenosine Triphosphate	Fundamental molecules used by cells for the storage and transfer of chemical energy for metabolic processes
ABC	ATP-Binding Cassette	Transmembrane protein that uses energy from ATP hydrolysis to transport substrates across the cellular membrane
HMM	Hidden Markov Models	Statistical model to predict a sequence of unknown events based on observed sequences.

1 INTRODUCTION

1.1 Development of Antimicrobial Drugs

The development of antimicrobial drugs in modern medicine started with the discovery of penicillin in 1928 by Alexander Fleming, which led to a wide range of antimicrobial agents. Over the last century, these antimicrobial agents have significantly transformed the treatment of infectious diseases into efficiently treatable and controllable bacterial infections. Prior to this turn, infectious diseases were linked to high morbidity and mortality rates due to the lack of therapeutic options or their extremely limited use [1].

However, the widespread use of antimicrobials across the global population has been introduced as a significant challenge, regarding their safety and appropriate use. Beyond this concern, adverse effects and drug toxicity have gradually emerged as a relevant public health threat, highlighting the necessity of balance the associated risks with the therapeutic benefits of antimicrobials. Consequently, understanding how antibiotics act at the cellular level is essential for optimizing their use and addressing these therapeutic challenges [2].

1.1.1 Antimicrobial Resistance

In this regard, some of the most important health problems facing today's world include antimicrobial resistance (AMR), which involves the ability of microorganisms to resist the exposure to therapeutic agents that in some cases were previously effective, leading to a reduction in treatment efficacy while increasing the risk of treatment failure. AMR can be classified into two types, the intrinsic resistance which is the natural insensitivity of certain microorganisms to specific antibiotics due to their physiological or structural characteristics. The acquired resistance is defined as a condition developed after the exposure to antimicrobial compounds and arises from genetic changes, that may occur through mutations in the bacterial genome during the replication process or through horizontal gene transfer mechanisms such as transformation, transduction, and conjugation.

Priority pathogens identified by the World Health Organization (WHO) face extremely limited effective antibiotic therapies and require innovative antimicrobial compounds for the treatment of infections because of the prevalence and spread of multidrug resistance. These bacterial organisms have been treated as a real threat to human health and are classified based on the requirement for developing novel antibiotic treatments: The critically important pathogens identified are *Acinetobacter baumannii, Pseudomonas aeruginosa,* and members of the *Enterobacteriaceae* family. The high-priority pathogens include *Enterococcus faecium, Staphylococcus aureus, Helicobacter pylori, Campylobacter, Salmonella spp., and Neisseria gonorrhoeae* while the medium-priority pathogens include *Streptococcus pneumoniae and Shigella spp* [3].

However, the clinical implications of AMR are significantly severe in life-threatening infections, even though the prevalence of these infections is less common than mild infections in developed countries. The major reason behind this problem is that a large proportion of antimicrobials prescribed occur within outpatient care settings, making its appropriate and judicious use important. Overprescription for unnecessary treatment does not provide any clinical advantage but may produce adverse effects while contributing to the spread of such resistant organisms. This issue is especially critical in severe infections caused by Gram-negative bacteria, which frequently exhibit high levels of resistance [4].

1.1.2 Mechanisms of Action of Antibiotics

Antibiotics exert their therapeutic effect by targeting essential cellular processes in bacteria, either inhibiting the growth or leading to cell death. Bacteria can be divided according to the structure of their cell envelope into Gram-positive bacteria which consist of a cytoplasmic membrane enclosed within a rigid peptidoglycan cell wall layer. As it is shown in Figure 1, Gram-negative bacteria expose a thin peptidoglycan layer that can be found between the cytoplasmic membrane and an additional lipid bilayer known as the outer membrane (OM). This OM acts as a protective barrier that prevents the entry of many antimicrobial agents into the cell, although it contains pores,(known as porin channels) which allow the entrance of small particles into the cell [5].

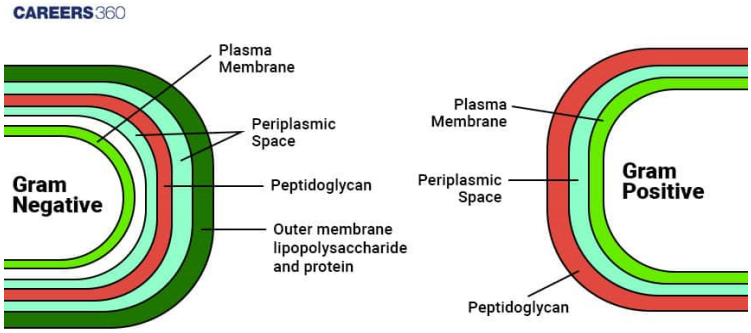


Figure 1. Structure of the bacterial cell envelope. Comparison between Gram-negative (left) and Gram-positive (right) cell walls, highlighting the multi-layered OM in Gram-negative bacteria [2].

The principal idea of the antibacterial mechanisms is to lead to the inhibition of cell wall synthesis and alterations of cell membrane integrity. These mechanisms of action are commonly grouped based on their pharmacological action. One of the oldest and most relevant antibiotics is penicillin, a member of β -lactam family that produces the inhibition of cell walls sythesis. This antibacterial family inhibits the formation of peptidoglycan by binding to penicillin-binding proteins (PBPs), causing a weakness in the bacterial structure or leading to cell lysis, as PBPs are essential enzymes involved in cell wall cross-linking [5, 6].

While the inhibition of essential metabolic functionalities and the interference with nucleic acid synthesis represent key antimicrobial pathways. The disruption of protein synthesis remains as a critical mechanism, where antibiotics act through the bind to bacterial ribosomes. Bacterial DNA carries coded information which is transcribed into Messenger Ribonucleic Acid (mRNA) which will be later translated through ribosomes into proteins. Ribosomes are composed of two subunits, 30S and 50S. The drugs inhibit this process by binding to these subunits and modifying translation. Classical antimicrobials such as aminoglycosides and tetracyclines inhibit the 30S subunit by interfering with mRNA reading, while chloramphenicol and macrolides inhibit the 50S subunit, inhibiting peptide bond formation [5, 6].

The increased threat of AMR makes necessary the design and development of new drugs, known as novel antimicrobials, that act on essential biological processes of the bacteria. There are various mechanisms that contribute to the generation of antimicrobial resistance, including enzymatic degradation of the antibiotics, modification of drug targets and activation of efflux systems (see Figure 2). Among these, one key mechanism is the inhibition of efflux systems, which pump the antimicrobial compounds out of the cytoplasm using specific membrane proteins called efflux pumps. Efflux mechanisms provide an important protective role for the bacteria that aim to reduce the intracellular antibiotic concentrations [7].

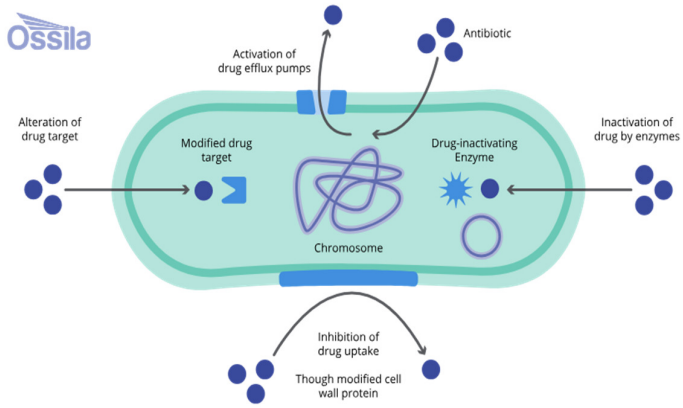


Figure 2. Schematic representation of antimicrobial resistance mechanisms [8]

Another emerging antimicrobial strategy focuses on essential bacterial pathways such as protein translocation pathways due to their cell functionality. Several stages of the translocation process can be targeted, including substrate recognition, the delivery of proteins to the OM, the insertion and assembly of membrane proteins, and the release and folding of newly synthesized proteins [9, 10].

1.1.3 Limitations of Traditional Drug Discovery

Following the discovery of penicillin, a wide variety of antimicrobial compounds were identified through natural product screening and observational studies. Additionally, many other agents were developed through chemical modification of existing compounds. These discoveries led to the establishment of major classes of antibacterial drugs, each characterized by distinct mechanisms of action. However, the discovery of entirely new antibiotic classes has significantly declined, and pharmaceutical development has largely focused on the introduction of new derivatives within existing antimicrobial classes to improve their efficacy and overcome the resistance observed in previously used agents[11].

There are several reasons why antibiotic research and development remains costly and highly complex. Strong competition with existing drugs requires new compounds to demonstrate clear advantages in terms of efficacy, safety or resistance. In addition, the development of a new drug typically requires 10–15 years and substantial financial investment. Experimental screening of large chemical libraries is also limited in scale, restricting the number of compounds that can be evaluated and reducing overall efficiency [12].

The risk that a new drug will fail to demonstrate efficacy compared to existing therapies is considerable, particularly during the late stages of clinical development. This challenge is increased by the rapid ability of bacteria to evolve and acquire mutations, increasing their AMR by reducing their long-term effectiveness. As a result, newly developed drugs may lose efficacy, increasing the risk of failure.

The success rate of drug development is also extremely low. It is estimated that approximately 90% of the compounds entering Phase I clinical trials fail to reach regulatory approval, reflecting the high level of uncertainty associated with this process. Consequently, for the small proportion of drugs that are ultimately approved, the overall development cost per compound is estimated to be around USD 1.4 billion [13].

Such high costs, along with the relatively low financial profit associated with antibiotics compared to those for chronic diseases, have led to a reduction in the efforts made by pharmaceutical companies in investing in new antimicrobial research. This means that the development of novel antimicrobial compounds has slowed considerably in the past decades, creating an urgent need for alternative and innovative approaches to the development of personalized antimicrobial agents.

Here lie some of the reasons why computational methods have emerged as powerful tools, as they have assisted in the acceleration of drug selection, reductionof costs and improvement of the efficiency and precision of candidate target selection.

1.1.4 Computational Drug Design Paradigm

Computational drug discovery has emerged because of the advances in computer hardware and software, enabling the efficient identification and prioritization of candidate molecules from large chemical and biological datasets. Virtual screening techniques allow the exploration of chemical libraries to identify potential drug candidates based on their interaction with target proteins, reducing the number of compounds subjected to in vitro experiments and consequently reducing critical aspects, such as time and cost [14].

A key concept in modern drug discovery is the use of integrated computational workflows, in which multiple bioinformatics techniques are applied sequentially. These workflows implement successive filtering steps to identify compounds with high biological activity and favorable pharmacological properties, enabling the early elimination of unsuitable targets and increasing the efficiency of the drug discovery process. Within this framework, this project adopts an integrated computational workflow to prioritize candidate compounds, following a data-driven strategy for the identification of potential antimicrobial agents.

1.2 Antimicrobial Targets

Antimicrobial drug targets refer to biological structures or essential processes required for the survival or proliferation of microorganisms that can be selectively inhibited by therapeutic agents. Proteins remain the major target class in antimicrobial drugs, due to their key role in many biological processes and their interactions with small molecules through binding sites. The selection of an appropriate target is a crucial step, as it directly determines the efficacy, selectivity, and safety of the resulting drug. Suitable targets should be essential for microbial survival, sufficiently different from host organisms in order to minimize toxicity, and exhibit suitable druggability, which refers to the ability of a target molecule to bind to drug-like compounds [15].

The methodological advances described in section 1.1.4 have significantly boosted drug discovery, complementing traditional wet-lab approaches. Approaches implemented on drug design and virtual screening have accelerated the overall discovery process by the identification of candidate protein molecules in relation to how they interact with macromolecular targets [16].

1.2.1 Protein Targeting

Protein-based drug discovery involves various stages, which include target identification, selection and validation, the usage of in vitro and in vivo experiments to determine leads and candidates, preclinical testing in animals and finally testing the drug's safety and efficacy in human clinical trials.

Apart from druggability, certain characteristics should be also considered in target selection and analysis. For instance, essentiality indicates whether a protein is necessary for the survival or growth of microorganisms, suggesting that the inhibition of this target may produce the desired bactericidal response. Specificity and selectivity are another important criterion for evaluating a target, in the sense that inhibits the bacterial cell without affecting the host or its microbiome. Finally, the biological relevance and functionality of the target during the pathogen’s metabolic and infection stages must be considered.

Once these suitable targets have been identified, their activity can be modulated through different mechanisms (see Figure 3). In antimicrobial drug discovery, the most employed is the enzymatic inhibition in antimicrobial drug discovery, where small molecules bind to the target protein and disrupt essential biological processes required for pathogen survival. Other mechanisms of modulation include the receptor activation or inhibition (agonist or antagonist), as well as modulation of ion channels that regulate the flow of ions across cell membranes [17, 18].

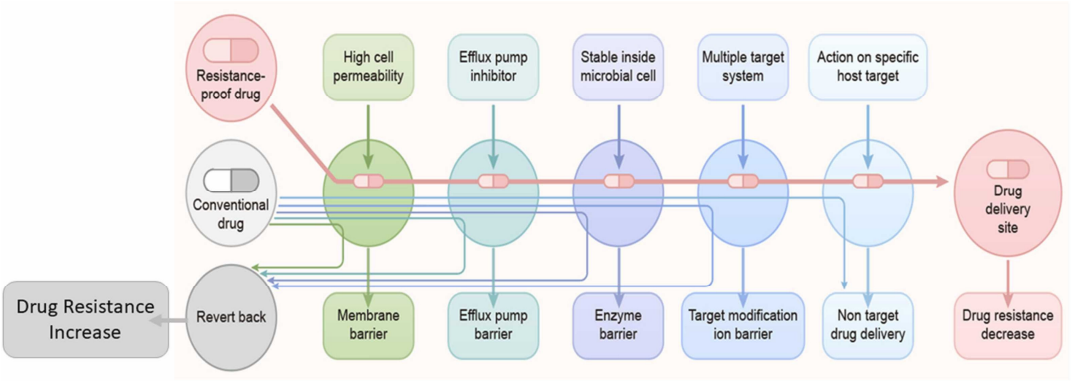


Figure 3. Representation of mechanisms of modulation [4].

The consideration of these features contributes to minimizing potential side effects while enhancing the probability of identification of effective antimicrobial therapies. A clear understanding of the target properties, their functionalities and their mechanisms of modulation are crucial for drug development. The identification and selection of specific protein targets represent a key role that influences the success of computational approaches.

1.2.2 Transmembrane Proteins as Targets

Transmembrane proteins (TM) are the proteins which are embedded within the lipid bilayer of cell membranes, providing the organism with transport roles, signal transduction functions and membrane integrity upkeep, making them necessary for the survival of bacteria. These properties make TM protein structure an excellent target for antimicrobial drug development. Membrane proteins are divided into two major classes, depending on their secondary structure's stability within the membrane: α-helical which are predominant among TM proteins and are mostly found in the inner membrane of bacteria, where they participate in transport and energy conversion; And β -barrel proteins typically localized within the OM, and composed of antiparallel β-strands, forming a cylindrical structure and providing pore-like channels (see Figure 4). It should be noted that α-helical proteins appear in Gram-negative and Gram-positive bacteria as well as in eukaryotes, whereas β-barrel proteins are mainly in OM of Gram-negative bacteria [19].

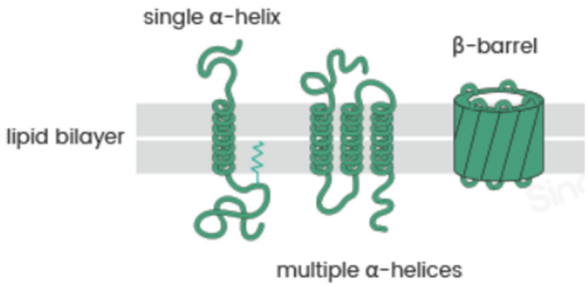


Figure 4. Representation of TM protein's structure [19].

The envelope of Gram-negative bacteria (see Figure 1) shows a complexity of structure which represents a challenging target in antimicrobial drug discovery, where the OM is exposed as a highly selective barrier that restricts the influx of antimicrobial compounds, thereby reducing drug permeability. Over this membrane β-barrel proteins like porins and α-helical as efflux pumps play an important role in the regulation of molecule transport. While porins promote the diffusion of small hydrophilic compounds in a passive manner, active expulsion is provided by efflux pumps which aid in pumping antibiotics from the cell, contributing to multidrug resistance. Consequently, targeting these proteins can offer new approaches to fight against Gram-negative resistant bacteria [20].

1.2.3 Subtractive Genomics for Target Identification

Experimental drug development is highly expensive and time-consuming. For this reason, complementary strategies such as bioinformatics and computational analyses have become essential for the identification of proteins as potential molecular targets. These approaches focus on the evaluation of potential proteins criterion like their localization, topology, essentiality or human homology, as it can be seen in Figure 5 [18].

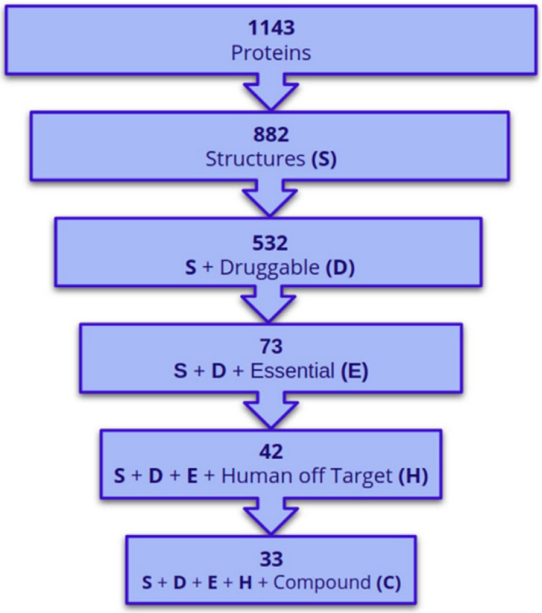


Figure 5. Sequential filtering of candidate targets based on the localization, topology, essentiality, and low similarity to human proteins for target prioritization [21].

Across these approaches, subtractive genomics technique has emerged as a powerful bioinformatics strategy for the identification of potential therapeutic targets. This approach distinguishes potential genes in pathogens from non-essential or host-homologous genes, reducing the initial dataset and enabling the identification of unique targets that are critical for pathogen survival. It has also facilitated the development of computational pipelines for target identification, representing a crucial step in drug discovery. This is particularly relevant in infectious diseases, where proteins involved in essential metabolic and cellular pathways that are unique to the pathogen can be identified, reducing the risk of host toxicity [22].

1.3 Objectives of the project

The principal objective of this project is to implement and optimize a computational workflow based on subtractive genomics strategy for systematic filtering prioritization and identification of potential protein targets in multidrug-resistant bacteria. The subtractive genomics approach applied in the workflow integrates multiple biological and pharmacological criteria for target filtering, which principally include essentiality, non-homology to the human reference proteome, subcellular localization and conservation across organisms. The sequential application of these filters allows the progressive reduction of candidate proteins that facilitate the identification of smaller set, enabling the prioritization of these potential targets on the workflow. Subsequently, potential candidates are ranked according to their functional relevance, considering their involvement in membrane-associated functions and essential biological processes required for pathogen survival, adaptation and virulence.

This workflow is based on a previously described bioinformatic strategy and is adapted into a reproducible and scalable pipeline to reduce large bacterial proteome datasets into smaller sets of potential therapeutic targets [23].Overall, this project aims to provide a reproducible multi-layered computational workflow for the identification and functional prioritization of proteins as potential antimicrobial targets, contributing to the development of innovative therapeutic strategies to address the growing challenge of antimicrobial resistance

2 MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Software environment

A Linux-based system (Ubuntu 22.04 LTS) using the Windows Subsystem for Linux (WSL) was used to perform this project. The computational workflow combined Bash and Python scripts. For the pipeline a Bash script was used to orchestrate and execute the different analysis steps through the call of different Python scripts that handle data processing, filtering, and sequence analysis.

The construction of the pipeline and scripts was conducted using Visual Studio Code, which is a cross-platform code editor. Most stages of this workflow implemented Python 3.10 in a specific virtual environment (**pipeline_env**), where dependencies were installed with fixed versions to minimize potential package conflicts. This Python version was chosen based on its compatibility with the required libraries as well as its stability on the Ubuntu 22.04 platform. A detailed list of the libraries used is provided in Table 1

Library	**Purpose in this study**
biopython	Retrieval and processing of biological sequence data. The Entrez module enabled access to National Center for Biotechnology Information (NCBI) databases for the download of FASTA files and proteomic datasets, while sequence handling tools were used for sequences reading, filtering and writing [24].
matplotlib	Generation of heatmaps and graphical representations to visualize the impact of each filtering step within the subtractive genomics approach [25].
NumPy	Core numerical library handling N-dimensional arrays and mathematical operations used for data visualization of each filtering step [26].

requests	Python library employed for online accessing of biological resources and obtention of external data required throughout the workflow [27].

Table 1. Implemented libraries in **pipeline_env.**

A separate virtual environment (**deeptmhmm_env**) was configured to isolate the execution of DeepTMHMM tool, using a local installed academic version of the software (DeepTMHMM 1.0 - Academic Version). This separation was necessary due to specific dependency constraints and helped to prevent conflicts with the main pipeline environment. Minor adjustments were made during the installation of the software to ensure proper integration with the pipeline and compatibility with the system. These modifications are documented in the setup section of the project README file to ensure this project reproducibility.

This environment includes additional dependencies required for deep learning-based protein sequence analysis. While the core pipeline libraries are summarized in Table 1, this environment incorporates specialized supplementary libraries for the *predict.py* execution of DeepTMHMM software. The complete list of additional dependencies and their purposes is provided in Table 2.

Library/models	Purpose in this study
PyTorch	Provides GPU and CPU execution of deep learning models used in DeepTMHMM for TM topology prediction [28].
fair-esm	Generation of protein sequence embeddings required by DeepTMHMM [29].
PeptideBuilder	Construction of peptide and protein structures for downstream structural analysis [30].

Table 2. Additional libraries implemented in **deeptmhmm_env.**

Lastly a third virtual environment (**eggnog_env**) was configured to isolate the execution of eggNOG-mapper, using a local installed software (version 2.1.13). This separation was necessary, as DeepTMHMM environment, to prevent conflicts with the main workflow due to specific dependencies. This isolated environment incorporates specialized supplementary libraries, for the execution of eggNOG-mapper. The complete list of additional dependencies is provided in Table 3.

Library/models	**Purpose in this study**
eggNOG-mapper	Used to assign functional annotation to proteins sequences
XlsxWriter	Used for the generation of spreadsheet reports containing the functional annotation by eggNOG-mapper tool

Table 3. Additional libraries in **eggnog_env.**

2.1.2 External Software and Tools

The computational workflow relied on various external bioinformatics tools to perform sequence analysis of the bacterial proteome. These tools were integrated at different stages of the pipeline to ensure accurate filtering, annotation, and prioritization of candidate targets following a subtractive genomics strategy. A summary of the external software employed, including their specific roles within the computational workflow, is provided in Table 4.

External Software	**Purpose in this study**
DeepTMHMM-Academic V 1.0	A modern bioinformatics tool for the categorization and prediction of TM helices and overall protein topology using deep neural networks. DeepTMHMM learns patterns in protein sequences and generates embeddings that help to determine the representation of each protein and whether it lies within the bacteria, while ensuring the prediction followed by real biological rules for alignments [31].
BLASTp-V 2.12.0+	Sequence similarity analyses from the NCBI BLAST+ suite were used to filter non-homologous and essential proteins.

	The software employs a heuristic approach to detect local alignments by identifying short matching regions between query and database amino acids sequences, which are then extended to form longer alignments. The statistical significance of each alignment is assessed using the e-value, which represents the expected number of matches with a similar score that could occur by chance in a database [32].
CD-HIT- **V 4.8.1**	A bioinformatics tool employed for the cluster of protein sequences based on sequence identity thresholds to generate non-redundant proteome datasets. CD-HIT employs a short amino acid (AA) word filtering to discard sequence pairs that are unlikely to be similar, computing identity only for those that pass this check. Sequences that exceed the identity threshold are grouped into the same cluster, retaining a single representative sequence per cluster to reduce redundancy. Sequences that fail to match any existing cluster initiate the formation of a new cluster [33].
eggNOG-mapper- **V 2.1.13**	A bioinformatics tool that performs large-scale functional annotations of protein sequences through an orthology-based approach using the eggNOG database, which contains evolutionary groups of proteins derived from different organisms that share common ancestral genes. eggNOG-mapper employs sequence alignment algorithms such as DIAMOND and MMseq2 to provide Gene Ontology (GO) terms, KEGG Orthology (KO) and functional categories (COG), based on the assumption that orthologous proteins retain similar functions across species, supporting the biological characterization [34] [35]

Table 4. External software

2.1 METHODS

2.1.1 Computational Workflow overview

The computational workflow implemented in this study was based on a previously published bioinformatic approach available on bioRxiv, titled: "*A bio-informatic approach to identify new drug targets in multidrug-resistant bacteria"* [23]. The original methodology described on the pipeline was adapted and optimized to meet the specific requirements and objectives of this project. These modifications were introduced to improve automation, ensure compatibility with updated databases, and integrate additional control and validation steps. These adjustments allowed the pipeline to operate more efficiently and to produce results paired with the biological context.

As illustrated in Figure 6, the complete pipeline was organized into a block structure composed of four main stages: (i) proteome sequence acquisition from NCBI database, (ii) protein filtering using subtractive genomics strategy, (iii) target identification and prioritization, and (iv) data analysis and validation. Each module uses the output generated by the previous one, gradually reducing the initial dataset through successive biological filters. These steps narrow the search space from thousands of proteins sequences to a refined and manageable subset of candidate protein targets, improving interpretability and enabling focused analyses.

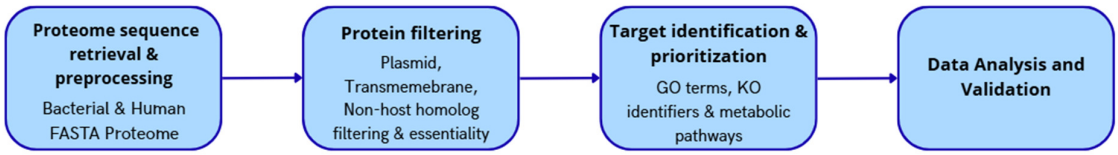


Figure 6. Visual pipeline scheme of algorithm development steps.

2.1.2 Proteome sequence acquisition

Bacterial proteome datasets were downloaded from the NCBI database, a public repository for molecular biology and genomic data, using the accession number corresponding to the selected organism. The use of accession numbers, rather than scientific names ensures the accurate identification of a specific genome assembly, avoiding ambiguity and improving reproducibility of the analysis. Since accession numbers are linked to genome assemblies that may be updated or reannotated over time, the version available at the time of data retrieval was used throughout this study [36].

Data retrieval was performed with the Python script **download_ncbi.py,** developed to retrieve sequences from the *RefSeq* collection, which provides curated and standardized datasets with high-quality annotations suitable for comparative studies.

Protein sequences from resistant bacteria organisms were downloaded from the NCBI database in compressed format (gz), decompressed and used in the downstream analyses within the pipeline in FASTA format. Alongside the bacterial proteomes, the human proteome (GCF_000001405.40) was consistently obtained from the NCBI database and used as a reference dataset for subsequent filtering steps. The 21 bacterial proteomes included in this study are summarized in Table 5, arranged alphabetically by species name.

Species Name	**Accession Number**
Acinetobacter baumannii	NZ_CP015121.1
Campylobacter jejuni	NC_002163.1
Enterobacter cloacae	NZ_CP009756.1
Enterococcus faecium	NZ_CP039729.1
Escherichia coli O157	NC_002695.2
Haemophilus influenzae	NZ_CP009610.1
Helicobacter pylor	NC_017379.1

Klebsiella pneumoniae	NC_016845.1
Morganella morganii	NZ_CP034944.1
Mycobacterium tuberculosis	NC_000962.3
Neisseria gonorrhoeae	NZ_CP012028.1
Salmonella typhimurium	NC_003197.2
Serratia marcescens	NZ_CP063354.1
Shigella dysenteriae	NZ_CP061527.1
Shigella flexneri	NC_004337.2
Staphylococcus aureus	NC_007795.1
Streptococcus pneumoniae	NZ_CP007593.1
Proteus mirabilis	NC_010554.1
Providencia rettgeri	NZ_CP029736.1
Providencia stuartii	NC_017731.1
Pseudomonas aeruginosa	NC_002516.2

Table 5. List of the 21 bacterial species analyzed in this project and their corresponding NCBI accession numbers.

2.1.3 Bacterial Proteome Filtering

This stage represents the core of the computational workflow and implements the subtractive genomics strategy to refine the complete bacterial proteome. The filtering process was designed as a sequential pipeline in which irrelevant proteins were progressively removed based on biological and functional criteria, allowing a reduction of the initial bacterial dataset toward a subset of meaningful target candidates.

As the first filtering step, proteins annotated as plasmid-derived in FASTA headers were excluded from the analysis to eliminate sequences associated with plasmid DNA. The decision was based on the biological characteristics of plasmid-encoded proteins, which can be easily gained or lost between strains and are therefore highly variable and not consistently present. In addition, these encoded proteins often originated through gene transfer, making them less suitable and reliable drug targets due to their instability. To implement this filtering step, the Python script named **plasmid_filter.py** was created to remove sequences annotated as plasmid-derived based on the "[*plasmid*]" label present in the FASTA headers. Since all datasets were obtained from the *RefSeq* collection, which provides curated proteome annotations, this label was considered a reliable indicator of plasmid origin. As a result, this filter ensures the retention of DNA chromosomally encoded proteins, while plasmid-associated sequences are excluded allowing the subsequent analyses to focus on conserved and biologically relevant targets.

Following the plasmid filtering step, DeepTMHMM was used to perform TM topology prediction on the remaining protein sequences. Although the original pipeline employed Phobius tool, its use was not feasible in this project due to the absence of an available API for automated large-scale predictions. In addition, the local version relies on outdated binaries that are not compatible with modern software systems, making its integration into an automated workflow impractical. Alternative tools, such as TMHMM and SignalP 4.0, were also evaluated; however, similar limitations were encountered, preventing their automated implementation in this study. Consequently, DeepTMHMM was selected as a suitable alternative, as it provides a modern, fully scriptable implementation capable of handling large datasets within an automated pipeline [31].

DeepTMHMM was executed in CPU mode due to hardware limitations. Although GPU execution is recommended for transformer-based and deep learning models, because of their reliance on parallel computations, this limitation only increased processing time without affecting prediction accuracy. Given the high memory requirements and initialization cost of transformer-based models, needed to employ DeepTMHMM prediction, the dataset was divided into manageable subsets of 150 sequences. This methodology ensured a balance between computational stability and RAM usage, while allowing partial re-execution of the pipeline without processing the complete proteome again. DeepTMHMM generates multiple output files, including *predicted_topologies.3line, deeptmhmm_results.md* and *TMRs.gff3*. Among these, the *predicted_topologies.3line* file was selected for analysis as it provides detailed AA-level topology predictions. In this file, each AA is assigned to a structural membrane localization. These annotations enable the classification of protein sequences into categories with labels such as *TM, SP, BETA, TM+SP*, or *GLOB*, according to their predicted structural features.

Additionally, when individual protein sequences are computed separately, DeepTMHMM generates graphical probability plots (see Figure 7), showing residue-level topology predictions and confidence scores across the AA sequence to facilitate visual interpretation.

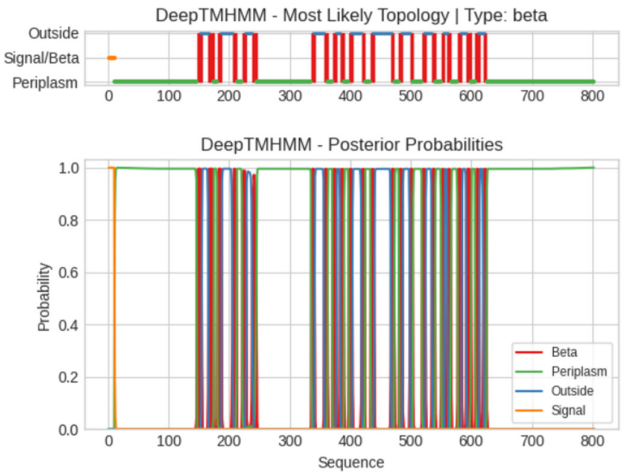


Figure 7. Example of a graphical probability plot generated by DeepTMHMM for a single protein sequence (RefSeq: NC_017731.1)

To further filter the dataset, the script **transmembrane_filter.py** was used to retain sequences containing TM labels based on DeepTMHMM predictions, while sequences labelled as signal peptides (SP) or combined categories such as *TM+SP, GLOB* or *BETA* were excluded. The reasoning behind this decision was based on the lowe reliability in SP predictions, as SP corresponds to short N-terminal that target proteins for secretion pathways or membrane-associated transport systems rather than representing true TM helices. This inclusion would be misleading since it might introduce false positive results in the determination of potential TM proteins for its analysis.

After TM proteins were selected, sequence similarity analysis was performed using BLASTp to determine the homology of the bacterial candidate proteins to the human proteome. This step was executed through the **blastp.py** script, which was designed as a reusable component within the computational workflow and applied at different filtering stages of the pipeline. In this specific filtering step, the script was used to identify proteins showing similarity to human proteins to exclude host homologues proteins in therapeutic target selection. For the execution of BLASTp, an E-value threshold of 0.001 was applied to retain only statistically meaningful alignments with an E-value lower than 0.001, reducing non-significant matches and ensuring the reliability of the software execution doing alignments.

Homology in proteins was determined using the established criteria defined in the published pipeline, where if it showed an E-value ≤ 0.001, already applied during BLASTp execution and a sequency identity of at least 35% the proteins is declared homologous. This sequence identity filtering step was implemented through the script **blastp_filter.py**, which automatically processed BLASTp output files to identify and extract the desired proteins. Proteins that met these criteria were classified as human homologous proteins, and only the non-homologous proteins were retained for later processing stages. This criterion is applied in accordance with the principles of subtractive genomics, as the proteins that share significant similarity with host proteins may lead to undesired off-target effects, brought by the interactions not only with bacterial proteins but also with homologous host proteins, causing toxicity or adverse side effects.

By applying this exclusion-based filtering strategy shown in Equation 1, the pipeline prioritizes targets that are unique to the pathogen, increasing the specificity and safety of potential therapeutic development.

$$Homologs = \{p \in B | \, identity\,(p, H) \geq 35\% \wedge E.value\,(p, H) \leq 0.001\}$$

$$Nonhomologs = B - Homologs$$

where B represents the total bacterial proteome, H the reference human host proteome and p denotes an individual protein sequence.

The essentiality analysis was performed by reusing the script **blastp.py** to evaluate sequence similarity between the remaining non-homologous proteins and the Database of Essential Genes (DEG) [37]. This analysis aimed to identify proteins associated with experimentally validated essential bacterial genes involved in the pathogen´s growth, or survival.

The DEG dataset was automatically retrieved from a public mirror repository which provides a reliable interface for accessing bacterial data from DEG version 15.2 [38]. This retrieval was handled by the custom script **download_deg.py,** integrating the database into the workflow as a reference. After this retrieval the **blast.py** script was executed for BLASTp alignment, using the DEG database as a reference to search for similarities against the filtered protein dataset, which served as the query.

Proteins were considered essential when they satisfied the same alignment constraints employed during the non- homology filtering step, which are a sequence identity of at least 35% and an E-value ≤ 0.001. This filtering step followed an inclusion-based strategy, implemented via the **blastp_filter.py** script**,** where proteins were retained only if they met the threshold similarity and E-value criteria against the DEG. This inclusion-based filtering can be seen in Equation 2, where *E* denotes the subset of proteins that successfully passed these constraints and are identified as essential genes.

$$E = \{p \in B | \, identity\,(p, DEG) \geq 35\% \wedge E.value\,(p, DEG) \leq 0.001\}$$

$$Potential\ Targets = Nonhomologs \cap E$$

This approach enables the prioritization of proteins that are essential for pathogen survival, as their inhibition is more likely to compromise the viability of the bacteria. These proteins are therefore considered particularly relevant therapeutic targets. This essentiality analysis complemented the previous filtering steps by introducing a functional criterion, ensuring that the selected candidates are not only pathogen-specific but also biologically indispensable. As a combination of the previous filtering steps, including the homology exclusion and essentiality inclusion, the dataset conducts towards a minor amount high-confidence target. Overall, this stepwise filtering strategy progressively reduced the initial proteome to a subset of candidate proteins with increased specificity and biological relevance, making them suitable for therapeutic exploration.

Finally, the subsets of essential, non-homologous transmembrane proteins were merged into a complete FASTA file and clustered using the Cluster Database at High Identity with Tolerance (CD-HIT) software through the custom **cd_hit.py** script to identify groups of highly similar protein sequences and reduce dataset redundancy. CD-HIT performs sequence clustering by comparing protein sequences according to their percentage of sequence identity. It can be seen in Equation 3, allowing the grouping of proteins that share high structural and evolutionary similarity while selecting representative sequences for each cluster.

$$Identity\ (\%) = \frac{Number\ of\ identical\ residues}{Alignment\ length}\ 100$$

A sequence identity threshold of 0.9 was applied, allowing the grouping of closely related proteins while preserving biologically relevant differences of distinct protein families. During clustering, only clusters containing five or more proteins were retained for subsequent sequence analysis, as smaller clusters were considered insufficiently representative for comparison.

For each cluster that it was retained, a filter was applied through **cluster_filter.py** to filter out those clusters that do not contain proteins from at least five different bacterial species. Each released cluster contained a representative sequence that was selected and merged into a FASTA file to evaluate the biological functionality This additional filtering step ensures that the selected sequences represent conserved genes across multiple organisms, rather than paralogous genes from a single proteome. The sequences that met all the requirements of the filtering block, shown in Figure 8, facilitated the prioritization and organization of large proteomic datasets, supporting the identification of suitable targets in many organisms.

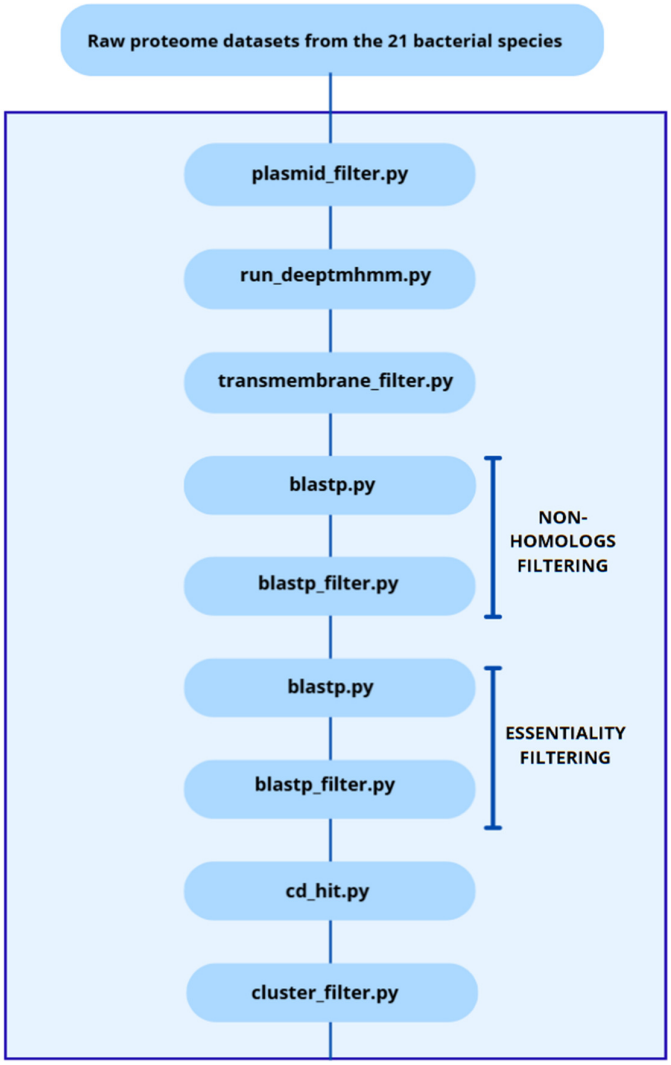


Figure 8. Schematic filtering block following subtractive genomics approach.

2.1.4 Target Identification and Prioritization

Finally, the representative sequences obtained after the complete bacterial filtering process were evaluated to determine their potential biological relevance and functional roles within the studied organisms. This final step is essential to validate the significance of the retained proteins as therapeutic targets, through their mapping to metabolic pathways and the assignment of GO scores. Although the original pipeline employed the BLASTKOALA server for biological annotation, its use was not feasible in this project due to the absence of an available API. Therefore, eggNOG-mapper software was used as an alternative [39].

This tool implements an ontology-based annotation strategy, enabling the functional characterization of proteins. To automate this process, the script **eggNOG_mapper.py** was developed to execute the software locally, using only bacterial taxa as the reference dataset for protein functional annotations. eggNOG-mapper generated multiple outputs, including ortholog assignments, and functional annotation tables. Among these, the file *eggnog_annotation.emapper.annotations* was selected, as it contains the functional annotations assigned to each protein, such as GO terms, KO identifiers, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and COG functional categories through orthology-based annotation. These results were processed using the script **prioritization.py** to score and prioritize potential antimicrobial target proteins according to their relevance.

The prioritization strategy reflected on the script was based on a previous manual functional assessment of the annotated protein dataset. As an initial exploratory step, the most abundant GO terms were identified and summarized using REVIGO, a tool to reduce redundancy and group terms into representative functional categories [36]. Subsequently, the biological significance of these representative terms was identified using QuickGO database, allowing their association with membrane localization, transmembrane transport, ion homeostasis and secretion systems processes [37].

Based on these analyses, a subset of representative GO terms was selected and used as the primary filtering criterion (see Table 7), assigning predefined scores to proteins according to their relevance as potential antimicrobial targets

GO terms	Biological Functionality	Score	Relevance as antimicrobial targets
GO:0055085	Transmembrane transport	+3	Essential for nutrient uptake, ion balance and antibiotic efflux mechanisms
GO:0005215	Transporter activity	+3	Responsible for substrate exchange and multidrug efflux systems associated with antimicrobial resistance mechanisms.
GO:0071944	Cell periphery component	+3	Highly accessible to antimicrobial compounds and often involved in host immune interactions.
GO:0015075	Monoatomic ion TM transport activity	+3	Ion transport regulates membrane potential required for maintenance and bacterial growth.
GO:0016020	Membrane components	+2	Accessible targets due to their localization on the surface and relation in transport and signaling.
GO:0005886	Plasma membrane component	+2	Participate in secretion and transport process as maintenance of membrane integrity.

Table 6. GO terms selected as primary filtering criterion on prioritization.py.

Proteins annotated with these GO terms received the indicated score due to their accessibility and essential role in bacteria. The pre-filter based on GO score was calculated by summing all the scores associated with each protein, and those proteins that exceed the minimum score threshold, introduced by the user, were retained for the final prioritization step.

The set of proteins that pass the primary GO-based filter criteria was subsequently prioritized through the incorporation of additional scores derived from KO and metabolomic pathway terms. The assignment of these scores was supported by the biological interpretation of the functional pathways identified with *KEGG Mapper-Seach* [42]. Highlighting the terms related to Adenosine Triphosphate (ATP)-Binding Cassette (ABC) transporters, as they mediate with the export of toxic compounds such as antimicrobial agents, protein export, bacterial secretion systems and quorum sensing which enables cell-to cell communication, coordinating collective behaviors as biofilm formation, which contribute to bacterial pathogenicity. The whole prioritization methodology can be seen in Figure 9.

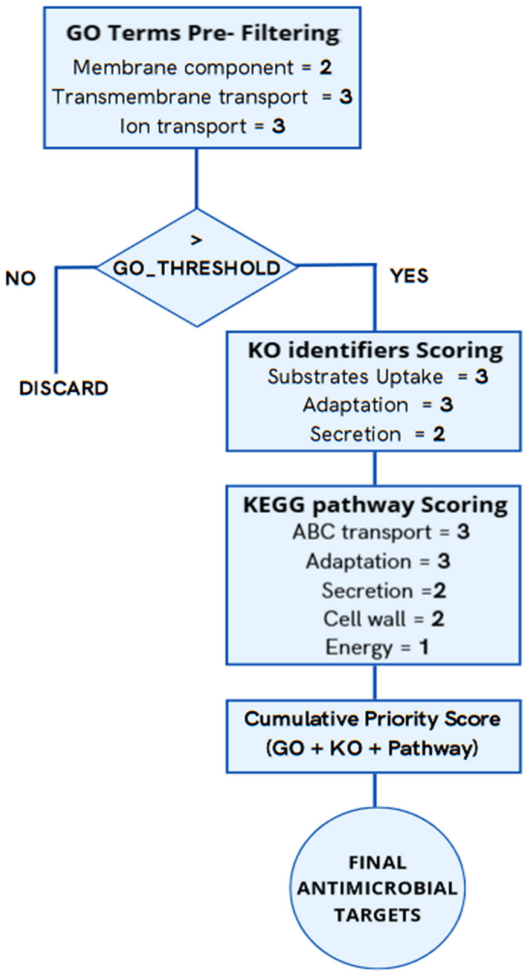


Figure 9. Functional prioritization strategy followed for antimicrobial proteins ranking.

This integrative approach combined the cumulative scores of GO terms, KO identifiers and KEGG pathways, and facilitated the identification of a reduced set of target proteins involved in key cellular processes such as metabolism, transport, signal transduction, and membrane-associated activities.

2.1.5 Data Analysis and Visualization

The final stage of the workflow previously described (see Figure 5) is focused on the data analysis and visualization of the results generated throughout the subtractive genomics approach. Candidate protein subsets obtained after each filtering stage were integrated and evaluated to assess that only robust and biologically meaningful targets progressed to the final prioritization stage after the implementation of the previous filtering criteria. To support these analyses, each filtering block generated JSON files containing relevant information associated with the processed protein subsets, including sequence identifiers, annotation results, filtering criteria and classification outcomes of each filtering stage. These files were subsequently processed using the **plot.py** script, developed for data statistical summarization and graphical visualization of the workflow results.

The script generated comparative bar plots and retention heatmaps to facilitate the interpretation of protein distributions and filtering efficiency across the different stages of the pipeline, including plasmid filtering, transmembrane filtering, non-homology filtering, and essentiality filtering. Bar plots were used to compare the total and the retained sequences subsets for each analyzed organism, whereas heatmaps represented protein retention percentages across the filtering stages. Together, these visualizations generated in “plots” folder, provided an overview of the computational workflow and supported the comparative analysis and prioritization of candidate therapeutic proteins identified through the subtractive genomics approach.

3 RESULTS

Initially, the complete bacterial dataset analyzed in this study consisted of 74.528 protein sequences distributed among the 21 selected bacterial organisms (see Table 5), which were directly downloaded from the NCBI database. Following the plasmid filtering stage, which was implemented to remove plasmid-derived sequences due to their high variability and lower suitability as therapeutic targets, a total of 73.613 protein sequences were retained. As it is observed in Figure 10, most of the analyzed organisms contained few or no plasmid-associated proteins, resulting in high sequence retention during this filtering step.

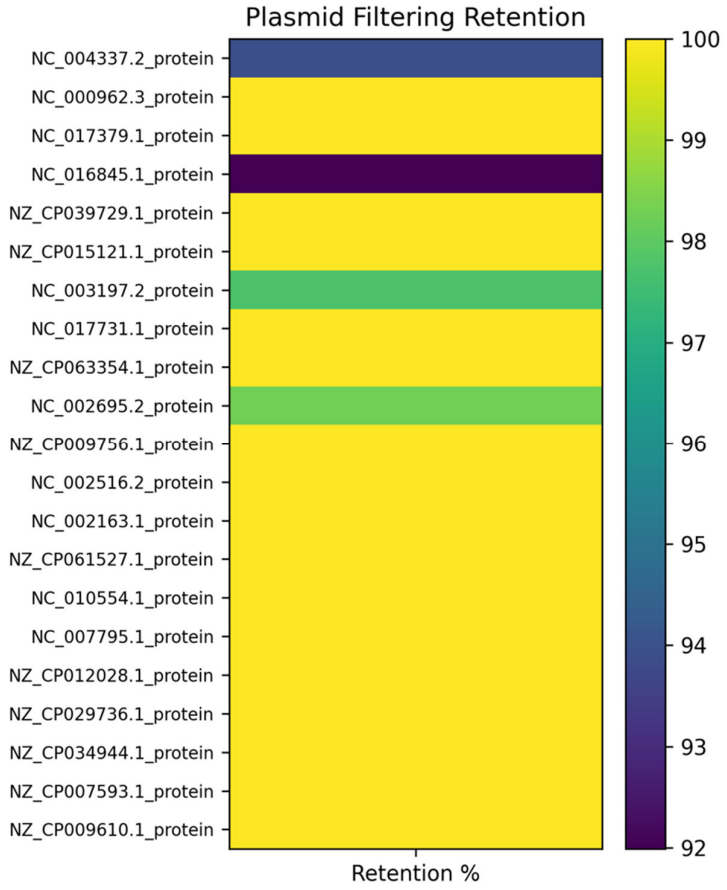


Figure 10. Heatmap of protein retention percentage after plasmid filtering

The subsequent filtering stage was applied after the protein topology prediction to retain only TM proteins because of their accessibility and relevance as potential drug

targets. As it is shown in Figure 11, this block represented one of the most restrictive filtering criteria within the subtractive genomics workflow, since approximately 82% of the protein sequences were discarded due to their predicted protein localization, keeping 13.899 TM protein sequences.

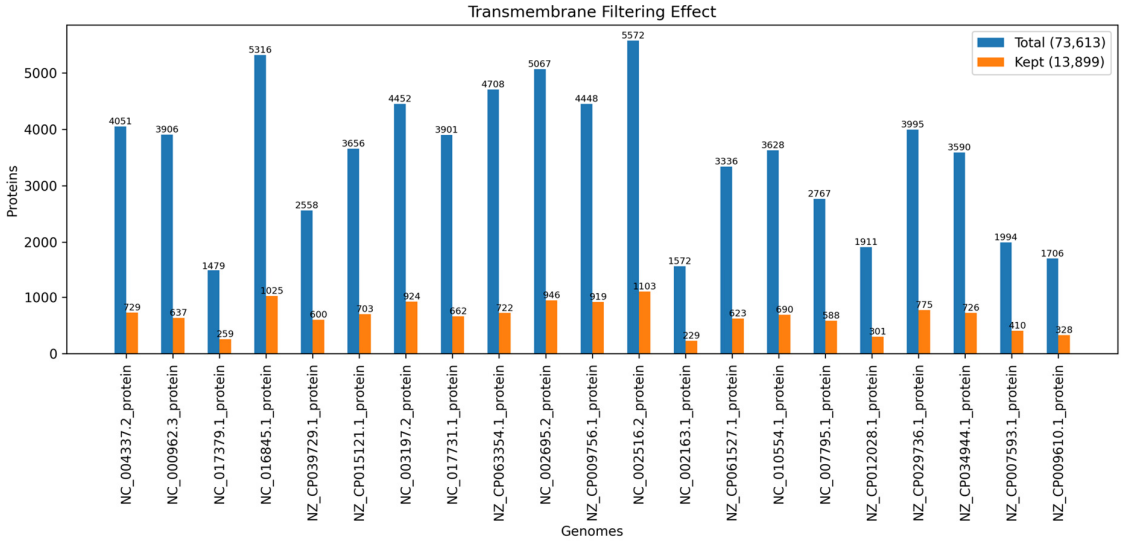


Figure 11. Bar plot of protein retention after DeepTMHMM TM topology filtering.

Once only TM proteins were retained, the host non-homology filtering stage was applied, to avoid sequence similarity to host proteins that may lead to undesired off-target interactions, toxicity or adverse side effects during therapeutic targeting. By applying this filtering step 13.083 sequences were retained as it is shown in Figure 12.

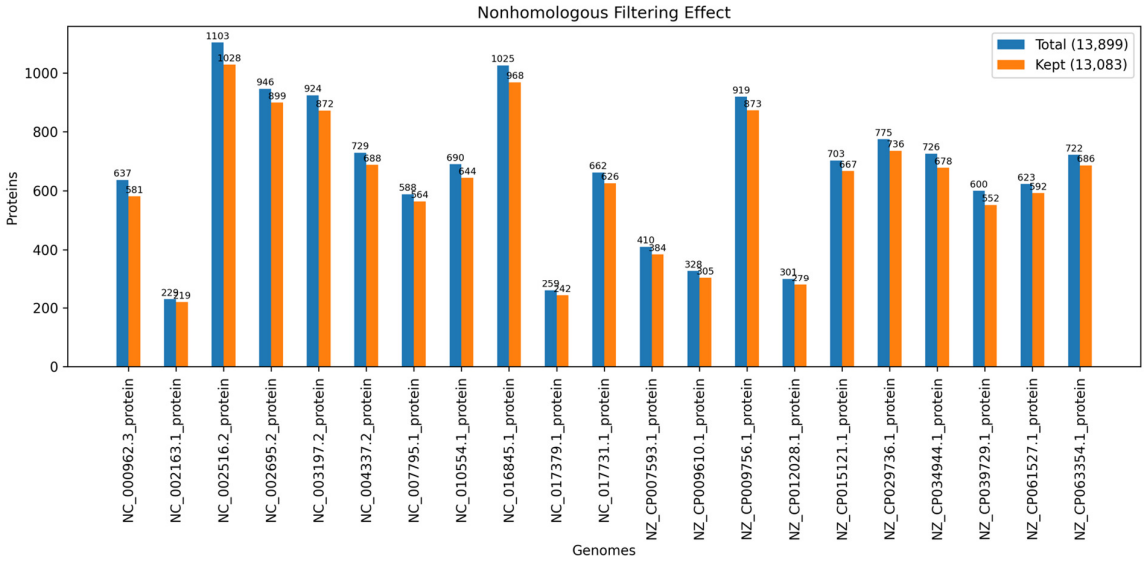


Figure 12. Bar plot of protein retention after non-homology filtering with BLASTp.

After non-homology filtering, the essentiality filtering module was employed to identify proteins essential for bacterial survival and viability, as the inhibition of these proteins is more likely to compromise the pathogen growth and persistence. Following this filtering stage, 5.256 protein sequences remained (see Figure 13) and were selected for the posterior procedures. See Figure 14 for the comparative analysis of TM, non-homologous and essentiality retention of bacterial proteins

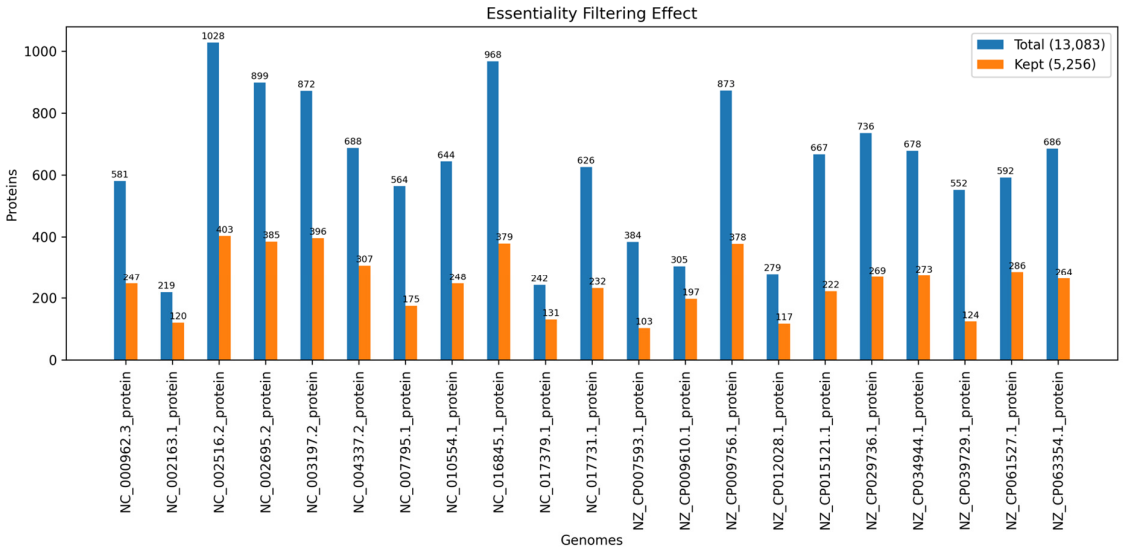


Figure 13. Bar plot of protein retention after essentiality filtering with BLASTp

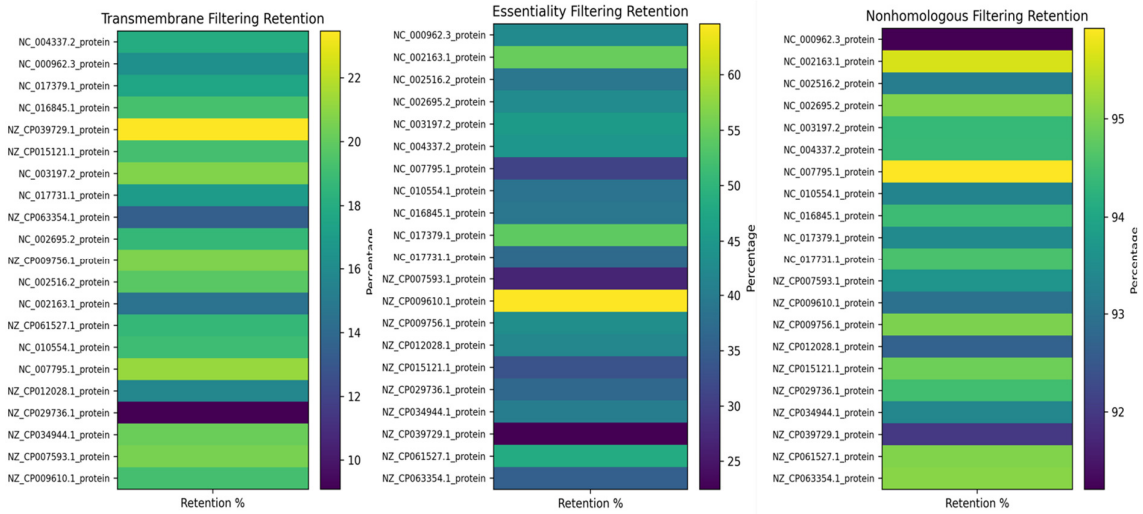


Figure 14. Protein retention across TM, essentiality, and non-homology filtering.

During the execution of CD-HIT, 114 protein clusters containing at least five sequences and a sequence similarity of 0.9, were identified, comprising a total of 647 protein sequences. Subsequently, an additional filtering procedure was performed such that clusters containing sequences from at least five different organisms were retained, ensuring evolutionary conservation among bacterial species. After this filtering process, 113 representative sequences were selected and functionally annotated using eggNOG-mapper, enabling the identification of GO terms and KEGG metabolic pathways associated with these final candidate proteins.

The grouping of GO terms revealed the predominance of membrane-associated biological processeses within the final dataset, principally related to transmembrane transport, secretion systems, ion exchange and membrane-associated process (see Figure 15). These functional categories were considered biologically relevant for antimicrobial target prioritization due to their accessibility and their essential role in bacterial survival and adaptation.

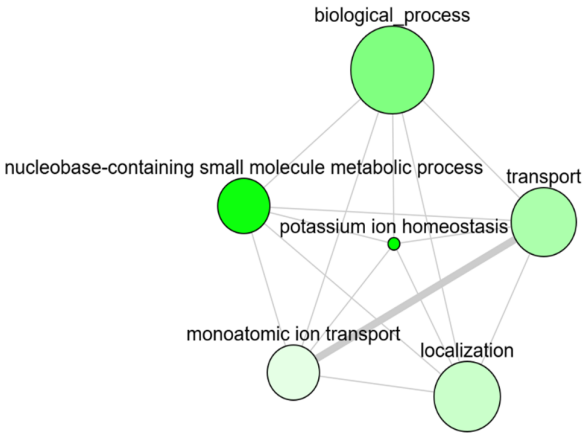


Figure 15. Semantic network and functional categorization of GO terms of the potential antimicrobial targets resulted from subtractive genomics using REVIGO [40].

After the application of the GO terms, KO identifiers and pathway-based prioritization criteria, the final subset, was reduced to nine proteins classified as “HIGH-priority” antimicrobial targets (see Table 7).

Accession Number|Protein	Global Score	Classification
NC_002695.2_protein|NP_308719.1	17	HIGH
NC_002695.2_protein|NP_312931.1	15	HIGH
NC_002695.2_protein|NP_312081.1	15	HIGH
NC_002695.2_protein|NP_312667.1	15	HIGH
NC_002695.2_protein|NP_312192.1	15	HIGH
NC_002695.2_protein|NP_311359.1	14	HIGH
NC_002695.2_protein|NP_313132.1	14	HIGH
NC_002695.2_protein|NP_312435.1	14	HIGH
NC_002695.2_protein|NP_309771.1	14	HIGH

Table 7. Final subset of the nine high-priority proteins and their global score, using a GO terms threshold of 5.

The functional annotations generated by eggNOG-mapper for these nine proteins were used as the basis for the biological interpretation. These annotations also contribute to the foundation for assessing their potential as antimicrobial targets, as presented in section 4.

3.1 Validation of DeepTMHMM Usage

To verify the correct implementation and execution of DeepTMHMM software within the developed computational pipeline, transmembrane topology predictions were performed for the 21 bacterial organisms previously described (see Table 5). To evaluate the consistency and reliability of the obtained predictions, the results generated through the workflow using DeepTMHMM tool were compared against manually supervised topology analyses performed using Phobius, mentioned in the published pipeline.

This validation strategy was performed to assess whether the automated implementation correctly reproduced biologically coherent topology classifications and transmembrane region predictions. Since the automation of Phobius was not feasible, its predictions were manually inspected, whereas DeepTMHMM predictions were automatically generated as part of bacterial filtering stage of the workflow. The comparison between both approaches is summarized in Table 9.

Accession Number	**DeepTMHMM**	**Phobius**
NZ_CP015121.1	19.23%	18.65%
NC_002163.1	14.57%	19.97%
NZ_CP009756.1	20.66%	18.86%
NZ_CP039729.1	23.46%	22.60%
NC_002695.2	18.67%	17.01%
NZ_CP009610.1	19.23%	17.58%
NC_017379.1	17.51%	17.11%
NC_016845.1	19.28%	17.44%
NZ_CP034944.1	20.22%	18.13%
NC_000962.3	16.31%	16.03%
NZ_CP012028.1	15.75%	13.76%
NC_003197.2	20.75%	19.16%
NZ_CP063354.1	15.34%	18.37%
NZ_CP061527.1	18.68%	17.12%
NC_004337.2	18.0%	17.25%
NC_007795.1	21.25%	21.68%

NZ_CP007593.1	20.56%	20.36%
NC_010554.1	19.02 %	18.11%
NZ_CP029736.1	19.40%	18.62%
NC_017731.1	16.97%	17.87%
NC_002516.2	19.8%	16.87%

Table 8. Comparison between Phobius and DeepTMHMM TM topology predictions

This comparison enabled the evaluation of consistency between manual topology predictions obtained with Phobius and the automated predictions generated through DeepTMHMM within the computational workflow, allowing the identification of differences in the number of TM predicted proteins, and variations in the classification of SP regions between both methodologies. Minor differences were observed between both tools; most bacterial proteomes differ less than 3% in the prediction of TM proteins. However, DeepTMHMM generally showed slightly higher retention percentages, which may be associated with methodological differences between both approaches. Phobius uses hidden Markov models (HMMs), based on statistical probabilities to predict TM regions and SP simultaneously, whereas DeepTMHMM applies deep learning models trained to recognize membrane topology patterns from protein sequences. Consequently, deep learning approaches may provide higher sensitivity than HMMs [43].

By performing this comparative validation, the robustness, reproducibility, and reliability of the automated transmembrane filtering stage were assessed. This validation supports the correct integration of DeepTMHMM into the computational pipeline and confirms that the generated topology predictions were sufficiently consistent for downstream subtractive genomics analyses, including transmembrane filtering, non-homology screening, essentiality analysis, and candidate target prioritization.

3.2 Impact of threshold parameters

It is important to distinguish between fixed methodological threshold parameters and configurable parameters by the user. Among the fixed thresholds a BLAST similarity identity cut-off of ≥ 35% was used to determine a relevant protein similarity or a biologically meaningful protein. This threshold was settled as it exhibits a significant degree of alignment considering the divergence through evolution and functionality. Lower identity values would be considered in the twilight zone, where a minimum similarity cannot be granted and a high number of false positives may be introduced. While higher cut-off values would ensue the similarity of those selected proteins but could lead to the loss of many evolutional divergent yet functionally similar proteins. The 35% threshold was applied during human non-homology and essentiality BLASTp analyses, keeping a balance between specificity and sensitivity, which are also consistent with previously published subtractive genomics studies following a conservative strategy to avoid the risk of off-target effects with the human host proteins, while simultaneously identifying sequences that share meaningful similarity with validated essential genes [44].

BLAST E-value follows a similar reasoning; a 0.001 threshold indicates one false alignment each 1000. Providing a lower E-value is extremely restrictive while a value close to 1 will lead to uncertainty, introducing the entrance of many false positives [45]. The ultimate fixed methodological threshold is CD-HIT similarity parameter, whose main objective is to eliminate the redundancy of sequences by clustering proteins based on sequence identity. The settled value of 0.9 allows the existence of minor mutations or sequence variants, which are very common among bacterial organisms. Reducing this value would lead to grouping divergent proteins into the same cluster, masking meaningful proteins.

3.2.1 User-configurable parameters

The user provides the value of three configurable parameters among the workflow: a minimum number of proteins per cluster during CD-HIT clustering, a minimum number of different organisms per cluster in the filtration stage of clusters to ensure conservation among species and the GO term score threshold as a prefilter on candidate proteins prioritization.

Unlike the fixed methodological thresholds, the optimal values for cluster size and number of organisms criteria depend on the size of the provided dataset and cannot be predefined. For the validation of this project dataset, which consisted of 21 bacterial species, a default value of five was selected. This five-value parameter indicates a representation of approximately 24% of analyzed bacterial dataset. A threshold value of six organisms (approximately 29% of the dataset) would also be reasonable and would not substantially alter the biological interpretation.

The workflow was also performed by increasing the minimum cluster size threshold from five to six proteins. These results showed a generation of 70 clusters out of the 114 clusters obtained with the default value, corresponding to a loss of approximately 39% of the clusters. This result suggests that a considerable proportion of clusters were represented by exactly five proteins. Additionally, increasing to six the minimum number of different organisms per cluster resulted in the removal of only one cluster, as with the default value threshold, indicating that most of the filtered clusters were already conserved across different bacterial species. As a conclusion, the workflow appears to be considerably more sensitive to changes in the minimum cluster size parameter than to variations in the organism conservation threshold.

Low values for this training dataset would allow the introduction of weakly represented clusters while the introduction of poor conservation among bacteria species, making the target protein less attractive. Whereas higher parameters could become restrictive, discarding meaningful candidates. However, different datasets may require alternative thresholds, and therefore these parameters were intentionally exposed to the user

Increasing the organism conservation threshold from 5 to 6 made the filtering slightly stricter, reducing the number of retained clusters. However, the final candidate set remained consistent, suggesting that the selected default threshold provides a reasonable balance between conservation and candidate retention.

Finally, the GO score threshold was also tested with a cumulative threshold of 6 obtaining a total of four high-candidate proteins representing the potential antimicrobial targets. All these four proteins were inherently included within the prioritization results obtained using the default GO value (see Table 9). This demonstrates that altering this parameter directly affects the volume of final candidate targets, increasing this value would lead to the reduction of high-priority classification proteins and lowering this threshold a substantial number of proteins would be introduced, increasing the user's workload due to the large amount of final target proteins.

Accession Number|Protein	Global Score	Classification
NC_002695.2_protein|NP_308719.1	17	HIGH
NC_002695.2_protein|NP_311359.1	14	HIGH
NC_002695.2_protein|NP_313132.1	14	HIGH
NC_002695.2_protein|NP_312435.1	14	HIGH

Table 9. Final subset of the four high-priority proteins and their global score, using a GO terms threshold of 6.

4 DISCUSSION AND CONCLUSIONS

Among the high prioritized targets, two proteins were involved in the Sec-dependent translocation machinery (NP_312081.1): specifically, SecY (NP_312192.1) and SecE (NP_312931.1), which constitute essential components of the SecYEG translocon, a membrane-embedded complex responsible for transporting recent synthesized proteins across the cytoplasmic membrane (see Figure 16). Sec Y forms the central translocation channel through which proteins are exported, whereas Sec E acts as a structural component that provides stabilization to SecYEG complex and maintains its proper function. These targets are functionally related to YidC (NP_312667.1), an essential membrane insertase that cooperates with the SecYEG translocon. While SecYEG provides the channel for protein translocation, YidC facilitates the insertion, folding and assembly of membrane proteins within the OM, as it is located near the lateral gate of SecY and ensures the correct assembly of proteins required for bacterial viability [46].

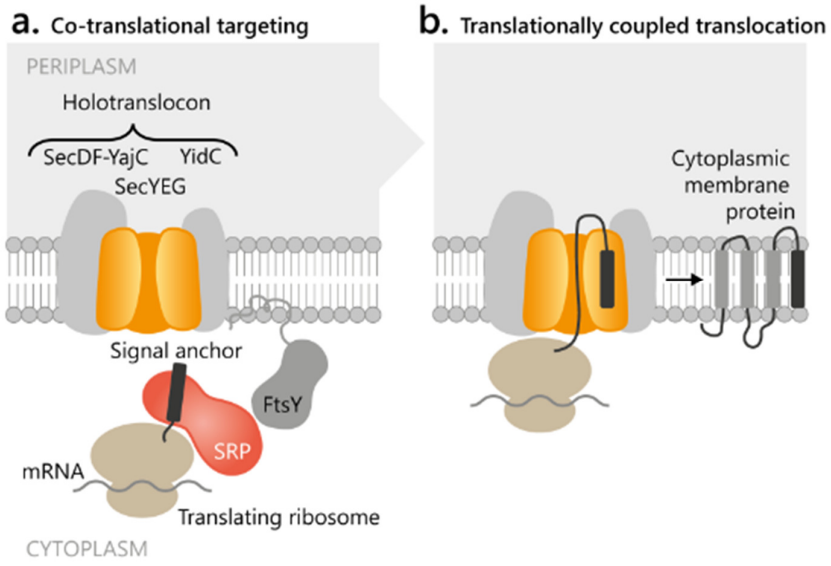


Figure 16. Representation of the bacterial Sec-dependent protein translocation machinery: (a) the SecYEG translocon complex and its interaction with YidC. (b) the lateral insertion and assembly within the membrane of a new synthesized protein [46].

Another prioritized protein corresponded to membrane-associated transport systems (NP_311359.1), annotated as efflux pump that actively export toxic compounds from the bacterial cell. Efflux systems contribute to pathogenicity and antimicrobial resistance mechanisms by reducing the intracellular drug concentrations, as it can be seen in Figure 17 [41,42]. The relevance of transport-related functions was further supported by the identification of the GltIJKL component of ABC transporters (NP_308719.1). As one of the largest families of membrane transport, ABC transporters mediate the transport of a wide variety of substrates through ATP hydrolysis. In particular, the GltIJKL complex is involved in the uptake of AAs, as glutamate and aspartate, required for bacterial nutrition, metabolism and cellular growth [49].

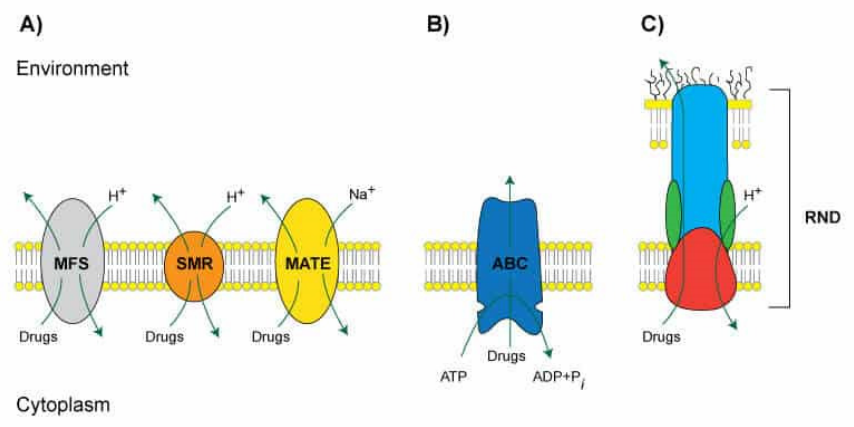


Figure 17. Representation of efflux pumps. A) Systems pumping drugs out while pumping H^+ or Na^+ into the cell. B) ABC transporter system powered by ATP. C) RND efflux system connect the inner membrane and OM in Gram-negative bacteria. [50].

However, targeting an individual efflux pump is limited by the functional redundancy of transport systems within bacterial proteomes, so a single efflux pump does not necessarily compromise the bacterial viability. Finally, two high-priority targets were associated with the uptake of C4-dicarboxylates (NP_312435.1 and NP_313132.1). These substrates such as malate, fumarate and aspartate are key intermediates in central carbon metabolism and fumarate respiration. Through them bacteria are able to conserve energy and sustain growth under diverse environmental conditions, contributing to bacterial adaptation and virulence [51].

4.1 Future Work

To enhance the usability, reproducibility and practical application of this computational workflow, a future development line that has been already proposed is the implementation of all the filtering criteria, the parameters settings and biological prioritization strategies within the Scipion- Chem platform developed by the National Center of Biotechnology. Scipion-Chem is an extension of the Scipion framework, an open-source software specifically designed for virtual drug screening and molecular dynamics workflows.

Future versions of the workflow could incorporate additional automated topology prediction tools to reduce computational and time costs associated with the identification of transmembrane proteins. Another line could be the integration of alternative biological resources and databases which allow the users to select different tools for each filtering step according to their specific research requirements. Finally, larger-scale analyses of multiple bacterial datasets could be conducted to evaluate the robustness of the proposed methodology and assess the prioritization criteria of the different functional annotations.

5 REFERENCES

[1] Z. Breijyeh and R. Karaman, "Design and Synthesis of Novel Antimicrobial Agents," *Antibiotics*, vol. 12, no. 3, p. 628, Mar. 2023, doi: 10.3390/antibiotics12030628.

[2] S. Mohsen, J. A. Dickinson, and R. Somayaji, "Update on the adverse effects of antimicrobial therapies in community practice," *Can. Fam. Physician*, vol. 66, no. 9, pp. 651–659, Sep. 2020.

[3] "WHO publishes list of bacteria for which new antibiotics are urgently needed." Accessed: Jun. 08, 2026. [Online]. Available: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

[4] L. A. Mandell, J. G. Bartlett, S. F. Dowell, T. M. File Jr., D. M. Musher, and C. Whitney, "Update of Practice Guidelines for the Management of Community-Acquired Pneumonia in Immunocompetent Adults," *Clin. Infect. Dis.*, vol. 37, no. 11, pp. 1405–1433, Dec. 2003, doi: 10.1086/380488.

[5] G. Kapoor, S. Saigal, and A. Elongavan, "Action and resistance mechanisms of antibiotics: A guide for clinicians," *J. Anaesthesiol. Clin. Pharmacol.*, vol. 33, no. 3, p. 300, Sep. 2017, doi: 10.4103/joacp.JOACP_349_15.

[6] A. Dowling, J. O'Dwyer, and C. C. Adley, "Antibiotics: Mode of action and mechanisms of resistance," in *Antimicrobial Research: Novel Bioknowledge and Educational Programs*, A. Méndez-Vilas, Ed. Badajoz, Spain: Formatex Research Center, 2013, p. 536.

[7] A. Upadhayay, J. Ling, D. Pal, Y. Xie, F.-F. Ping, and A. Kumar, "Resistance-proof antimicrobial drug discovery to combat global antimicrobial resistance threat," *Drug Resist. Updat.*, vol. 66, p. 100890, Jan. 2023, doi: 10.1016/j.drup.2022.100890.

[8] "Antimicrobial Resistance," Ossila. Accessed: Jun. 08, 2026. [Online]. Available: https://www.ossila.com/pages/mechanism-of-antimicrobial-resistance

[9] H. Elkady, I. Nasser Salman, and M. M. Khalifa, "Small-molecule strategies to combat antibiotic resistance: mechanisms, modifications, and contemporary approaches," *RSC Adv.*, vol. 15, no. 30, pp. 24450–24474, 2025, doi: 10.1039/D5RA04047G.

[10] V. Van Puyenbroeck and K. Vermeire, "Inhibitors of protein translocation across membranes of the secretory pathway: novel antimicrobial and anticancer agents," *Cell. Mol. Life Sci.*, vol. 75, no. 9, pp. 1541–1558, May 2018, doi: 10.1007/s00018-017-2743-2.

[11] J. H. Powers, "Antimicrobial drug development – the past, the present, and the future," *Clin. Microbiol. Infect.*, vol. 10, no. s4, pp. 23–31, 2004, doi: 10.1111/j.1465-0691.2004.1007.x.

[12] "New challenges in drug discovery," in *Novel Platforms for Drug Delivery Applications*, Woodhead Publishing, 2023, pp. 619–643. doi: 10.1016/B978-0-323-91376-8.00021-5.

[13] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *J. Health Econ.*, vol. 47, pp. 20–33, May 2016, doi: 10.1016/j.jhealeco.2016.01.012.

[14] B. K. Shoichet, "Virtual screening of chemical libraries," *Nature*, vol. 432, no. 7019, pp. 862–865, Dec. 2004, doi: 10.1038/nature03197.

[15] R. Santos *et al.*, "A comprehensive map of molecular drug targets," *Nat. Rev. Drug Discov.*, vol. 16, no. 1, pp. 19–34, Jan. 2017, doi: 10.1038/nrd.2016.230.

[16] X. Lin, X. Li, and X. Lin, "A Review on Applications of Computational Methods in Drug Screening and Design," *Molecules*, vol. 25, no. 6, p. 1375, Jan. 2020, doi: 10.3390/molecules25061375.

[17] N. van de Sande-Bruinsma *et al.*, "Antimicrobial Drug Use and Resistance in Europe," *Emerg. Infect. Dis.*, vol. 14, no. 11, pp. 1722–1730, Nov. 2008, doi: 10.3201/eid1411.070467.

[18] M. W. Y. Southey and M. Brunavs, "Introduction to small molecule drug discovery and preclinical development," *Front. Drug Discov.*, vol. 3, Nov. 2023, doi: 10.3389/fddsv.2023.1314077.

[19] "What are Transmembrane Proteins | Sino Biological." Accessed: Jun. 08, 2026. [Online]. Available: https://www.sinobiological.com/resource/protein-review/transmembrane-proteins

[20] L. Fernández and R. E. W. Hancock, "Adaptive and Mutational Resistance: Role of Porins and Efflux Pumps in Drug Resistance," *Clin. Microbiol. Rev.*, vol. 25, no. 4, pp. 661–681, Oct. 2012, doi: 10.1128/CMR.00043-12.

[21] F. Serral *et al.*, "From Genome to Drugs: New Approaches in Antimicrobial Discovery," *Front. Pharmacol.*, vol. 12, Jun. 2021, doi: 10.3389/fphar.2021.647060.

[22] B. Mudgal, D. Verma, D. Venugopal, S. V. Atram, D. Mitra, and S. Gupta, "Subtractive genomics approach: A guide to unveiling therapeutic targets across pathogens," *J. Microbiol. Methods*, vol. 232–234, p. 107127, Jul. 2025, doi: 10.1016/j.mimet.2025.107127.

[23] I. Ramsden, D. de Jong-Hoogland, A. Chiam, and M. B. Ulmschneider, "A bio-informatics approach to identify new drug targets in multidrug-resistant bacteria," Jun. 03, 2025, *bioRxiv*. doi: 10.1101/2025.05.31.657076.

[24] "Biopython Tutorial & Cookbook — Biopython 1.87 documentation." Accessed: Jun. 08, 2026. [Online]. Available: https://biopython.org/docs/latest/Tutorial

[25] "Matplotlib documentation — Matplotlib 3.10.9 documentation." Accessed: Jun. 08, 2026. [Online]. Available: https://matplotlib.org/stable/index.html

[26] "NumPy documentation — NumPy v2.3 Manual." Accessed: May 06, 2026. [Online]. Available: https://numpy.org/doc/2.3/

[27] "Requests: HTTP for Humans™ — Requests 2.33.1 documentation." Accessed: Jun. 08, 2026. [Online]. Available: https://requests.readthedocs.io/en/latest/

[28] "PyTorch documentation — PyTorch 2.11 documentation." Accessed: Jun. 08, 2026. [Online]. Available: https://docs.pytorch.org/docs/stable/index.html

[29] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023. Accessed: Jun. 08, 2026. [Online]. Available: https://github.com/facebookresearch/esm

[30] M. Z. Tien, D. K. Sydykova, A. G. Meyer, and C. O. Wilke, "PeptideBuilder: A simple Python library to generate model peptides," *PeerJ*, vol. 1, p. e80, May 2013. Accessed: Jun. 08, 2026. [Online]. Available: https://github.com/clauswilke/PeptideBuilder

[31] J. Hallgren *et al.*, "DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks," Apr. 10, 2022, *bioRxiv*. doi: 10.1101/2022.04.08.487609.

[32] "BLAST: Basic Local Alignment Search Tool." Accessed: Jun. 08, 2026. [Online]. Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi

[33] 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, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[34] C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale", Accessed: Jun. 08, 2026. [Online]. Available: https://dx.doi.org/10.1093/molbev/msab293

[35] J. Huerta-Cepas *et al.*, "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D309–D314, Jan. 2019, doi: 10.1093/nar/gky1085.

[36] "National Center for Biotechnology Information." Accessed: Jun. 08, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[37] R. Zhang, H.-Y. Ou, and C.-T. Zhang, "DEG: a database of essential genes," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D271–D272, Jan. 2004, doi: 10.1093/nar/gkh024.

[38] "DEG, database of essential genes." Accessed: Jun. 08, 2026. [Online]. Available: https://tubic.org/deg_bak/download.php

[39] C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale," *Mol. Biol. Evol.*, vol. 38, no. 12, pp. 5825–5829, Dec. 2021, doi: 10.1093/molbev/msab293.

[40] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms," *PLoS ONE*, vol. 6, no. 7, p. e21800, Jul. 2011, doi: 10.1371/journal.pone.0021800.

[41] "QuickGO::Annotation List." Accessed: Jun. 08, 2026. [Online]. Available: https://www.ebi.ac.uk/QuickGO/annotations

[42] "KEGG Mapper Search." Accessed: Jun. 08, 2026. [Online]. Available: https://www.genome.jp/kegg/mapper/search.html

[43] "Phobius." Accessed: Jun. 08, 2026. [Online]. Available: https://phobius.sbc.su.se/

[44] B. Rost, "Twilight zone of protein sequence alignments," *Protein Eng.*, vol. 12, no. 2, pp. 85–94, Feb. 1999, doi: 10.1093/protein/12.2.85.

[45] Y. Y. Lu, W. S. Noble, and U. Keich, "A BLAST from the past: revisiting blastp's E-value," *Bioinformatics*, vol. 40, no. 12, p. btae729, Dec. 2024, doi: 10.1093/bioinformatics/btae729.

[46] T. Salter, "The bacterial Sec-machinery as an antibiotic target," Ph.D. dissertation, Dept. of Cellular and Molecular Medicine, University of Bristol, Bristol, U.K., 2023, pp. 5–15.

[47] L. Huang *et al.*, "Bacterial Multidrug Efflux Pumps at the Frontline of Antimicrobial Resistance: An Overview," *Antibiotics*, vol. 11, no. 4, p. 520, Apr. 2022, doi: 10.3390/antibiotics11040520.

[48] Y. Zeng and A. O. Charkowski, "The Role of ATP-Binding Cassette Transporters in Bacterial Phytopathogenesis," *Phytopathology®*, vol. 111, no. 4, p. 215, Apr. 2021, doi: 10.1094/PHYTO-06-20-0212-RVW.

[49] "UniProt," UniProt. Accessed: Jun. 08, 2026. [Online]. Available: https://www.uniprot.org/uniprotkb/P0AAG3/entry

[50] "Screening for Inhibitors of Bacterial Efflux Pumps." Accessed: Jun. 08, 2026. [Online]. Available: https://emerypharma.com/blog/screening-bacterial-efflux-pump-inhibitors/

[51] C. Schubert and G. Unden, "C4-Dicarboxylates as Growth Substrates and Signaling Molecules for Commensal and Pathogenic Enteric Bacteria in Mammalian Intestine," *J. Bacteriol.*, vol. 204, no. 4, pp. e00545-21, Mar. 2022, doi: 10.1128/jb.00545-21.

6 ANNEX: Repositories

All source codes, including python and bash scripts, as well as the different previously described Python virtual environments (venv), are available in a public GitHub repository, ensuring transparency, facilitating reproducibility across different users and supporting future development as maintenance of the project.

The source code and documentation can be found in GitHub through: https://github.com/Noeliaauba/Antimicrobial_Target_Workflow

6.1.1 Auxiliary resources

During the design and implementation of the workflow algorithms, conversational artificial intelligence systems such as CHATGPT and Claude were used as supporting tools for code debugging, troubleshooting, and conceptual clarification CHATGPT was occasionally consulted as source of brainstorming, to provide alternative design decisions or implementation functional strategies. Finally, artificial intelligence tools, as Claude and Copilot, were used to improve the readability and academic style of the written report through language revision and text refinement.

However, the software development, biological decisions and result interpretation were evaluated by the author and the project tutor. In clarification, artificial intelligence was employed as a supportive tool and not as a substitute for the author's judgment.