University of Minho
School of Engineering

PhageLysin: Phage Endolysin Finder

TIAGO BAPTISTA[1], HUGO OLIVEIRA[2], AND ÓSCAR DIAS[2]

[1] UNIVERSITY OF MINHO, SCHOOL OF ENGINEERING (EEUM), PORTUGAL

[2] UNIVERSITY OF MINHO, CENTER OF BIOLOGICAL ENGINEERING (CEB), PORTUGAL

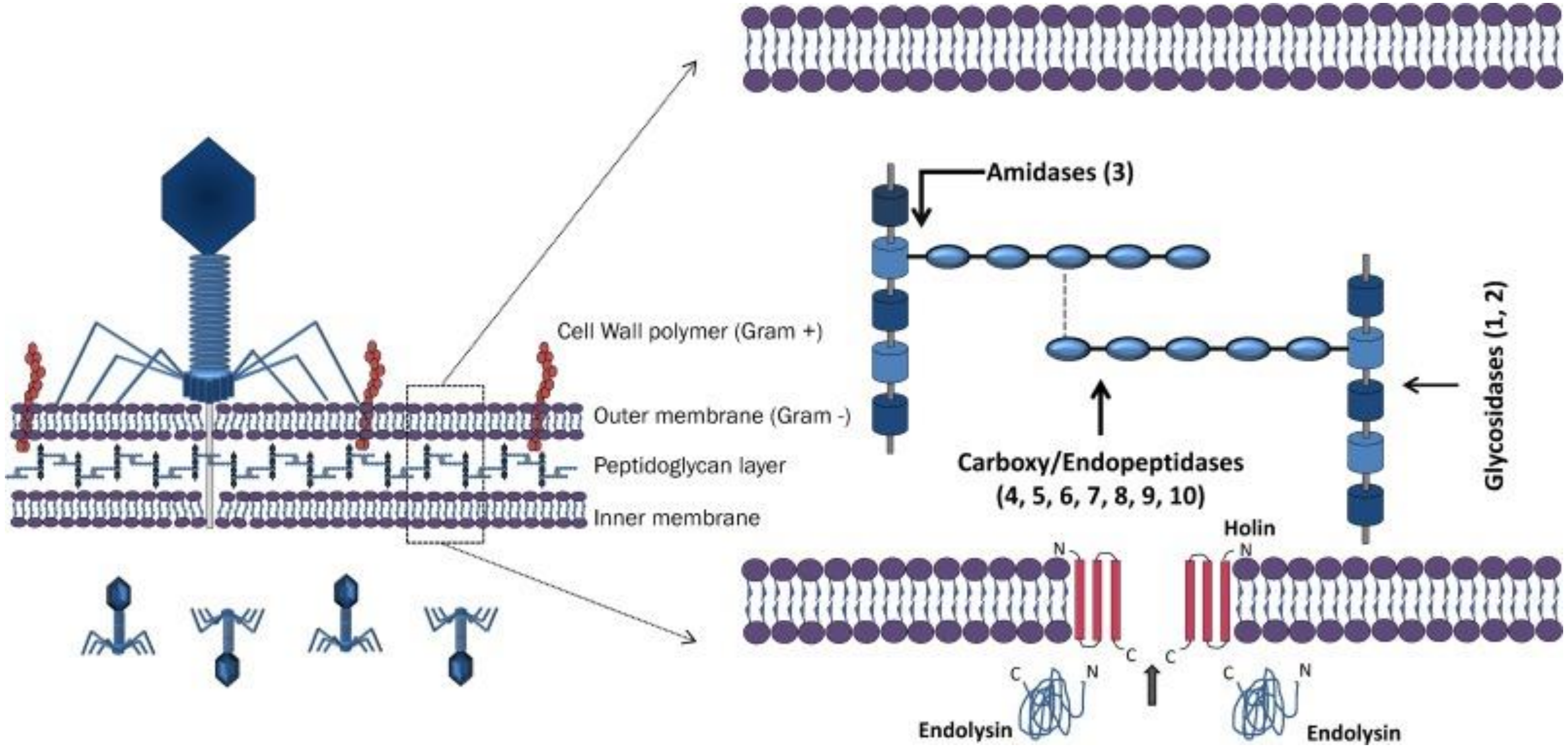
Contextualization, Aims, and Tasks

Bacterial multidrug resistance problem → Phage endolysins as an alternative to antibiotics

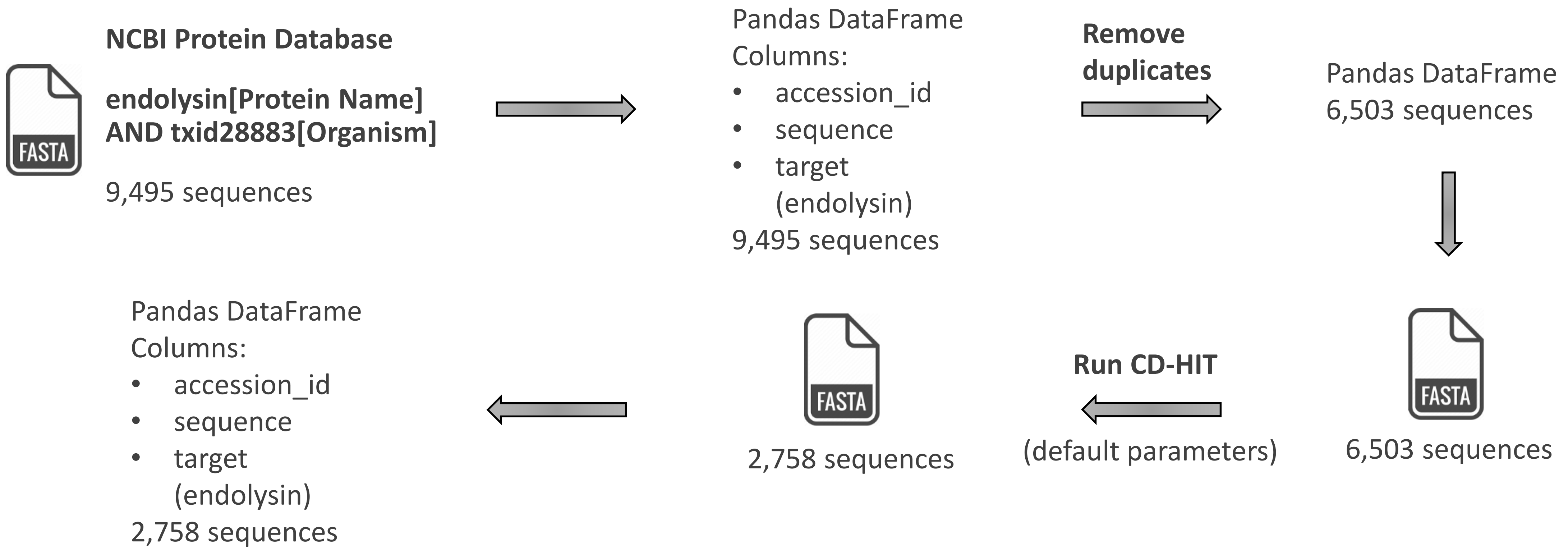
- Review public datasets containing metagenomic data;
- Review annotation tools applied to phage genomes on metagenomic data;
- Build a positive dataset of phage endolysins and a negative dataset of proteins;
- Explore Machine Learning (ML) approaches to improve the accuracy of endolysins prediction;

- End goal: Development of a **machine learning (ML) tool** to provide a quick and reliable way to **identify endolysins** in new phage genomes and metagenomic sequencing data.

What are phage endolysins?



Positive dataset



NCBI Protein Database

endolysin[Protein Name] AND txid28883[Organism]

9,495 sequences

Pandas DataFrame
Columns:

- accession_id
- sequence
- target (endolysin)

9,495 sequences

Remove duplicates

Pandas DataFrame
6,503 sequences

6,503 sequences

Run CD-HIT

(default parameters)

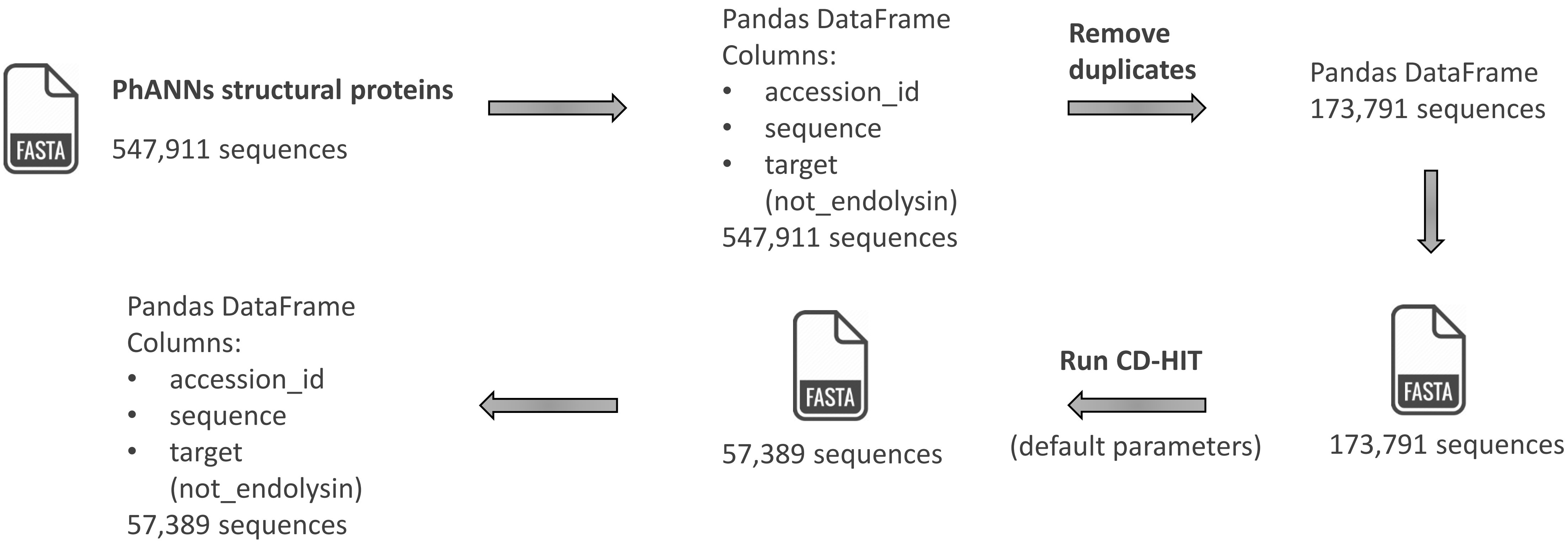
2,758 sequences

Pandas DataFrame
Columns:

- accession_id
- sequence
- target (endolysin)

2,758 sequences

Negative datasets



PhANNs structural proteins

547,911 sequences

Pandas DataFrame
Columns:

- accession_id
- sequence
- target (not_endolysin)

547,911 sequences

Remove duplicates

Pandas DataFrame
173,791 sequences

173,791 sequences

Run CD-HIT

(default parameters)

57,389 sequences

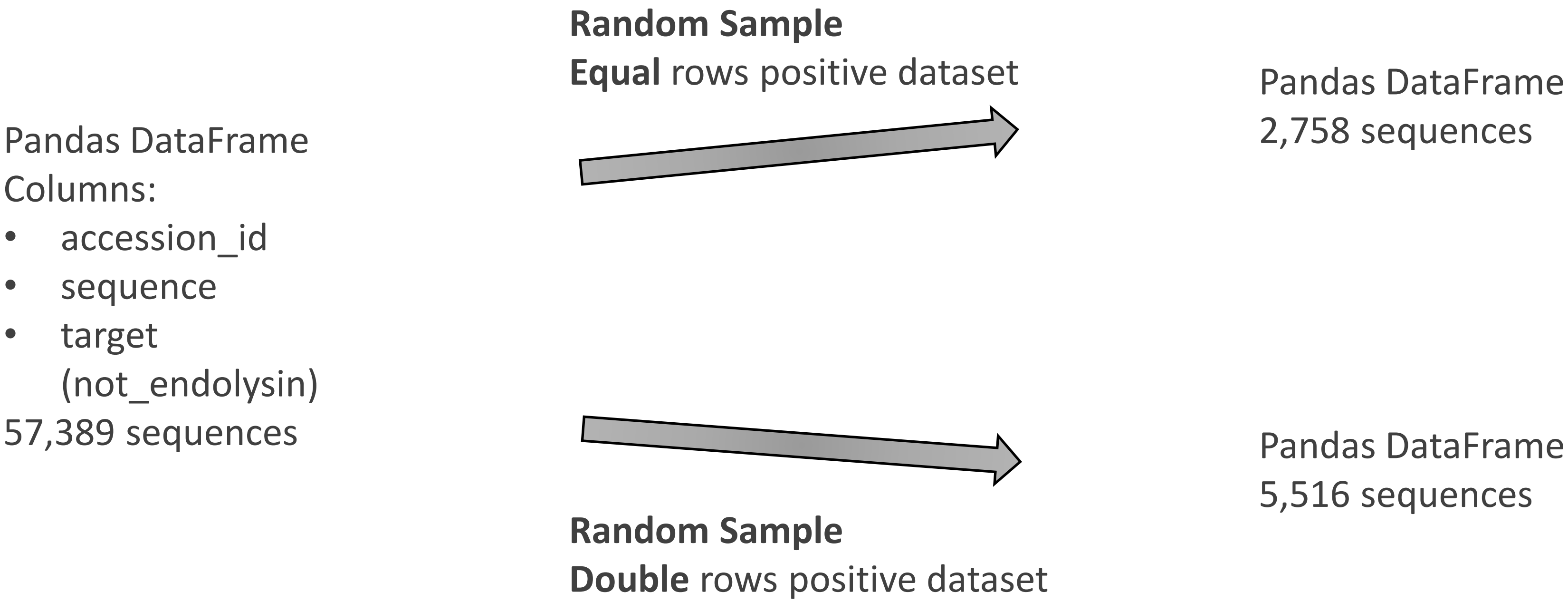
Pandas DataFrame
Columns:

- accession_id
- sequence
- target (not_endolysin)

57,389 sequences

Negative datasets

Positive Dataset:
2,758 sequences



Pandas DataFrame
Columns:

- accession_id
- sequence
- target (not_endolysin)

57,389 sequences

Random Sample
Equal rows positive dataset

Pandas DataFrame
2,758 sequences

Random Sample
Double rows positive dataset

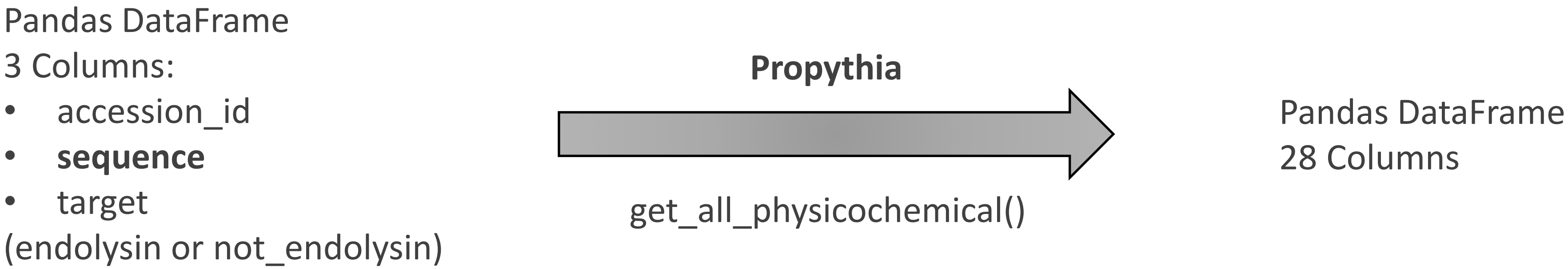
Pandas DataFrame
5,516 sequences

Generate Physicochemical Descriptors with Propythia

Positive Dataset
2,758 sequences

Negative dataset (equal)
2,758 sequences

Negative dataset (double)
5,516 sequences



Pandas DataFrame
3 Columns:

- accession_id
- **sequence**
- target

(endolysin or not_endolysin)

Propythia

get_all_physicochemical()

Pandas DataFrame
28 Columns

Propythia Physicochemical Descriptors

Descriptor	Description
length	length
charge	charge
chargedensity	charge density
formula	calculates number of C, H, N, O and S of the protein sequence
bond	total number of hydrogen, single, double and aromatic bonds
mw	molecular weight
gravy	gravy from a sequence (accordingly to biopython)
aromacity	aromacity (accordingly to biopython)
isoelectric point	isoelectric (accordingly to biopython)
instability index	instability (accordingly to biopython)
secondary structure	fraction of aa that tend to be in helix, turn or sheet
molar extinction coefficient	value of reduced cysteins and oxidized (with disulfid bridges)
flexibility	flexibility according to Vihinen, 1994 (return proteinsequencelenght-9 values) from biopython
aliphatic index	aliphatic index of sequence (1 value) from modlamp
boman index	boman index of sequence (1 value) from modlamp
hydrophobic ratio	hydrophobic ratio from modlamp

Dataset for ML

Positive Dataset
28 columns
2,758 sequences

+

Negative dataset (equal)
28 columns
2,758 sequences

→

Dataset (equal)
28 columns
5,516 sequences

Positive Dataset
28 columns
2,758 sequences

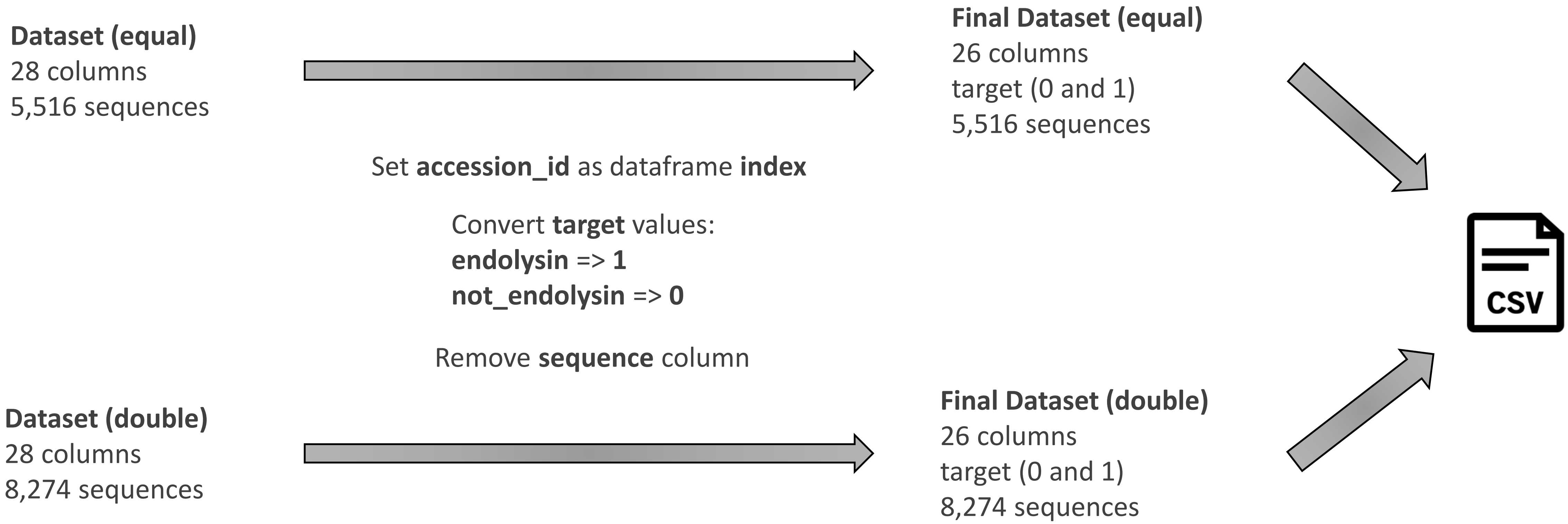
+

Negative dataset (double)
28 columns
5,516 sequences

→

Dataset (double)
28 columns
8,274 sequences

Dataset for ML



Dataset (equal)
28 columns
5,516 sequences

Dataset (double)
28 columns
8,274 sequences

Set **accession_id** as dataframe **index**

Convert **target** values:
endolysin => **1**
not_endolysin => **0**

Remove **sequence** column

Final Dataset (equal)
26 columns
target (0 and 1)
5,516 sequences

Final Dataset (double)
26 columns
target (0 and 1)
8,274 sequences

Train and Test sets split

Dataset (equal)
26 columns
target (0 and 1)
5,516 sequences

Dataset (double)
26 columns
target (0 and 1)
8,274 sequences

	other 25 columns (X)	**target (y)**
Train (70%)	X_train set 3,861 rows (equal) 5,791 rows (double)	y_train set 3,861 values (equal) 5,791 values (double)
Test (30%)	X_test set 1,655 rows (equal) 2,483 rows (double)	y_test set 1,655 values (equal) 2,483 values (double)

Cross-Validation and Hyperparameter Tuning

Model	Selector	Estimator
Logistic Regression (LogR)	**15** features	penalty: **l2**
Random Forest (RF)	**25** features	n_estimators: **100**
Support Vector Machine (SVM)	**20** features	C: **1.0**
Decision Tree (DT)	**20** features	criterion: **gini**
Artificial Neural Network (ANN)	**15** features	hidden_layer_sizes: **(10,)**

Dataset **equal**

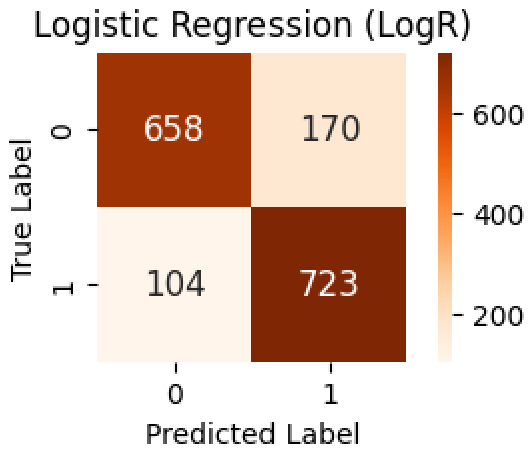
Model	Selector	Estimator
Logistic Regression (LogR)	**20** features	penalty: **l2**
Random Forest (RF)	**25** features	n_estimators: **50**
Support Vector Machine (SVM)	**20** features	C: **1.0**
Decision Tree (DT)	**20** features	criterion: **entropy**
Artificial Neural Network (ANN)	**25** features	hidden_layer_sizes: **(100,)**

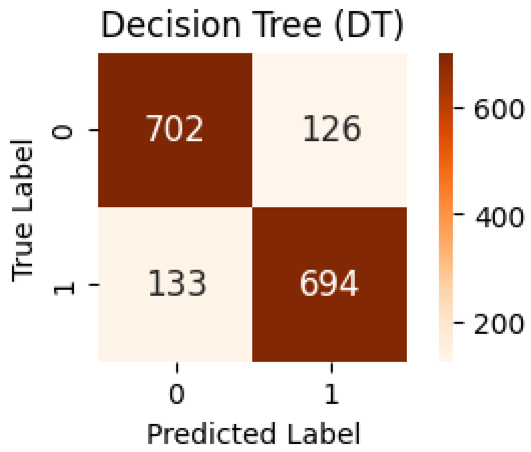
Dataset **double**

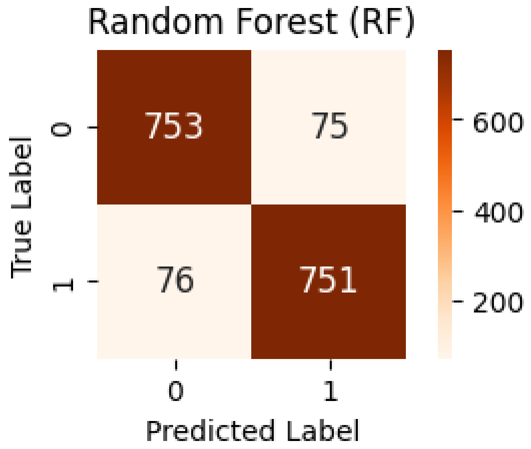
Confusion Matrices

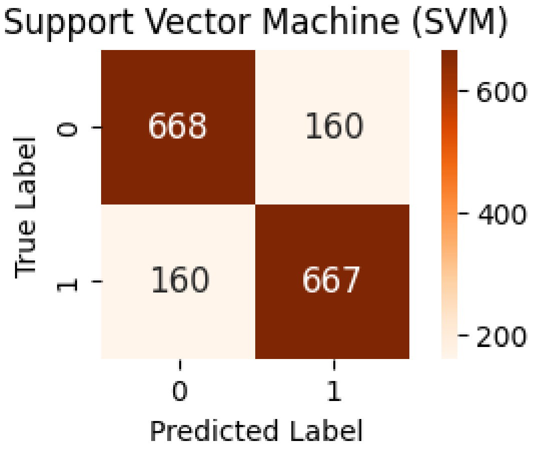
y_test set (**equal**)

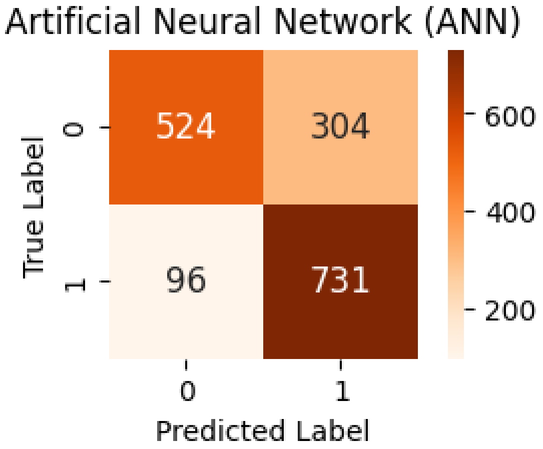
1,655 values


Logistic Regression (LogR)


Decision Tree (DT)


Random Forest (RF)

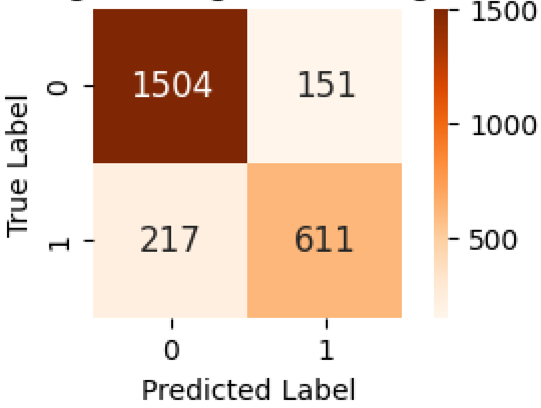

Support Vector Machine (SVM)


Artificial Neural Network (ANN)

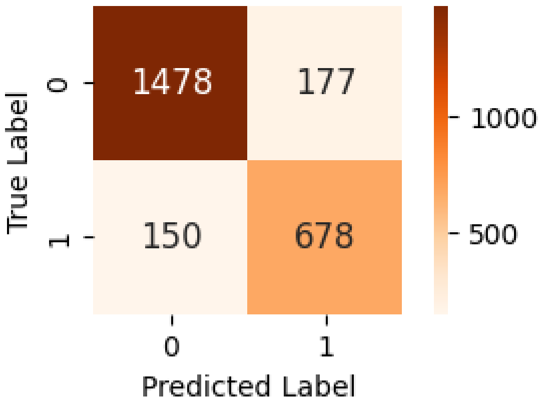
Confusion Matrices

y_test set (**double**)

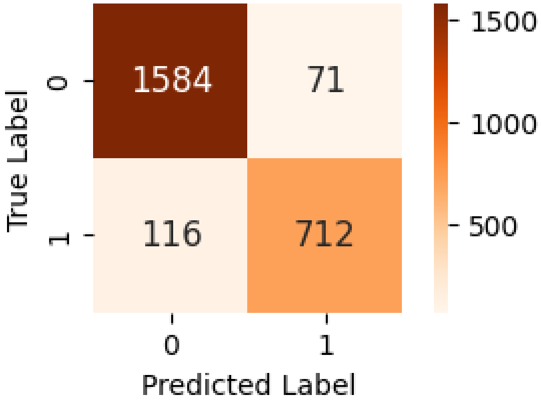
2,483 values



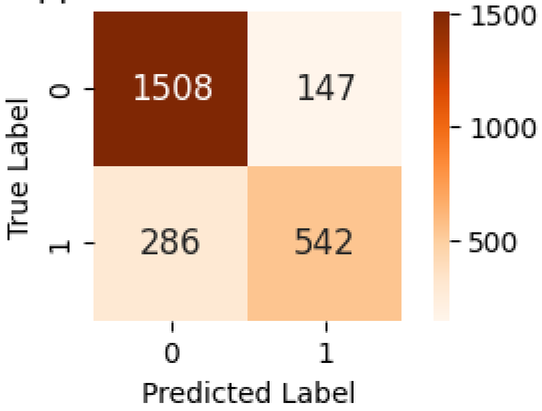
Logistic Regression (LogR)



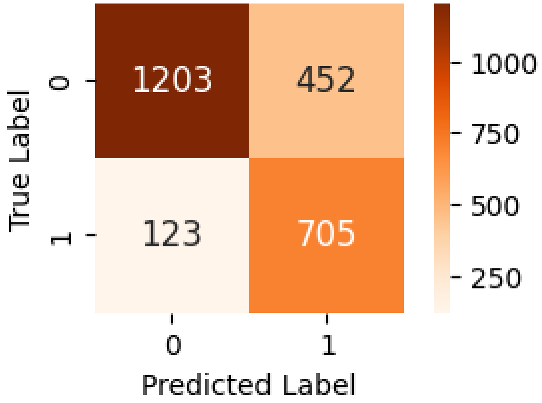
Decision Tree (DT)



Random Forest (RF)



Support Vector Machine (SVM)



Artificial Neural Network (ANN)

Score Metrics

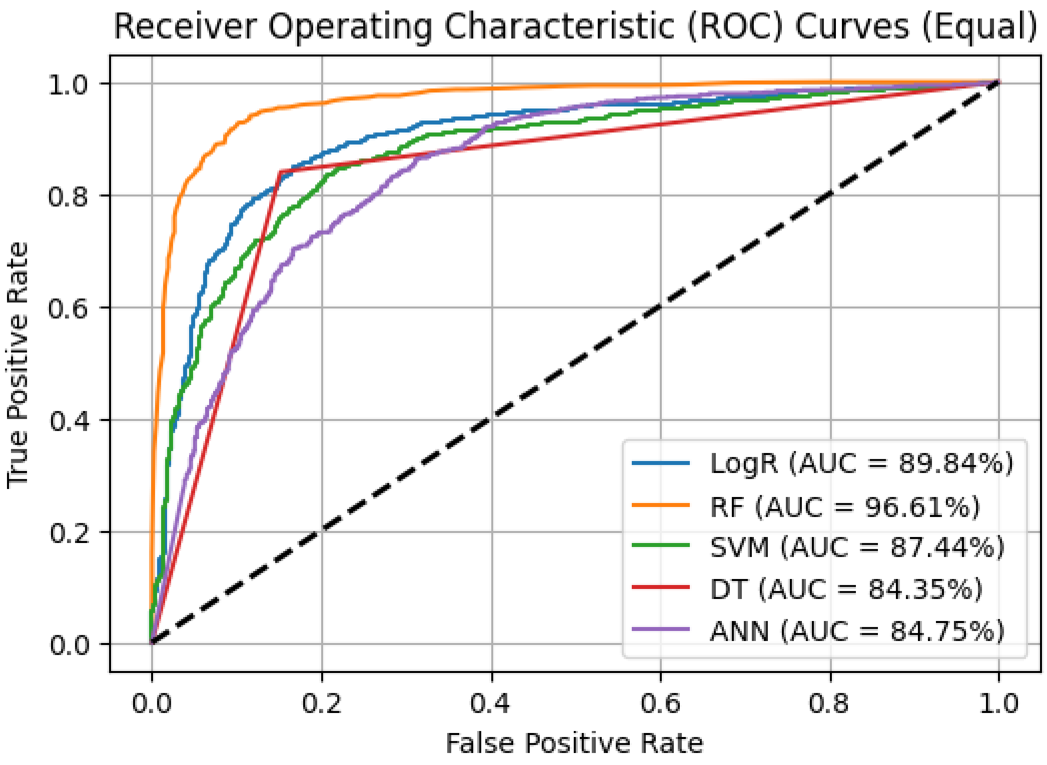
Model	Accuracy	Precision	Recall	F1 Score
LogR	83%	81%	87%	84%
RF	91%	91%	91%	91%
SVM	81%	81%	81%	81%
DT	84%	85%	84%	84%
ANN	76%	71%	88%	79%

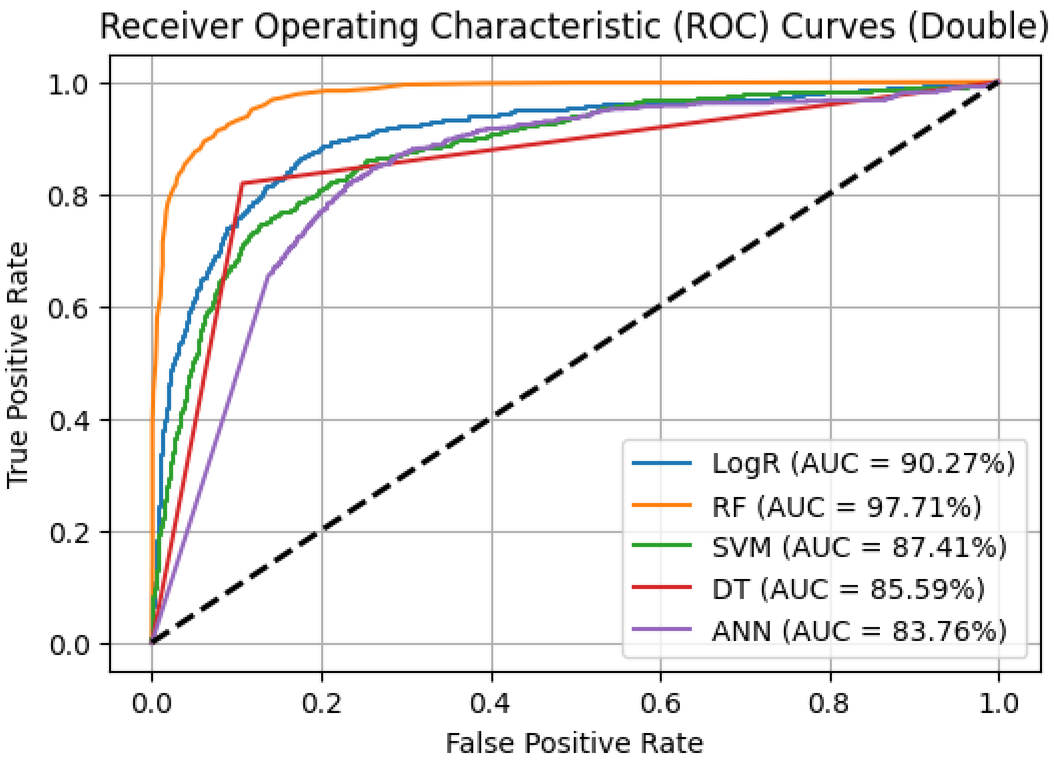
Dataset
equal

Model	Accuracy	Precision	Recall	F1 Score
LogR	85%	80%	74%	77%
RF	92%	91%	86%	88%
SVM	83%	79%	65%	71%
DT	87%	79%	82%	81%
ANN	77%	61%	85%	71%

Dataset
double

Receiver Operating Characteristic (ROC) Curves and Area Under the Curves (AUCs)





Final Remarks and Future Perspectives

- Improved and more consistent results for the **“equal” dataset**;
- **Random Forest** was the **best** performing model, while **ANN** as the **worst** performing one.

- Generate **all Propythia descriptors**;
- **Tuning** of **more hyperparameters**;
- Use **phage lytic protein** and **enzybiotics databases** (PhaLP, PhalydDB, phiBIOTICS, EnzyBase, GMEnzy);
- Use **other proteins** to build the **negative dataset**;
- Incorporate a **differentiation** between **gram-positive** and **gram-negative** targeting endolysins;
- Explore approaches to identify endolysins in genes in **new phage genomes** and **metagenomic data**;
- Development of a **ML tool** to **identify endolysins**.