In silico prediction of phage-bacteria infection networks as a tool to implement personalized phage therapy—INPHINITY

Diogo Leite[1,2], Grégory Resch[3], Yok-Ai Que[4], Carlos Peña[1] and Aitana Neves[1]

[1]*School of Business and Engineering Vaud (HEIG-VD), University of Applied Sciences Western Switzerland (HES-SO), Switzerland & Swiss Institute of Bioinformatics (SIB)*
[2]*School of Engineering of Polytechnic Institute of Porto (ISEP), Porto, Portugal*
[3]*Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland.*
[4]*Department of Intensive Care Medicine, Bern University Hospital (Inselspital), Bern, Switzerland*

SIB | heig-vd HAUTE ÉCOLE D'INGÉNIERIE ET DE GESTION DU CANTON DE VAUD www.heig-vd.ch | HEALTH ENGINEERING ECONOMICS | INSELSPITAL UNIVERSITÄTSSPITAL BERN HOPITAL UNIVERSITAIRE DE BERNE BERN UNIVERSITY HOSPITAL | Unil UNIL | Université de Lausanne

1—Abstract

The emergence and rapid dissemination of antibiotic resistance worldwide threatens medical progress and calls for innovative approaches for the management of multidrug resistant infections. Phage therapy might represent such an alternative. This re-emerging therapy uses viruses that specifically infect and kill bacteria during their life cycle to reduce/eliminate bacterial load and cure infections.

The success of phage therapy however relies on the exact matching between both the target pathogenic bacteria and the therapeutic phage. Therefore, having access to a fully-characterized phage library is necessary to start with phage therapy. An essential second step to conceive personalized phage therapy treatments is the capacity to predict the interactions between the target pathogen and its potential phage. To address this, we aim at developing predictive *in silico* models of phage-bacteria infection networks, using genomic features from sequenced phages and bacteria, and taking advantage of bioinformatics and machine learning techniques.

Using the publicly available information from Genbank and phagesdb.org, we were able to construct a dataset containing +1000 known phage-bacteria interactions with corresponding sequenced genomes. An equal amount of potential negative interactions were added to the dataset by considering the specificity of phage-bacteria interactions. We are currently extracting features from the genomes to build quantitative datasets to train machine learning models. These features include distribution of predicted protein-protein interaction scores, as well as proteins' amino acids frequency and chemical composition. Future work will focus on the development of ensemble machine learning models to optimize the predictive power of our methodology.

2—Context

Phage-therapy is an alternative to antibiotics

Causes of antibiotics resistance

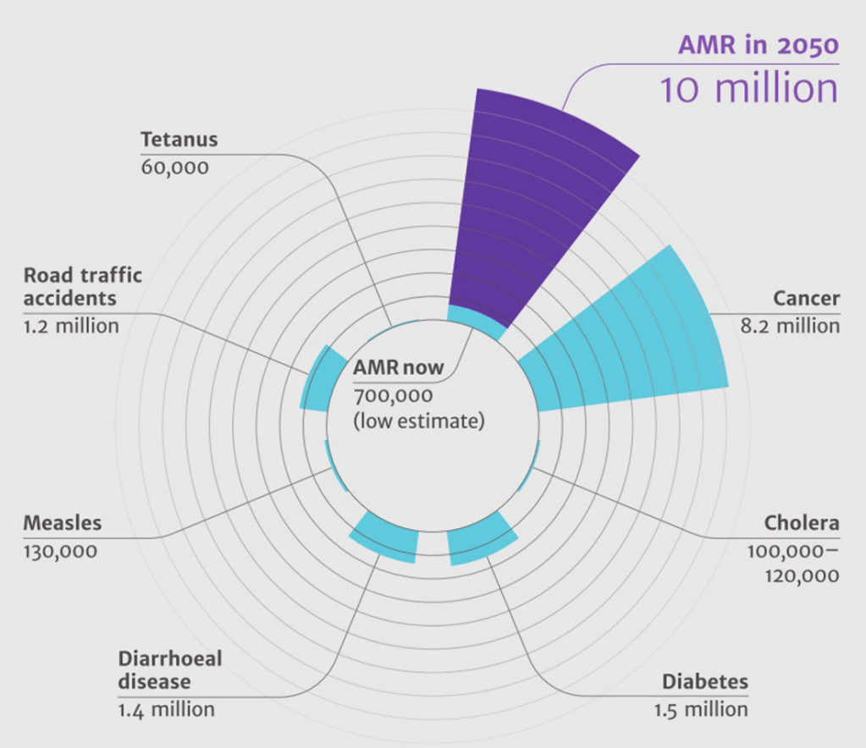
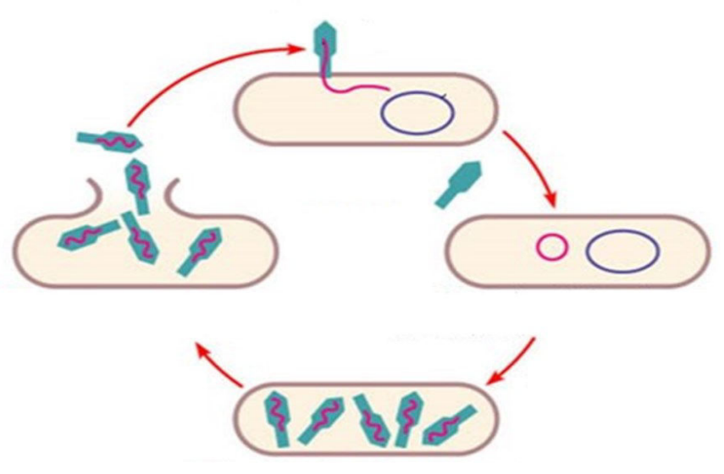


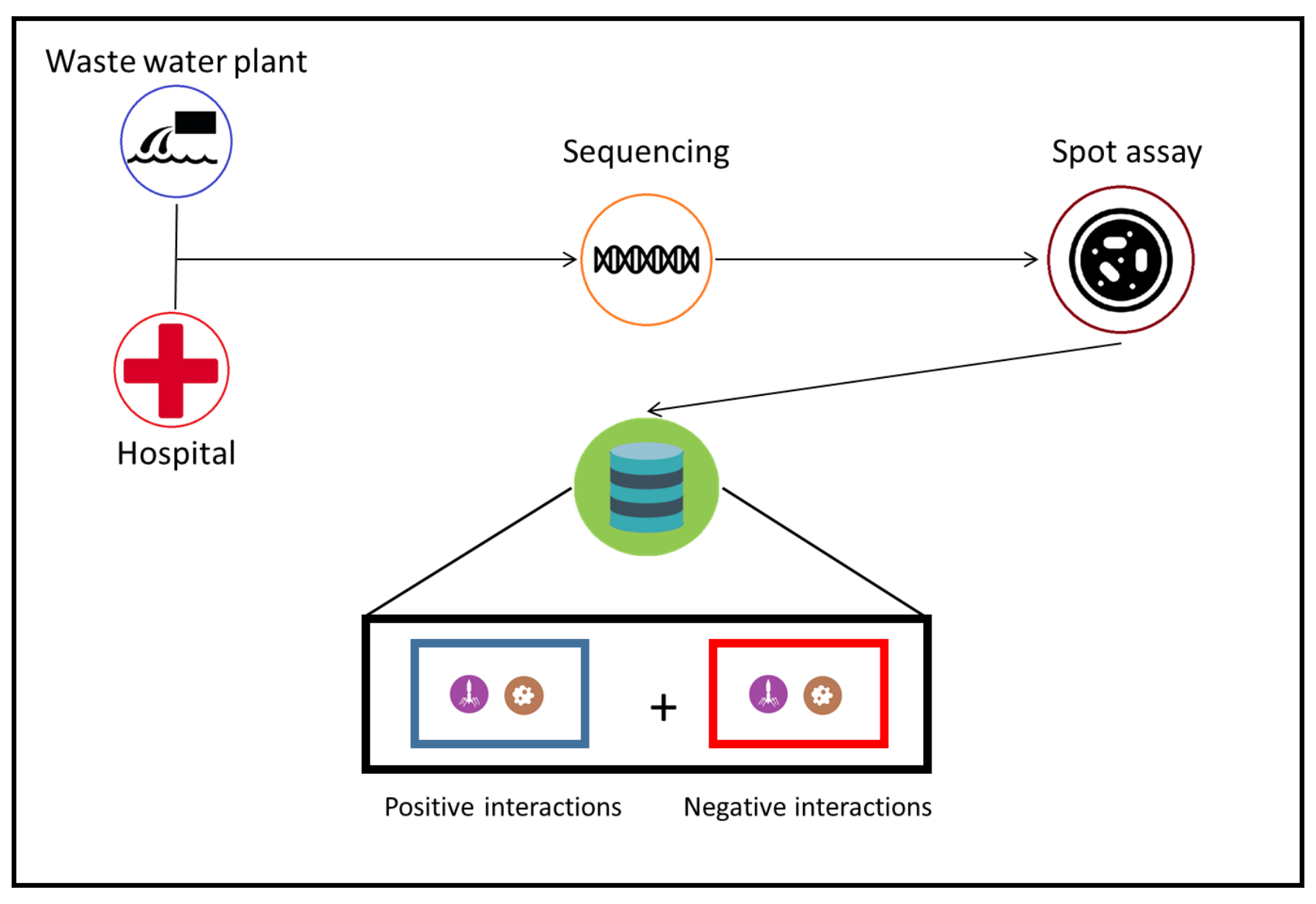
Image from [1]

Cycle of infection of a Lytic phage (results in bacterial death)

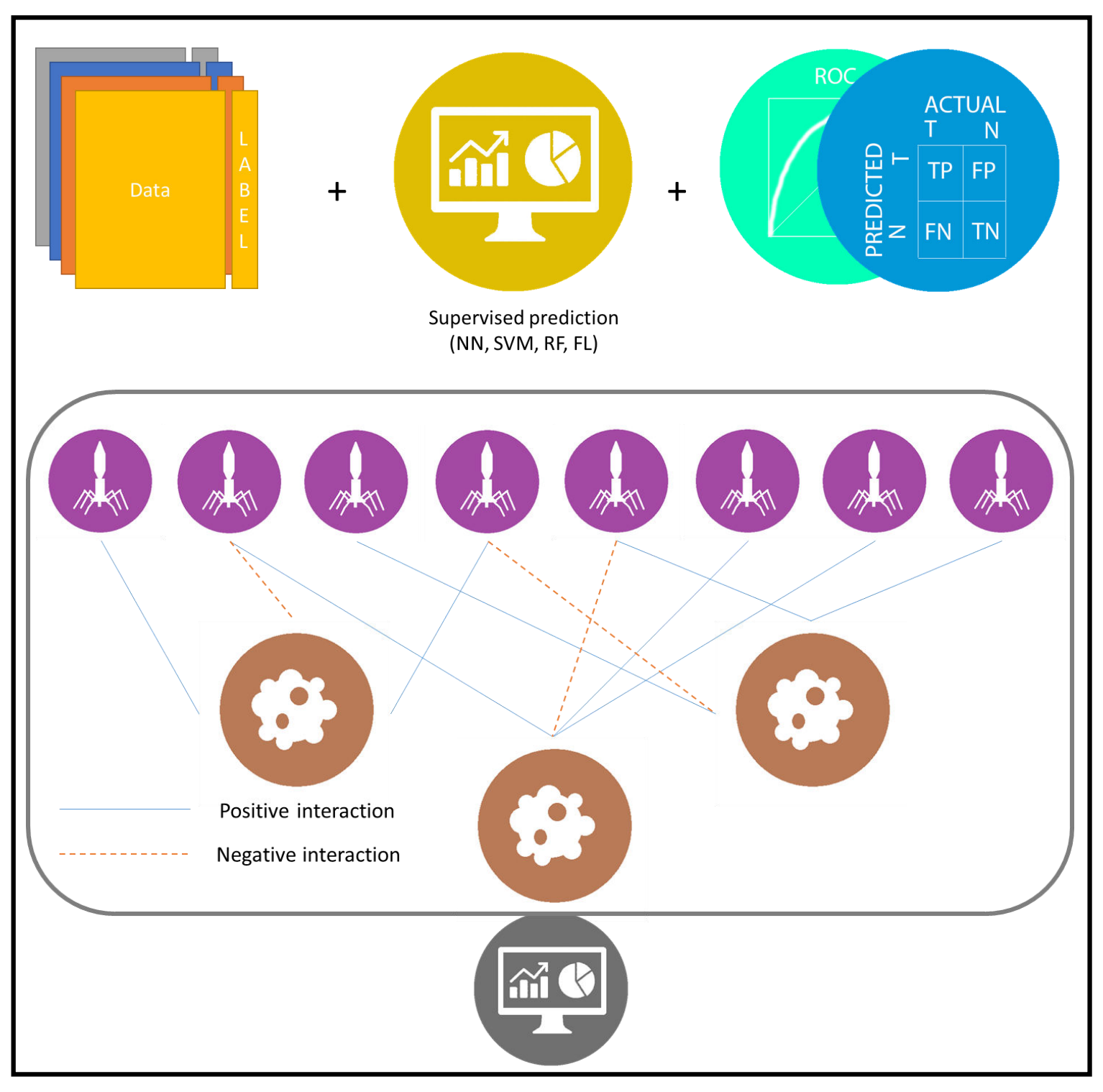


3—Overview of the project

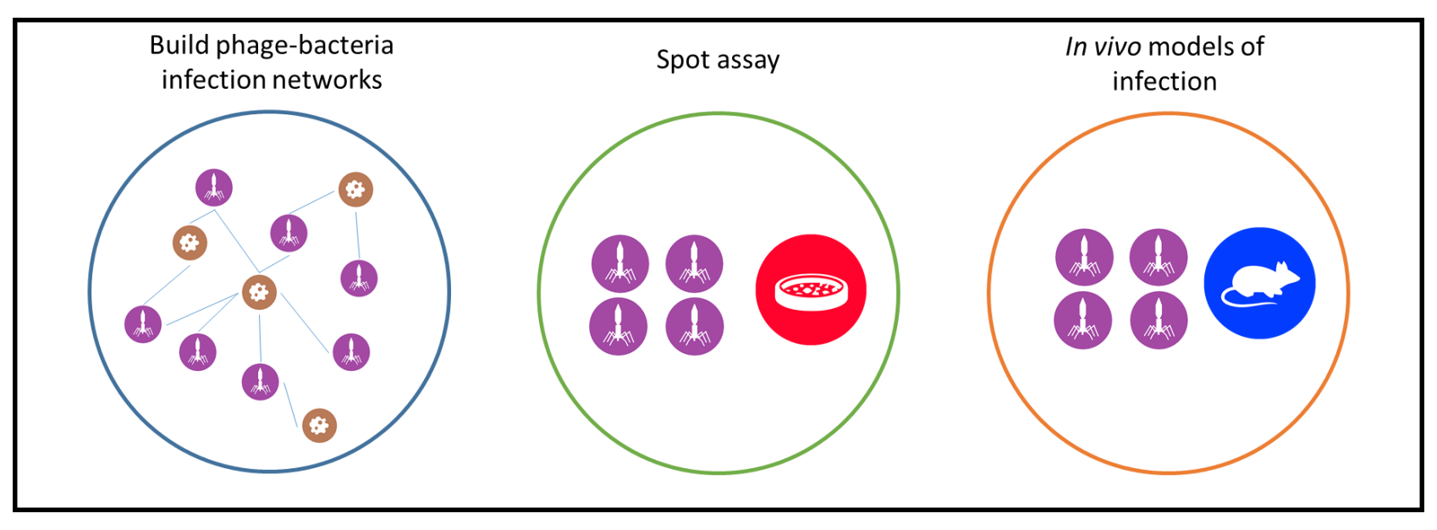
A—Bacteria and phages collection, sequencing and labeling.



B—*In silico* predictions of phage-bacteria interactions. (neural network, SVM, fuzzy logic, random forest)



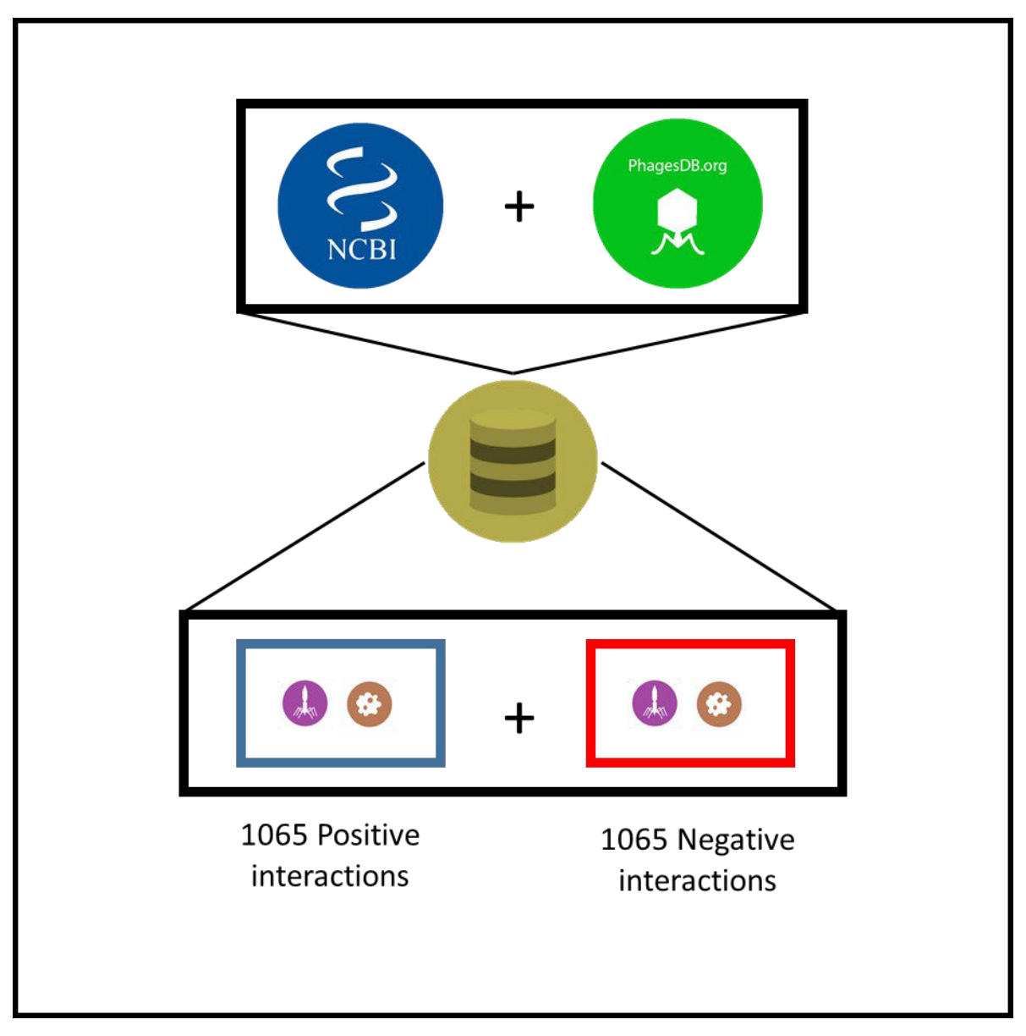
C—Prospective and validation study.


Build phage-bacteria infection networks
Spot assay
In vivo models of infection

Computational objectives

1—**features extraction** -> 2—features selection -> 3—predictive model -> 4— test the model

4—Data used to start working on the computational models



- Collect all publicy available phage genoms from NCBI [2] and Phagesdb.org [3];
- If the host bacteria is sequenced (NCBI[2]), retrieve genome as well;
- Predict genes using GeneMarkS [4] if necessary;
- constitute positive dataset of 1065 entries. A negative dataset of equal size was built artificially by matching phages with bacteria from a different species than their know host.

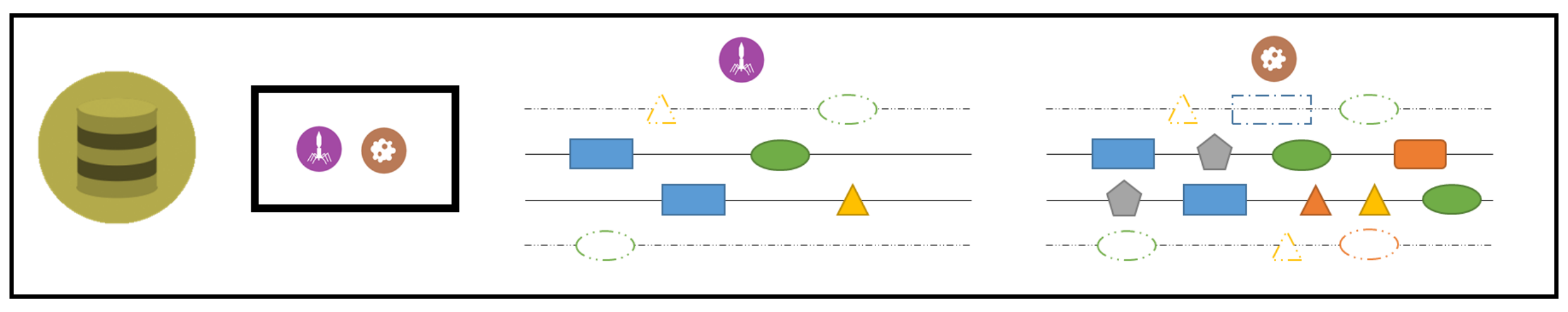
5—Data Sets

Create two type of data sets:

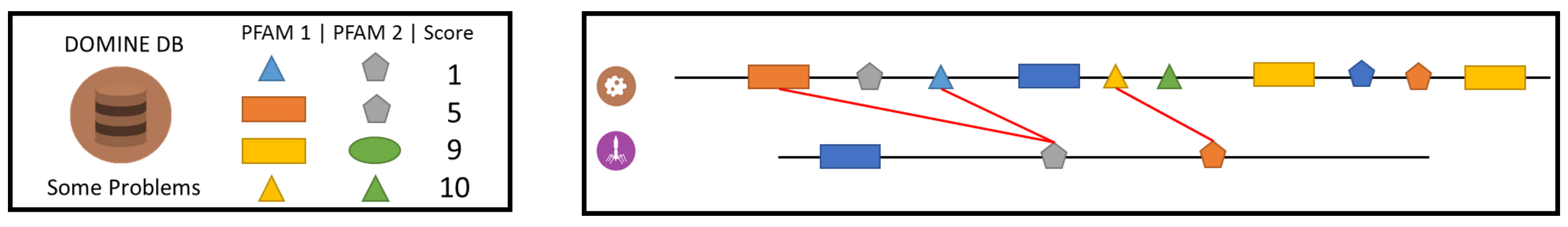
- Based on domains interactions (5[a])
- Based on proteins sequence (5[b])

5[a]—Features extraction with protein domains

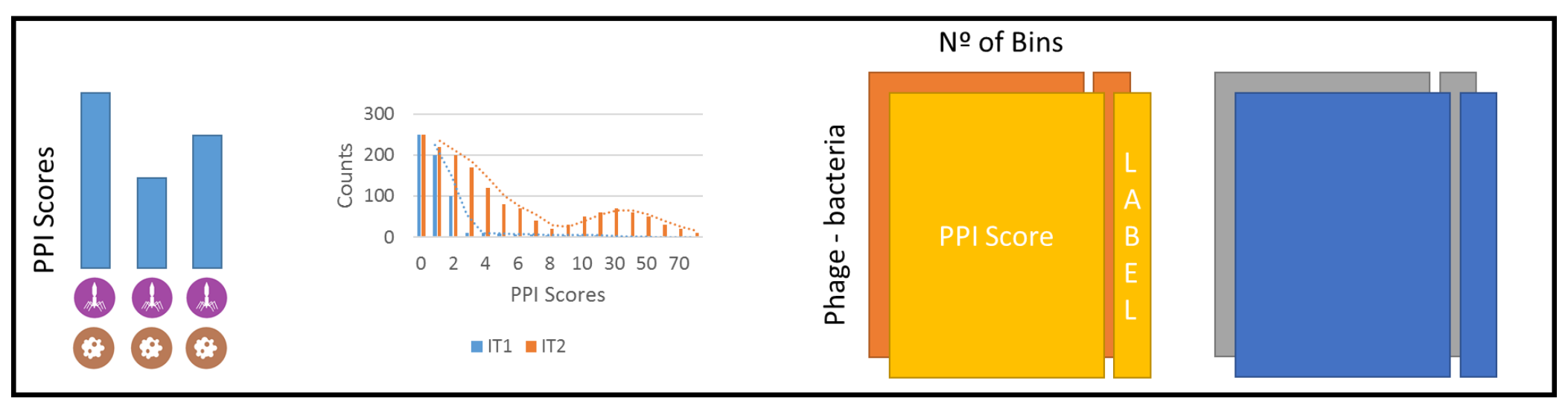
Predict domains using HMMER-PFAM [5] for all proteins of all phages and bacteria in the database



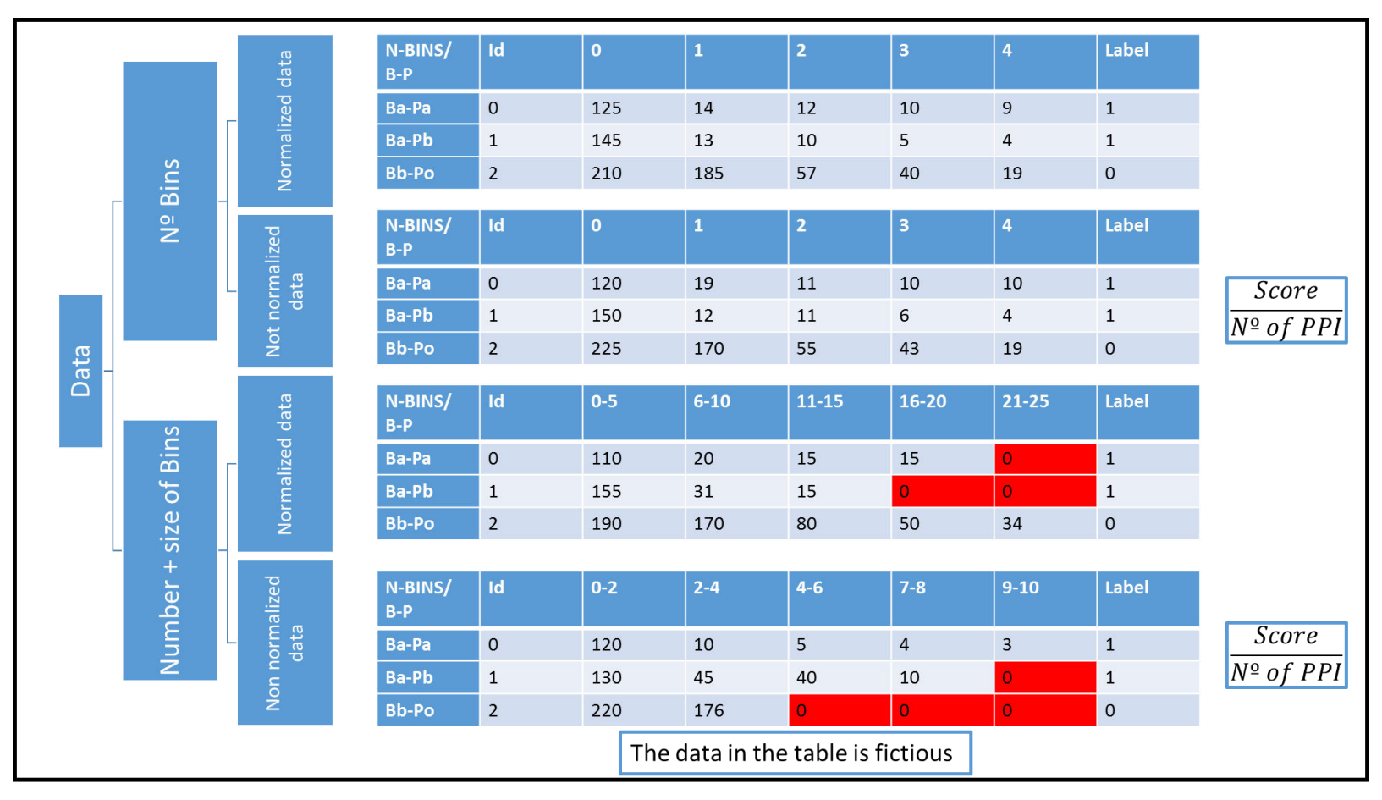
Identify interacting domains (DOMINE [6])



Build data set from PPI scores: use histogram of scores to define a binning and get the same number of features for each pair of phage-bacteria



Example of a fictious dataset obtained whit this procedure


Data
Nº Bins
Number + size of Bins
Normalized data
Not normalized data
Normalized data
Non normalized data

N-BINS/ B-P	Id	0	1	2	3	4	Label
Ba-Pa	0	125	14	12	10	9	1
Ba-Pb	1	145	13	10	5	4	1
Bb-Po	2	210	185	57	40	19	0

N-BINS/ B-P	Id	0	1	2	3	4	Label
Ba-Pa	0	120	19	11	10	10	1
Ba-Pb	1	150	12	11	6	4	1
Bb-Po	2	225	170	55	43	19	0

Score / Nº of PPI

N-BINS/ B-P	Id	0-5	6-10	11-15	16-20	21-25	Label
Ba-Pa	0	110	20	15	15	0	1
Ba-Pb	1	155	31	15	0	0	1
Bb-Po	2	190	170	80	50	34	0

N-BINS/ B-P	Id	0-2	2-4	4-6	7-8	9-10	Label
Ba-Pa	0	120	10	5	4	3	1
Ba-Pb	1	130	45	40	10	0	1
Bb-Po	2	220	176	0	0	0	0

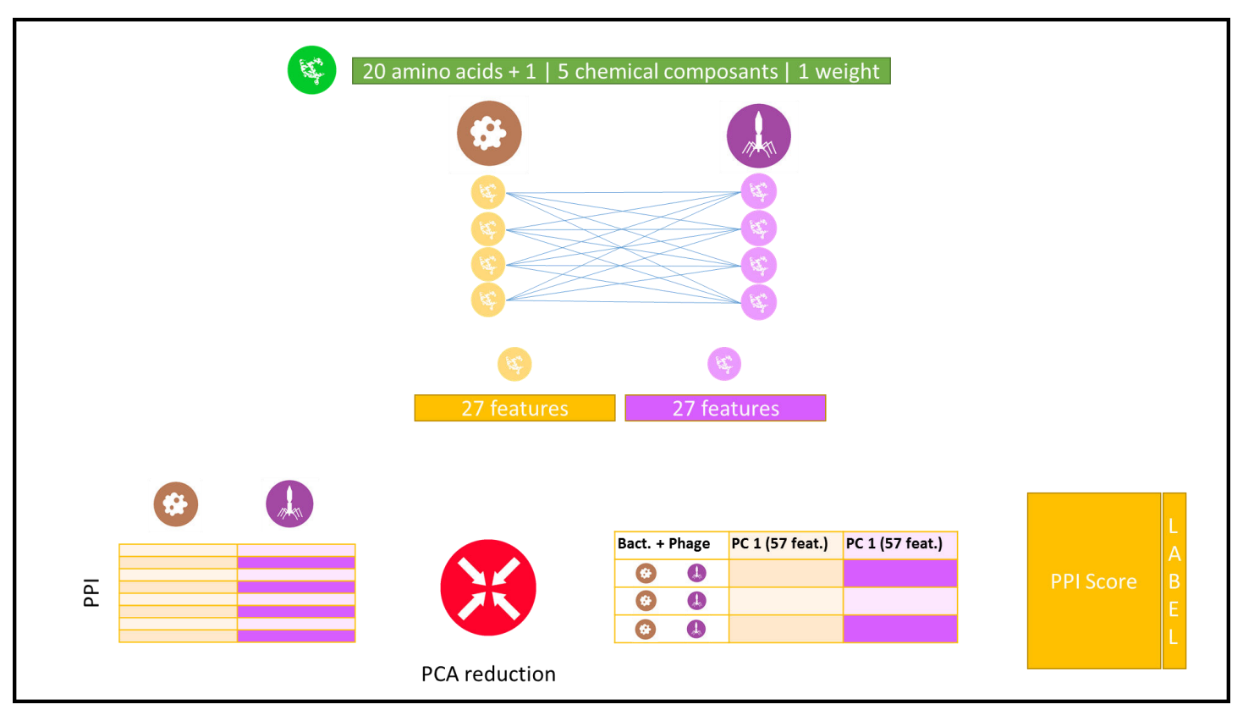
Score / Nº of PPI

The data in the table is fictious

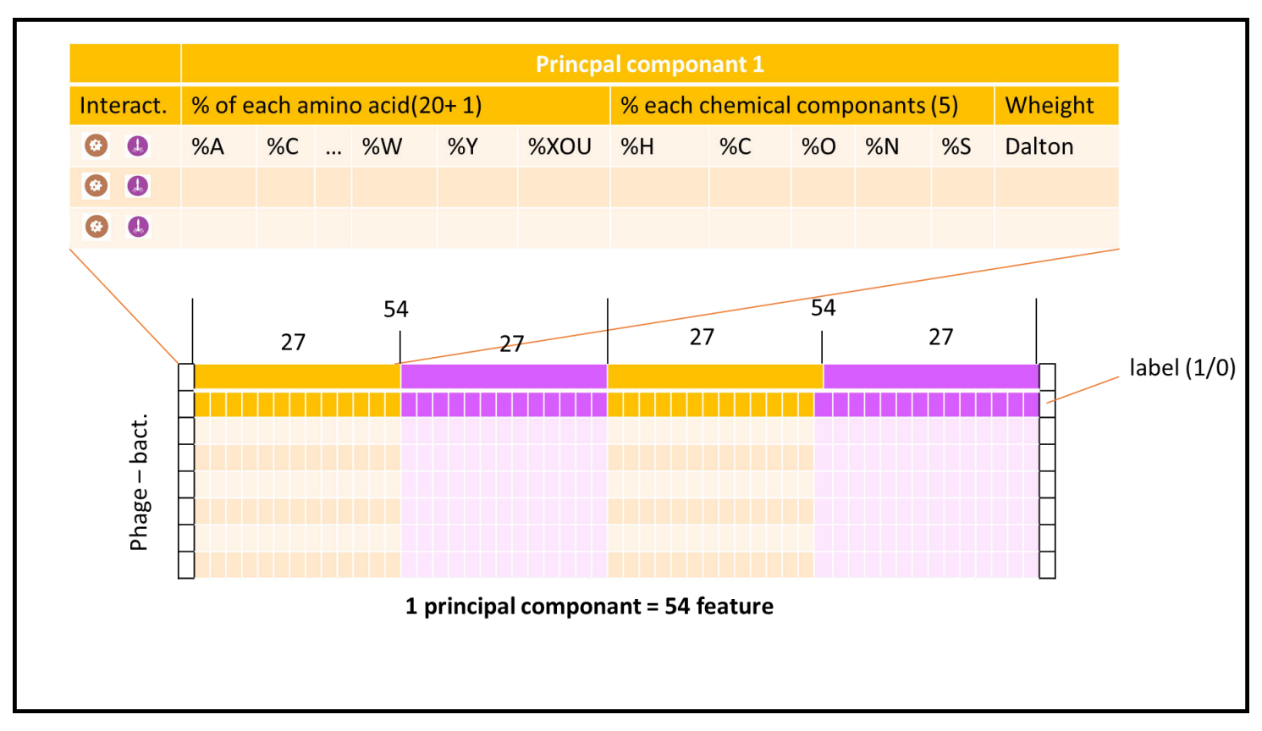
5[b]—Data Set by Protein sequence

Combination of all proteins between each bacteria and phage

- Extract % of amino acid, % chemical components (C,H,O,N,S) and weight -> 27 features
- Concatenate both vectors to get a PPI -> 27 + 27 = 54 features
- For a given pair phage-bacteria we obtain a matrix of size -> Nº PPI × 54
- We use PCA reduction to reduce the number of features per phage-bacteria to ≈ 100 features



Structure of the data set



6—Future work and References

Future work:

- Select the data set to use in machine learning;
- In each data set, remove the redundant/correlated variables and those that provide little information gain;
- Select the tests to assess model performance;
- Select the machine-learning models to build a predictive model;
- Assess model performance.

References:
[1] Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations; [2] http://www.ncbi.nlm.nih.gov/; [3] http://phagesdb.org/; [4]http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi; [5] http://www.ebi.ac.uk/Tools/hmmer/search/hmmscan ; [6] http://domine.utdallas.edu/cgi-bin/Domine; [Features extraction based on article by E. Coelho] Computational prediction of the human-microbial oral interactome