Host Prediction

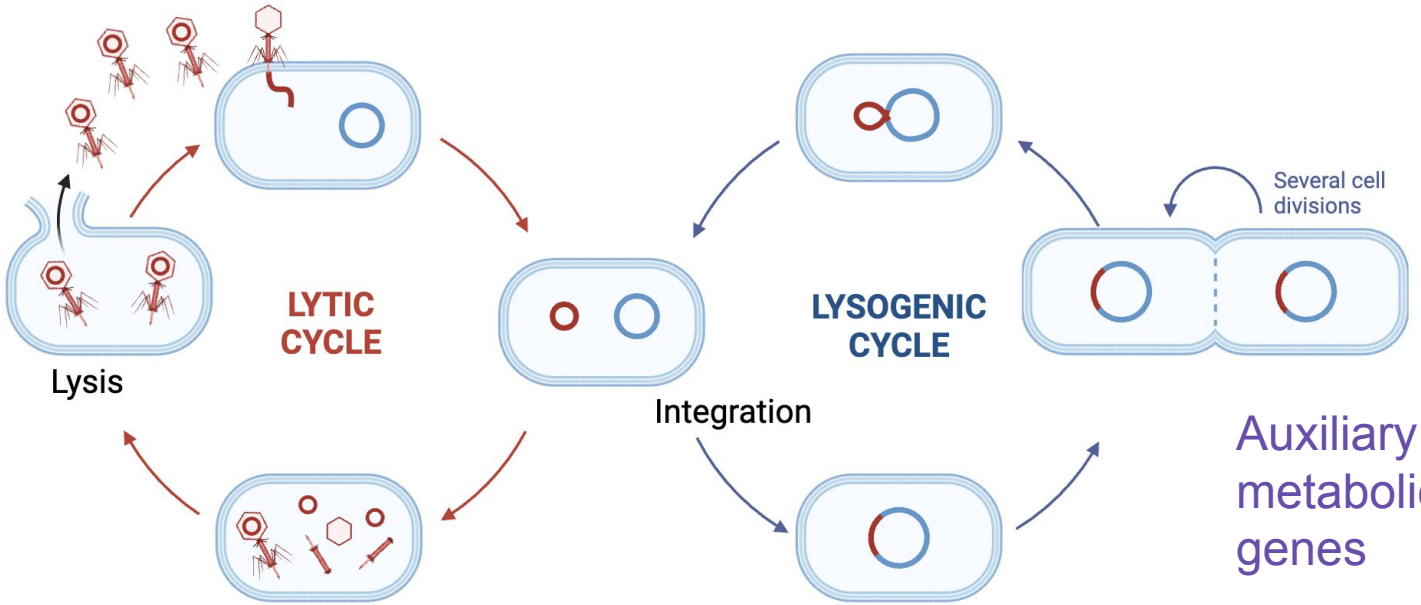
By Malte and Varada

23.09.2024

Background

Viruses affect microbial communities and therefore their environments THROUGH their hosts.

Release organic matter



Auxiliary metabolic genes

Background

Ideally, you would have an isolate bacteria that you test phages on… but we just have our data

These “signals” are based on **biological interactions**

What are some biological interactions?

Adsorption - attachment

Insertion of the genome into the cell

Horizontal gene transfer

Defense/anti-defense mechanisms

- Crispr
- Restriction/modification

Using cellular machinery

- tRNA
- Ribosome binding sites
- Regulatory RNAs
- Auxiliary metabolic genes
- Codon usage
- Modifying stress response

LYSIS

Homology Based

Prophages /viral genes

tRNAs

Transduction

Auxiliary Metabolic Genes

Viral genes inside host genome

Host genes inside viral genome

CRISPR spacers

Non-Homology Based

Host Defense Mechanisms

Using cell machinery

Ecological Dynamics

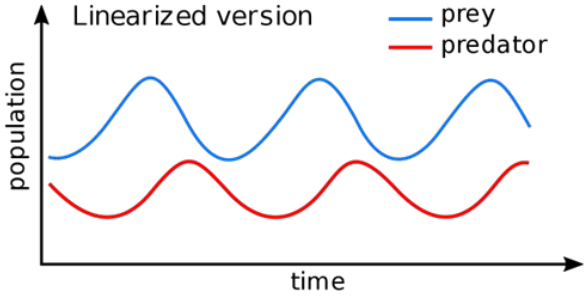
K-mer profiles

Matching Coverage Profiles

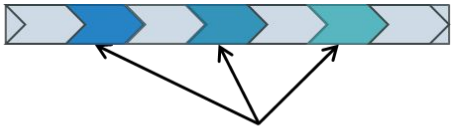
Host genome

Metabolic gene?

Viral genome



ATGC, GCTT, TACC etc



Viral genome snippets (25-35 bp)

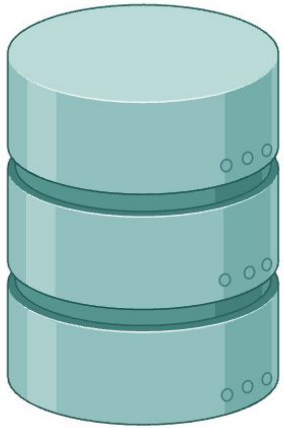
Where to find the hosts?

Database

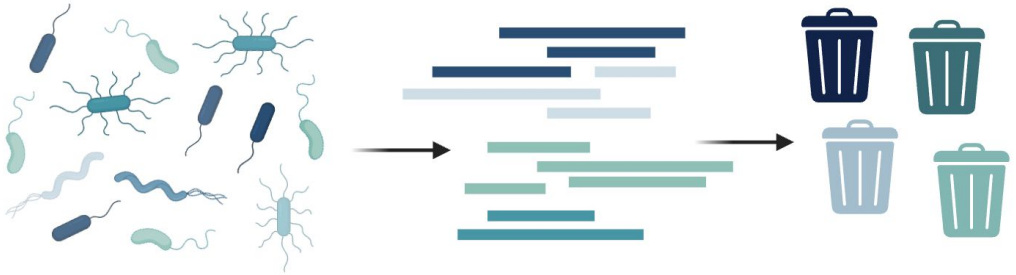
CRISPR spacers DB

Refseq

IMGVR/ Mgnify



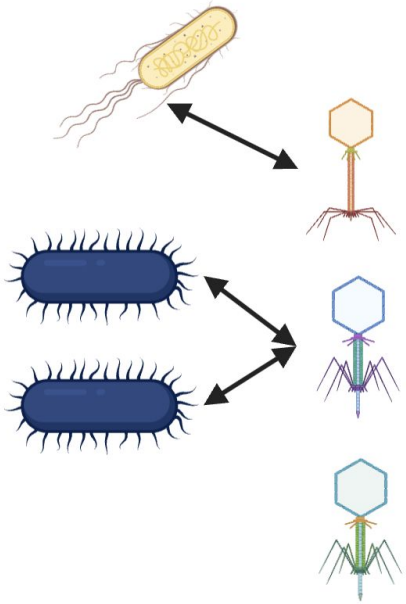
Prokaryotic fraction of your metagenome



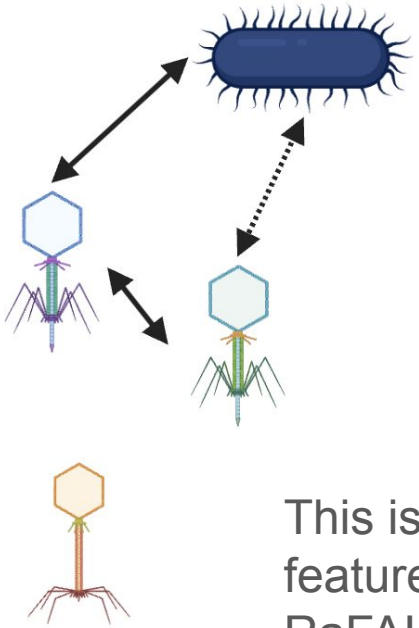
“Bins” or MAGs

“Host-based” vs. “Phage-based”

Phage-host



Phage-phage



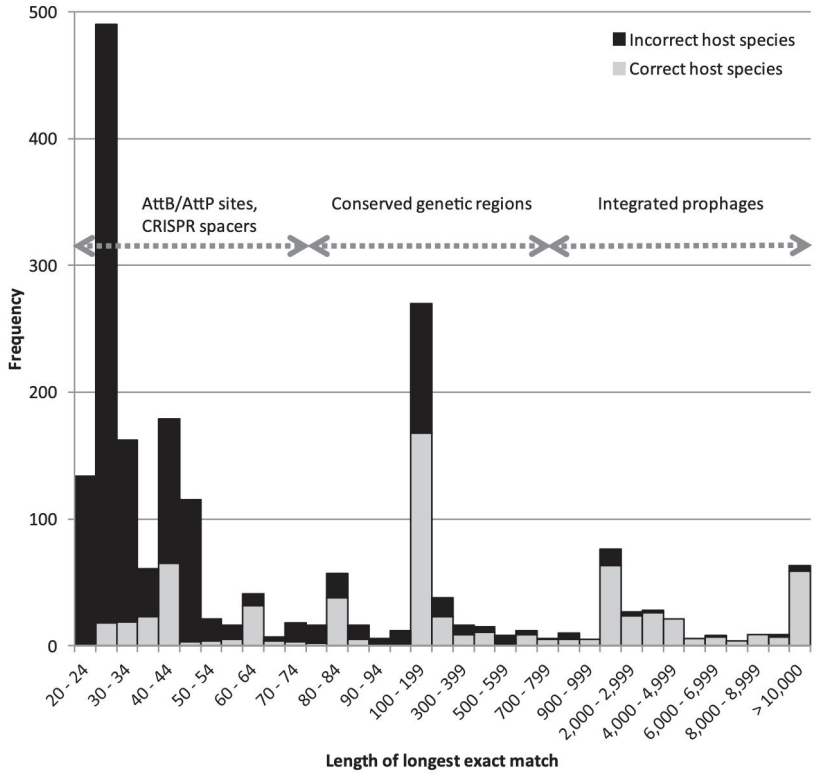
This is a feature of RaFAH also!

Disadvantages - homology-based

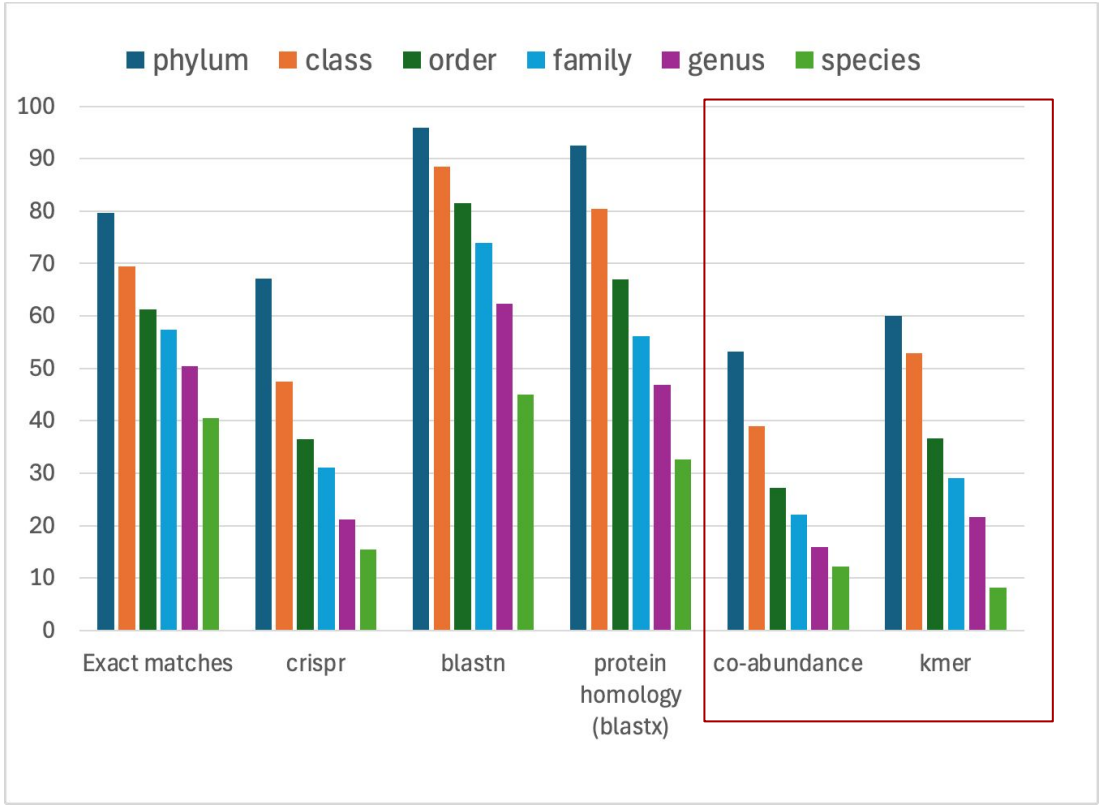
1. A recall/sensitivity tradeoff
2. Simply not enough matches
3. No CRISPR arrays found?

One does not simply

blast a host



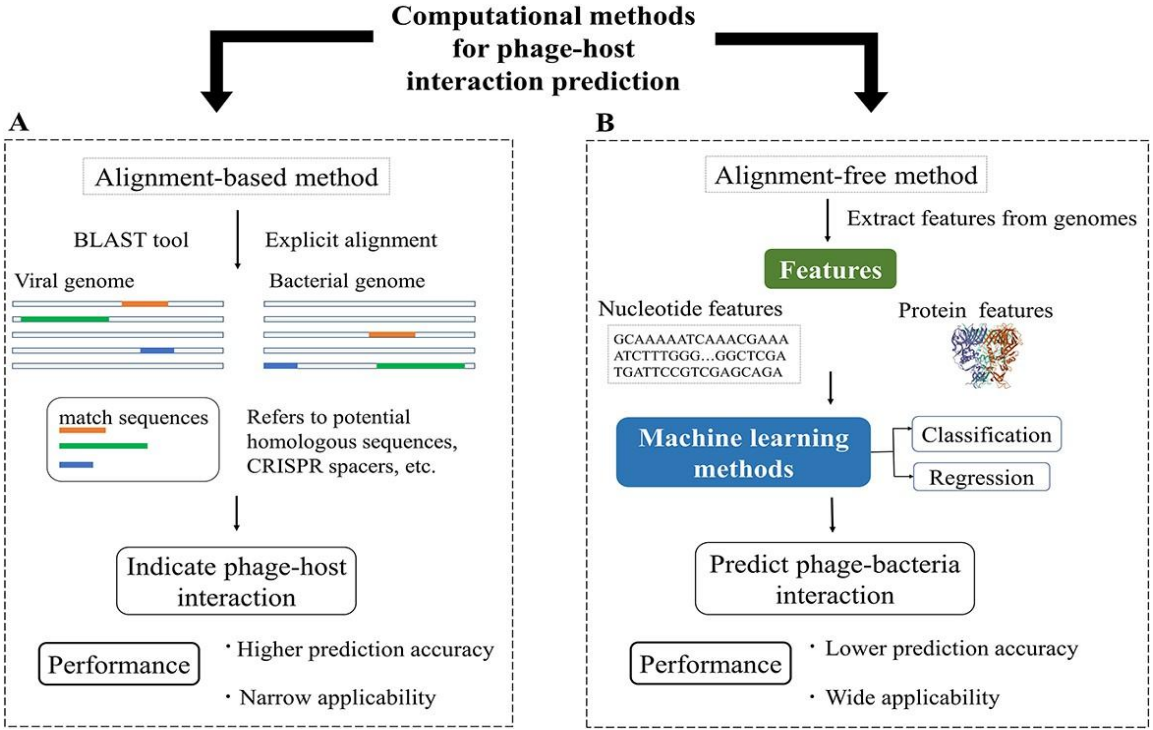
Disadvantages - non-homology



High recall, but matches many hosts!

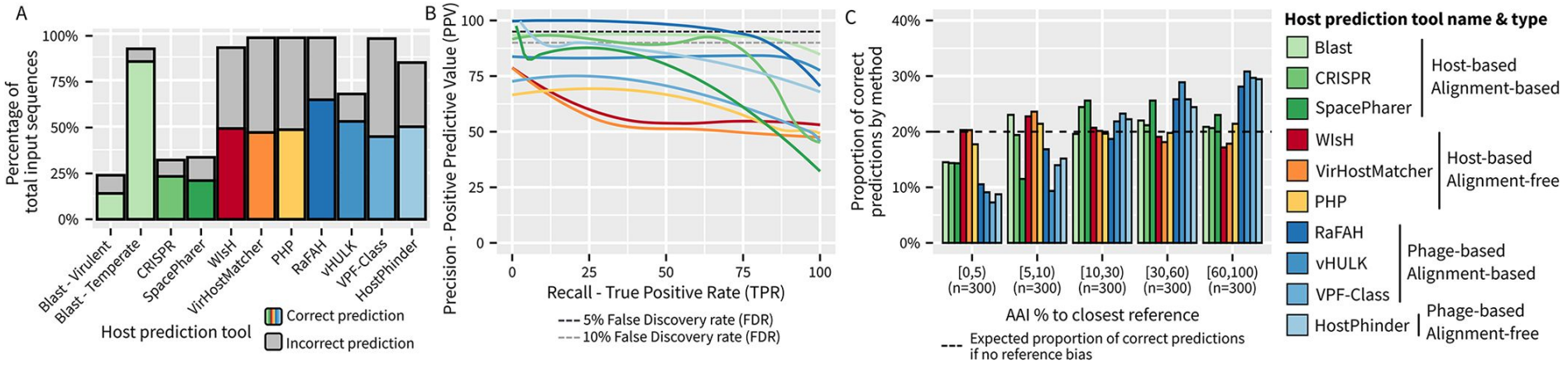
Machine learning methods for phage-host interactions

Many methods – all have biases



Nie et al., *Briefings in Bioinformatics*, 2024

Many methods – all have biases



Roux et al., *PLOS Biology*, 2023

ML methods for phage host interactions

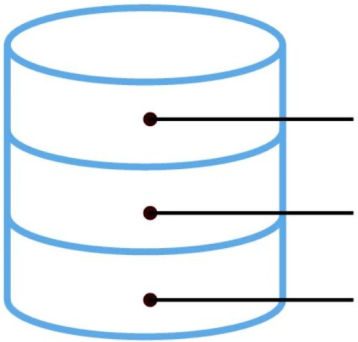
- Which features are informative for PHI
- Training data and some related caveats
- Which ML algorithms are used to predict PHI

Informative features for PHI

- WIsH: 8th order Markov models of host genomes (k-mers)
- RaFAH: viral proteins mapped to protein families
- Boeckaerts et al.: sequence and structure of receptor-binding proteins

Informative features for PHI

A Database construction



UniProtKB: manual searches for RBPs

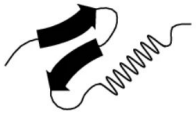
UniRef: clusters from mapped RBPs in UniProt

NCBI: collected phage genomes from MillardLab

B Feature construction

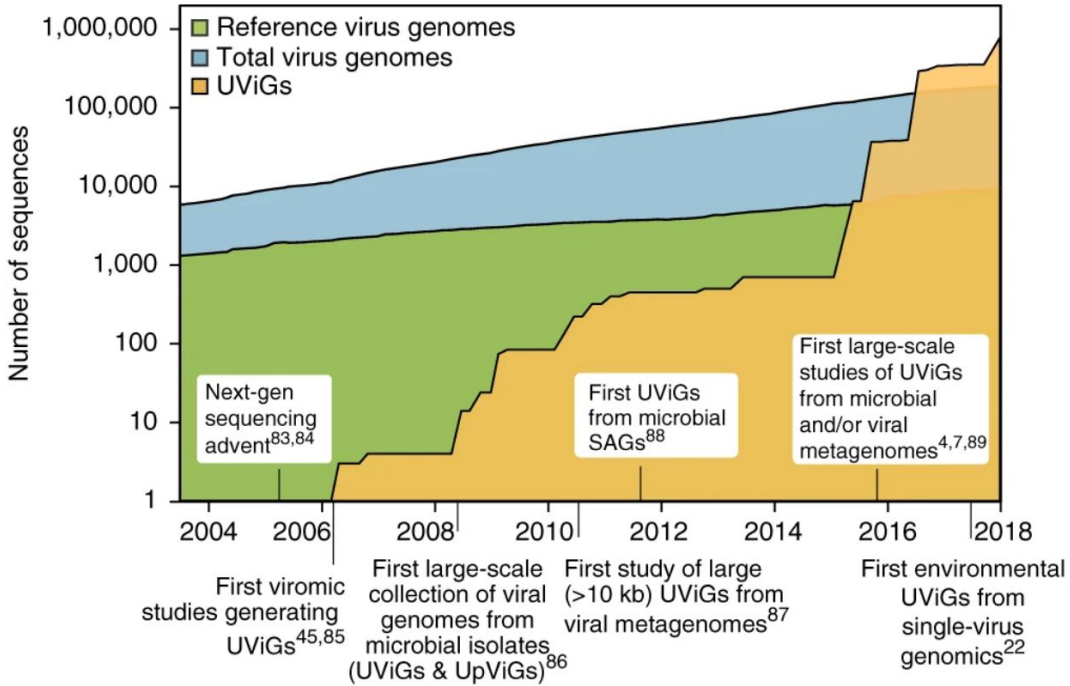
CDS ATGTCCACGA…
- nucleotide composition
- codon composition
- codon usage bias

Protein sequence MSTITQFPS…
- amino acid composition
- amino acid characteristics

Protein structure 
- fraction predicted α-helices
- fraction predicted β-sheets
- fraction predicted turns

Boeckaerts et al., *Sci Rep*, 2021

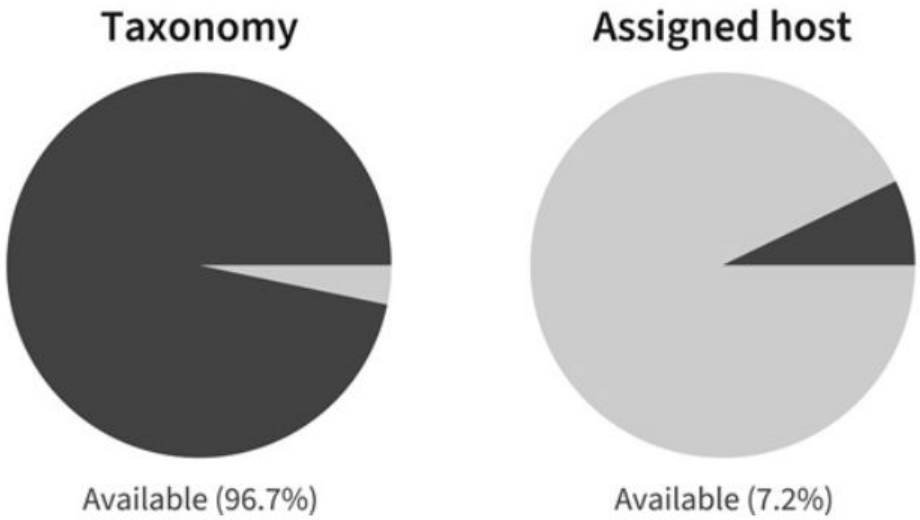
Training data – the good



Roux et al., *Nat Biotechnol*, 2019

Training data – the bad

IMG/VR 4 database



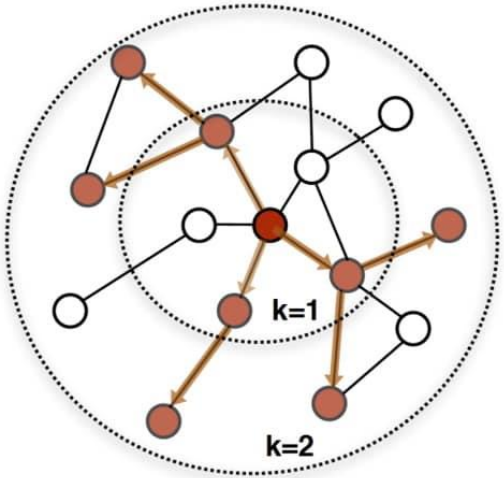
Camargo et al., *NAR*, 2023

Training data – the ugly

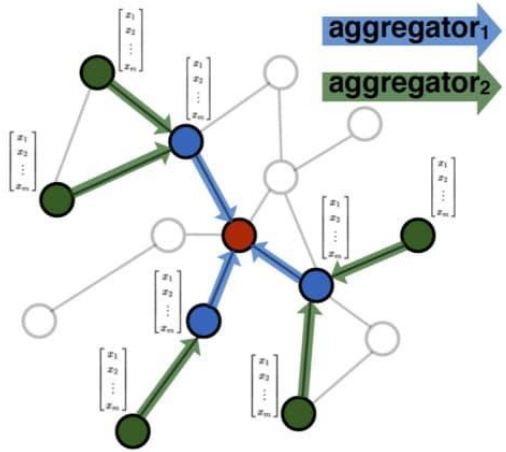
- Very skewed datasets (most phages concentrated on few hosts)
  -> subsample large datasets

- No negative examples
  -> random sampling of hosts distant to known hosts
  -> model-based sampling

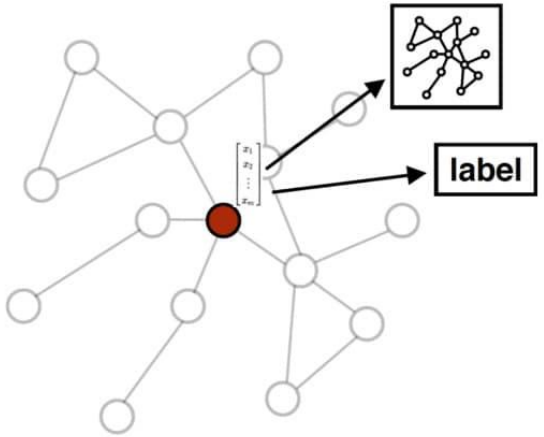
Machine learning algorithms - GNN



1. Sample neighborhood



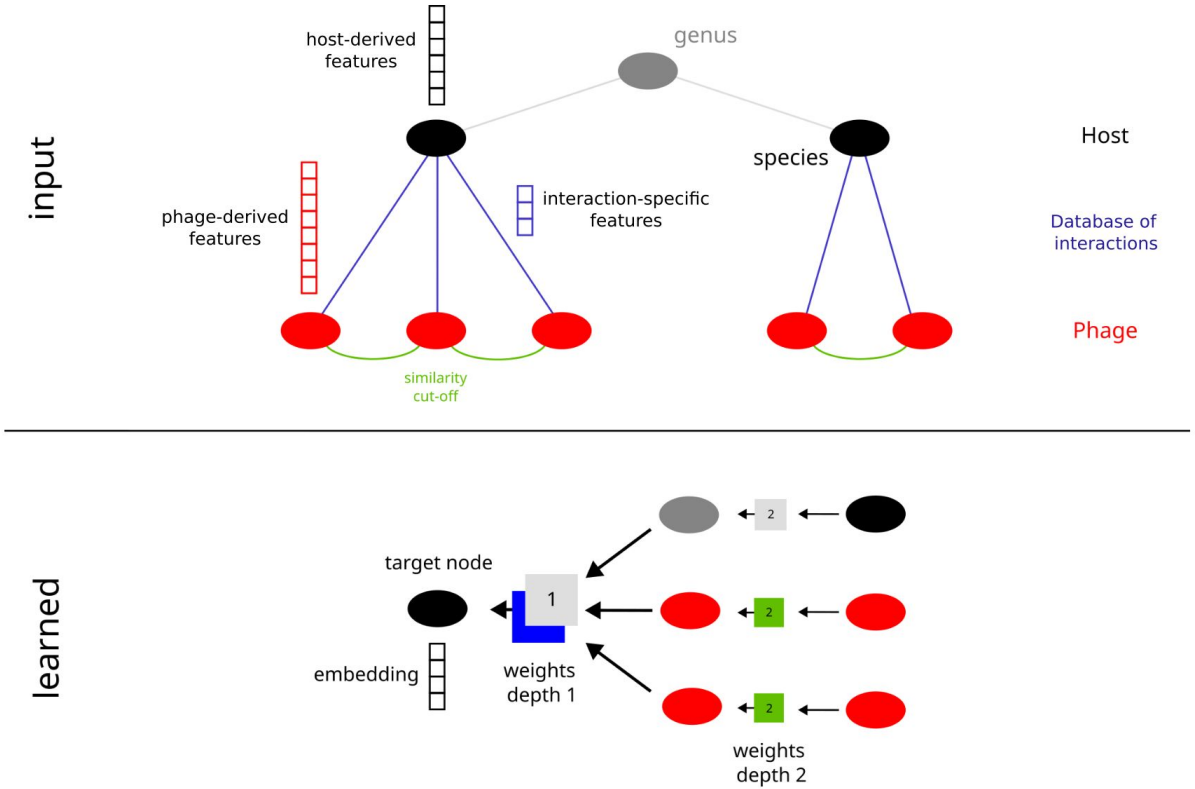
2. Aggregate feature information from neighbors



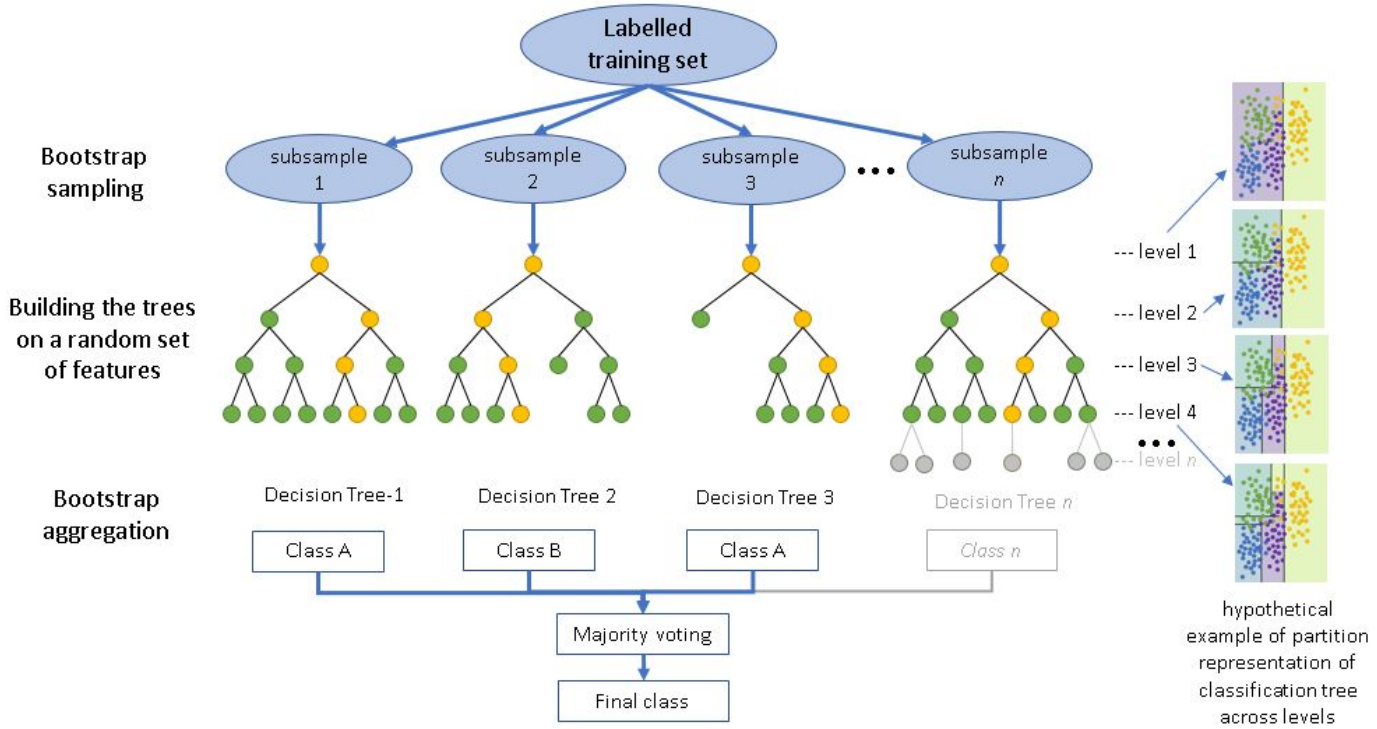
3. Predict graph context and label using aggregated information

Hamilton et al., *arxiv*, 2017

Machine learning algorithms - GNN



Machine learning algorithms - Random Forest



https://catalyst.earth/catalyst-system-files/help/concepts/focus_c/oa_classif_intro_rt.html