



**HAL**  
open science

# VirHunter: a deep learning-based method for detection of novel viruses in plant sequencing data

Grigorii Sukhorukov, Macha Nikolski

## ► To cite this version:

Grigorii Sukhorukov, Macha Nikolski. VirHunter: a deep learning-based method for detection of novel viruses in plant sequencing data. PlantGen 2021, Jun 2021, Novosibirsk, Russia. hal-03271424

**HAL Id: hal-03271424**

**<https://hal.science/hal-03271424>**

Submitted on 25 Jun 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

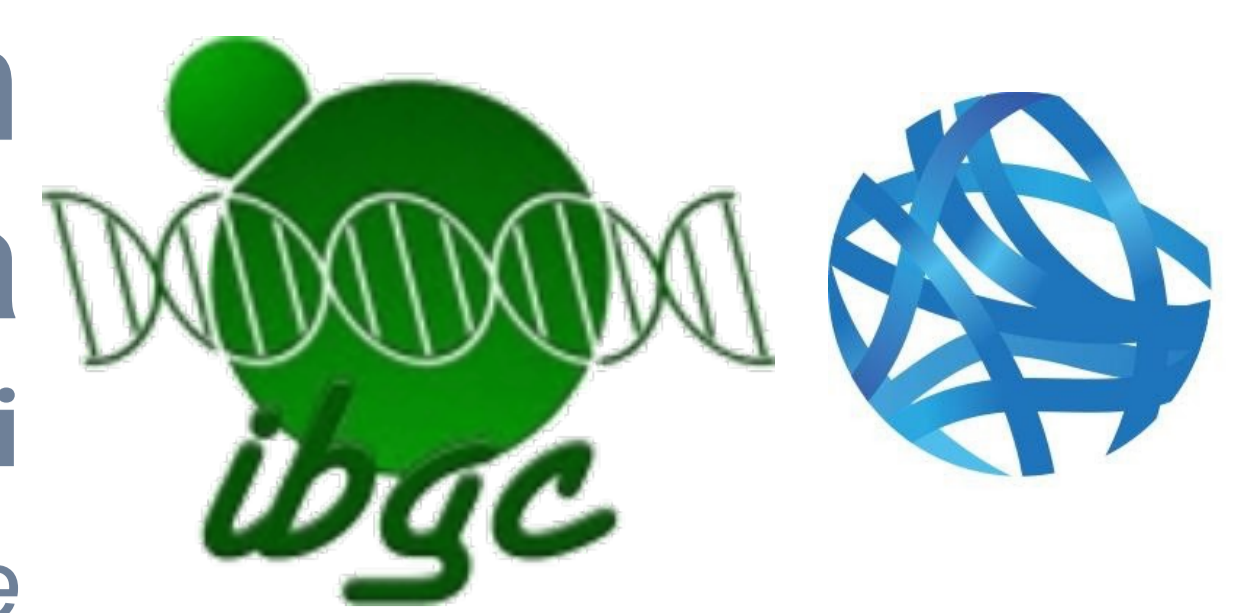
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# VirHunter: a deep learning-based method for detection of novel viruses in plant sequencing data

Sukhorukov Grigori, Macha Nikolski

IBGC/CNRS and CBiB, Bordeaux University, France

✉ grigori.sukhorukov@u-bordeaux.fr



cbib

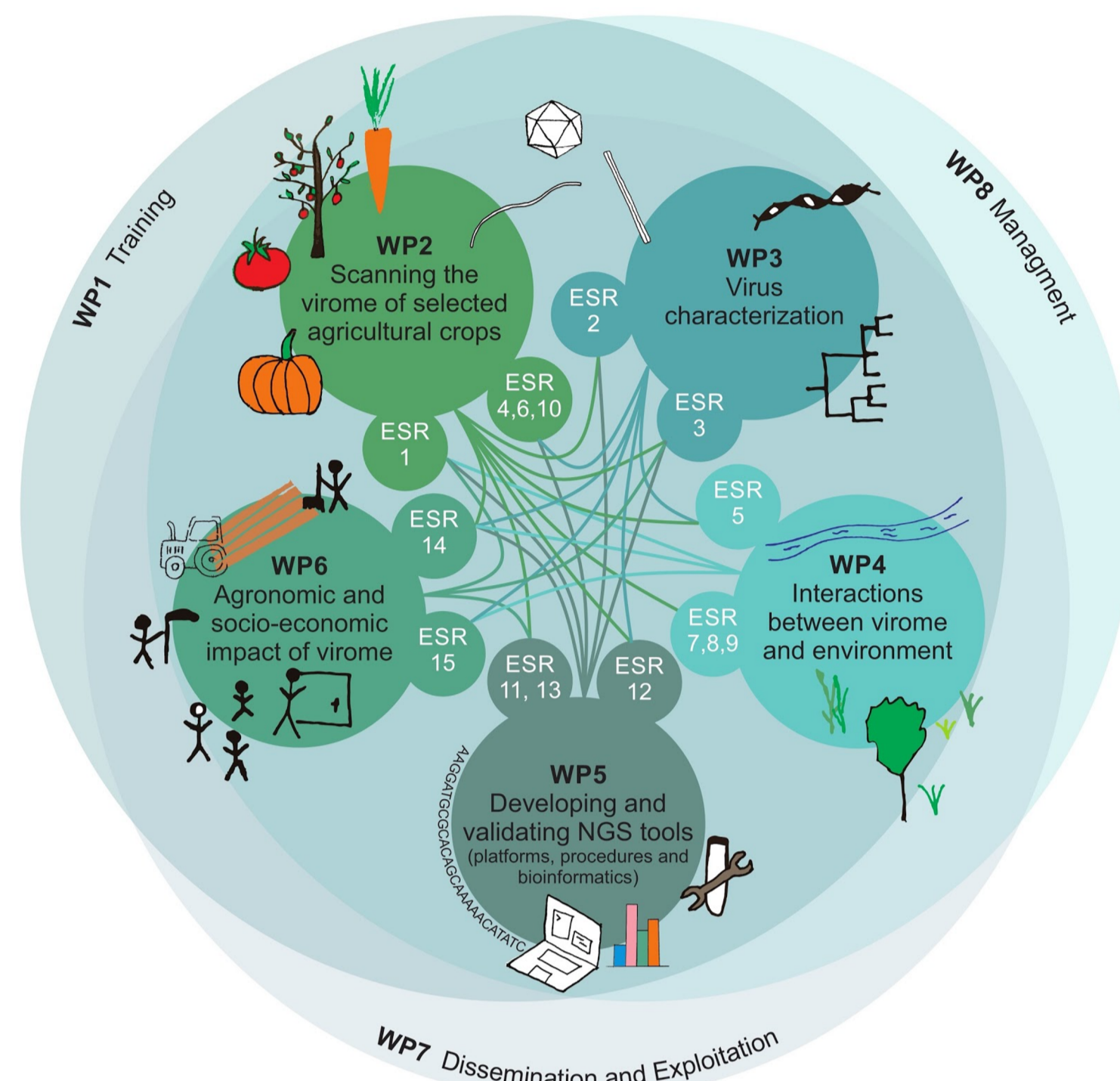
BIOINFORMATIQUE

## Plant viruses in our lives

- **Plant Viruses cause 50% of the emerging plant diseases** and pose an important threat to many agricultural crops worldwide, **they are responsible for production losses estimated at € 15 to 45 billion per year.**

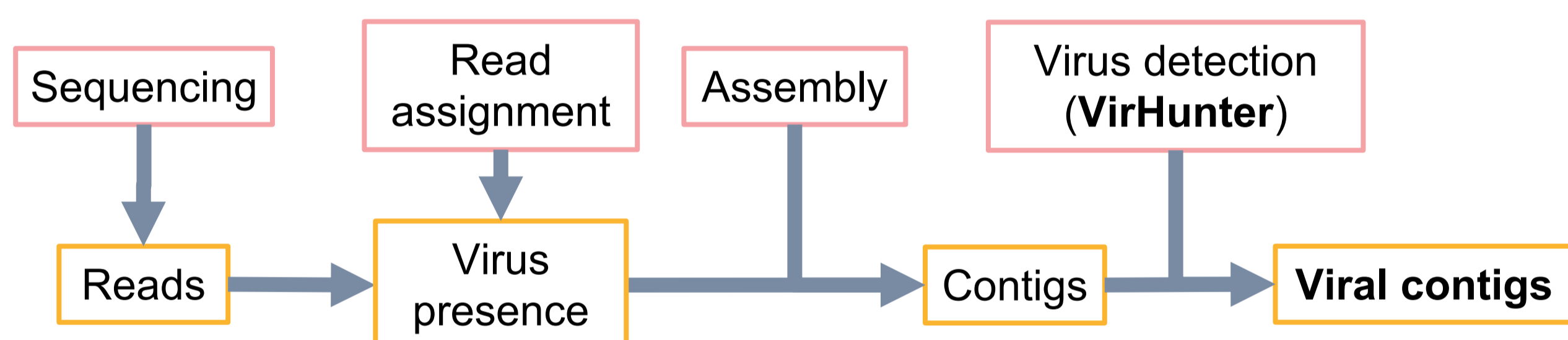
## MSCA ITN INEXTVIR

- Network for Next Generation Training and Sequencing of Virome classification
- Our research: Bioinformatics pipelines development for virome classification



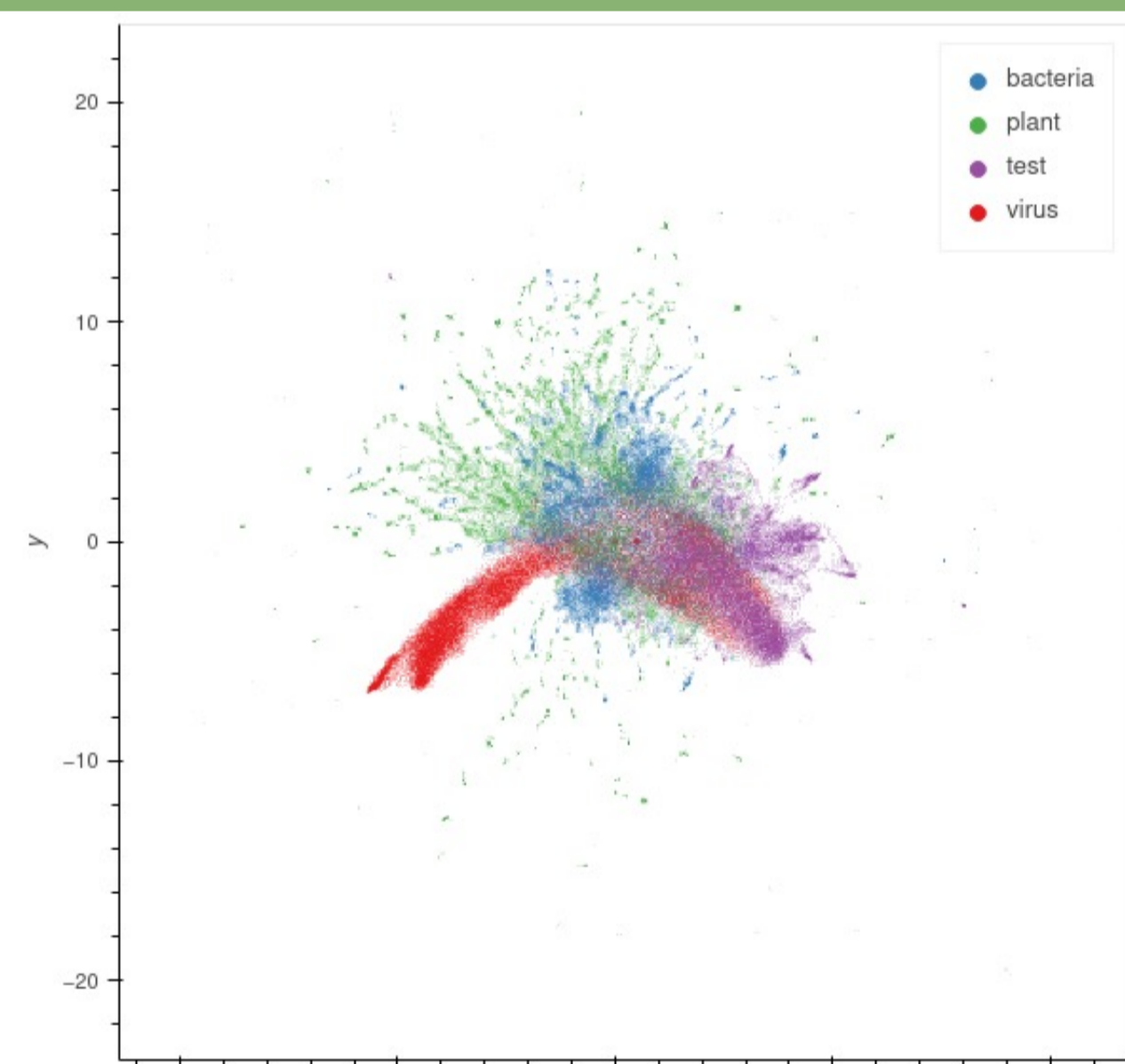
INEXTVIR project landscape

## Novel virus detection workflow



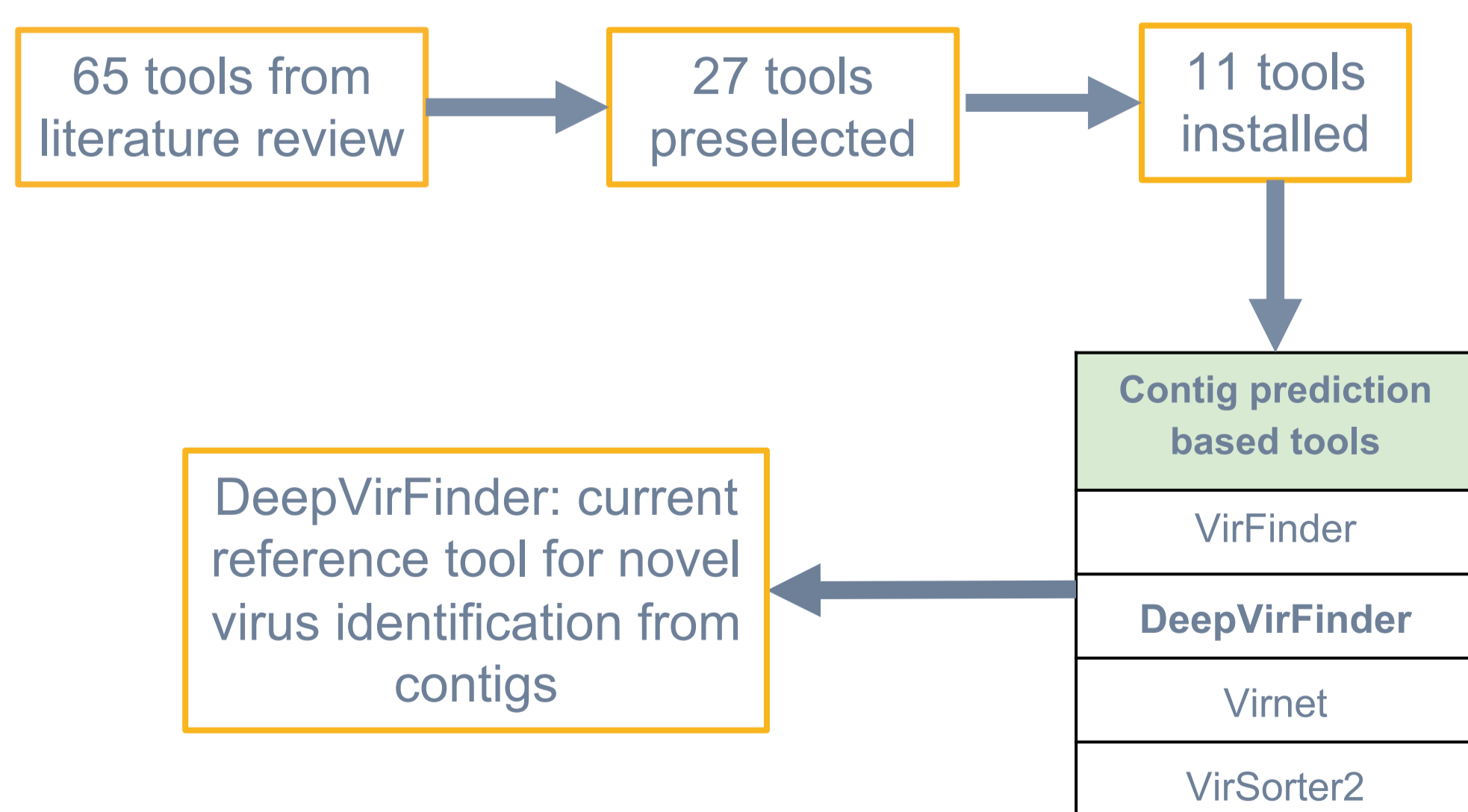
## Challenges of novel virus detection

- Virome samples are usually contaminated with bacterial or plant genetic material
- Sequencing technology may disfavor detection of some viruses
- Plant virus families are very dissimilar
- Much of plant viral diversity is not yet described



Umap clustering of sequence fragments sampled from different organisms

## Existing tools\*

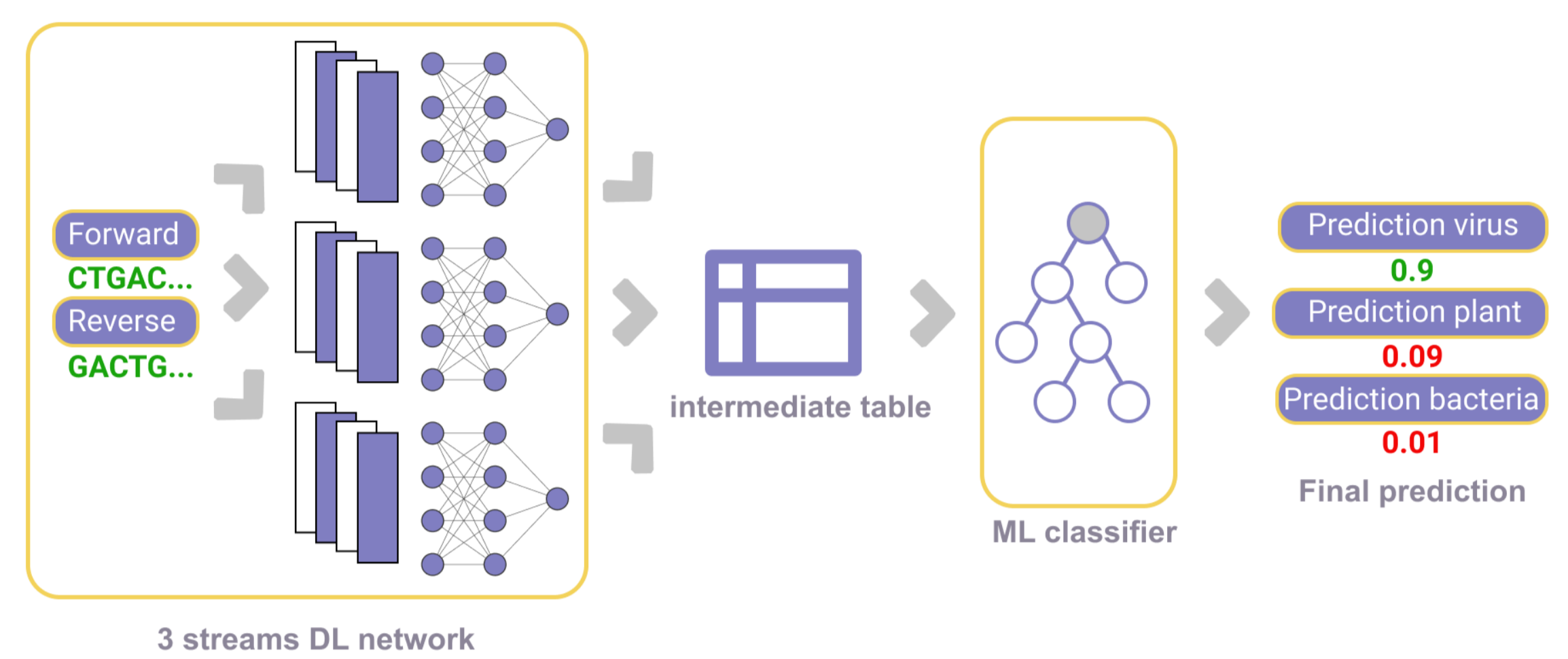


\* work done together in collaboration with: University of Ljubljana, Faculty Of Computer And Information Science, Ljubljana, Slovenia; Plant Pathology Laboratory, TERRA, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium

## VirHunter idea

- Develop deep learning approach to detect novel plant viruses from assembled contigs of plant virome sequencing data
- Leverage multiple models that rely on different k-mer sizes
- Enable multi-class approach to deal with contamination problem (virus, plant, bacteria)
- Validate the approach on leave-out datasets (family, genus, specie levels)

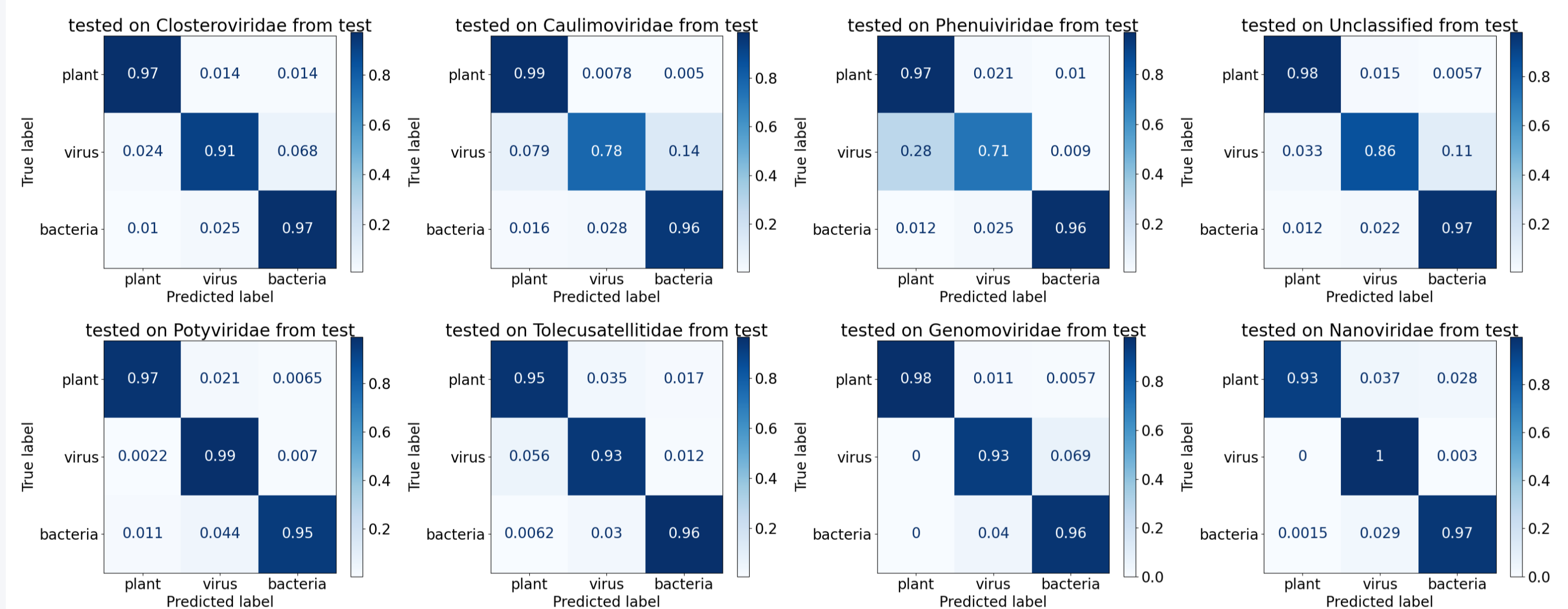
## Network architecture



## What makes VirHunter special

- Highly accurate novel virus identification
- Deals with contamination
- Can be easily integrated in standard pipelines
- Uses novel hybrid architecture (deep learning + standard machine learning)

## Accuracy of novel virus detection

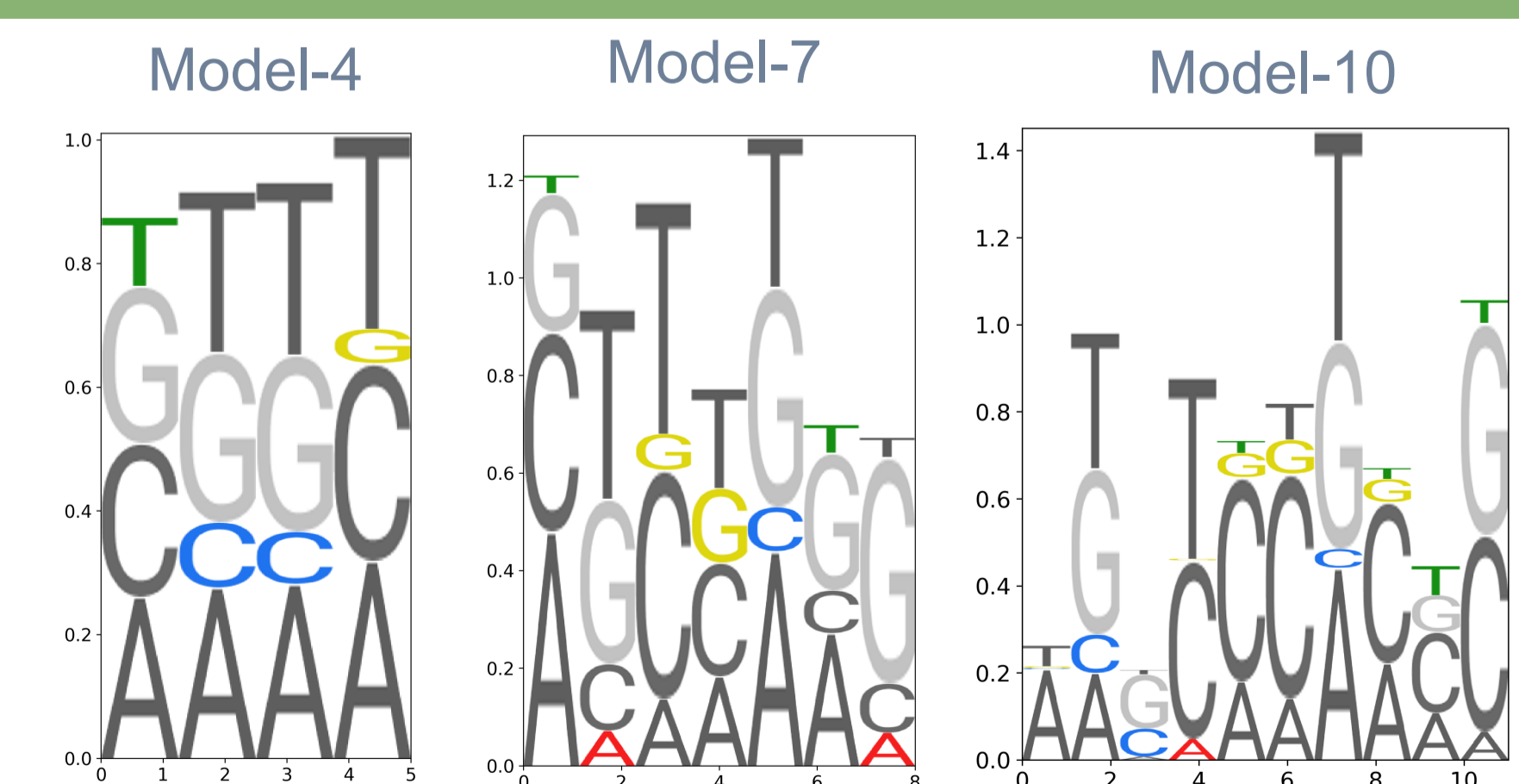


Confusion matrices for viral family leave-out datasets

Tool	% of correctly identified viruses	Weighted average
VirHunter	(69-91%)	85.6%
DeepVirFinder	(57-83%)	72.5%

VirHunter outperforms DeepVirFinder!

## Model interpretability



Examples of motifs learnt by the models. Letters in color are rewarded, letters in gray are penalized