Detection of Divergent Viral Sequences in Respiratory Metagenomes using Deep Learning Models with Uncertainty Quantification

Racim BENFOUGHALI
racim.benfoughali@etu.univ-amu.fr

Sana SELLAMI
sana.sellami@lis-lab.fr

Frédéric FLOUVAT
frederic.flouvat@lis-lab.fr

Philippe COLSON
philippe.COLSON@univ-amu.fr

Clinical and Biological Context

- Respiratory viral infections remain a major cause of morbidity and mortality
- Recent DNA/RNA-sequencing methods are fast but can miss off-panel or highly divergent agents
- Metagenomes may recover additional signals, but respiratory samples contain low viral signals, abundant host background, and many sequences remain unassigned (Dark matter)

Study Objective

*Assess whether a transformer-based model (**LucaProt**) can serve as a reproducible and biologically meaningful baseline for detecting divergent viral signatures before transfer to clinical data*

What is LucaProt?

- Transformer-based binary classifier designed to detect viral RdRP proteins
- Combines sequence information and predicted structural information
- Inputs consist of both tokenized aa sequences with a transformer encoder, and ESM-derived structural/contextual representations
- Final output is a binary classification

Adaptation Workflow for Our Use Case

Reads → Proteins → Prediction (→ Future UQ)

Practical Validation Strategy

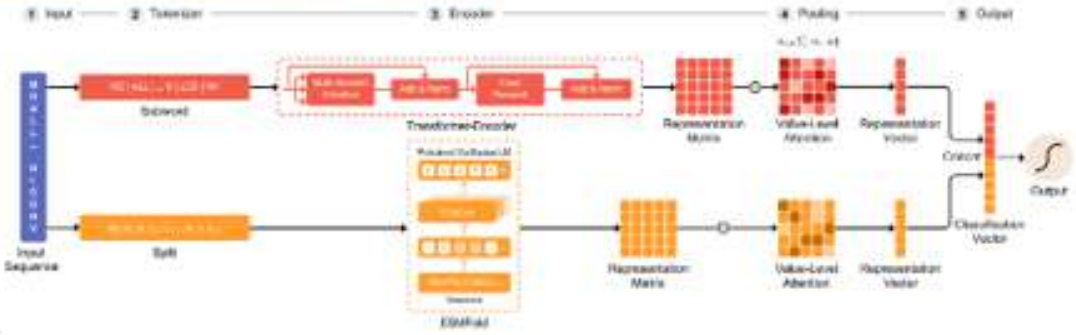
- Reproduced the original demo inference end-to-end
- Tested the model on external control datasets
- Mid-term objective : Explored both public protein sets and curated tutor-provided controls

Preliminary Results

Sample	Total sequences	RdRP+	RdRP-
SARS-CoV-2	2199	2199	/
Hepatitis C Virus	1	1	/
Staph. Aureus	1600	/	1600
Poxvirus	1689	59	1630
Many DNA pol.	162	/	162
Many RNA pol.	143	/	143
HIV	8	/	8

Works well with RdRP+ but struggles with RdRP-

→ XAI (SHAP, LIME...)



Next Steps :

- Respiratory datasets have not yet been fully processed through the complete pipeline
- Uncertainty estimation/calibration is still to be added