Zoonosis Prediction Using Language Models

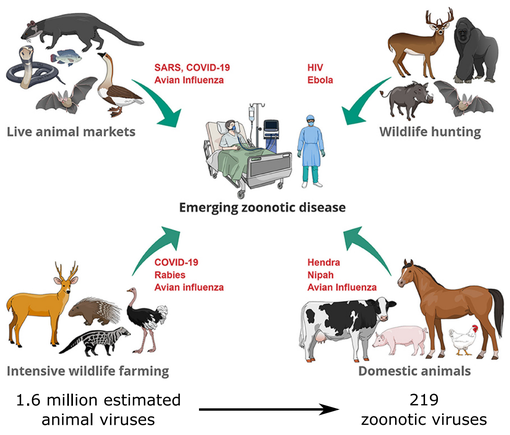
Blessy Antony[1], **Jie Bu**[1], **Andrew Chan**[1], **Anuj Karpatne**[1], **T. M. Murali**[1]

[1]Department of Computer Science, Virginia Tech, Blacksburg, USA

Motivation

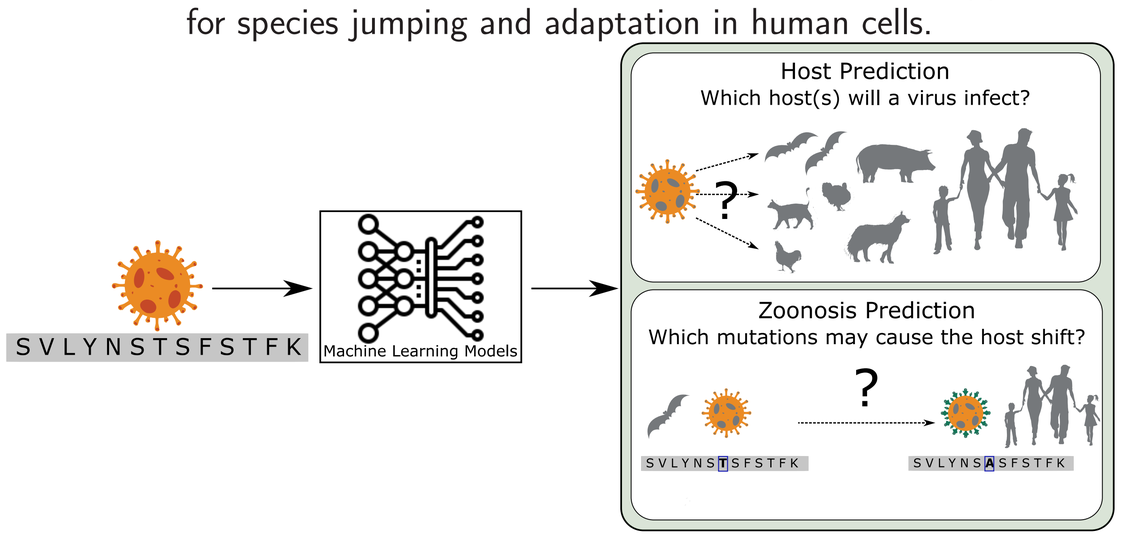
- Zoonosis is an infectious disease that has jumped from an animal to humans. Examples: H5N1 Avian Influenza, Ebola virus disease, COVID-19.
- As of April 2019, there are 1.6 million known animal viruses in nature, but only ∼0.01% of the animal viruses are known to infect humans.[1]

Mutations in a virus enable them to switch hosts, evade the immune system, and infect, adapt, and replicate in the new host.



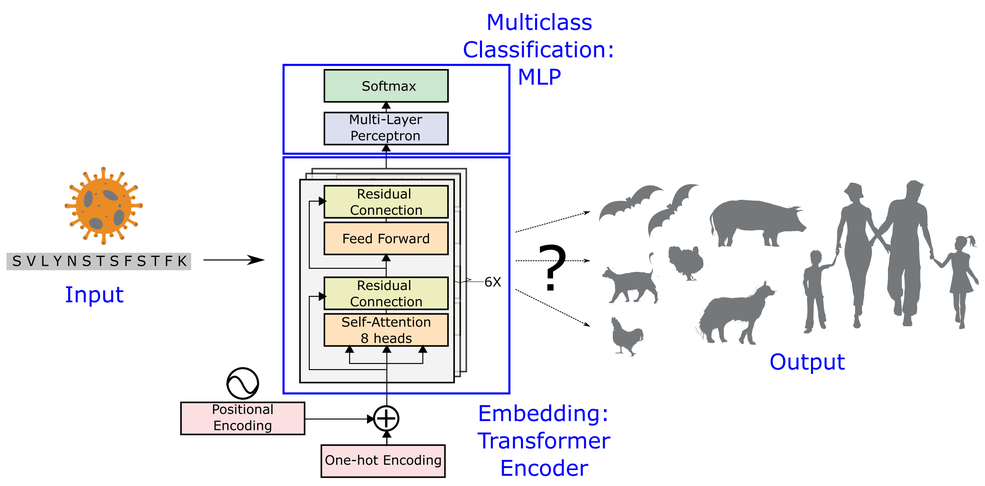
Goal

Develop machine learning models to predict the viral genetic element(s) responsible for species jumping and adaptation in human cells.



- Host Prediction: Given the protein sequence of a virus, predict which host(s) the virus will infect?
- Zoonosis Prediction: Given the sequences of a virus that infects both animals and humans, determine which mutations may cause the host shift.

Approach - Host Prediction



- Use language models based on the analogy that the protein sequences follow grammatical rules like natural languages.[2]
- Learn embeddings for protein sequences of viruses using the Encoder of a Transformer.
- Classify the learned embeddings using Multi Layer Perceptron and predict the host of a given viral protein sequence.
- Fit the model to solve the multi-class classification problem of host prediction.
- Learn using Focal loss to tackle the class-imbalance in the dataset.

Dataset

- **UniRef90**: Clusters of protein sequences from UniProt with atleast 90% similarity.
- Protein sequences of viruses known to infect mammals or aves.
- Included sequences from hosts with atleast 1% prevalence in the dataset.
- 19,093 sequences
- 97 viruses

Host	Prevalence
Human	77.80%
Desert warthog	5.63%
Lesser bandicoot rat	3.69%
Goat	2.41%
Horse	2.33%
Red junglefowl	1.62%
Wood mouse	1.52%
Cattle	1.00%

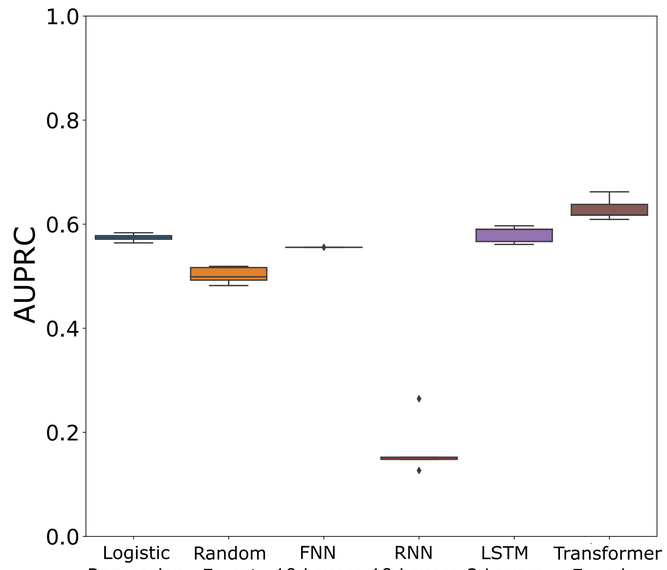
Acknowledgements

UL1TR003015

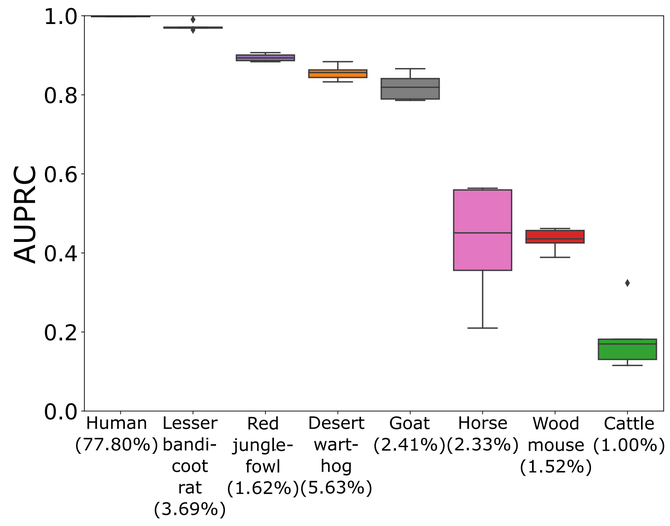
CCF 2200045

Host Prediction Results

Self-attention and long term memory yield better host prediction performance.



The prediction performance of each host class improves with its prevalence in the dataset.



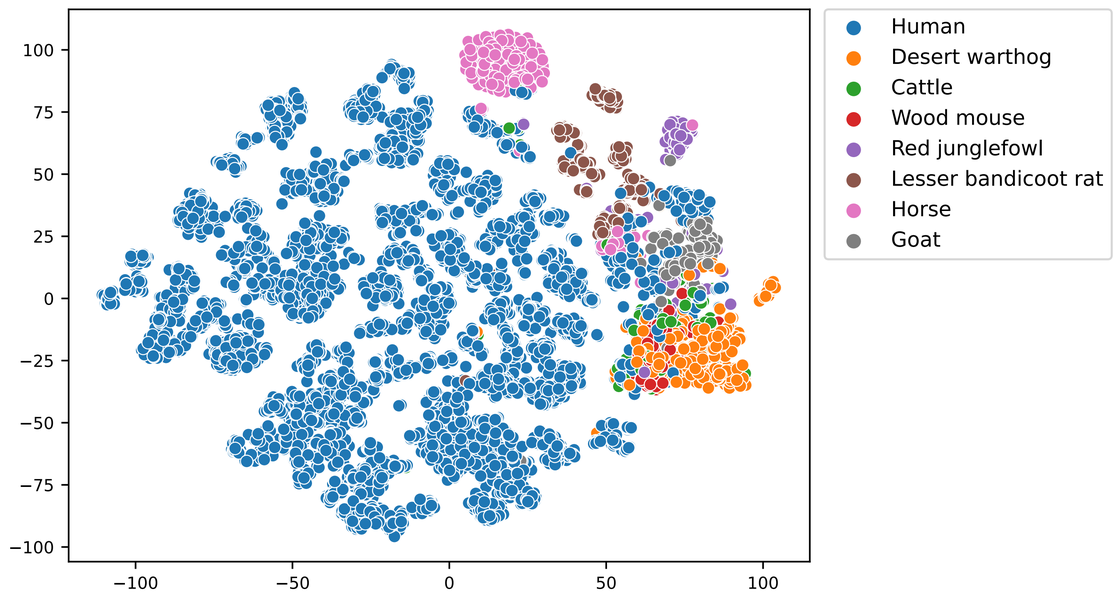
Approach - Zoonosis Prediction

Analysis of self-attention values for one SARS-CoV-2 Spike protein sequence sampled from humans.

- For each amino acid in a given position, compute the average attention paid by all other amino acids in the sequence.
- Three of the top-ten important positions are involved in binding with the human ACE2 receptor protein to initiate the human infection. 12% of the amino acids in spike protein belong to the Receptor Binding Domain.

Embeddings Suggest Sequences Underlying Zoonosis

- Low dimensional visualization of embeddings of protein sequences learned using Transformer-Encoder model.
- Dimensional reduction and visualization using TSNE.
- Overlapping clusters suggest sequences indicative of zoonosis.



On-going and Future Work

- Use saliency maps for interpretation of transformer models to identify important amino acid tokens for host-prediction.
- Use models pre-trained on protein sequences.[3]
- Leverage the structural information of proteins.

References

1. I. Magouras *et al.*, "Emerging Zoonotic Diseases: Should We Rethink the Animal–Human Interface?", *Frontiers in Veterinary Science*, (2020).
2. Hie *et al.*, "Learning the language of viral evolution and escape", *Science*, (2021)
3. Brandes *et al.*, "ProteinBERT: a universal deep-learning model of protein sequence and function", *Bioinformatics*, (2022)