Predicting Human Infectivity of Influenza A Sequences Using Masked Language Modeling

Vedant Hathalia[1]; Praneeth Gangavarapu[2]; Karthik Gangavarapu[2], PhD; Kristian Andersen, PhD[2]
[1]Bellarmine College Preparatory, San Jose, CA
[2]The Scripps Research Institute, La Jolla, CA

Introduction

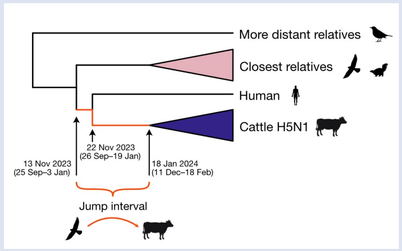

Figure 1: Schematic depicting time period when H5N1 likely spilled over into cattle.

Reproduced from Worobey et al., 2024.

- Emergence of human cases near this timeline highlights growing risks of mammalian adaptation.
- Understanding which mutations enable such jumps is critical for early detection and pandemic prevention.
- Influenza evolves via antigenic drift (gradual mutations) and antigenic shift (reassortment), both of which can increase zoonotic risk.[2]
- PB2 segment is important for viral replication and host adaptation, with mutations like D701N and S714R linked to mammalian adaptation.[3]

Dataset and Modeling Approach

- 97.5% of sequences infect one host, but 2.5% span multiple species.
- Since sequences can infect more hosts than labeled, we use Masked Language Modeling to learn infectivity signals directly from sequence patterns, enabling early detection of emerging threats.

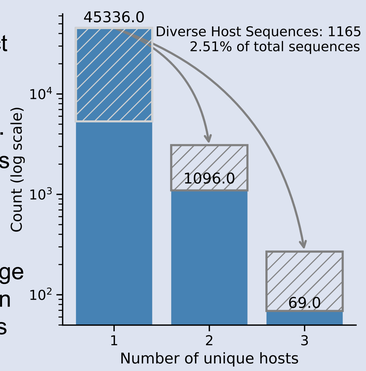

Figure 2: Bar plot of PB2 sequences by number of unique hosts.

- PB2 sequences were retrieved from NCBI and clustered with MMseqs2 at 80% identity to avoid redundancy and data leakage.
- Model was trained solely on human-infective sequences, learning patterns linked to human infectivity.

Overlapping Log-Likelihoods Reveal Zoonotic Risk

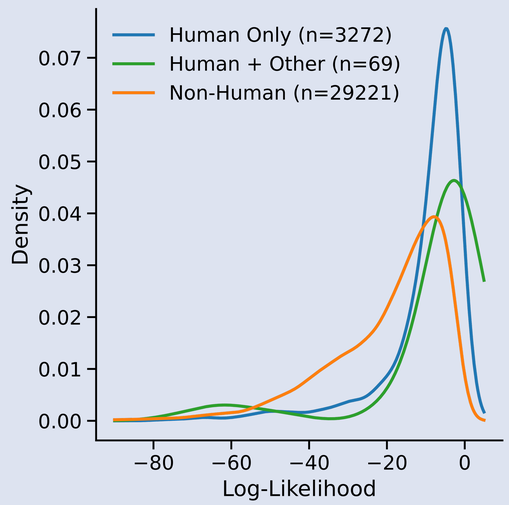

Figure 3: Log-likelihood distributions from human-trained model.

- Higher log-likelihoods represent how closely sequences show human-infecting patterns.
- Non-human sequences with log-likelihoods similar to those of human sequences suggests that these overlapping sequences could infect humans.

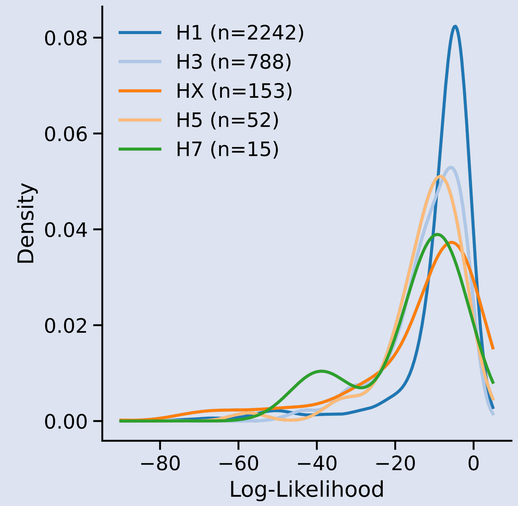

Figure 4: Log-likelihood distributions by HA subtype.

- More prevalent subtypes like H1 and H3, commonly associated with human infection, show high log-likelihoods.
- Model better understands well-represented, real-world strains.

Finetuned Model Outperforms Base

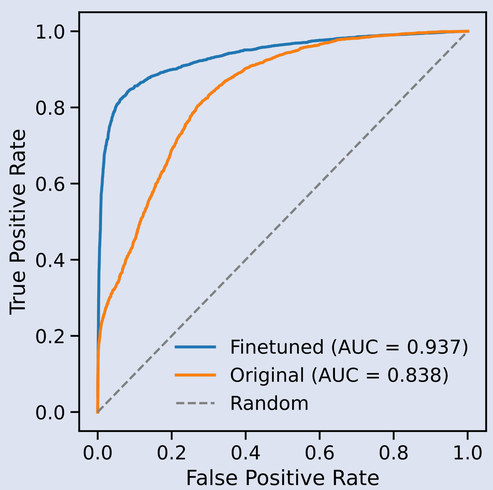

Figure 5: XGBoost classifier trained on sequence embeddings shows improved infectivity prediction.

Human Adaptive Mutations Found in Non-Human Sequences

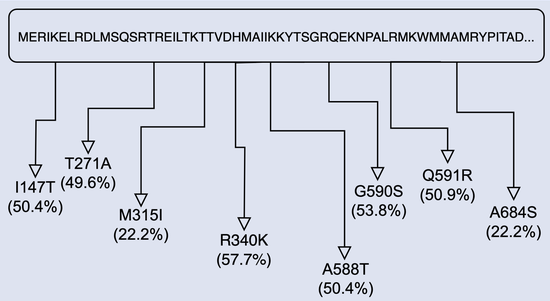

Figure 6: Human-adaptive mutations (e.g., G590S, T271A, I147T) and their frequency in non-human sequences with human-like log-likelihoods.

- These H1 and H5 subtype mutations are known to enhance polymerase activity or replication in human cells.[4,5,6]
- Occurring in 53.8% of overlapping sequences, G590S increases polymerase activity at lower temperatures, facilitating efficient replication in the human upper respiratory track.[4]

Future Work

- Train longer context models like Evo-2
- Model multiple influenza segments beyond PB2
- Train on nucleotide sequences

References

1) Worobey, M., Gangavarapu, K., Pekar, J. E., Joy, J. B., Moncla, L., Kraemer, M. U. G., Dudas, G., Goldhill, D., Ruis, C., Malpica Serrano, L., Ji, X., Andersen, K. G., Wertheim, J. O., Lemey, P., Suchard, M. A., Rasmussen, A. L., Chand, M., Groves, N., Pybus, O. G., Peacock, T. P., Rambaut, A., & Nelson, M. I. (2024). *Preliminary report on genomic epidemiology of the 2024 H5N1 influenza A virus outbreak in U.S. cattle (Part 1)* [Report]. *Virological.org*.
2) Aryal, S. (2022, August 10). *Antigenic shift and antigenic drift*. *MicrobiologyInfo*. Retrieved July 16, 2025.
3) Czudai-Matwich, V., Otte, A., Matrosovich, M., Gabriel, G., & Klenk, H.-D. (2014). PB2 mutations D701N and S714R promote adaptation of an influenza H5N1 virus to a mammalian host. *Journal of Virology, 88*(16), 8735–8742.
4) Mehle, A., & Doudna, J. A. (2009). Adaptive strategies of the influenza virus polymerase for replication in humans. *Proceedings of the National Academy of Sciences, 106*(50), 21312–21316.
5) Peacock, T. P., Sheppard, C. M., Lister, M. G., Staller, E., Frise, R., Swann, O. C., Goldhill, D. H., Long, J. S., & Barclay, W. S. (2023). Mammalian ANP32A and ANP32B proteins drive differential polymerase adaptations in avian influenza virus. *Journal of Virology, 97*(5).
6) Suttie, A., Deng, Y.-M., Greenhill, A. R., Dussart, P., Horwood, P. F., & Karlsson, E. A. (2019). Inventory of molecular markers affecting biological characteristics of avian influenza A viruses. *Virus Genes, 55*(6), 739–768.

Acknowledgements

This research was conducted as part of the SURI Internship Program at Scripps Translational Institute and supported in some capacity by philanthropic gifts and endowments from John & Susie Diekman. Special thanks to Praneeth Gangavarapu for his guidance and the Andersen Lab for making this project possible.