HYBRID TRANSFORMER, VARIATIONAL AUTOENCODER, AND GENETIC ALGORITHM APPROACH FOR PREDICTING CROSS-SPECIES TRANSMISSION IN INFLUENZA A VIRUS DRIVEN BY CLIMATE FACTORS.

Anonymous authors
Paper under double-blind review

1 BACKGROUND

Influenza remains one of the most significant respiratory diseases, with influenza A viruses posing a substantial risk due to their rapid mutations, genetic reassortment, and ability to infect multiple hosts. Their segmented genome facilitates reassortment, enabling genetic exchange between different viral strains, which can lead to the emergence of novel variants with pandemic potential. This increases the risk of zoonotic spillover and the emergence of highly transmissible and virulent strains [Morse et al. (2012)]. The highly pathogenic avian influenza (H5N1) has already caused over 860 confirmed human infections and more than 450 deaths, with outbreaks in poultry leading to billions of dollars in economic losses. Similarly, the 2009 H1N1 swine flu pandemic resulted in many fatalities worldwide, highlighting the devastating consequences of cross-species transmission. Accurate prediction of cross-species transmission is therefore critical for pandemic preparedness and outbreak prevention.

1.1 CHALLENGE AND INNOVATION

Despite these risks, traditional phylogenetic-based reassortment detection methods lack predictive power, while purely machine learning-based approaches often fail to incorporate biological constraints. Additionally, the limited availability of labeled genomic data constrains fully supervised learning approaches. To address these challenges, we propose a semi-supervised learning framework that integrates Transformers, Variational Autoencoders (VAEs), and Genetic Algorithms (GAs) to enhance feature extraction and simulate novel reassortment events. Our hybrid framework incorporates Transformers for robust sequence feature extraction, VAEs for structured latent-space representations, and GAs for biologically plausible reassortment simulations. Furthermore, considering the seasonal nature of influenza and its dependence on environmental conditions, we integrate climate factors—including temperature, rainfall, and humidity—into our model to better understand climate-driven mutations [He et al. (2023)].

1.2 METHODS

We begin by retrieving genomic sequences of influenza A viruses (IAVs) from online databases such as GenBank, alongside epidemiological metadata that includes the host species they infect [Ren et al. (2016)]. These sequences are categorized into positive samples (known to infect multiple hosts, such as humans, birds, and pigs) and negative samples (restricted to a single host). In parallel, we calculate critical climate parameters from ERA5 climate data using an in-house AI pipeline. Additionally, we conduct routine environmental surveillance, collecting unlabeled influenza A virus samples from locations such as poultry farms, wetlands, lakes, and wildlife sanctuaries. These environmental samples, along with their corresponding climate data, form the unlabeled dataset. The objective is to predict whether any of these environmental virus samples have the potential to gain cross-species transmissibility through genetic reassortment and climate-driven mutations.
To achieve this, we first perform bioinformatics analyses, extracting features such as GC content, known single nucleotide polymorphisms (SNPs), and receptor-binding site mutations. Specifically, we analyze the HA protein for mutations at Q226L (which shifts receptor preference from α 2,3 to α 2,6 sialic acids) and G228S (which enhances human receptor binding), as well as PB2 pro-

tein mutations E627K and T271A (which enhance replication and polymerase activity in mammals) [Yan et al. (2023)]. These analyses are performed for both the labeled and unlabeled sequences. In addition, we leverage a Transformer-based model, DNABERT, to capture deeper sequence relationships. DNABERT is fine-tuned on the labeled influenza dataset to learn influenza-specific sequence contexts and patterns. After fine-tuning, the same DNABERT model is used to process unlabeled sequences, generating embeddings that encode both global sequence motifs and nuanced influenza-specific features indicative of reassortment potential. At this stage, we obtain three feature sets: manually extracted bioinformatics features, climate factors, and high-level sequence embeddings from DNABERT.

Next, we integrate these diverse features using a Variational Autoencoder (VAE). The VAE facilitates feature fusion, dimensionality reduction, and structured representation learning. Initially, it is trained in an unsupervised manner using only labeled samples, compressing the concatenated feature representations into a lower-dimensional latent space. The decoder then reconstructs the original input, ensuring that critical biological information is retained. Subsequently, a classification branch is added to the VAE, fine-tuning it using labeled data to distinguish between high-risk and low-risk reassortment potentials. This structured latent space allows similar reassortment risks to cluster together. Once trained, the VAE is used to map the unlabeled environmental samples into this structured latent space, capturing essential biological and environmental characteristics relevant to reassortment potential.

To explore reassortment dynamics, we employ a Genetic Algorithm (GA) to simulate novel reassortant strains within the latent space learned by the VAE. The GA begins by initializing a population of latent representations derived from unlabeled environmental samples. It then applies crossover and mutation operations, introducing biologically plausible genetic exchanges while maintaining constraints—such as ensuring that reassortment occurs only between sequences from different host species. Additionally, mutations are introduced based on climate-driven patterns, aligning with observed influences of environmental conditions on viral evolution. The GA operates within the latent space, using a fitness function that evaluates reassortants based on biological feasibility (proximity to known transmissible sequences) and climate compatibility (mutations aligning with observed environmental influences). The fittest reassortant candidates are then reconstructed back into feature vectors using the VAE decoder, followed by sequence reconstruction. Throughout this process, the GA maintains a mapping between parent sequences and their reassortant progeny, enabling the identification of high-risk parental populations.

1.3 Results and Impact

This methodology serves as an early warning system, aiding in pandemic preparedness, strengthening healthcare systems (HSS), and enabling data-driven decision-making. By incorporating climate parameters, we track how the virus evolves across seasons and anticipate emerging threats. Through this interdisciplinary approach, we provide actionable insights for pandemic prevention, helping policymakers and health organizations mitigate the risks posed by influenza A virus.

Meaningfulness Statement

We leverage the RNA sequence of influenza virus in our study. This meaningful representation of life contains crucial information on viral segmentation, reassortment potential, and structure. We have used Transformers and VAEs to capture these relationships, and GAs, which mimic natural evolution, mutation, and selection, to simulate reassortments in nature. Additionally, other modalities like climate data also influence the nucleotide sequence, and our model aims to deduce this relationship, enhancing our understanding of the virus cycle in nature.

References

Yucong He, William J. Liu, Na Jia, Sol Richardson, and Cunrui Huang. Viral respiratory infections in a rapidly changing climate: the need to prepare for the next pandemic. *eBioMedicine*, 93: 104593, 2023. ISSN 2352-3964. doi: https://doi.org/10.1016/j.ebiom.2023.104593.

Stephen S Morse, Jonna AK Mazet, Mark Woolhouse, Colin R Parrish, Dennis Carroll, William B Karesh, Carlos Zambrana-Torrelio, W Ian Lipkin, and Peter Daszak. Prediction and prevention of the next pandemic zoonosis. *The Lancet*, 380(9857):1956–1965, 2012.

Hongguang Ren, Yuan Jin, Mingda Hu, Jing Zhou, Ting Song, Zhisong Huang, Beiping Li, Kaiwu Li, Wei Zhou, Hongmei Dai, et al. Ecological dynamics of influenza a viruses: cross-species transmission and global migration. *Scientific reports*, 6(1):36839, 2016.

Zhanfei Yan, You Li, Shujian Huang, and Feng Wen. Global distribution, receptor binding, and cross-species transmission of h6 influenza viruses: risks and implications for humans. *Journal of Virology*, 97(11):e01370–23, 2023. doi: 10.1128/jvi.01370-23.