Influ-BERT: An Interpretable Model for Enhancing Low-Frequency Influenza A virus Subtype Recognition

Rongye Ye[1,2,3,#], Lun Li[1,2,#], Shuhui Song[1,2,3,*]

[1] National Genomics Data Center, China National Center for Bioinformation, Beijing 100049, China, [2] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100049, China, [3] University of Chinese Academy of Sciences, Beijing 100049, China

[#]These authors contributed equally.

*To whom correspondence should be addressed.

Abstract

Influenza A Virus (IAV) poses a continuous threat to global public health due to its wide host adaptability, high-frequency antigenic variation, and potential for cross-species transmission. Accurate recognition of IAV subtypes is crucial for the early pandemic warning. Here, we propose Influ-BERT, a domain-adaptive pretraining model based on the transformer architecture. Optimized from DNABERT-2, Influ-BERT constructed a dedicated corpus of approximately 900,000 influenza genome sequences, developed a custom Byte Pair Encoding (BPE) tokenizer, and employ a two-stage training strategy involving domain-adaptive pretraining followed by task-specific fine-tuning. This approach significantly enhanced recognition performance for low-frequency subtypes. Experimental results demonstrate that Influ-BERT outperforms traditional machine learning methods and general genomic language models (DNABERT-2, MegaDNA) in subtype recognition, achieving a substantial improvement in F1-score, particularly for subtypes H5N8, H5N1, H7N9, H9N2. Furthermore, sliding window perturbation analysis revealed the model's specific focus on key regions of the IAV genome, providing interpretable evidence supporting the observed performance gains.

Availability: Source code is written in PyTorch and available at https://github.com/oooo111/Influenza-BERT and https://huggingface.co/rongye1/Influenza_BERT under the MIT license.
Contact: songshh@big.ac.cn (Song S).
Key words: IAV Subtype Recognition; Influenza Virus; Long-tail Distribution; Influ-BERT; Genomic Language Model

Biographical note:
Rongye Ye is currently a master's student at the Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation.
Lun Li is an assistant professor at Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation.
Shuhui Song is a professor at Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation.

Introduction

Influenza viruses are among the most epidemiologically significant respiratory pathogens globally, posing a long-standing and significant threat to global public health [1,2]. Among these, Influenza viruses (IAV) constitutes the primary pathogen responsible for human influenza pandemics due to its broad host range and antigenic variation characteristics. IAV is an RNA virus, and its genome is composed of eight single-stranded negative-sense RNA segments [3]. Specially, Segment 4 encodes hemagglutinin (HA), and Segment 6 encodes neuraminidase (NA), which determine viral subtype characteristics due to their roles in mediating viral entry into host cells and promoting viral release from host cells, and their recognition by the host immune system. The generation of new HA-NA combinations through gene reassortment is the molecular basis for IAV causing cross-species transmission and pandemics [4]. For example, the highly pathogenic avian influenza virus of the H5N1 subtype is one of the most significant pandemic threats currently identified. It has demonstrated the ability to transmit among multiple mammalian species, indicating a potential for cross-species transmission to humans [5]. The widespread prevalence of H9N2 in live bird markets increases the zoonotic risk [6]. H7N9, as a highly pathogenic avian influenza virus, poses a considerable threat to the poultry industry and public health, and human infection carries a high fatality risk [7].

Traditional IAV subtype recognition mainly relies on molecular biology techniques like reverse transcription PCR (RT-PCR) for accurate recognition of common subtypes [8,9]. Serological methods, including hemagglutination inhibition (HAI) assays and neutralization assays are also widely employed. HAI detects subtype-specific antibodies by their ability to prevent hemagglutination [10], while neutralization assays identify subtypes by measuring antibody-mediated inhibition of viral infection [11]. However, these conventional methods have limited throughput, posing challenges for large-scale automated sample analysis. Advances in high-throughput sequencing have enabled genome-based methods for HA/NA subtyping, primarily through sequence alignment against reference databases [12]. These approaches recognize subtypes based on sequence similarity and analyze evolutionary relationships through phylogenetic trees [13,14]. Tools like NextClade [15] are particularly valuable for detecting novel subtypes or recombinant variants. However, alignment-based strategies face two key challenges, one is IAV high mutation rate requires constant database updates, making pre-built references quickly obsolete, and the other one is large-scale analyses suffer from high computational complexity during alignment and phylogenetic tree construction.

The advent of artificial intelligence has revolutionized influenza subtype recognition, with machine learning methods increasingly employed for this purpose. Current approaches typically use k-mer frequencies as features to train recognition models [16-18]. However, the performance of these methods heavily relies on manually designed features (e.g., k-mer length selection), making them sensitive to feature extraction quality. Besides, these traditional models struggle to effectively capture contextual dependencies in viral genome sequences. Recent breakthroughs in Transformer architecture have created new opportunities for biological sequence analysis [19-22]. Leveraging the symbolic similarity between nucleic acid sequences and natural language, researchers have developed numerous pretrained genomic language models [23,24], such as DNABERT [25,26], MegaDNA [27], and RNA-specific models like LucaProt [28] which integrating RdRp amino acid sequences with predicted protein structures in a multimodal Transformer framework, enhanced virus discovery and viral "dark matter" exploration.

However, long-tail data distribution is prevalent across various real-world applications. This distribution pattern is characterized by a severe class imbalance, where most samples are concentrated in a small number of categories, while the numerous remaining categories contain relatively few samples. In the context of IAV, subtypes that have exhibited sustained and widespread transmission among humans (e.g., H1N1 and H3N2) are classified as frequent subtypes. In our dataset, subtypes with an occurrence frequency below 5% and only limited documented human infection cases were classified as low-frequency subtypes, despite their potential importance in zoonotic transmission or pandemic risk. Affected by the long-tail distribution of IAV data, despite the models demonstrating strong recognition capabilities for prevalent subtypes (H1N1, H3N2), performance significantly attenuates for low-frequency subtypes, which are equally crucial for epidemiological surveillance [29-31]. Therefore, establishing efficient subtype prediction models is of significant value for early warning and prevention for influenza pandemics.

Here, we propose Influ-BERT, a model inspired by DNABERT-2 architecture, our framework is specifically optimized to enhance recognition capability for low-frequency IAV subtypes (**Figure 1a**). We evaluated Influ-BERT against k-mer-based machine learning methods (Random Forest, SVM, Logistic Regression), the masked pretrained model DNABERT-2, and the generative model MegaDNA. Results show that Influ-BERT outperforms existing methods in recognition low-frequency subtypes while maintaining strong overall metrics. Furthermore, we introduced sequence perturbation analysis based on a sliding window approach, which revealed differences between Influ-BERT and DNABERT-2 in feature recognition patterns, providing a preliminary explanation for Influ-BERT's performance improvement.

Materials and methods

Dataset

We systematically collected approximately 900,000 influenza virus genome sequences, including more than 100 naturally occurring HA–NA combinations that have been detected to date from NCBI Nucleotide database [32]. There are approximately 800,000 IAV sequences, ~100,000 influenza B virus (IBV) sequences, and over 4,000 influenza C and D virus (ICV and IDV) sequences. Notably, the data we obtained encompass various segments without targeting specific segments. This dataset was utilized in two key steps: (1) construction of a BPE vocabulary specific to influenza virus sequences; and (2) domain-adaptive pretraining of the language model.

To comprehensively evaluate the performance of the adapted model, we designed three biologically meaningful recognition tasks with increasing complexity. Task 1 involves recognition of five major IAV subtypes; Task 2 expands the recognition granularity to ten IAV subtypes; and Task 3 focuses on cross-type recognition across influenza A, B, C, and D viruses. For Tasks 1 and 2, we utilized a dataset of 40,000 independently collected IAV sequences, representing 66 subtype combinations, which were excluded from the domain-adapive pretraining data. The dataset exhibits a long-tail distribution (**Figure 1b),** and the length distribution of sequences in the domain adaptive pretraining corpus is shown in (**Figure 1c)**. Task 3 incorporated 20,000 IAV sequences, 20,000 IBV sequences, and all available ICV and IDV sequences, type A and type B data show no sequence overlap with those used in previous tasks. All datasets were partitioned into training, validation, and test sets using a ratio of 8:0.5:1.5. Model performance was comprehensively evaluated using precision, recall, and F1-score as the primary metrics, providing a comprehensive quantitative assessment of recognition accuracy.

Experiment Setting

Software Configuration: During the domain-adaptive pretraining phase, BPE vocabulary was constructed with a size of 512. All other training hyperparameters were kept consistent with those used in DNABERT-2. In the fine-tuning stage, the number of training epochs was set to 10 for MegaDNA, DNABERT-2, and Influ-BERT. The batch size for Influ-BERT was set to 8. All experiments were conducted using the PyTorch framework (version 2.5.1), with model training and evaluation implemented via the Hugging Face Transformers library (version 4.49.0).

Baseline Model: This experiment compares five models, including three machine learning models based on k-mer (3-mer) encoding, the DNABERT-2 model based on masked pretraining, and the MegaDNA model based on generative pretraining: 1) **Random Forest (RF) [33]**: An ensemble learning model that improves model generalization by combining the predictions of multiple decision trees. It is widely used in classification and regression tasks; 2) **Support Vector Machine (SVM) [34]**: A supervised discriminative model whose core goal is to find an optimal hyperplane that separates samples of different classes as much as possible; 3) **Logistic Regression (LR) [35]**: A commonly used classification machine learning model that outputs a probability representing the likelihood of a sample belonging to a certain class. It is widely applied in classification tasks; 4) **DNABERT-2 [26]**: A self-supervised pretrained DNA language model that uses BPE encoding and is pretrained on a large-scale multi-genome dataset containing 32.5 billion nucleotide bases and covering 135 species. This model is capable of capturing complex dependencies in DNA sequences and has demonstrated superior performance in multiple biological tasks; 5) **MegaDNA [27]**: A generative pretrained long-sequence DNA model that uses nucleotide-level tokenization. This model can handle genomic sequences up to 96,000 base pairs and has the capability for sequence generation and classification of unannotated sequences.

Mask Language Model

Inspired by the success of natural language processing (NLP), masked pretraining has emerged as an effective paradigm for learning representations from genomic sequences. As a classical self-supervised learning strategy, masked pretraining operates by randomly masking a subset of input tokens during training and requiring the model to predict the original tokens based on their surrounding context. This approach enables the model to capture deep semantic representations of sequences. The objective function is designed to maximize the likelihood of the masked tokens, which is formally defined as follows:

$$\mathcal{L}_{mlm} = -\sum_{i\epsilon\mathcal{M}} logP(x_i|X^{mask}), \tag{1}$$

where x_i represents the token at position i in the original sequence, and X^{mask} refers to the masked input sequence. $P(x_i|X^{mask})$ denotes the model's probability of reconstructing x_i given the masked input X^{mask}. The Influ-BERT model adopts the same architectural framework as DNABERT-2, utilizing the Transformer attention mechanism as its core component for sequence representation learning. The attention function [14] is computed as follows:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})\text{V}, \tag{2}$$

where Q denotes the query matrix, representing the current input query; K denotes the key matrix, which typically encodes the features of candidate elements; V is the value matrix, containing the information associated with each key; and d_k is a scaling factor used to normalize the dot product of queries and keys. Within the Transformer architecture, knowledge representations across different subspaces are typically learned using a multi-head attention mechanism. The computation is defined as follows:

$$\begin{gathered}\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(head_1, \dots, head_h)W_O, \\ head_j = Attention\left(QW_Q^j, KW_K^j, VW_V^j\right),\end{gathered} \tag{3}$$

where W denotes the learnable projection matrices applied during linear transformations, and $head_j$ refers to the j-th attention head among a total of h heads in the multi-head attention mechanism. Finally, during the fine-tuning phase, the task-specific loss function is defined as the cross-entropy loss, which is formulated as follows:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(p_i)\,, \tag{4}$$

where, C denotes the total number of classes, y_i represents the ground-truth label for the sample, and p_i denotes the predicted probability that the sample belongs to class i.

Transfer Knowledge from Domain-adaptive model

Although genome language models pretrained on cross-species genomic data have achieved remarkable performance across a variety of biological sequence analysis tasks, directly applying them to influenza virus genotyping remains a significant challenge. The intrinsic characteristics of influenza viruses—namely, long-tailed data distribution, high mutation rates—constrain the fine-tuning performance of models such as DNABERT-2. To address this limitation, we first adopt DNABERT-2 as the foundational architecture, which has been pretrained on a large-scale dataset comprising 32.5 billion bases from 135 species, including bacteria, fungi, and mammals. Subsequently, we train a BPE tokenizer using an influenza-specific corpus and perform domain-adaptive pretraining on this corpus to enhance the model's capacity for representing influenza viral sequences. To accommodate the short sequence lengths characteristic of influenza viruses, we set the vocabulary size from 4096 to 512, while keeping all other pretraining hyperparameters consistent with those of DNABERT-2. For model adaptation, we employ a two-stage learning strategy: (1) domain-adaptive pretraining to enable the model to learn general representations of influenza sequences; and (2) transfer learning by fine-tuning the domain-adaptive pretraining model on downstream tasks with the addition of task-specific classification heads. During fine-tuning, we apply a supervised learning strategy with full parameter updates on Influenza sequences, optimizing the model using the cross-entropy loss function (Equation 5). This approach effectively combines the model's ability to capture general sequence features of influenza viruses with task-specific learning [36], thereby improving its performance in identifying low-frequency subtypes.

Sequence Perturbation Analysis

The core idea of sequence perturbation analysis is to systematically mask local regions of the input sequence and then feed the masked sequences into the model to observe the resulting changes in output. This allows for a quantitative assessment of each sequence fragment's contribution to the model's decision. Specifically,

for an influenza viral genome sequence $S = (s_1, s_2, \dots, s_n)$, and a set of class labels $C = \{c_1, c_2, \dots, c_m\}$, we define a window size w, and first compute the model's prediction on the original input sequence S:

$$P(S) = [p_1, p_2, \dots, p_m], \tag{5}$$

where $\sum_{i=1}^{m} p_i = 1$. Subsequently, for a window starting at position j, we set the $w = 5$, and apply masking to this region. The perturbed sequence is defined as:

$$Mask_j(S) = \left(s_1, \dots, s_{j-1}, N * w, s_{j+w}, \dots, s_n\right), \tag{6}$$

Here, N denotes the masking operation. Finally, for a sequence labeled as class c, the importance score of the window is computed as:

$$Score_{j,w} = |p_c - p_c^{j,w}|, \tag{7}$$

where, p_c denotes the model-predicted probability of class c for the original input sequence, and $p_c^{j,w}$ denotes the predicted probability of class c after masking a window of size w starting at position j in the original sequence.

Results

Overview of Influ-BERT

As shown in **Figure 2**, the Influ-BERT framework comprises by the domain-adaptive pretraining module and sequence recognition fine-tuning module. We first construct a vocabulary tailored for influenza-specific corpus sequences using BPE algorithm. Then, all influenza virus sequences in the corpus were tokenized using this vocabulary. The pre-trained model was initialized with the DNABERT-2 pretrained weights and conducted for a domain adaptive pretraining using masked language model (MLM) objective. For subtype recognition downstream tasks, different input sequence datasets are also tokenized using the influenza-specific BPE vocabulary. Following the transfer learning paradigm, we use the domain-adaptive model to extract features and attach task-specific classification heads for end-to-end fine-tuning.

Performance Evaluation for IAV Subtype Recognition

This study systematically evaluated Influ-BERT performance on the task of IAV subtype recognition for H3N2, H1N1, and other low frequency (frequency < 5%) subtypes. Among which, H5N1, H7N9, and H9N2 subtypes, with frequency 0.6%, 0.5%, and 1%, were prioritized for analysis due to their relatively significant impact on health as reported in previous studies. We assessed performance using three metrics—precision, recall, and F1-score, and found Influ-BERT outperformed all other models across overall dataset performance metrics (As shown in **Table 1**). Next, we conduct a detailed analysis of the specific subtype identification performance.

In terms of F1-score **(Figure 3a-b)**, Influ-BERT demonstrated a significant advantage compared with the other five models. In the five-class(five-subtype) recognition task, the F1-scores of influ-BERT achieved 97% and 94% for H5N1 and H9N2, respectively, which is significantly improved when compared to the other models. While in the ten-class(ten-subtype) task, Influ-BERT also outperformed all other models on all low-

frequency subtypes, especially for H5N1, H9N2, H5N8, and H7N9. Notably, in the H5N8 subtype recognition task, most other models failed to recognize this subtype, while Influ-BERT achieved an F1-score of 87%, providing strong evidence for the effectiveness of Influ-BERT. Besides, Influ-BERT demonstrated consistent performance between recall and F1-score metrics, while exhibiting superior capability in low-frequency subtype recognition compared to other benchmark models.

In terms of precision, common subtypes like H1N1 and H3N2, achieved high precision for all models. Influ-BERT did not consistently achieve the best performance across all subtypes. In the five-class recognition task, Influ-BERT demonstrated superior performance only on H3N2 and H1N1 subtypes compared to other models, while in the ten-class recognition scenario, its advantage was solely observed on the H13N6 subtype. This phenomenon may stem from the significant under-representation of low-frequency subtypes in the dataset. Such class imbalance compels comparative models to adopt more conservative prediction strategies, whereby samples are classified as low-frequency subtypes only when the prediction confidence exceeds a predetermined threshold.

In all, these experiments found that directly applying the cross-species pre-trained DNABERT-2 model in its original form yielded poor performance in recognizing low-frequency subtypes, and traditional machine learning models also underperformed on low-frequency subtype recognition. It is noteworthy, when compared to conventional machine learning methods, DNABERT-2 exhibited inferior performance, further highlighting the limitations of existing methods in handling extremely imbalanced class distributions. However, Influ-BERT (the modified DNABERT-2 model), after optimization using the two-stage training strategy proposed in this study, achieved significant improvements in identifying both common and low-frequency subtypes, and exhibited consistent improvements across both five-class and ten-class tasks, demonstrating that Influ-BERT retains a relative advantage in low-frequency subtype recognition even as the task scale expands.

Interpretability Analysis by Perturbation

To investigate how the model captures informative features within influenza sequences, we designed a masking-based sliding window perturbation analysis. By systematically masking input segments and observing variations in the model's predictions, we reveal differences in feature attention patterns between DNABERT-2 and Influ-BERT (**Figure 4**), providing preliminary insights into the latter's improved performance in low-frequency subtype recognition. Taking three influenza viral genome sequences of H5N1 subtype as example (**Figure 4**), both models exhibited varying degrees of attention to the initial nucleotide regions of the sequences. However, Influ-BERT demonstrated relatively consistent overall attention distribution patterns and consistently attended to specific genomic regions near positions 650-700 nt, 1100 nt, 1500 nt, and 1750 nt (**Figure 4**), indicating robust global modeling capabilities. In contrast, DNABERT-2 displayed irregular and substantially divergent attention distributions across sequences of the same subtype (**Figure 4**), which correlates with its failure to recognize all these three sequences. We further conducted perturbation analyses on three subtypes H13N8**,** and H3N2. The results demonstrating patterns essentially consistent with those observed in H5N1(**Figure S1-S2**). These findings suggest that Influ-BERT could identify discriminative sequence features that may underlie its superior robustness and accuracy in subtype recognition. These identified features or key regions may have practical implications for the model's decision-making process.

Conclusion and Discussion

Although masked pre-trained DNABERT-2 models based on cross-species datasets and generative MegaDNA models have demonstrated excellent performance in various biological sequence analysis tasks, their application in the specific scenario of IAV subtype recognition presents limitations due to significant long-tail data distribution properties and high-frequency mutation rates. Notably, DNABERT-2 underperforms conventional machine learning models in recognition low-frequency subtypes. To address these challenges, We propose Influ-BERT, which significantly improves F1-score and recall for low-frequency subtype recognition while outperforming existing benchmark models in influenza virus type classification (A/B/C/D).
The novelty of Influ-BERT is primarily reflected in two aspects: Firstly, we constructed a dedicated influenza corpus comprising approximately 900,000 influenza genome sequences and developed a BPE encoding strategy specifically adapted to the characteristics of influenza virus sequences; Secondly, we adopted a two-stage optimization paradigm—domain adaptation was first conducted on the influenza sequence corpus, followed by task-specific fine-tuning via transfer learning. Influ-BERT is capable of accurately recognition subtypes based solely on genome sequences, without relying on specific HA or NA segments. This intriguing phenomenon may be attributed to the co-evolutionary relationships among the viral gene segments, whereby discriminative information related to HA or NA may be implicitly encoded in other segments.

Despite the effective progress achieved by the Influ-BERT model in influenza virus subtype recognition, some limitations remain. Firstly, the fine-tuning phase of the current model relies on supervised closed-set learning, which prevents the recognition of novel subtypes not present in the training data. Future work will incorporate confidence and uncertainty estimation methods to extend the model to an open-set recognition mode. Secondly, due to the model's end-to-end learning paradigm, which does not require a reference genome, it cannot directly provide users with viral gene annotation information. Future work will incorporate multi-modal learning to endow the model with sequence annotation capabilities.

Key Points

- We constructed an Influenza-specific corpus and adopted a two-stage pretraining strategy, significantly improving the model's ability to identify low-frequency subtypes and addressing the limitations of general models under long-tail distribution.
- We introduced a sliding-window perturbation analysis to enhance model interpretability, enabling the visualization of subtype-discriminative regions and revealing attention patterns critical for accurate recognition.
- Influ-BERT achieves robust subtype recognition while demonstrating strong generalizability, enabling its potential extension to other viral typing tasks.

Acknowledgements

We thank Yiji Yang for valuable help and insightful discussions.

Funding

This work was supported by the National Key Research & Development Program of China (2025YFF1207901 to S.H.S., 2023YFC2604400 to F.C., 2024YFC2311303 to Z.Z.), the Key Collaborative Research Program of the Alliance of National and International Science Organizations for the Belt and Road

Regions (ANSO-CR-KP-2022-09 to S.H.S.), the National Natural Science Foundation of China (32270718 to S.H.S.).

References

1. Nair, Harish, et al. "Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis." The Lancet 378.9807 (2011): 1917-1930.

2. Iuliano, A. Danielle, et al. "Estimates of global seasonal influenza-associated respiratory mortality: a modelling study." The Lancet 391.10127 (2018): 1285-1300.

3. Li, Xiuli, et al. "Packaging signal of influenza A virus." Virology Journal 18.1 (2021): 36.

4. Taubenberger, Jeffery K., and John C. Kash. "Influenza virus evolution, host adaptation, and pandemic formation." Cell host & microbe 7.6 (2010): 440-451.

5. Peacock, Thomas P., et al. "The global H5N1 influenza panzootic in mammals." Nature 637.8045 (2025): 304-313.

6. Zhang, Jiahao, et al. "H9N2 avian influenza viruses: challenges and the way forward." The Lancet Microbe 4.2 (2023): e70-e71.

7. Shi, Jianzhong, et al. "Rapid evolution of H7N9 highly pathogenic viruses that emerged in China in 2017." Cell host & microbe 24.4 (2018):

8. Eisfeld, Amie J., Gabriele Neumann, and Yoshihiro Kawaoka. "Influenza A virus isolation, culture and identification." Nature protocols 9.11 (2014): 2663-2681.

9. Lee, Ming-Shiuh, et al. "Identification and subtyping of avian influenza viruses by reverse transcription-PCR." Journal of virological methods 97.1-2 (2001): 13-22.

10. Trombetta, Claudia Maria, et al. "Overview of serological techniques for influenza vaccine evaluation: past, present and future." Vaccines 2.4 (2014): 707-734.

11. Gauger, Phillip C., and Amy L. Vincent. "Serum virus neutralization assay for detection and quantitation of serum neutralizing antibodies to influenza A virus in swine." Animal Influenza Virus: Methods and Protocols. New York, NY: Springer US, 2020. 321-333.

12. Ghedin, Elodie, et al. "Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution." Nature 437.7062 (2005): 1162-1166.

13. Patrono, Livia V., et al. "Archival influenza virus genomes from Europe reveal genomic variability during the 1918 pandemic." Nature communications 13.1 (2022): 2314.

14. Poon, Art F. Y. "Prospects for a sequence-based taxonomy of influenza A virus subtypes." Virus Evolution 10.1 (2024): veae064.

15. Aksamentov, Ivan, et al. "Nextclade: clade assignment, mutation calling and quality control for viral genomes." Journal of open source software 6.67 (2021): 3773.

16. Humayun, Fahad, et al. "Computational method for classification of avian influenza A virus using DNA sequence information and physicochemical properties." Frontiers in Genetics 12 (2021): 599321.

17. Cacciabue, Marco, and Débora N. Marcone. "INFINITy: A fast machine learning‐based application for human influenza A and B virus subtyping." Influenza and Other Respiratory Viruses 17.1 (2023): e13096.

18. Borkenhagen, Laura K., Martin W. Allen, and Jonathan A. Runstadler. "Influenza virus genotype to phenotype predictions through machine learning: a systematic review: computational prediction of influenza phenotype." Emerging microbes & infections 10.1 (2021): 1896-1907.

19. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

20. Theodoris, Christina V., et al. "Transfer learning enables predictions in network biology." Nature 618.7965 (2023): 616-624.

21. Sapoval, Nicolae, et al. "Current progress and open challenges for applying deep learning across the biosciences." Nature Communications 13.1 (2022): 1728.

22. Peng, Cheng, et al. "ViraLM: empowering virus discovery through the genome foundation model." Bioinformatics 40.12 (2024): btae704.

23. Consens, Micaela E., et al. "Transformers and genome language models." Nature Machine Intelligence (2025): 1-17..

24. Nguyen, Eric, et al. "Sequence modeling and design from molecular to genome scale with Evo." Science 386.6723 (2024): eado9336.

25. Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." Bioinformatics 37.15 (2021): 2112-2120.

26. Zhou, Zhihan, et al. "Dnabert-2: Efficient foundation model and benchmark for multi-species genome." arXiv preprint arXiv:2306.15006 (2023).

27. Shao, Bin, and Jiawei Yan. "A long-context language model for deciphering and generating bacteriophage genomes." Nature Communications 15.1 (2024): 9392.

28. Hou, Xin, et al. "Using artificial intelligence to document the hidden RNA virosphere." Cell 187.24 (2024): 6929-6942.

29. Krammer, Florian, and Stacey Schultz-Cherry. "We need to keep an eye on avian influenza." Nature Reviews Immunology 23.5 (2023): 267-268.
30. Petrova, Velislava N., and Colin A. Russell. "The evolution of seasonal influenza viruses." Nature Reviews Microbiology 16.1 (2018): 47-60.
31. Caserta, Leonardo C., et al. "Spillover of highly pathogenic avian influenza H5N1 virus to dairy cattle." Nature 634.8034 (2024): 669-676.
32. Federhen, Scott. "The NCBI taxonomy database." Nucleic acids research 40.D1 (2012): D136-D143.
33. Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.
34. Pisner, Derek A., and David M. Schnyer. "Support vector machine." Machine learning. Academic Press, 2020. 101-121.
35. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. John Wiley & Sons, 2013.
36. Niu, Shuteng, et al. "A decade survey of transfer learning (2010–2020)." IEEE Transactions on Artificial Intelligence 1.2 (2021): 151-166.

Figures Legend

Figure 1 Influenza viruses (IAV) subtype recognition and distribution. a) Diagram of specific subtypes recognition task. The circle size represents the ability to recognize specific subtypes, where the baseline model demonstrates limited performance for some subtypes (left panel) and our research is to improve subtype recognition ability; b) The frequency distribution of the top 20 subtypes in the downstream task dataset; c) Length distribution of segments within the corpus used during the domain-adaptive pretraining phase;

Figure 2 Schematic overview of the Influ-BERT framework. a) A BPE tokenizer is trained on a curated influenza-specific corpus to construct a domain-specific vocabulary, which is used for domain-adaptive pretraining based on the DNABERT-2 architecture; b) Sequence recognition fine-tuning workflow: After domain-adaptive pretraining, downstream subtyping tasks are performed via transfer learning combined with task-specific fine-tuning; c) Interpretability analysis: A masking-based perturbation strategy is employed to identify attention-sensitive regions, enabling interpretation of model predictions.

Figure 3 Experimental results of downstream subtyping tasks. a) Performance evaluation across five common subtypes; b) Performance evaluation across ten subtypes.

Figure 4 Illustration of perturbation analysis for three randomly selected H5N1 sequences. The left panel shows the results of DNABERT-2, while the right panel corresponds to Influ-BERT.

Supplementary figures legend

Supplementary Figure S1 H3N2 Sensitivity Analysis. Influ-BERT demonstrates relatively consistent sensitivity patterns across sequences, whereas DNABERT-2 exhibits greater variability. Notably, Influ-BERT consistently highlights region-specific sensitivity at positions 1300–1350 nt, 1200 nt, 1050–1100 nt, and 350 nt, a pattern not observed in DNABERT-2.

Supplementary Figure S2 H13N8 Sensitivity Analysis. Influ-BERT demonstrates a higher consistency in sensitivity distribution, whereas DNABERT-2 shows almost no discernible pattern. DNABERT-2 exhibits a more widespread sensitivity across regions, making it difficult for the model to identify important areas. In contrast, Influ-BERT reveals a more systematic distribution of sensitive regions, indicating that sequences with such a distribution are more likely to be classified as the H13N8 subtype.

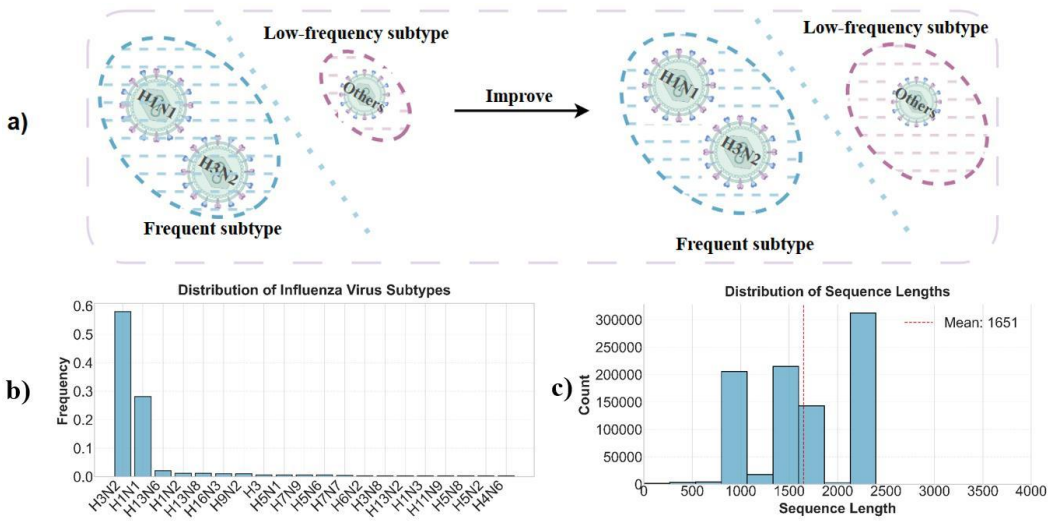

Figure 1

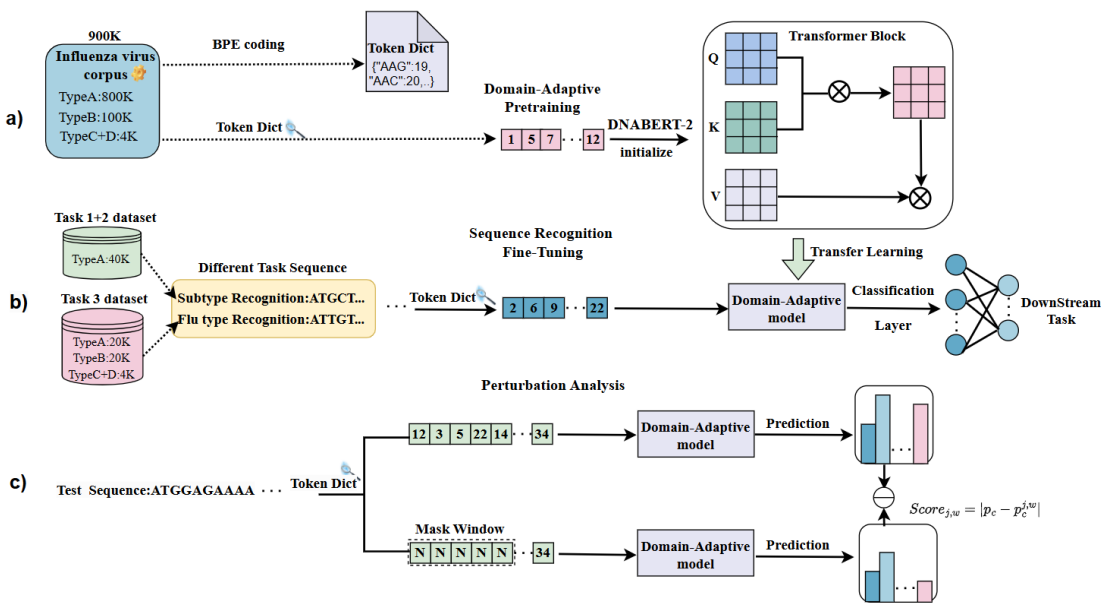

Figure 2

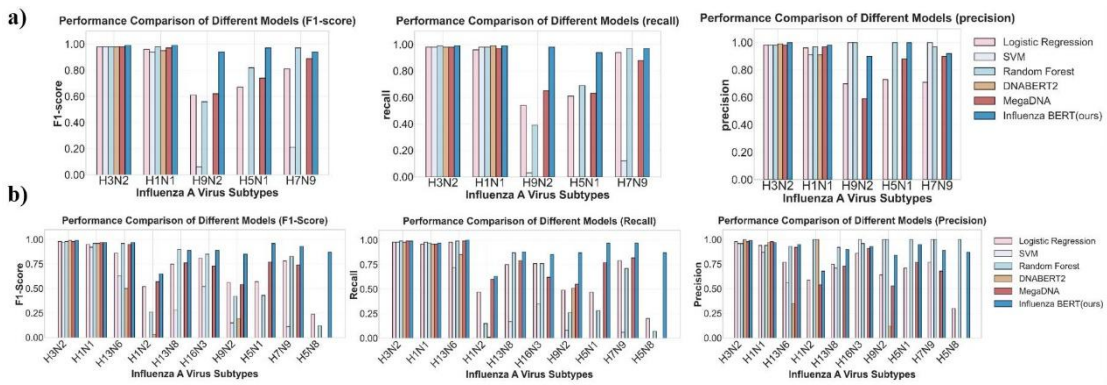

Figure 3

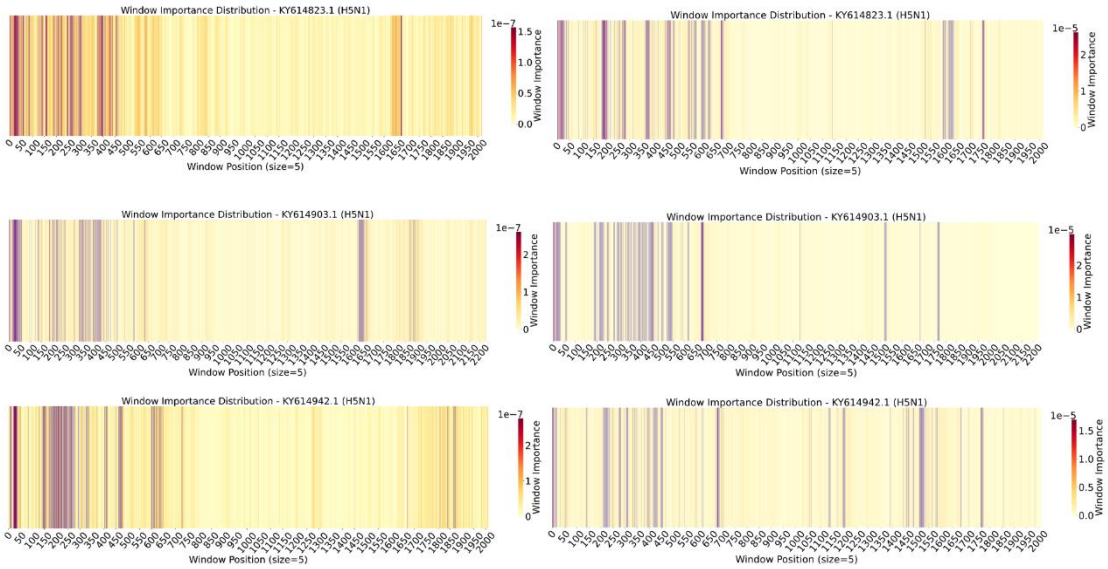

Figure 4

Table 1. Influenza BERT was compared with baseline models on three downstream tasks.

Task	Model	Accuracy	Precision	Recall	F1_Score
IAV Subtype 5	LR	0.9677	0.9672	0.9677	0.9672
	SVM	0.9565	0.9517	0.9565	0.9462
	RF	0.9774	0.9776	0.9774	0.9757
	DNABERT2	0.9578	0.9356	0.9578	0.9461
	MegaDNA	0.9785	0.9785	0.9785	0.9785
	Influenza BERT	**0.9899**	**0.9901**	**0.9899**	**0.9900**
IAV Subtype 10	LR	0.9490	0.9473	0.9490	0.9474
	SVM	0.9210	0.9038	0.9210	0.9006
	RF	0.9533	0.9546	0.9533	0.9444
	DNABERT2	0.9114	0.9267	0.9114	0.9090
	MegaDNA	0.9613	0.9599	0.9613	0.9601
	Influenza BERT	**0.9756**	**0.9753**	**0.9756**	**0.9754**
FLU typing	DNABERT2	0.9874	0.9877	0.9874	0.9868
	MegaDNA	0.9978	0.9978	0.9978	0.9978
	Influenza BERT	**0.9982**	**0.9982**	**0.9982**	**0.9982**

Note: "IAV Subtype 5" and "IAV Subtype 10" refer to recognition tasks involving five and ten influenza A subtypes, respectively. "FLU Typing" denotes the recognition of influenza virus types (A/B/C/D).