Development a computational framework for risk evaluation of emerging flu viruses based on genomic sequences and AI

Huiru Li[1,#], Yang Feng[1,#], Congyu Lu[2], Ping Fu[1], Xinglian Wang[1], Lei Yang[3], Yuelong Shu[4], Taijiao Jiang[2,*], Dayan Wang[3,*], Yousong Peng[1,*]

[1]Bioinformatics Center, College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha 410082, China

[2]Guangzhou National Laboratory, Guangzhou 510005, China

[3]National Institute for Viral Disease Control and Prevention, China CDC, Beijing, 102206, China

[4]Key Laboratory of Pathogen Infection Prevention and Control (MOE), State Key Laboratory of Respiratory Health and Multimorbidity, National Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 102629, China

[#] These authors contributed equally to this work

* Corresponding author. Email: pys2013@hnu.edu.cn (YP), dayanwang@cnic.org.cn (DW), jiang_taijiao@gzlab.ac.cn (TJ)

Abstract

The recent resurgence of highly pathogenic avian influenza H5N1 viruses in North America and Europe has heightened global concerns regarding potential influenza pandemics. Despite significant progress in the surveillance and prevention of emerging influenza viruses, effective tools for rapid and accurate risk assessment remain limited. Here, we present FluRisk, an innovative computational framework that integrates viral genomic data with artificial intelligence (AI) to enable rapid and comprehensive risk evaluation of emerging influenza strains. FluRisk incorporates a curated database of over 1,000 experimentally validated molecular markers linked to key viral phenotypes, including human adaptation, mammalian virulence, human-to-human transmission, receptor binding specificity, and antiviral drug resistance. Leveraging these markers, we

developed three state-of-the-art machine learning models to predict human adaptation, mammalian virulence, and human receptor-binding potential, all of which demonstrated superior performance compared to traditional approaches such as BLAST, prior models, and baseline classifiers. In addition, a reference-based method was implemented to provide preliminary estimates of transmissibility and resistance to six commonly used antiviral drugs. To facilitate broad accessibility and practical application, we developed a user-friendly web server that integrates both the molecular marker atlas and predictive tools for influenza virus phenotyping (available at: http://www.computationalbiology.cn/FluRisk/#/). This computational platform offers a valuable resource for the timely risk assessment of emerging influenza viruses and supports global influenza surveillance efforts.

Introduction

The influenza virus is a negative-sense, single-stranded RNA virus with segmented genomes, classified into four types: A, B, C, and D[1]. Among these, influenza A viruses (IAVs) exhibit the broadest host range, infecting diverse avian species as well as mammals, including humans[2]. IAVs are further categorized into subtypes—such as H1N1, H3N2, and H7N9—based on their surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA)[1]. Over the past two decades, emerging IAVs have posed persistent threats to global public health, exemplified by outbreaks of H7N9 and H5N1[3–6]. Notably, the highly pathogenic avian influenza (HPAI) H5N1 virus has demonstrated alarming cross-species transmission, infecting over 50 mammalian species such as cattle, sheep, pig, cats, etc[7–10]. This expanding host range has intensified global concerns about its pandemic potential. Effective prevention and control of emerging IAVs remain a critical priority for public health worldwide.

Risk evaluation of emerging influenza viruses is a prerequisite for effective control of the virus. The World Health Organization (WHO) has developed the Tool for Influenza Pandemic Risk Assessment (TIPRA) to evaluate the potential public health impact of

emerging influenza viruses[11]. TIPRA relies on comprehensive data, including virological, epidemiological, and clinical information, to assess the likelihood of pandemics caused by emerging influenza virus[11]. However, obtaining detailed phenotypic information such as host, virulence and transmission ability for newly emerging viruses remains a significant challenge, as it often requires extensive laboratory testing and epidemiological studies. The inherent complexity and time-consuming nature of these processes mean that critical information may not be available in a timely manner, potentially delaying effective and accurate evaluation of pandemic risk of emerging influenza viruses.

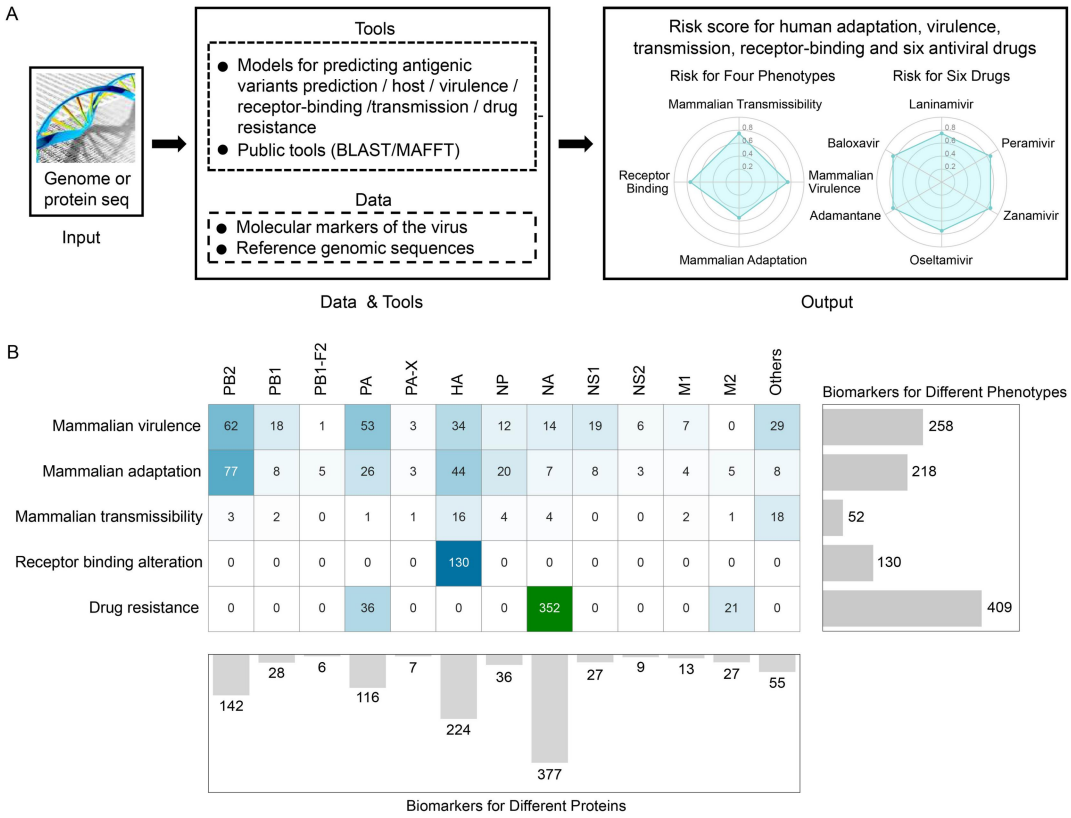
Development of computational methods for rapid determination of virus phenotypes may help the risk assessment of influenza viruses, given the low cost and ease of obtaining influenza virus genomes, and the development of powerful artificial intelligence (AI) algorithms such as deep-learning algorithms and large language models. For example, the GISAID database has stored more than 3 million genomic sequences from 637,120 virus strains as of April 15th, 2025[12]. Lots of computational methods have been developed to infer virus phenotypes such as virus host, virulence and antigenic variation directly from genomic sequences based on AI algorithms[13–15]. For example, we developed the naive bayes method named as PREDAC to predict antigenic variation of influenza viruses based on HA protein sequences[13]. Yin et al. developed a general framework for mammalian virulence prediction of influenza viruses with prior viral knowledge using genomic sequences[14]. Borkenhagen et al. conducted a systematic review of machine-learning (ML) approaches for predicting influenza virus phenotypes[16]. However, despite these advancements, the widespread adoption of these models in influenza virus surveillance and risk assessment has been limited, partly due to inadequate model performance and a lack of user-friendly tools.

Molecular markers are defined as specific amino acids at a given site that are associated with influenza virus phenotypes. For example, PB2-627K is associated with increased pathogenicity of influenza viruses in mammals[17,18]. Numerous molecular markers associated with multiple phenotypes of influenza viruses including host adaptation, virulence, antigenic variation, drug-resistance, and human transmissibility have been

87 identified by experiments[19–23]. These markers offer a potential shortcut for roughly
88 inferring viral phenotypes, thereby facilitating rapid risk assessments in influenza virus
89 surveillance. However, most of molecular markers are determined with one or a few viral
90 strains. Whether they can be used in other viral strains are not known, given the epistasis
91 among amino acid sites. Besides, there is still a lack of a unified resource for these
92 molecular markers.

93 Previously, we developed FluPhenotype, a one-stop platform for inferring the influenza
94 virus phenotypes of influenza A viruses[24]. It only provides functions of predicting viral
95 host and antigenicity. In this study, we presents FluRisk, an innovative platform that
96 leverages genomic sequences and AI to provide rapid and comprehensive risk assessments
97 of emerging influenza viruses. FluRisk contains an up-to-date atlas of molecular markers
98 of influenza viruses and several ML models for predicting five important phenotypes of
99 influenza viruses including human adaptation, mammalian virulence,
100 human-receptor-binding ability, human transmission and drug-resistance. A online
101 web-server was further built for FluRisk. It would be a useful tool for risk assessment of
102 emerging influenza viruses.
103
104
105 **Results**
106 **Overview of FluRisk**
107 The overview of FluRisk is illustrated in Figure 1A. It only takes influenza virus genomic or
108 protein sequences as input. FluRisk stored a database of molecular markers associated with
109 phenotypes including human adaptation, mammalian virulence, human receptor-binding
110 preference, human transmission and resistance to six commonly used anti-flu virus drugs. Based
111 on the database, FluRisk automatically extract molecular markers associated with phenotypes
112 from genomic or protein sequences. Then, several computational models were developed to
113 predict the aforementioned phenotypes and antigenic characteristics based on genomic or protein
114 sequences. The output of FluRisk includes the molecular markers associated with phenotypes
115 identified from the inputted genomic or protein sequences and the predicted phenotypes which

116 were presented as risk scores for early warnings of emerging flu viruses. The details about the
117 molecular markers and computational methods were described in the following texts.
118



119
120 **Figure 1** Overview of FluRisk (A) and Summary of molecular markers of important influenza
121 virus phenotypes (B)
122
123 **Summary of molecular markers associated with flu phenotypes**
124 A database of molecular markers associated with flu phenotypes were firstly built in FluRisk. It
125 includes a total of 1067 molecular markers that were manually curated and checked from public
126 literatures (see Methods). For the phenotype of human adaptation, there were 218 molecular
127 markers that were located on all viral proteins (Figure 1B). PB2 contained most molecular
128 markers (77) associated with human adaptation, such as PB2-627K and 701N, followed by HA,
129 PA and NP proteins that contained 44, 26 and 20 molecular markers, respectively. Other proteins
130 contained less than 10 molecular markers. For the phenotype of mammalian virulence, there were
131 258 molecular markers that were located on all viral proteins except M2. Similar to the human
132 adaptation, PB2 contained the largest number of molecular markers (62), followed by PA and HA

133 proteins. Besides, NS1 and PB1 also contained near 20 molecular markers. For the phenotype of
134 human receptor-binding preference, there were 130 molecular markers that were located on the
135 HA protein of 7 HA subtypes. For the phenotype of human transmission, there were a total of 52
136 molecular markers that were located on multiple proteins. Nearly 1/3 of molecular markers were
137 located on the HA protein. Interestingly, there were 1/3 molecular markers that were located on
138 two or more proteins, such as H3-225G & N2-315N. For the phenotype of drug-resistance, there
139 were 208, 197, 144, 71, 37, and 11 molecular markers for anti-virus drugs of oseltamivir,
140 zanamivir, peramivir, laninamivir, baloxavir, and adamantane, respectively. They were classified
141 into four groups based on the IC50 that measures the extent of drug resistance: low, moderate,
142 high and unknown. Most of these molecular markers (352) were located on the NA protein of 3
143 NA subtypes that were associated with the NA inhibitor. The remaining 57 molecular markers
144 were located on PA and M2 proteins.
145 When analyzed by protein, the NA protein contained the largest number of molecular markers and
146 most of which were related to the drug-resistance; the HA protein contained the second largest
147 number of molecular markers and half of which were related to the human receptor-binding
148 preference; the PB2 protein contained the third largest number of molecular markers that were
149 mainly related to the human adaptation and mammalian virulence. Besides, the PA protein also
150 contained more than 100 molecular markers that were related to the human adaptation,
151 mammalian virulence and drug-resistance.
152
153
154 **Influenza virus host prediction with machine-learning algorithms**
155 Then we built several computational methods to predict the flu virus phenotypes based on the
156 molecular markers associated with flu virus phenotypes. Firstly, we attempted to predict the flu
157 virus host based on the molecular markers associated with human adaptation. Analysis of the
158 number of molecular markers associated with human adaptations in flu virus genomes showed that
159 the human influenza virus has much larger number of molecular markers associated with human
160 adaptation than the avian influenza virus (a median of 84 *vs* 18) (Figure 2A). When analyzing the
161 molecular markers by protein, the results showed that nearly all proteins of human influenza
162 viruses had significant larger number of molecular markers than that of avian influenza viruses

(Figure 2B). The PB2 protein of human influenza viruses had the largest number of molecular markers among all proteins, with a median of 40, followed by NP and PA protein that have a median of 17 and 12, respectively.

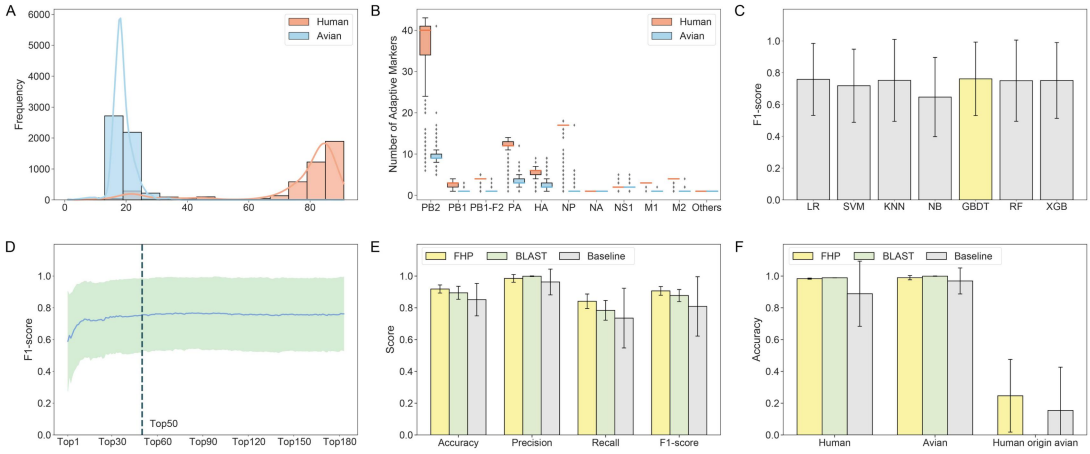


Figure 2 Flu Host Prediction based on molecular markers of human adaptation. Comparison of the number of human-adaptation molecular markers in human and avian influenza viruses in the whole genome (A) and each protein (B). (C) and (D) showed the selection of machine-learning algorithms and feature sets for FHP. (E) Performance comparison of FHP, BLAST and the baseline method on the independent test dataset. (F) Performance comparison of three methods by flu virus host.

Based on the analysis mentioned above, we then attempted to build a machine learning model (named as FluHostPredictor (FHP)) to discriminate human and avian influenza viruses based on the number of molecular markers associated with human adaptation. The workflow of model construction was shown in Figure S1 and described in detail in the Methods section. As shown in Figure 2C, the Gradient Boosting Decision Tree algorithm (GBDT) had the largest F1 score (0.762) among multiple algorithms on the validation dataset when using all features and was selected as the optimal algorithm in FHP. To reduce over-fitting, the top 50 most important features were selected when balancing the number of feature and the model performance (Figure 2D). When testing on the testing dataset, the model had an average F1 score of 0.906 (Figure 2E). Specifically, the accuracy, precision and recall was 0.918, 0.985 and 0.841, respectively. For comparison, we also evaluated the performance of the traditional method of BLAST which takes the host of the best hit as the predicted host of the query virus, and a baseline method which determined the host

186 the influenza virus based on the total number of human-adaptation molecular markers in the virus
187 genome (see Methods). Both the BLAST and baseline methods had high precision of near 100%,
188 however, they had much lower recall rate than the FHP. Analysis of the model performance by
189 virus host showed that all three methods had high accuracy in predicting avian influenza viruses
190 and human influenza viruses (Figure 2F). However, all of them had poor performance in
191 predicting the human-origin avian influenza viruses. FHP had a median accuracy of 0.444, while
192 the BLAST and baseline method had little ability in identifying human-origin avian influenza
193 viruses.
194
195 **Mammalian virulence prediction of influenza viruses with machine-learning algorithms**
196 Then, we attempted to predict the mammalian virulence of influenza viruses based on molecular
197 markers of mammalian virulence. Firstly, a dataset of 659 influenza virus strains with
198 experimentally-determined virulence to mammals was manually curated and checked from public
199 literatures. It contains 198 and 461 high and low-virulence influenza viral strains, respectively,
200 from 21 subtypes (Figure 3A). Among them, the H5N1 had the largest number of viral strains,
201 with 101 and 143 high and low-virulence strains, respectively, followed by the H1N1 subtype
202 which had 41 and 120 high and low-virulence strains, respectively. Other subtypes had low than
203 100 viral strains. Analysis of the number of molecular markers associated with mammalian
204 virulence showed that the high-virulence strains had larger number of molecular markers than the
205 low-virulence strains (p-value = 6.60e-04) in the wilcox-test) (Figure 3B). The former had a
206 median of 45 molecular markers, while the latter had a median of 43 molecular markers. Analysis
207 of molecular markers by protein showed that the high-virulence viral strains had larger number of
208 molecular markers than the low-virulence viral strains on PB2 and HA proteins (Figure 3C).
209

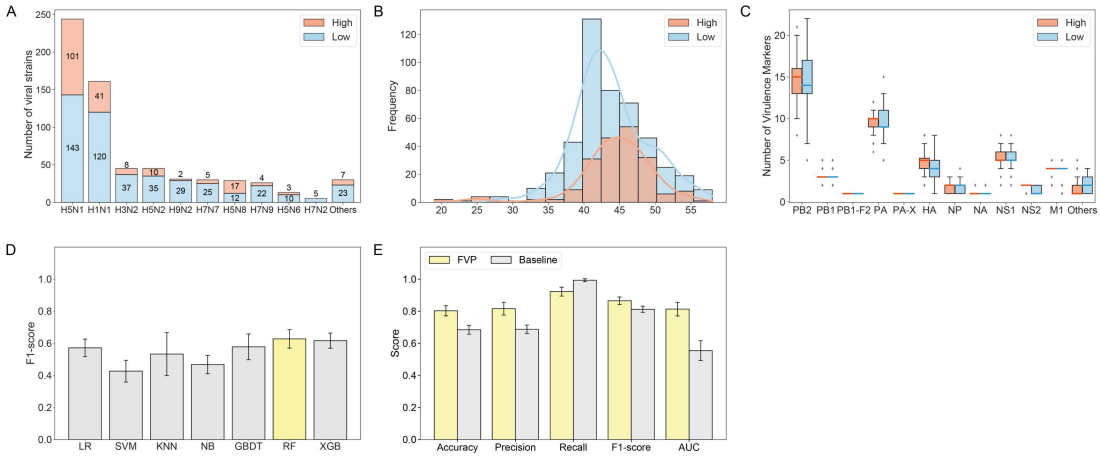


Figure 3 Mammalian virulence prediction of influenza viruses. (A) The number of high and low-virulence influenza virus strains in different influenza subtypes. (B) and (C) shows the comparison of the number of mammalian-virulence molecular markers in high and low-virulence influenza viruses in the whole genome and each protein, respectively. (D) Selection of machine-learning algorithms for FVP. (E) Performance comparison of FVP and the baseline method on the independent test dataset.

Based on the analysis, we built a computational method named FluVirulencePredictor (FVP) based on molecular markers associated with mammalian virulence. The workflow of FVP was similar to that of FHP. The random-forest (RF) algorithm had the best performance (F1-score) among several algorithms including LR, SVM, KNN, NB, GBDT and XGBoost, and was selected for FVP. To reduce over-fitting, the method of VSURF was used in the feature selection (see Methods). Evaluation of FVP on the testing dataset showed that FVP had a overall F1-score and AUC of 0.814 and 0.866, respectively. Specifically, the accuracy, recall and precision for FVP were 0.803, 0.922 and 0.816, respectively. For comparison, we evaluated the baseline method that discriminates the high and low-virulence viral strains only based on the total number of molecular markers in the viral genome. As show in in Figure 3E, FVP outperformed the baseline method on all metrics except the recall rate (0.922 *vs* 0.993). Specifically, FVP had an average accuracy, precision, F1-score and AUC of 0.803, 0.816, 0.867, 0.814, respectively.

Besides, we also compared FVP to two methods developed previously by Yin et al. in predicting mammalian virulence on their dataset since we cannot repeat their methods due to unavailability of several important parameter values. The first method is ViPal that incorporated genomic

233 sequences, discrete prior viral mutation and reassortment information using a general
234 machine-learning framework[14]. The second method is VirPreNet that used a weighted ensemble
235 convolutional neural network (CNN) with ProtVec encoding to predict mammalian virulence[25]. As
236 show in in Table S1, FVP outperformed these methods on all metrics except the recall rate for
237 which FVP performed comparably with VirPreNet.
238
239 **Prediction of receptor-binding preference of influenza virus with machine-learning**
240 **algorithms**
241 Then, we attempted to build a computational method to predict the receptor-binding preferences of
242 influenza viruses. Firstly, a dataset of 620 influenza A virus strains with
243 experimentally-determined human or avian receptor-binding preference was manually curated and
244 checked from public databases and literatures. The receptor-binding data of more than half of viral
245 strains (58.1%) were obtained using the glycan array (Figure 4A); 39.0% of viral strains were
246 determined using ELISA; the remaining viral strains were determined using BLI (1.9%) and SPR
247 (1.0%). The dataset contains 321 and 299 influenza virus strains with human and avian-receptor
248 preferences, respectively, from 33 influenza A virus subtypes. Among them, the H1N1 had the
249 largest number of viral strains, with 103 and 64 strains having human and avian-receptor
250 preference, respectively, followed by the H3N2 and H9N2 subtypes that both had 100 or more
251 strains.
252 Analysis of molecular markers associated with human-receptor-binding preference in viral strains
253 showed that the viral strain with human-receptor preference had larger number of molecular
254 markers than those with avian-receptor preference (p-value = 9.01e-31 in the wilcox-test). The
255 former had a median of 9 molecular markers associated with human-receptor-binding preferences,
256 while the latter had a median of 6 molecular markers (Figure 4B). Based on the analysis, a
257 computational method named Flu Receptor-binding preference Predictor (FRP) was built based on
258 the molecular markers associated with human-receptor-binding preference. FRP used a similar
259 workflow as FHP (Figure S1). The SVM algorithm had the best performance (F1-score) among
260 several algorithms including LR, KNN, NB, GBDT, RF and XGBoost, and was selected in FHP
261 (Figure 4C). To reduce over-fitting, the top 30 most important features were selected when
262 balancing the number of feature and the model performance (Figure 4D). When testing on the

testing dataset, the model had a F1-score and AUC of 0.834 and 0.870, respectively (Figure 4E). Specifically, the accuracy, precision and recall was 0.830, 0.832 and 0.841, respectively. For comparison, we also evaluated the performance of the baseline method which determined the receptor-binding preference based on the number of human-receptor-binding-related molecular markers in the virus genome. FRP outperformed the baseline method on all metrics except the recall (Figure 4E).

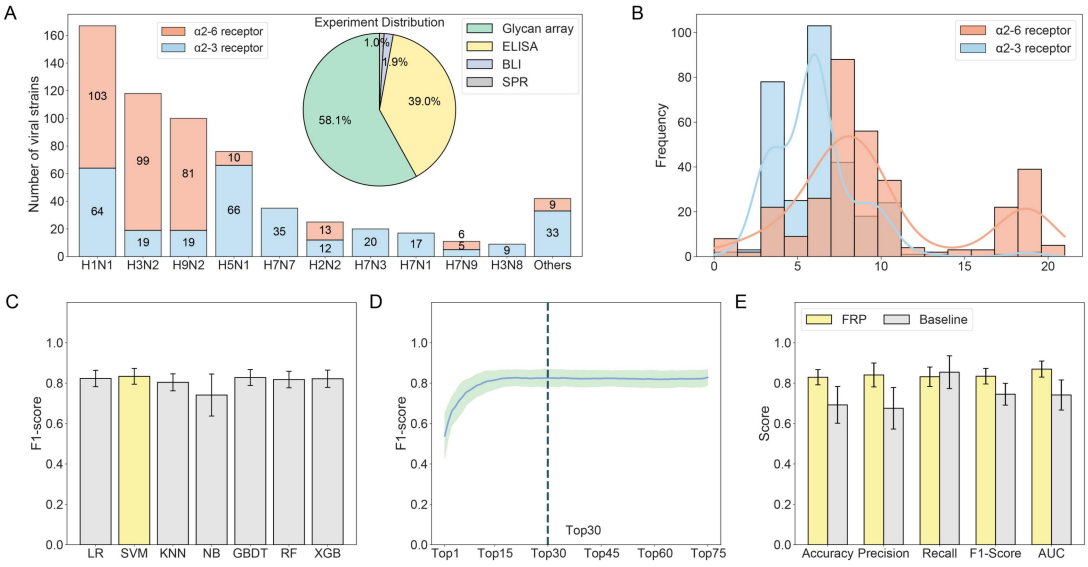


Figure 4 Prediction of influenza virus receptor-binding preference based on molecular markers associated with human-receptor-binding preference. (A) The number of viral strains with either human (α2-6-linked) or avian (α2-3-linked) receptor preference in different subtypes. The pie-chart on the top-right refers to the composition of experiments used in determining the receptor-binding ability of influenza viruses. (B) Comparison of the number of human-receptor-binding related molecular markers in viral strains with either human or avian receptor preferences. (C) and (D) showed the selection of machine-learning algorithms and feature sets for FRP. (E) Performance comparison of FRP and the baseline method on the independent test dataset.

Estimating the human transmission and drug-resistance of influenza viruses using a reference-based method

To estimate the human transmission abilities of influenza virus, we developed a reference-based method since there is a lack of enough data for building the machine-learning model. Firstly, a

library of reference strains which includes 500 representative human and avian virus strains each was built. As shown in Figure 5A, the human influenza viruses had much larger number of human-transmission-related molecular markers (orange) than the avian influenza virus (light blue) (p-value = 3.16e-135 in wilcox-test), with the former having a median of 5 and the latter having a median of 2, suggesting that the virus with more molecular markers of human-transmission have a higher ability of human transmission. Then, a cumulative distribution of the number of molecular-markers associated with human transmission in virus genomes was built using all reference strains (right panel in Figure 5A). To estimate the ability of human transmission for a given virus, the number of molecular markers associated with human transmission in the virus genome was obtained and then transformed to the ranking ratio in the cumulative distribution (see the red dashed line in right panel of Figure 5A) which is considered as the human transmission ability of the given virus.

The reference-based method was also used to estimate the drug-resistance of influenza viruses. A thousand representative human virus strains were selected from public databases and taken as the reference strains. The number of molecular markers associated with six common anti-flu drugs was obtained from the reference strains and a cumulative distribution was built for each drug. As shown in Figure 5 and S2, the reference strains had a median of 11 anti-Oseltamivir molecular markers (Figure 5B), while for other anti-flu drugs such as Zanamivir and Peramivir, most reference strains had less than ten molecular markers associated with the resistance to these drugs (Figure 5C&D).

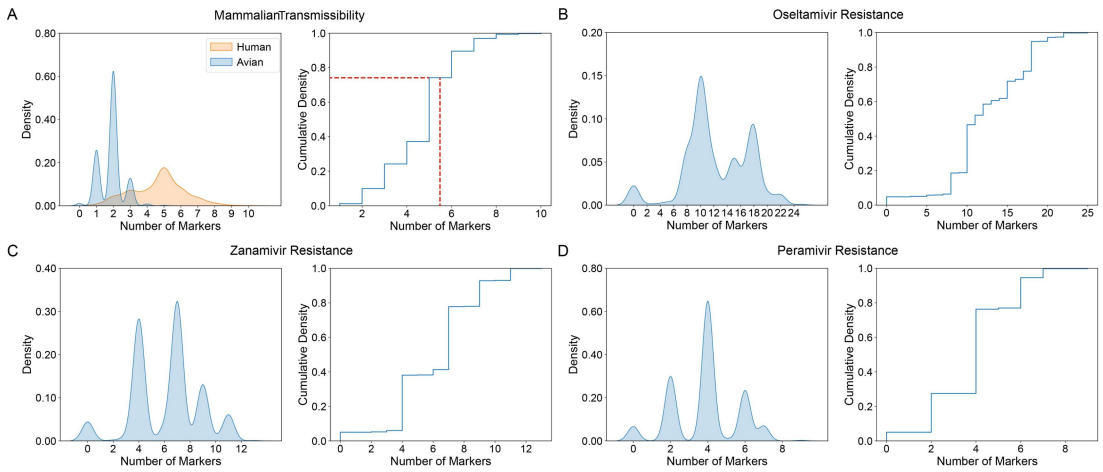


Figure 5 Estimation of human transmission and drug-resistance of influenza viruses based on molecular markers using a reference-based method. (A) The left panel shows the comparison of

the number of molecular markers associated with human transmission in reference human (orange) and avian (blue) influenza virus strains, while the right panel shows the cumulative distribution of the number of molecular-markers associated with human transmission in virus genomes using all reference strains. The red dashed line shows an example of estimating the ability of human transmission for a given virus. (B)-(D) shows the distribution of the number of molecular markers and the cumulative distribution of the number of molecular markers associated with Oseltamivir, Zanamivir, and Peramivir, respectively.

Development of online server of FluRisk

To make the database and computational methods available to the public, a user-friendly web server of FluRisk was built (Figure 6A) and was publicly available at http://47.90.94.155/FluRisk. Besides the tools mentioned above, the antigenicity of influenza viruses was predicted using the PREDAC methods that were developed in our previous studies. The main function of FluRisk is to evaluate the risk of emerging flu virus based on genomic sequences. It only takes genome or proteome sequences of influenza viruses as input, and outputs the molecular markers associated with flu phenotypes in the virus genome and the risk of the virus for five phenotypes. Besides, FluRisk provides the functions of searching and browsing molecular markers either by phenotype or protein. The detailed information about molecular markers including the functional consequences, source literature and the structure view were provided. Users can also submit newly discovered molecular markers to the server. In addition, several useful tools were incorporated in the FluRisk which included the antigenic variant prediction based on protein sequences, the amino acid numbering conversion of HA and NA subtypes, the FHP and FVP.

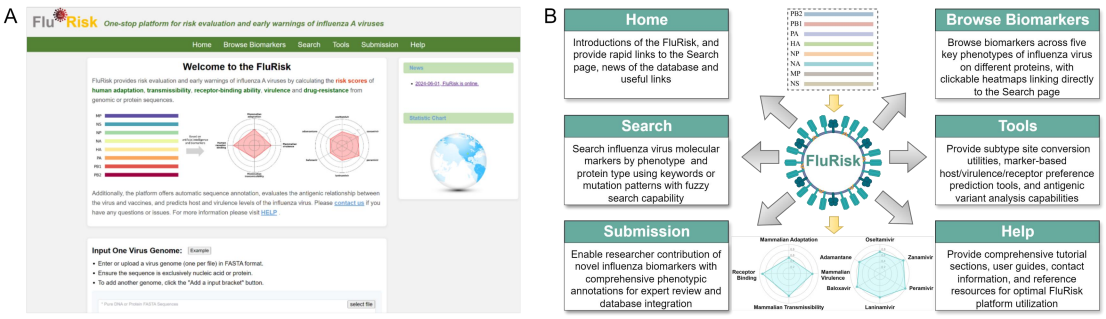


Figure 6 Construction of the web-server of FluRisk. (A) The homepage of the FluRisk web-server. (B) The structure of the FluRisk web-server.

Discussions

In this study, a comprehensive atlas of molecular markers associated with five important phenotypes of influenza viruses were built. It can not only provide a valuable resource for annotation of influenza virus variants, but also facilitate rapid inference of phenotypes for emerging influenza viruses. Based on the atlas of molecular markers, three up-to-date machine-learning models were built to predict the flu virus host, mammalian virulence and receptor-binding preferences using the genomic sequence. A reference-based method was also developed to estimate the human transmission ability and drug-resistance of influenza viruses based on the molecular markers. The rapid prediction of phenotypes can greatly facilitate the risk evaluation and early warnings of emerging influenza viruses. We further developed the FluRisk web server to integrate all these computational methods and the library of molecular markers， which make it a valuable tool in early warning of emerging viruses in practice.

WHO has built the TIPRA framework for evaluating the risk of flu pandemics based on essential virological, epidemiological, and clinical information[11]. Although it is an effective system for pandemic warnings, it may be not practical since there is limited data about the emerging influenza virus. Determination of biological properties of influenza viruses rely on experiments that are lab-intensive and time-consuming. Moreover, the emerging viruses may be dangerous and experiments on the virus need to be conducted in high-level biosafety labs, which further delay the risk assessment of the virus. Thus, large-scale screening of emerging viruses based on experiments is not feasible. Computational methods based on influenza virus genomic sequences may be more effective, given the low cost and ease of viral genome sequencing. Several methods have been developed to infer the influenza virus phenotypes directly from genomic sequences. Borkenhagen et al. summarized these methods in their review[16]. This study has two advantages compared to previous studies. Firstly, our models used molecular markers associated with phenotypes as features in the ML modeling. All these molecular markers have been experimentally-validated and may be reliable than the genomic features used in previous studies. Secondly, the datasets used in modeling the mammalian virulence and receptor-binding preference were larger than those in previous studies. They can be taken as benchmark datasets for further development of

362 computational methods.

363

364 Most phenotypes of influenza viruses such as virulence, host adaptation and human transmission
365 are determined by multiple genes and there are complex relationship between genotypes and
366 phenotypes[26,27]. Decoding such relationship relies on both advanced statistical methods and big
367 data on both virus genome and phenotypes. Unfortunately, the data of influenza virus phenotypes
368 are still limited. For example, only less than 1000 viral strains were used in modeling the
369 mammalian virulence and receptor-binding ability of influenza viruses. This on one hand asks for
370 high-through-put methods to determine the flu virus phenotypes, on the other hand, it asks for
371 development of advanced computational methods to decode the flu virus genomes since the virus
372 phenotypes are to a large extent determined by the genome. In recent years, several large language
373 models (LLMs) have been developed in virus research and they had been demonstrated to
374 outperform traditional ML algorithms in predicting biological characteristics of viruses[28–30]. For
375 example, Nie et al. developed the E2VD framework for predicting effect of mutations on binding
376 affinity, expression, and antibody escape of SARS-CoV-2 based on a customized pretrained
377 language model[30]. These models have been built based on the deep mutational scanning (DMS)
378 data. Although they had excellent performances in predicting the functional or structural influence
379 of single or multiple mutations, their performance decreased much when used in viral strains that
380 have different genetic background with the original viral strains used in the DMS experiment,
381 such as different variant of SARS-CoV-2[30]. In flu phenotype inference, the emerging viral strains
382 may differ much in genetic background as they may come from different influenza subtypes such
383 as H5N1, H7N9, H10N8, etc. Current LLMs developed based on DMS data may not be suitable in
384 inferring the flu phenotypes of influenza viruses. As the accumulation of DMS data on influenza
385 viruses, it is possible to build a LLM that can accurately capture the complex relationship between
386 virus genomes and phenotypes.

387 The study has several limitations. Firstly, although we have attempted to build a comprehensive
388 atlas of molecular markers of viral phenotypes, they may be still incomplete as novel mutations
389 are continuous generated. Therefore, the database should be continuously updated. Users of
390 FluRisk web server can also help update the database by submitting the newly-discovered
391 molecular markers to us. Secondly, the computational models of predicting flu phenotypes still

needs improvement, although they outperformed the competitor. More advanced statistical methods such as LLMs and more phenotype data of influenza viruses may help improve these models. Thirdly, our framework of risk evaluation of emerging flu viruses only used genomic sequences of influenza viruses. Incorporation of epidemiological, clinical and ecology factors as the WHO did can help much in risk evaluation of emerging flu viruses.

Conclusion

The study built a comprehensive atlas of molecular markers associated with five important influenza virus phenotypes which can be a valuable resource for flu virus research and surveillance. Several computational methods for inferring flu virus phenotypes were further built based on genomic sequences that can help rapid determination of viral phenotypes. Finally, a user-friendly web-server was built to integrate both computational methods and molecular markers which make it a valuable tool in early warning of emerging viruses.

Methods

Collection of molecular markers associated with phenotypes of influenza viruses

The molecular markers of influenza viruses, i.e., specific amino acids at a given site that are associated with influenza virus phenotypes, were mainly curated from two sources: the first is 397 molecular markers used in the FluPhenotype that was developed in our previous study to infer influenza virus phenotypes from genomic sequences[24]; the second is the literature from PubMed (https://pubmed.ncbi.nlm.nih.gov/) database. For the latter, we firstly obtained the abstract of literatures published between January 1st, 2019 to December 31th, 2022 since FluPhenotype has collected molecular markers from public literatures published before 2019. Only 8238 papers that were related to the influenza virus were kept. They were further filtered by a custom python script to keep only the literatures with mutations, which led to 1431 papers. Then, each paper was manually read to identify molecular markers associated with five phenotypes including human adaptation, mammalian virulence, human-receptor-binding ability, human transmission and resistance to six commonly used anti-flu virus drugs including amantadine, oseltamivir, zanamivir,

422 baloxavir marboxil, peramivir and laninamivir. For the drug-resistance molecular markers, the
423 IC50 (if available) that measure the extent of drug-resistance was also recorded. The
424 drug-resistance molecular markers were classified into four groups based on fold-change of IC50
425 (IC50-FC) after mutations: low-resistance group (2< IC50-FC <=10), middle-resistance group
426 (10< IC50-FC <=100), high-resistance group (IC50-FC > 100) and unknown group (IC50
427 unavailable).
428 The molecular markers curated from two sources mentioned above were integrated together,
429 which resulted in 1067 molecular markers associated with five important phenotypes of influenza
430 viruses. The were publicly available on the web server of FluRisk
431 (http://www.computationalbiology.cn/FluRisk/#/).
432
433 **The data of influenza virus genomic sequences**
434 The influenza virus genomic sequences were downloaded from GISAID database on May 20th,
435 2023. Only 18966 complete genomic sequences were kept. Then, the CD-HIT (version 4.8.1)[31]
436 was used to remove sequence redundancy at the level of 0.999, which led to 5257, 4155, 349
437 genomic sequences of avian, human and human-origin avian influenza viruses, respectively. They
438 were used in building the models for predicting influenza virus host. Among them, 1000 genomic
439 sequences of human influenza viruses were randomly selected and were taken as reference
440 sequences for inferring the risk of resistance to anti-flu virus drugs; 500 genomic sequences for
441 both avian and human influenza virus were randomly selected and were taken as the reference
442 sequences for inferring the risk of human transmission of the influenza virus.
443
444 **The data of influenza virus mammalian virulence**
445 The data of mammalian virulence of influenza virus were mainly curated from the studies of Yin
446 et al. and Peng et al[14,32]. The influenza virus strains with mammalian virulence from these two
447 studies were combined. To ensure consistence between two studies, viral strains that cause a high
448 mortality with a low dose or have a LD50 ≤ $10^{\wedge 3}$ 50% egg infectious doses were considered as
449 high-virulence; otherwise, low-virulence. This led to a dataset of 659 viral strains including 198
450 and 461 high and low-virulence strains, respectively. The genomic sequences of these viral strains
451 were obtained from these two studies and the GISAID database.

452

453 **The data of influenza virus binding to human and avian receptors**

454 The binding ability of influenza virus to α2-3-linked and α2-6-linked glycan receptors were
455 obtained from two sources: the first is the CFG (Consortium for Functional Glycomics) database
456 that stored the binding ability of proteins to glycans in mammals using the glycan array[33]. A total
457 of 606 influenza virus-related records were obtained from the database on December 18th, 2024.
458 The hemagglutinin (HA) protein sequences of these viral strains were obtained from the CFG
459 database or the GISAID database. After removing the viral strains without HA protein sequences,
460 we obtained 360 viral strains with the binding ability to both α2-3-linked and α2-6-linked glycans
461 and with HA protein sequences. The second source is the literatures that reported the
462 receptor-binding data of influenza viruses. Except the glycan array, three kinds of methods were
463 commonly used in investigating the binding ability of influenza viruses to the glycan: the surface
464 plasmon resonance (SPR) and the biolayer interferometry (BLI) that measure the receptor-binding
465 ability of influenza viruses by the equilibrium dissociation constants (K_d) based on the changes in
466 refractive index and the wavelength shifts resulting from biomolecular interactions on the
467 biosensor surface, respectively, and the enzyme-linked immunosorbent assay (ELISA) which
468 quantifies the virus-receptor binding through absorbance values, K_d or equilibrium association
469 constants (K_a) (reciprocal of K_d)[34–36]. The larger of the absorbance or K_a refers to the stronger
470 binding ability of viruses to glycans. Taken together, a total of 620 viral strains with
471 experimentally-validated binding ability to both α2-3-linked and α2-6-linked glycans were
472 obtained. Viral strains with higher binding ability to α2-6-linked glycans than that to α2-3-linked
473 glycans were considered to have preference to human receptors, otherwise, the avian receptors.

474

475 **Framework for predicting the flu host, virulence and receptor-binding preference using**
476 **machine learning algorithms**

477 Figure S1 showed the workflow of predicting the flu host, mammalian virulence and
478 receptor-binding preference using machine learning algorithms based on genomic sequences.
479 Firstly, molecular markers associated with a given phenotype such as human adaptation were
480 extracted from the genomic or protein sequences of influenza viruses. They were taken as the
481 features in the machine learning modeling. Then, the dataset was randomly separated into two

482 parts with a ratio of about 8:2. The first part was further separated into the training and validation
483 datasets with a ratio of 8:2, while the second part was taken as an independent testing dataset. The
484 training and validation dataset were used to select the optimal algorithm and feature set, and then
485 used to train the machine learning models, while the testing data was used to evaluate the
486 performance of the model. This process was repeated 100 times to reduce randomness in dataset
487 separation.
488 Considering that the dataset for flu host prediction is large and contains lots of HA and NA
489 subtypes, the classification of training, validation and testing dataset was based on the HA and NA
490 subtype of influenza viruses, which ensured that there were no overlap of influenza virus subtypes
491 in the training, validation and testing dataset and thus can accurately reflect the ability of the
492 model in predicting host for novel flu viruses.
493
494 **Machine-learning algorithms and feature selection for predicting flu virus host, virulence**
495 **and receptor-binding preference**
496 Several traditional machine-learning algorithms including Logistic Regression (LR), Support
497 Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes (NB), Gradient Boosting
498 Decsion Tree (GBDT), Random Forest (RF), and XGBoost were used for predicting flu virus
499 phenotypes. They were achieved using the scikit-learn (version 3.5.1)[37] package in Python
500 (version 3.7.12). Specifically, the SVM was achieved using the function of SVC; RF was achieved
501 using RandomForestClassifier; LR was implemented with LogisticRegression; KNN was
502 implemented using KNeighborsClassifier; NB was achieved using GaussianNB; GBDT was
503 implemented with GradientBoostingClassifier, and XGBoost was achieved using XGBClassifier.
504 In selection of optimal machine learning algorithms in predicting the virus host, mammalian
505 virulence or receptor-binding preference, the default parameters were used.
506 In predicting flu virus host and receptor-binding preference, the optimal ML algorithm was GBDT
507 and SVM respectively. The method of forward feature selection based on permutation importance
508 was used to select the best feature set. Specifically, features were firstly sorted in descending order
509 based on the importance scores that were calculated using permutation_importance function on the
510 training set and then they were sequentially added in modeling. For each iteration, the
511 performance of the current feature subset was evaluated using F1-score on the validation set. The

512 feature subset that yielded the maximum F1-score on the validation set was identified as the
513 optimal selection. In predicting flu virus mammalian virulence, the RF was selected as the optimal
514 algorithm. Thus, the RF-related feature selection method VSURF (Variable Selection Using
515 Random Forests)[38] was used to select the best feature set.
516
517 **The BLAST method for inferring flu virus host**
518 To infer the host for a given flu virus strain in the testing dataset, firstly, a library of influenza
519 virus protein sequences was built using the influenza viruses in the training and validation dataset;
520 then, BLASTp (version 2.13.0+)[39] was used to search the library using the protein sequences of
521 the given flu virus. The final host prediction was determined by a voting method based on the
522 prediction results of individual proteins.
523
524 **The baseline method for predicting the influenza virus host, mammalian virulence and**
525 **receptor-binding preference**
526 The baseline method predicted a given flu phenotype based on only the total number of molecular
527 markers associated with the phenotype. The threshold was determined using the training and
528 validation datasets.
529
530 **Build the FluRisk online web-server**
531 Tornado, a Python web framework and asynchronous networking library, was used to implement
532 web application framework of FluRisk. Celery, a simple, flexible, and reliable distributed system
533 to process vast amounts of messages, was used to support asynchronous notification and the
534 computing task scheduling and execution. The JavaScript Object Notation (JSON) format was
535 employed to implement the communication between the client and server. The JavaScript library
536 JQuery and the web framework bootstrap were employed for producing dynamic and interactive
537 data visualization in the web interface. In addition, the DataTables library was applied to construct
538 tables in web pages.
539
540 **Data availability**
541 All data used in the study, including the atlas of molecular markers, genomic sequences for FHP

542 and reference strains, mammalian virulence and receptor-binding ability of viral strains used in
543 FVP and FRP, were public available at web server of FluRisk.
544
545 **Acknowledgements**
546 We thank members in PengLab for helpful discussions on the manuscript. This work was
547 supported by National Key Plan for Scientific Research and Development of China
548 (2022YFC2303802), the R&D Program of Guangzhou National Laboratory (GZNL2024A01002),
549 National Natural Science Foundation of China (32170651 & 32370700), and Hunan Provincial
550 Natural Science Foundation of China (2024JJ2015).
551
552
553 **Competing interests**
554 The authors declare that they have no competing interests.
555
556 **References**
557 1. Krammer, F. *et al.* Influenza. *Nat Rev Dis Primers* **4**, 3 (2018).
558 2. Short, K. R. *et al.* One health, multiple challenges: The inter-species transmission
559 of influenza A virus. *One Health* **1**, 1–13 (2015).
560 3. Subbarao, K. Avian influenza H7N9 viruses: a rare second warning. *Cell Res* **28**,
561 1–2 (2018).
562 4. Yang, L. *et al.* Genesis and Spread of Newly Emerged Highly Pathogenic H7N9
563 Avian Viruses in Mainland China. *Journal of Virology* **91**, e01277-17 (2017).
564 5. Ison, M. G. & Marrazzo, J. The Emerging Threat of H5N1 to Human Health. *N*
565 *Engl J Med* **392**, 916–918 (2025).
566 6. Peacock, T. P. *et al.* The global H5N1 influenza panzootic in mammals. *Nature*
567 **637**, 304–313 (2025).
568 7. Puryear, W. B. & Runstadler, J. A. High-pathogenicity avian influenza in wildlife:
569 a changing disease dynamic that is expanding in wild birds and having an
570 increasing impact on a growing number of mammals. *javma* **262**, 601–609 (2024).
571 8. Sah, R. *et al.* Concerns on H5N1 avian influenza given the outbreak in U.S. dairy

572 cattle. *The Lancet Regional Health - Americas* **35**, 100785 (2024).
573 9. Nidom, C. A. *et al.* Influenza A (H5N1) Viruses from Pigs, Indonesia. *Emerg*
574 *Infect Dis* **16**, 1515–1523 (2010).
575 10. Domańska-Blicharz, K. *et al.* Outbreak of highly pathogenic avian influenza
576 A(H5N1) clade 2.3.4.4b virus in cats, Poland, June to July 2023. *Euro Surveill* **28**,
577 2300366 (2023).
578 11. WHO. Tool for Influenza Pandemic Risk Assessment (TIPRA).
579 https://www.who.int/teams/global-influenza-programme/avian-influenza/tool-for-i
580 nfluenza-pandemic-risk-assessment-(tipra) (2016).
581 12. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data –
582 from vision to reality. *Euro Surveill* **22**, 30494 (2017).
583 13. Peng, Y. *et al.* Automated recommendation of the seasonal influenza vaccine
584 strain with PREDAC. *Biosafety and Health* **02**, 117–119 (2020).
585 14. Yin, R. *et al.* ViPal: A framework for virulence prediction of influenza viruses
586 with prior viral knowledge using genomic sequences. *Journal of Biomedical*
587 *Informatics* **142**, 104388 (2023).
588 15. Mock, F., Viehweger, A., Barth, E. & Marz, M. VIDHOP, viral host prediction
589 with deep learning. *Bioinformatics* **37**, 318–325 (2021).
590 16. Borkenhagen, L. K., Allen, M. W. & Runstadler, J. A. Influenza virus genotype
591 to phenotype predictions through machine learning: a systematic review. *Emerg*
592 *Microbes Infect* **10**, 1896–1907 (2021).
593 17. Mok, C. K. P. *et al.* Amino Acid Substitutions in Polymerase Basic Protein 2
594 Gene Contribute to the Pathogenicity of the Novel A/H7N9 Influenza Virus in
595 Mammalian Hosts. *Journal of Virology* **88**, 3568–3576 (2014).
596 18. Zhang, H. *et al.* The PB2 E627K mutation contributes to the high polymerase
597 activity and enhanced replication of H7N9 influenza virus. *Journal of General*
598 *Virology* **95**, 779–786 (2014).
599 19. Xu, G. *et al.* Prevailing PA Mutation K356R in Avian Influenza H9N2 Virus
600 Increases Mammalian Replication and Pathogenicity. *Journal of Virology* **90**,
601 8105–8114 (2016).

602 20. Zhang, X. *et al.* PB1 S524G mutation of wild bird-origin H3N8 influenza A virus
603 enhances virulence and fitness for transmission in mammals. *Emerging Microbes*
604 *& Infections* **10**, 1038–1051 (2021).
605 21. Lin, T.-H. *et al.* A single mutation in bovine influenza H5N1 hemagglutinin
606 switches specificity to human receptors. *Science* **386**, 1128–1134 (2024).
607 22. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive Secondary Mutations
608 Enable the Evolution of Influenza Oseltamivir Resistance. *Science* **328**, 1272–1275
609 (2010).
610 23. Zhou, L. *et al.* A single N342D substitution in Influenza B Virus NA protein
611 determines viral pathogenicity in mice. *Emerging Microbes & Infections* (2020).
612 24. Lu, C. *et al.* FluPhenotype—a one-stop platform for early warnings of the
613 influenza A virus. *Bioinformatics* **36**, 3251–3253 (2020).
614 25. Yin, R., Luo, Z., Zhuang, P., Lin, Z. & Kwoh, C. K. VirPreNet: a weighted
615 ensemble convolutional neural network for the virulence prediction of influenza A
616 virus using all eight segments. *Bioinformatics* **37**, 737–743 (2021).
617 26. Schrauwen, E. J. & Fouchier, R. A. Host adaptation and transmission of influenza
618 A viruses in mammals. *Emerg Microbes Infect* **3**, 1–10 (2014).
619 27. Long, J. S., Mistry, B., Haslam, S. M. & Barclay, W. S. Host and viral
620 determinants of influenza A virus species specificity. *Nat Rev Microbiol* **17**, 67–81
621 (2019).
622 28. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral
623 evolution and escape. *Science* **371**, 284–288 (2021).
624 29. Wang, G. *et al.* Deep-learning-enabled protein–protein interaction analysis for
625 prediction of SARS-CoV-2 infectivity and variant evolution. *Nat Med* **29**,
626 2007–2018 (2023).
627 30. Nie, Z. *et al.* A unified evolution-driven deep learning framework for virus
628 variation driver prediction. *Nat Mach Intell* **7**, 131–144 (2025).
629 31. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the
630 next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
631 32. Peng, Y. *et al.* Identification of genome-wide nucleotide sites associated with

632 mammalian virulence in influenza A viruses. *Biosafety and Health* **2**, 32–38
633 (2020).
634 33. Raman, R. *et al.* Advancing glycomics: Implementation strategies at the
635 Consortium for Functional Glycomics. *Glycobiology* **16**, 82R-90R (2006).
636 34. Suenaga, E., Mizuno, H. & Kumar, P. K. R. Influenza virus surveillance using
637 surface plasmon resonance. *Virulence* **3**, 464–470 (2012).
638 35. Verzijl, D., Riedl, T., Parren, P. W. H. I. & Gerritsen, A. F. A novel label-free
639 cell-based assay technology using biolayer interferometry. *Biosensors and*
640 *Bioelectronics* **87**, 388–395 (2017).
641 36. Aydin, S. A short history, principles, and types of ELISA, and our laboratory
642 experience with peptide/protein analyses using ELISA. *Peptides* **72**, 4–15 (2015).
643 37. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn.*
644 *Res.* **12**, 2825–2830 (2011).
645 38. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. VSURF: An R Package for Variable
646 Selection Using Random Forests. *The R Journal* **7**, 19–33 (2015).
647 39. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*
648 **10**, 1–9 (2009).
649