Improving influenza A vaccine strain selection through deep evolutionary models

Wenxian Shi[1*], Jeremy Wohlwend[1], Menghua Wu[1], Regina Barzilay[1]

[1*]Department of Computer Science, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, 02139, MA, USA.

*Corresponding author(s). E-mail(s): wxsh@mit.edu;
Contributing authors: jwohlwend@csail.mit.edu; rmwu@mit.edu; regina@csail.mit.edu;

Abstract

Even though vaccines have the potential to significantly alleviate the disease burden of epidemics such as the seasonal flu, current influenza vaccines offer limited protection. According to the Centers for Disease Control and Prevention (CDC), vaccine effectiveness has hovered below 50% for the past decade. Identifying the optimal strains to use in a vaccine is central to increasing its efficacy. However, this task is challenging due to the antigenic drift that occurs during the flu season. In this paper, we propose to select vaccines based on their escapability score, a metric that quantifies the antigenic similarity of vaccine strains with future dominant strains and demonstrates a strong correlation with clinical vaccine effectiveness. We introduce a deep learning-based approach that predicts both the antigenic properties of vaccine strains and the dominance of future circulating viruses, enabling efficient virtual screening of a large number of vaccine compositions. We utilized historical antigenic analysis data from the World Health Organization (WHO) to demonstrate that our model selects vaccine strains that reliably improve over the recommended ones.

Keywords: vaccine selection, influenza vaccine, virus evolution, deep learning

1 Introduction

Overview

The genetic composition of an influenza vaccine is one of the critical drivers of its effectiveness. Every year, vaccine strains are selected in advance of the upcoming flu season by a panel of experts from the World Health Organization. Their goal is to identify virus strains that will induce antibodies effective against future prevalent circulating strains. The effectiveness of an inactivated influenza vaccine can typically reach 40% to 60% when the vaccine strains are antigenically well-matched with the circulating viruses. However, despite decades of research in prevention and surveillance, current influenza vaccines provide limited protection: the Centers for Disease Control and Prevention (CDC) report that vaccine effectiveness in the United States has dropped below 40% in five of the past ten years [1–5]. In fact, during the 2014-2015 winter season, vaccine effectiveness was only 19% [1].

Two aspects limit the effectiveness of vaccines in the current selection process: the emergence of new mutations during the upcoming season, and the experimental burden of producing and validating vaccine strains. First, vaccine strains are selected in February and September for the Northern and Southern Hemisphere seasons respectively, given the lead time of six to nine months for egg-based production [6]. During these months between the design and deployment of the vaccine, the distribution of existing influenza strains may change significantly [7–9], as illustrated in Fig. S3, leading to antigenic drift. In order to estimate the effectiveness of vaccines in the next season, it is therefore crucial to model the distribution of variants that are expected to occur. Second, the need to empirically validate each candidate vaccine's effectiveness against all circulating viruses results in a significant volume of combinatorial experiments. The burden of vaccine production and the availability of virus specimens for study also constrain the number of potential vaccines and viral strains that can be examined. Therefore, conventional experimental methods, such as the hemagglutination inhibition (HI) test [10], are difficult to scale to a large number of candidate vaccines. In practice, only a limited number of vaccine candidates, typically fewer than ten, may be considered [11].

To address these limitations, we develop deep-learning models capable of estimating the distribution of future strains and predicting the inhibition capacity of vaccines against these strains *in silico*. More specifically, we first present a dominance predictor that forecasts the distribution of variants in the upcoming season. Second, we introduce an antigenicity predictor that accurately predicts antigenic similarity between vaccine and virus strains, exemplified as the hemagglutination inhibition (HI) test results, and enables the screening of a massive number of candidate vaccines virtually. We leverage both of these predictors to produce an "escapability score," which is used to rank and select the optimal vaccine strains. There exist computational methods that model viral evolution [12–16] or antigenicity [17–20] as independent factors. However, to the best of our knowledge, we are the first to connect these aspects into a unified computational pipeline for vaccine selection.

We use historical epidemiology data and antigenic analysis data from WHO to retrospectively evaluate our model. Our results show that our model selects vaccine

strains with better antigenic properties with respect to future dominant strains, based on a subset of information available to human experts. This approach can also be applied to any other circulating pathogen with sufficient antigenic and evolutionary data.

Results

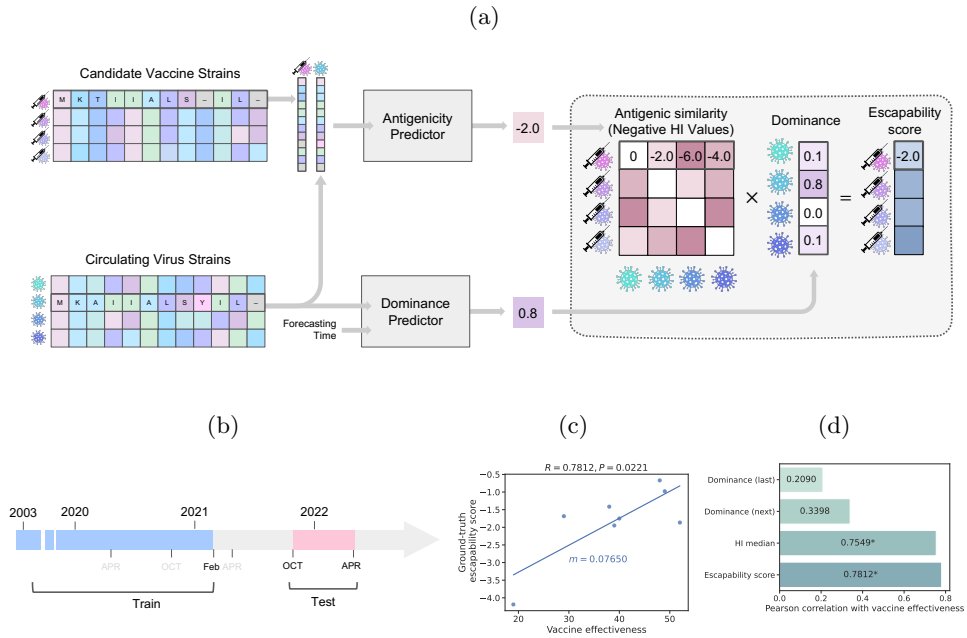

Fig. 1: **Deep learning models to calculate the escapability scores for vaccine selection.** (**a**) Model overview, and definition of escapability score. (**b**) Training and testing splits, based on collection time. (**c**) The ground-truth escapability score exhibits a strong correlation with the U.S. flu vaccine effectiveness from the CDC (Pearson correlation $R = 0.7812$, p-value $P = 0.0221$, linear regression slope $m = 0.07650$). (**d**) The Pearson correlation between vaccine effectiveness and various scoring strategies for the vaccine strains. The star mark (*) indicates $P < 0.05$. Our escapability score shows the strongest correlation.

Overview of the algorithm

Clinically determining vaccine effectiveness in a population is expensive and inefficient. To evaluate the effectiveness of the vaccine efficiently and accurately, we propose the **escapability score**, a computational metric that quantifies the antigenicity of vaccine

strains with respect to future dominant viruses. Specifically, as illustrated in Fig. 1a, we define the escapability score ES of a vaccine v at season $t+1$ as the weighted average of the negative normalized HI value h between the vaccine strain v and circulating virus strains x, where circulating strains are weighted by predicted dominance in the next season $p_{t+1}(x)$:

$$ES_{t+1}(v) = \sum_x p_{t+1}(x)\left(-h(v,x)\right). \tag{1}$$

Intuitively, the escapability score measures the overall antigenic similarity of vaccine strains with respect to future virus strains. The negative normalized HI value quantifies the antigenic similarity between a vaccine strain and a potentially circulating virus, whereas the dominance indicates the probability of that virus strain circulating in the upcoming season. The escapability score is obtained by taking the expectation over all possible circulating strains. A higher escapability score indicates that the antibodies induced by the vaccine are expected to be more effective against the dominant strains in the next season.

The escapability score is well correlated with historical vaccine effectiveness and associated clinical outcomes. We compare the ground-truth escapability scores with the United States flu vaccine effectiveness from the CDC[1] for the recommended vaccines during the North Hemisphere flu seasons from 2012 to 2020. For the ground-truth scores, the HI values are obtained from experimental HI tests, and dominance values adopt the empirical frequencies calculated from genomics epidemiology data[2]. As shown in Fig. 1c, there is a significant correlation between the escapability score and vaccine effectiveness, with a Pearson correlation coefficient of 0.7812 ($P = 0.02$). Fitting a linear regression model results in a slope of $m = 0.07650$, which suggests that a 0.765 increase in escapability score corresponds to a 10% improvement in vaccine effectiveness. We also observed positive correlations between the escapability score and estimated reduction of disease burden [21, 22] (Fig. S8). The Pearson correlation coefficients between the escapability score and the estimated number of averted symptomatic illnesses, medical visits, and hospitalizations are 0.73, 0.76, and 0.64 respectively. Thus, optimizing the escapability score would enable us to improve vaccine effectiveness and reduce the disease burden caused by influenza.

We further illustrate the importance of considering both virus dominance and antigenic properties in the escapability score in Fig 1d. Here, the dominance of vaccine strains correlates poorly with vaccine effectiveness, even if we consider the dominance in the future season. Thus, selecting the vaccine strains merely based on their dominance is insufficient. In contrast, antigenic properties have a substantial impact on the assessment of vaccine effectiveness: the median of experimental HI values over the set of expert-selected testing viruses shows a positive correlation with the vaccine effectiveness. Finally, our escapability score, which considers both the antigenic properties and the dominance of circulating viruses in the next season, exhibits the highest correlation with vaccine effectiveness.

Now that we have motivated the escapability score as a valid objective for vaccine selection, we propose an approach to compute the escapability score based on deep

[1] https://www.cdc.gov/flu/vaccines-work/2021-2022.html
[2] https://gisaid.org/

learning models, as summarized in Figure 1a. The inputs to the models are the hemagglutinin (HA) protein sequences from both candidate vaccine strains and circulating virus strains. Given a vaccine-virus pair, an *antigenicity predictor* estimates their HI test result (antigenic similarity), and a *dominance predictor* estimates the likelihood that the virus will circulate in the upcoming flu season (dominance). These two scores are summarized into the escapability score for each candidate vaccine. Candidate vaccine strains are ranked by the escapability score, and the highest-scoring (best) strains will be recommended by our algorithm for further validation.

The antigenicity predictor is a regression model based on the MSA-Transformer[23] architecture, which takes as input the aligned HA sequences from vaccines and virus strains, and predicts their antigenic similarity, i.e., the negative normalized HI value. We train the antigenicity predictor on hemagglutination inhibition titer (HI test) results from published reports[3] prepared for the WHO annual consultation on the composition of influenza vaccines since 2003. The data processing pipeline and the model architecture are described in detail in Methods.

The dominance predictor is a dynamic protein language model that outputs the likelihood a given HA protein sequence will be prevalent during any given season. Based on a GPT-2 [24] architecture, our model considers epistasis constraints during evolution [25], and it is tailored to capture the distributional changes of HA sequence mutations. We train the dominance predictor based on decades of influenza surveillance data with protein sequences and collection times from GISAID [26]. The formulation of our dominance model and description of data processing are provided in Methods.

Retrospective evaluation for escapability score prediction

We evaluate the accuracy of escapability score prediction retrospectively for influenza A/H1N1 and A/H3N2 data collected from the 2015 to 2021 winter seasons. As illustrated in Fig. 1b, the training data for dominance predictor and antigenicity predictor includes only strains collected at least eight months before the start of the testing winter season, matching the time frame of the WHO's recommendation schedule. For example, suppose we would like to evaluate the vaccines in the 2021-2022 winter season (from October 2021 to March 2022). The antigenicity predictor is trained on available HI test experimental data before February 2021, and the dominance predictor is trained on HA protein sequences collected before February 2021.

During inference, the candidate vaccine set is constructed by including the HA sequences circulating in the past three years, for example between February 2018 and February 2021, with greater than five occurrences. The testing virus set is the same as the one considered for the candidate vaccine set. We convert the dominance of these testing virus sequences into probabilities and compute the escapability scores as the expected negative HI predicted values based on these probabilities.

We evaluate the accuracy of predicting the escapability score of our model by comparing it with the ground-truth escapability scores. In the computation of the ground-truth escapability score, we utilize the actual frequencies of circulating viruses

[3] https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports

during the testing season as the dominance, and incorporate the experimental HI values available in the WHO reports. In practice, the ground-truth escapability scores are computed not only across circulating virus sequences (sequence-level) but also across circulating virus clades[4] (clade-level). The definition of clade-level escapability score is analogous to the equation described in Eq. 1, with the distinction that the variable x represents the virus clade instead of the sequence. The dominance of each clade is the summation of the dominance of each sequence belonging to that clade. The HI value for each clade is the average of the HI value of each sequence belonging to that clade.

To make sure that the ground-truth escapability scores are accurate, we only consider vaccines with sufficient experimental HI values in our evaluation. Precisely, vaccines are included if their experimental coverage, the proportion of circulating viruses covered by the experimental HI data related to that vaccine, exceeds 80%. A total of 54 candidate vaccines from A/H3N2 and 51 candidate vaccines from A/H1N1 are subjected to comparison, whose predicted and ground-truth escapability scores are shown in Fig. 2a-2b (clade-level) and Fig. S4 (sequence-level). There is a strong correlation between our predicted scores and the ground-truth scores for all tested vaccines, demonstrating the capacity of our model to accurately predict escapability scores.

Evaluation of dominance predictor

Compared to three state-of-the-art baselines, our dominance predictor achieves the highest correlation with ground truth escapability score, in conjunction with our antigenicity predictor (Fig. 2c and Fig. 2d). We postulate that this is due to its consideration of antigenic drift, evolving fitness, and epistatic interactions. Furthermore, as shown in Fig. S5 and Fig. S6, the combination of our antigenicity predictor and our dominance predictor achieves the best performance among all other combinations of antigenicity and dominance prediction methods.

To quantify the impact of antigenic drift on vaccine selection, we first compare with a baseline that defines dominance as the empirical frequencies calculated from the previous winter season as the dominance (**Last**). This baseline assumes that the distribution of variants does not change between seasons, and its inferior performance implies that it is important to consider the antigenic drift between seasons.

We further conduct a comparison with three protein language models, **LM** [27], **CSCS** [12] and **EVEscape** [14], which have shown great potential in predicting virus fitness based on their protein sequences [12–14]. LM uses the same GPT-2 architecture as our dominance model but is trained as a vanilla auto-regressive language model without incorporating time-related information. CSCS is a masked language model built on a bi-directional LSTM [28], which adopts the mutation probability and the embedding distance from wild-type as the protein fitness. EVEscape builds a language model based on variational auto-encoders [29], and defines the protein fitness by a likelihood lower-bound and extra structural and biological information. All these protein language models are trained using the same sequences as our model, but they exhibit lower performance in dominance prediction, because they are not designed to

[4]The clades are defined by Nextclade(https://clades.nextstrain.org/)

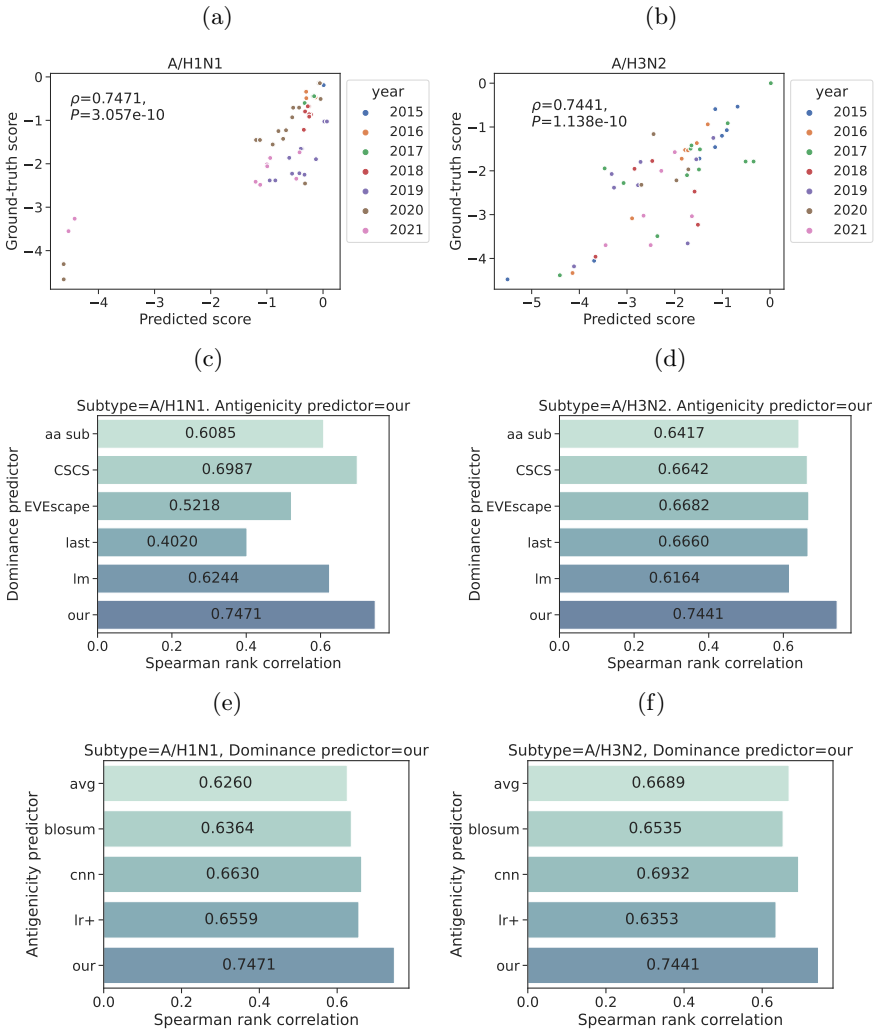

Fig. 2: **Performance of escapability score prediction.** (**a, b**) Spearman rank correlation ρ between the predicted escapability score and ground-truth score (clade-level) for A/H1N1 and A/H3N2, from 2015 to 2021 seasons. $P < 0.005$. (**c, d**) Spearman rank correlation between the ground-truth escapability score and predicted escapability score based on different dominance models for A/H1N1 and A/H3N2 ($P < 0.005$). Given our antigenicity predictor, our dominance model achieves the best performance. (**e, f**) Spearman rank correlation between the ground-truth escapability score and predicted escapability score based on different antigenicity predictor for A/H1N1 and A/H3N2 ($P < 0.005$). Given our dominance predictor, our antigenicity predictor achieves the best performance.

model the evolving dynamics of the fitness landscape. In contrast, our model's ability to capture changes in the prevalence of variants is beneficial for selecting better vaccine strains.

To emphasize the importance of modeling epistasis, our model is compared with an **AA sub** model, which models the frequency evolution of single amino acid mutations. Our model performs significantly better by incorporating epistatic interactions within HA protein sequences. A detailed description of the baseline models can be found in Methods.

Evaluation of antigenicity predictor

We compare our antigenicity predictor with various baseline methods. As illustrated in Fig. 2e and Fig. 2f, our antigenicity predictor outperforms all other approaches when combined with our dominance predictor to calculate escapability scores.

Given the sparsity of experimental results, we first show the importance of inferring absent HI values for untested vaccine-virus combinations for scoring vaccines. We compare with a baseline, annotated as **AVG**, using experimental data collected at least eight months prior to the forecasted influenza season. For vaccine-virus pairs that are untested, the absent HI value is filled by the average of all experimental HI values. A sequence similarity-based imputation method is also compared, annotated as **blosum**. This method imputes the HI values for untested vaccine-virus pairs, by the results from the most similar vaccine-virus pairs that have been tested. The protein similarity is determined by the BLOSUM62 [30] matrix. The performance decreases for both baselines indicate the experimental results could not cover a sufficient number of circulating viruses for accurate escapability score estimation.

We also compare our antigenicity predictor with previous research applying traditional machine learning methods [17, 18] and other deep learning models [19, 20] to this problem. We compare our model with two representative models. One is the linear regression model, denoted as **LR+**[18]. It is constructed using features such as amino acid substitutions between vaccine and virus sequences, the avidity of the virus, and the potency of antiserum. The other one utilizes convolutional neural networks [31] (**CNN**) built on aligned amino acid HA sequences of viruses and vaccines. Our model outperformed both the LR+ and CNN models by effectively capturing the correlations between amino acids within and between vaccines and virus strains.

Our antigenicity predictor model was able to reconstruct the escaping mutations identified in the mutational antigenic profiling assay [32] concerning specific antibodies, even without integrating the knowledge of antibodies into our training procedure. The results are discussed in Supplementary Results.

Optimal vaccine selection based on escapability scores

Our model is able to select better vaccine strains than the WHO's past recommendations, as measured by the ground-truth escapability score. As shown in Fig. 3a and Fig. 3b, the strains our model selected performed better than the strains selected by WHO in five of seven years for A/H1N1 and in six of seven years for A/H3N2. Furthermore, in four years for both A/H1N1 and A/H3N2, our model recommends the strain

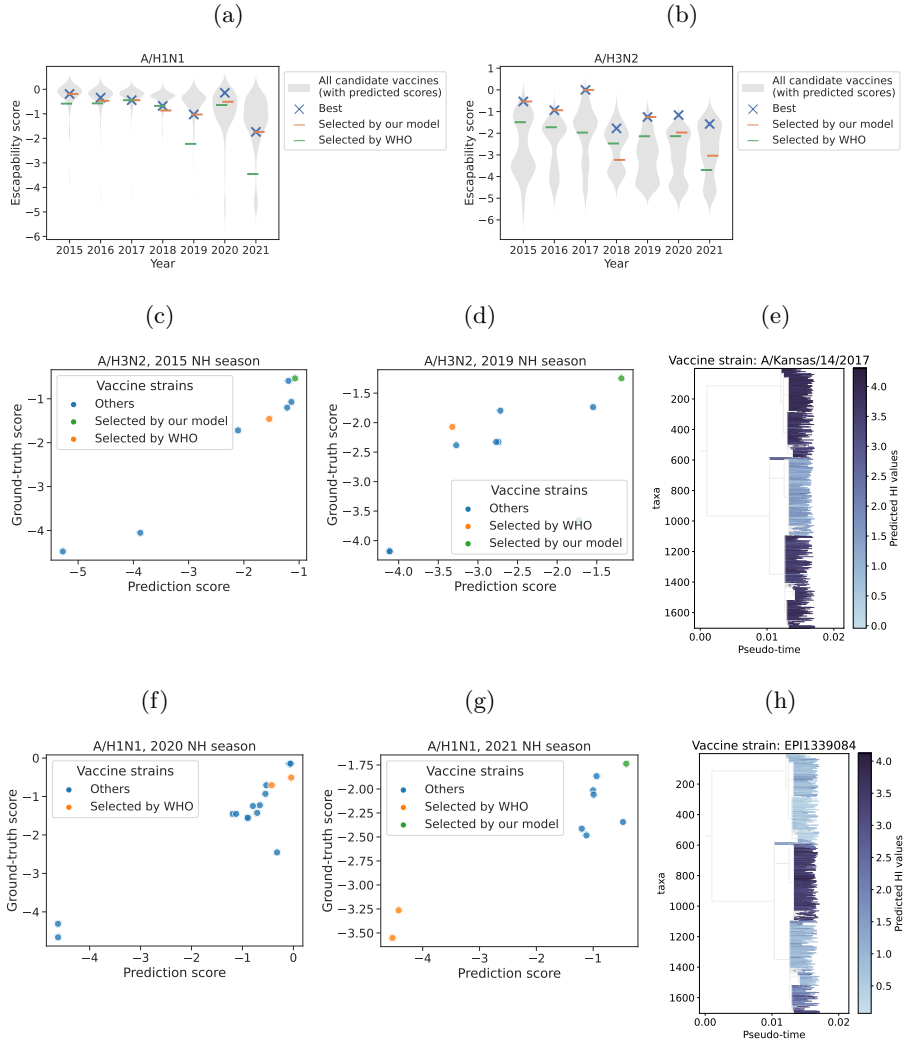

Fig. 3: **Escapability score of strains selected by our model, WHO, and the strain with the highest (best) score.** (**a, b**) Ground-truth escapability score of strains selected by our model, WHO, and the strain with the highest (best) score. Violin plots represent all candidate vaccines with their escapability score predicted by our model (including those that have not been assessed experimentally). (**c, d, f, g**) Escapability score of candidate vaccines given by our model (x-axis) and the ground truth (y-axis). The orange dots are the recommended strains by WHO. The green dots are the recommended strains by our model. (**e, h**) The predicted HI values of circulating A/H3N2 strains in the 2019 NH season, with respect to the vaccine (**e**) recommended by WHO (A/Kansas/14/2017) and (**h**) our model (EPI339084).

with the highest ground-truth escapability score among all candidate vaccine strains tested by WHO, denoted "best". It is important to note that the strains labeled as "selected by our model" in Fig. 3 are the vaccine strains with the highest predicted scores among those whose antigenic properties were tested for over 80% circulating viruses by experts from the WHO. The distribution of predicted escapability scores over all candidate vaccine strains (violin plots in Fig. 3a and Fig. 3b) includes strains with even higher predicted escapability scores, that were not selected to be tested by experts. This highlights the possibility that there exist even more efficacious vaccine strains to be discovered.

Fig. 3c-3d and Fig. 3f-3g present four specific case studies: the 2015 and 2019 seasons for A/H3N2 and the 2020 and 2021 seasons for A/H1N1. Each dot represents a candidate vaccine with available experimental HI results that cover more than 80% circulating viruses. The orange dots are the recommended strains from WHO, while the green ones are the recommended strains from our model. These figures illustrate the strong correlation between the scores predicted by our model (x-axis) and the ground-truth scores (y-axis). The optimal strains with the highest ground-truth escapability scores are also recommended by our model according to the predicted escapability scores. In particular, in Fig. 3d, though the recommended strain by the WHO, A/Kansas/14/2017, has a relatively high escapability score, an alternative strain collected earliest on 2018-12-24, A/Hong Kong/45/2019, exhibits an even better escapability score. Intriguingly, A/Hong Kong/45/2019 is the recommended strain for the subsequent winter season.

To visualize the effective breadth of different vaccine strains, we present the predicted HI values of A/H3N2 strains circulating in the 2019 Northern Hemisphere (NH) season, comparing the vaccines recommended by WHO (Fig.3e) with our model's recommendation (Fig.3h). The strain recommended by the WHO only aligns with one circulating clade. In contrast, our model's recommended strain offers a complementary antigenic characterization, and effectively covers a majority of the circulating clades, resulting in a much stronger escapability score.

Predicted escapability score correlated with vaccine effectiveness and disease burden prevention

We first demonstrate that there is a significant correlation between our predicted escapability score and vaccine effectiveness, as shown in Fig. 4a with a Pearson correlation of 0.76 ($P = 0.028$). We aggregated the escapability scores for A/H3N2 and A/H1N1 according to the ratio of samples collected during the respective flu seasons. In Fig. 4a, the line represents a linear regression model fit of the data and m is the slope of the fitting line. Roughly speaking, enhancing the predicted escapability scores of recommended vaccines by approximately 0.47 is projected to result in a 10% improvement in vaccine effectiveness. As illustrated in Fig. S7, the predicted escapability score also shows a stronger correlation than other scoring metrics considering only HI values or dominance.

We further analyze the association between predicted escapability score with the hospitalization rates and the prevention of disease burden. The hospitalization rates (Fig. 4b) are obtained from the Influenza Hospitalization Surveillance Network

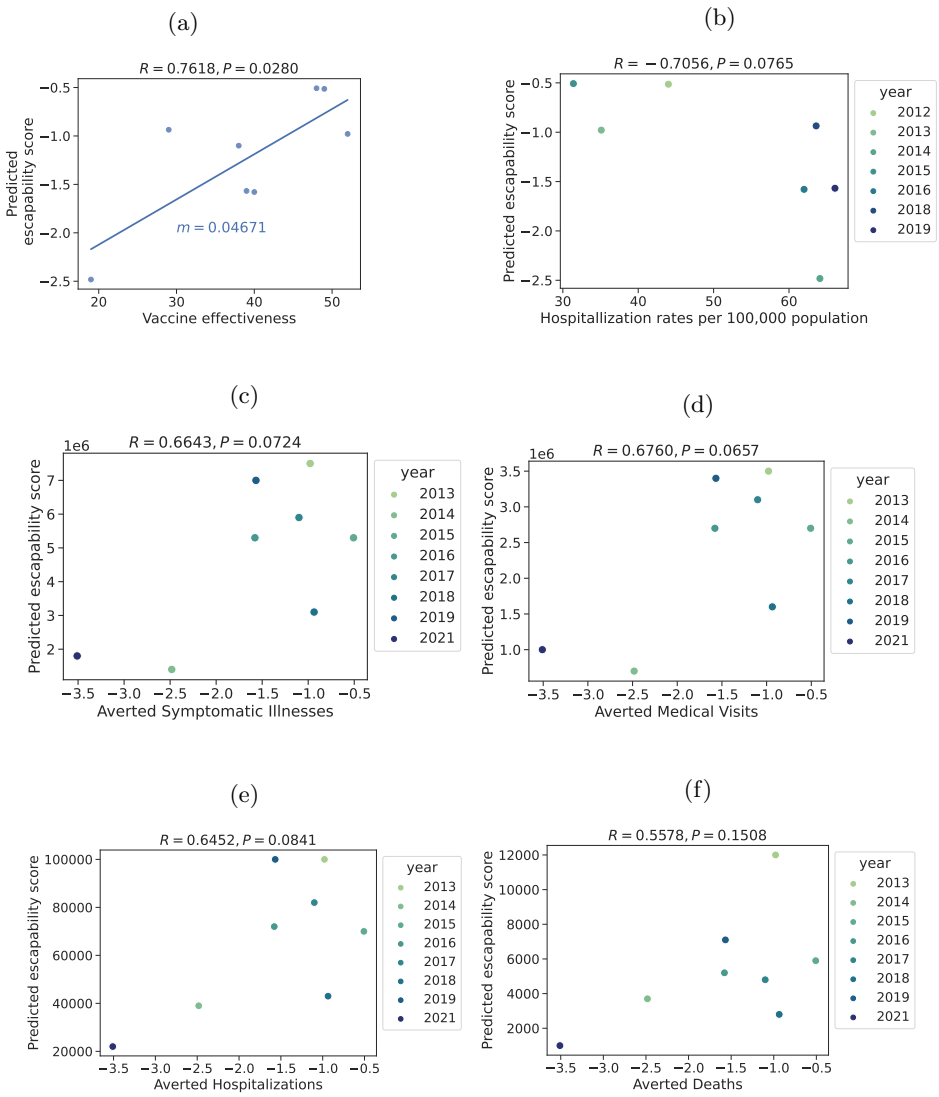

Fig. 4: **The predicted escapability score is correlated with vaccine effectiveness and disease burden prevention.** (**a**) Predicted escapability score and effectiveness of vaccines are highly correlated. A linear regression fitting is depicted with the annotated slope m. (**b**) Predicted escapability score and hospitalization rates show a negative correlation. (**c-f**) Correlation between the predicted escapability score and the estimated number of influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination. Pearson correlations (R) and corresponding P-values (P) are illustrated in the text.

(FluSurv-NET)[5]. We discard the 2017 flu season with extremely high hospitalization rates and complicated hospitalization [33]. The 2020 and 2021 flu seasons are also removed with extremely low hospitalization rates because of the SARS-CoV-2 pandemic. The prevention of disease burden is estimated in four aspects from CDC [21, 22]: the number of influenza illnesses (Fig. 4c), medical visits (Fig. 4d), hospitalizations (Fig. 4e), and deaths (Fig. 4f) averted by vaccination[6].

As expected, our predicted escapability scores show a negative correlation with the hospitalization rates, as shown in Fig. 4b. Setting aside the influence of other factors on hospitalization rates, enhancing escapability scores exhibits substantial potential for reducing hospitalization rates. As shown in Fig. 4c-4f, our predicted escapability scores also show positive alignment with the number of averted symptomatic illnesses, medical visits and hospitalization, with a Pearson correlation of 0.66, 0.68, and 0.65 respectively. Adopting vaccine strains with a higher escapability score could potentially reduce the symptomatic illnesses, medical visits, and hospitalization associated with influenza. Even though the escapability scores show a weaker correlation with the reduction in deaths ($R = 0.56$), they still maintain a positive correlation. Positive correlations with the reduction of disease burden are also observed for vaccine effectiveness (Fig. S9).

Discussion

Vaccines remain one of our most important defenses against infectious diseases. However, against continually-evolving viruses, it is challenging both to forecast the future landscape of viral strains, and to assess the antigenicity of candidate vaccines at scale. In this work, we proposed a deep learning model that can select optimal vaccine strains in advance of the annual flu season by scoring vast numbers of candidate vaccine strains *in silico*. Our model is able to accurately predict both the inhibition ability of specific vaccines to testing viruses and the dominance of virus strains in the upcoming flu season, based on the HA sequences. By jointly considering these two aspects, our model ranks vaccine strains based on how well they inhibit virus strains expected to be dominant in the upcoming season. Retrospective experiments have shown our escapability score is well correlated with vaccine effectiveness, and our recommended vaccine strains are superior to those recommended by the WHO in the majority of years evaluated.

Though our experiments were conducted on influenza surveillance data and hemagglutination inhibition assays, our approach is applicable to any neutralization assays, vaccine types, and viruses, as long as there exists data relating the genotypes of viruses and vaccines to their neutralization behavior. For influenza, our model can utilize other assays such as HINT [34], or other vaccine types, like mRNA vaccines. After all, since neutralization is directly inferred from the vaccine and virus genotypes, our model does not require knowledge of specific antibodies involved or neutralization mechanisms. In fact, our model can also be directly applied to other pandemic viruses such as

[5] https://gis.cdc.gov/GRASP/Fluview/FluHospRates.html
[6] https://www.cdc.gov/flu/vaccines-work/burden-averted.htm

SARS-CoV-2. For example, the dominance prediction model can be trained with protein sequence data (e.g. Spike protein) with annotated collection times. Likewise, the antigenicity prediction model can be trained on neutralization assays such as ELISA.

Compared to existing, wet-lab methods for validating vaccine candidates, our model is able to explore a much larger vaccine space in a more effective way. For example, our model can score 100 vaccines against $\sim$500 testing viruses within thirty minutes on a single A100 GPU. Furthermore, the antigenicity prediction model relies solely on the HA protein sequences, eliminating the need for cumbersome testing of their crystal structures or antibody profiles. This allows us to compute the escapability score for *any* possible vaccine strains, including those that have not yet been isolated, or do not exist in nature. Thus, our model can also be used as a tool for functional optimization and *de-novo* vaccine design.

There are multiple future directions that can further improve this study. First, unseen HA sequences including novel mutations or novel combinations of existing mutations count over 40% sequences in the future season. The absence of these novel sequences in our circulating virus set might bias the predicted escapability score. We could sample virus sequences from the future distribution given by the dominance model, and calculate the inhibition ability of vaccines over the generated viruses set. Second, proteins such as neuraminidase (NA) and even the non-coding segment of the virus genome might also have an impact on both virus fitness and antigenicity, making them potential components for inclusion in our model.

This study has several limitations. First, it focuses solely on the antigenicity of vaccines. Here, we do not consider other aspects that influence the effectiveness or practicality of vaccines, such as host genetic background, vaccine platform, egg-adaptations in the production process, and the subject's history of previous vaccines [35]. Furthermore, only viruses collected from human hosts are used in our training process. The potential impact of viruses in animal hosts on the vaccine efficacy, and the effectiveness of the vaccines to animals remains unexplored. As a deep learning model, our approach might not perform well on viruses that have limited data available, such as newly emerged viruses or less well-studied viruses. Sufficient data is crucial to train a robust dominance prediction model and antigenicity prediction model. Finally, we only evaluate the performance of our algorithm in vaccines with available HI experimental data. Potential vaccine strains with lower escapability scores but without available experimental data are required for further validation.

In conclusion, we proposed a computational solution to accelerate the discovery of more effective vaccines. Our approach is extensible to arbitrary viruses and vaccine types, and our retrospective experiments on influenza have demonstrated significant promise. Though there are many biological and epidemiological elements that could be incorporated in future iterations, this work presents a proof-of-concept for efficient and accurate vaccine selection.

Data availability. The hemagglutination sequences and their metadata including collection time and strain name are from the GISAID (https://gisaid.org/). The hemagglutination inhibition titer results are collected from the reports of WHO CC (https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports).

The human influenza vaccine composition is found from https://gisaid.org/resources/human-influenza-vaccine-composition/. The vaccine effectiveness in U.S. is from the study of CDC https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm. The estimated influenza illnesses, medical visits, and hospitalizations averted by vaccination are found from https://www.cdc.gov/flu/vaccines-work/burden-averted.htm.

Supplementary materials

Methods

Antigenicity predictor

Model architecture

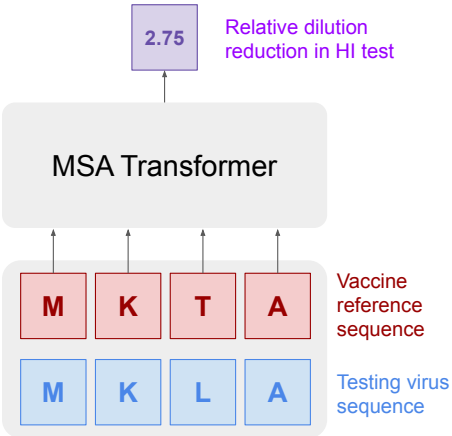

Fig. 1: The architecture of the antigenicity predictor.

Figure 1 depicts the high-level architecture of this model, which is built on top of the MSA-Transformer [23]. The HA protein sequences of the vaccine and testing circulating virus are aligned by the MMSeqs2 [36].

Data processing

The hemagglutination inhibition titer (HI test) data is extracted from published reports[7] prepared for the WHO annual consultation on the composition of influenza vaccines since 2003. The HI test data comprises the names of viruses and vaccine strains, along with the dilution of antibodies leading to hemagglutination inhibition. We retrieve the HA sequences with respect to strain names from GISAID. Strains

[7] https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports

for which HA sequences cannot be found are excluded from the dataset. If there are multiple HA sequences corresponding to one strain name, we will enumerate all the possible HA sequences. For a pair of vaccine-virus sequences, if there are multiple HI test values, we use the geometric mean of them.

Following previous work [18], we quantify antigenic distance by relative reduction of dilution, $h(v, x)$:

$$h(v, x) = \max(\log \mathrm{HI}(v, v) - \log \mathrm{HI}(v, x), 0). \tag{1}$$

where $\mathrm{HI}(v, x)$ is the dilution of antibodies from vaccine v against testing strain x. A higher relative reduction of dilution indicates lower antibody efficacy and lower antigenic similarity between the vaccine and testing strains. We obtained the HI titer results for 58,209 vaccine-virus pairs for A/H1N1 and 69,622 vaccine-virus pairs for A/H3N2.

Baselines

AVG. For the vaccine-virus pairs with available experimental HI results eight months before the winter season starts, we use their experimental results. For the vaccine-virus pairs without available experimental HI results, we use the average HI values of the whole dataset as their HI values.

Blosum. For the vaccine-virus pairs without available experimental HI results, we search for the most similar vaccine-virus pairs that have tested the HI values in the dataset and use the average of them as the estimated HI results. The similarity is calculated from the BLOSUM62 matrix, and the similarity between two virus-vaccine pairs is the summation of the similarity between virus sequences and the similarity between vaccine sequences.

LR+. A linear regression model, whose input features are the amino acid substitution of virus and vaccine protein sequences, as well as their amino acids in each position.

CNN. A Convolution Neural Network consists of two 1D convolution layers with 64 channels and kernel sizes equal to 15 intervening by two one-dimensional max pooling layers. The input is the concatenation of two one-hot representations of HA protein sequences from vaccine and virus.

Dominance predictor

Model definition

Given an amino acid sequence $x \in V^L$ with length L (where V is the set of all possible amino acids, typically $|V| = 20$), its probability of occurrence in season t is defined as

$$p_t(x) \propto \exp(R(x)t + b(x)), \tag{2}$$

in which $R(x) \in \mathbb{R}$ defines the growth rate (fitness) of sequence x, while $b(x) \in \mathbb{R}$ describes when strain x emerges. To normalize the probability over the combinatorial

space of potential sequences, we factorize the probability $p_t(x)$ autoregressively and model the conditional probability with a protein language model

$$p_t(x_i|x_{<i};\theta) = \frac{\exp(R(x_i|x_{<i};\theta)t + b(x_i|x_{<i};\theta))}{\sum_{x'_i \in V} \exp(R(x'_i|x_{<i};\theta)t + b(x'_i|x_{<i};\theta))}. \tag{3}$$

Both $R(x_i|x_{<i};\theta)$ and $b(x_i|x_{<i};\theta)$ are parameterized by Transformers-based deep neural networks. During the training process, we sample pairs of protein sequences and their collection times. The parameters θ are optimized based on the maximum likelihood estimation objective:

$$L(\theta) = \mathbb{E}_{(x,t)\sim p_{train}} \log p_t(x;\theta). \tag{4}$$

Data processing

The training corpora of the dominance prediction model were obtained from the GISAID. We downloaded 394,090 HA sequences submitted before 2023-03-02. We retained only HA amino acid sequences from human hosts and with full length (with a minimum length of 553 amino acids). Starting from 2003-10, we discretized every six months into one season. Two subtypes, A/H1N1 and A/H3N2 are considered separately. For each subtype, the seasons with less than 100 HA samples are discarded. After the pre-processing, 28,546 non-repeated HA sequences are obtained for A/H3N2, and 23,736 non-repeated HA sequences are obtained for A/H1N1.

The sequences were further split into training and testing sets based on collection time. Specifically, to predict the dominance of sequences in the winter season of a particular year (e.g. test set 2018-10 to 2019-04), we train on sequence data collected before the February of that year (e.g. before 2018-02).

Baselines

AA Sub models the dominance of sequences based on the dynamic likelihood of amino-acid substitutions. The probability of amino acid sequences x occurring in season t is the product of the probabilities of amino acid at each position, considered independently.

$$p_t(x) = \prod_{i=1}^{L} p_t(x_i) = \frac{\exp(R_i(x)t + b_i(x))}{\sum_{x'} \exp(R_i(x')t + b_i(x'))}, \tag{5}$$

in which $R_i(x)$ and $b_i(x)$ are learnable scalar parameters describing the growth rate and emerging time for amino acid x at site i.

LM models the dominance of sequences in a static manner. Similarly to Eq. 3, the likelihood of amino acid sequence x is factorized autoregressively without using the collection time information:

$$p(x_i|x_{<i};\theta) = \frac{\exp(F(x_i|x_{<i};\theta))}{\sum_{x'_i \in V} \exp(F(x'_i|x_{<i};\theta))}. \tag{6}$$

Supplementary Results

Interpreting antigenicity predictor with validated escaping mutations

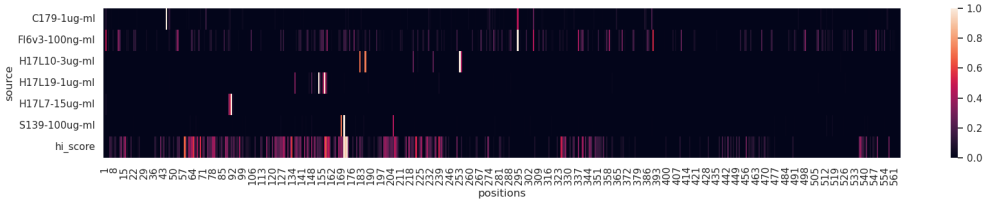

Fig. 2: Escapable effects of single amino acid mutations given by antigenicity predictor and mutational antigenic profiling assays relative to different antibodies. The x-axis is the position in reference virus strain A/WSN/1933 of A/H1N1. The mutational antigenic profiling results are from [32].

Our antigenicity predictor could identify the escaping mutations observed in other neutralization assays with respect to specific antibodies, without incorporating any of this information into our training process. The mutational antigenic profiling assay [32] investigates the effect of single amino acid mutations on escaping broad and narrow antibodies targeting HA. We predict the HI value between the same wild-type vaccine strain A/WSN/1933 used in their experiment and the same virus strains with single amino acid mutations to the wild-type strain. The quantified escape effect given by the mutational antigenic profiling assays and our model is shown in Fig. 2. Our model predicts that the mutations on site 171 allow for escape from the antibodies produced by A/WSN/1933. This is consistent with the discovery in mutational antigenic profiling results that mutations in site 171 allow for escape from the broadly neutralizing antibody S139 [32]. Our model also predicts that mutations on site 158 escape from the vaccine, which is consistent with the escaping mutations from antibody H17L19 given by mutational antigenic profiling assays [32]. Finally, our model assigned high HI values to some additional sites, but mutations at these points have not been observed to escape from the antibodies screened thus far. Additional experiments are required to elucidate these sites, which might escape antibodies that have not yet been identified, or act through unknown immune escape mechanisms.

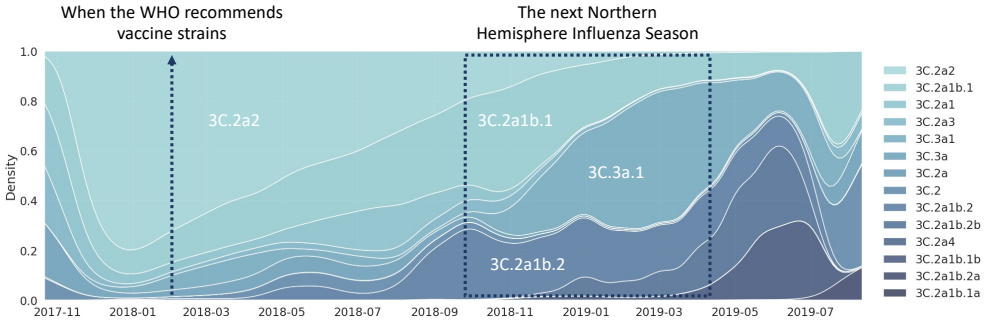

Fig. 3: Antigenic drift of A/H3N2 occurs between the period of vaccine selection and the following Northern Hemisphere Influenza Season. Prior to the vaccine strain selection, the prevailing clade was 3C.2a2. However, in the subsequent influenza season, the dominant clades shift to 3C.2a1b.1, 3C.3a.1, and 3C.2a1b.2.

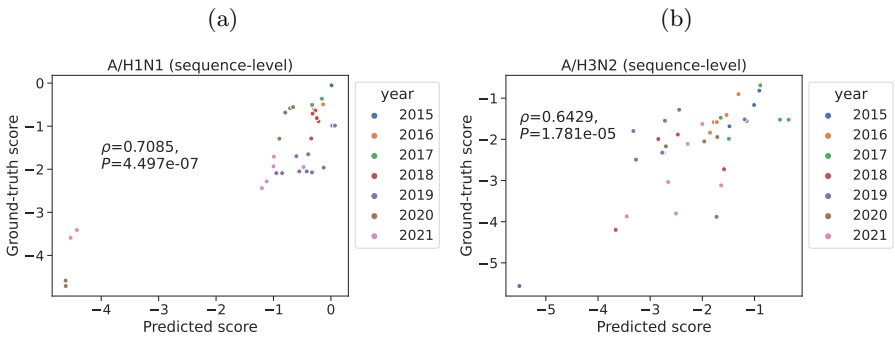

Fig. 4: **Performance of escapability score prediction.** Spearman rank correlation ρ between the predicted escapability score and ground-truth score (sequence-level) for (**a**) A/H1N1 and (**b**) A/H3N2, from 2015 to 2021 seasons. $P < 0.005$.

References

[1] Zimmerman, R.K., Nowalk, M.P., Chung, J., Jackson, M.L., Jackson, L.A., Petrie, J.G., Monto, A.S., McLean, H.Q., Belongia, E.A., Gaglani, M., et al.: 2014–2015 influenza vaccine effectiveness in the united states by vaccine type. Clinical Infectious Diseases, 635 (2016)

[2] Rolfes, M.A., Flannery, B., Chung, J.R., O'Halloran, A., Garg, S., Belongia, E.A., Gaglani, M., Zimmerman, R.K., Jackson, M.L., Monto, A.S., *et al.*: Effects of influenza vaccination in the united states during the 2017–2018 influenza season. Clinical Infectious Diseases **69**(11), 1845–1853 (2019)

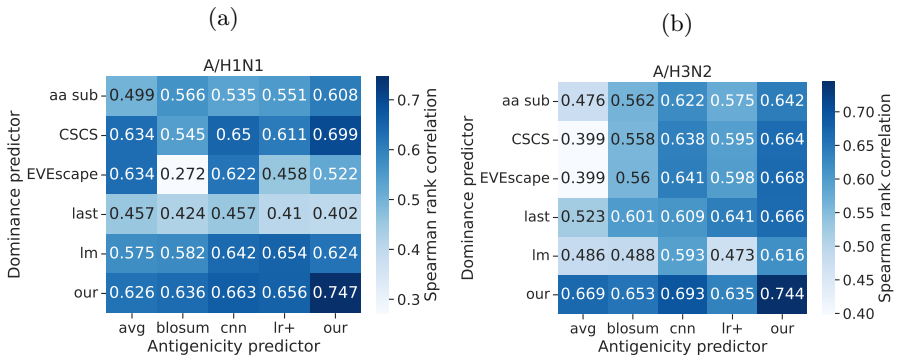

Fig. 5: Spearman rank correlation ρ between the predicted escapability score and ground-truth score (clade-level) for A/H3N2 and A/H1N1, from 2015 to 2021 seasons. $P < 0.005$. Vaccines with HI data covering more than 80% circulating clades are compared.

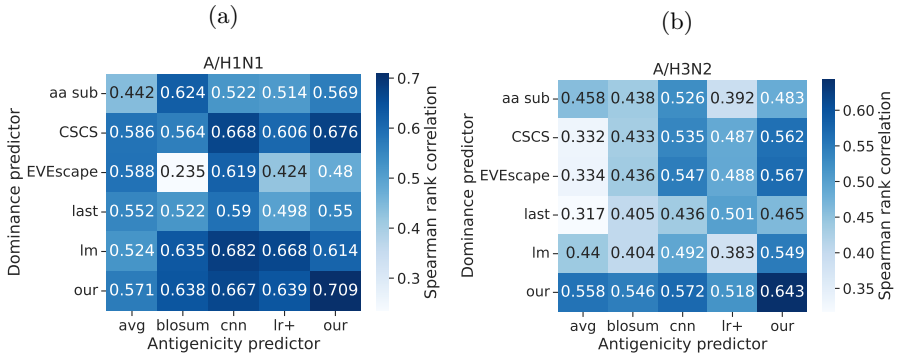

Fig. 6: Spearman rank correlation ρ between the predicted escapability score and ground-truth score (sequence-level) for A/H3N2 and A/H1N1, from 2015 to 2021 seasons. $P < 0.005$. Vaccines with HI data covering more than 40% circulating sequences are compared.

[3] Flannery, B., Kondor, R.J.G., Chung, J.R., Gaglani, M., Reis, M., Zimmerman, R.K., Nowalk, M.P., Jackson, M.L., Jackson, L.A., Monto, A.S., *et al.*: Spread of antigenically drifted influenza a (h3n2) viruses and vaccine effectiveness in the united states during the 2018–2019 season. The Journal of infectious diseases **221**(1), 8–15 (2020)

[4] Tenforde, M.W., Kondor, R.J.G., Chung, J.R., Zimmerman, R.K., Nowalk, M.P., Jackson, M.L., Jackson, L.A., Monto, A.S., Martin, E.T., Belongia, E.A., *et*

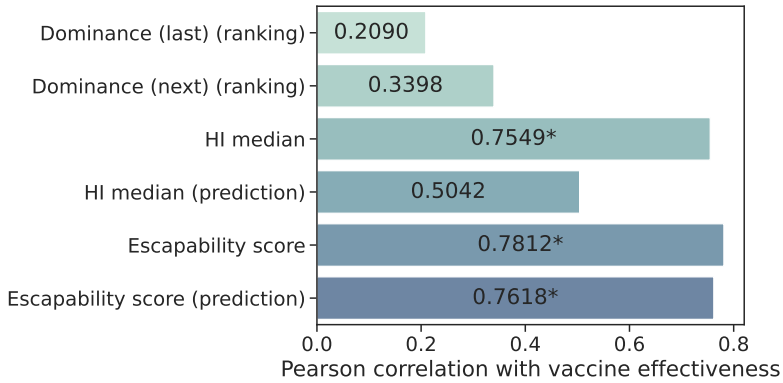

Fig. 7: The Pearson correlation between vaccine effectiveness and various scoring strategies for the vaccine strains. The star mark (*) means that the $P < 0.05$. Both the ground-truth and prediction escapability scores show the strongest correlations.

al.: Effect of antigenic drift on influenza vaccine effectiveness in the united states—2019–2020. Clinical Infectious Diseases **73**(11), 4244–4250 (2021)

[5] Price, A.M., Flannery, B., Talbot, H.K., Grijalva, C.G., Wernli, K.J., Phillips, C.H., Monto, A.S., Martin, E.T., Belongia, E.A., McLean, H.Q., *et al.*: Influenza vaccine effectiveness against influenza a (h3n2)-related illness in the united states during the 2021–2022 influenza season. Clinical Infectious Diseases **76**(8), 1358–1363 (2023)

[6] Gerdil, C.: The annual production cycle for influenza vaccine. Vaccine **21**(16), 1776–1779 (2003)

[7] Smith, D.J., Lapedes, A.S., De Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D., Fouchier, R.A.: Mapping the antigenic and genetic evolution of influenza virus. science **305**(5682), 371–376 (2004)

[8] Gouma, S., Weirick, M., Hensley, S.E.: Antigenic assessment of the h3n2 component of the 2019-2020 northern hemisphere influenza vaccine. Nature Communications **11**(1), 2445 (2020)

[9] Flannery, B., Kondor, R.J.G., Chung, J.R., Gaglani, M., Reis, M., Zimmerman, R.K., Nowalk, M.P., Jackson, M.L., Jackson, L.A., Monto, A.S., *et al.*: Spread of antigenically drifted influenza a (h3n2) viruses and vaccine effectiveness in the united states during the 2018–2019 season. The Journal of infectious diseases **221**(1), 8–15 (2020)

[10] Hirst, G.K.: Studies of antigenic differences among strains of influenza a by means of red cell agglutination. The Journal of experimental medicine **78**(5), 407–423 (1943)

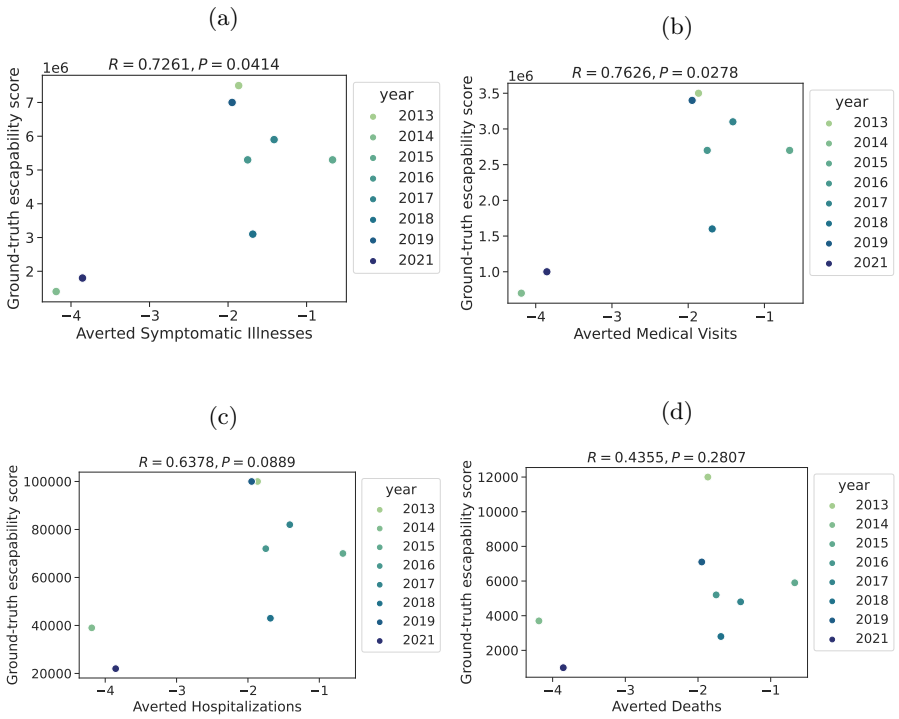

Fig. 8: **The ground-truth escapability score is correlated with disease burden prevention.** Ground-truth escapability scores and the estimated number of influenza illnesses (**a**), medical visits (**b**), hospitalizations (**c**), and deaths (**d**) averted by vaccination. Pearson correlations (R) and corresponding P-values (P) are illustrated in the text.

[11] Barr, I.G., Russell, C., Besselaar, T.G., Cox, N.J., Daniels, R.S., Donis, R., Engelhardt, O.G., Grohmann, G., Itamura, S., Kelso, A., *et al.*: Who recommendations for the viruses used in the 2013–2014 northern hemisphere influenza vaccine: Epidemiology, antigenic and genetic characteristics of influenza a (h1n1) pdm09, a (h3n2) and b influenza viruses collected from october 2012 to january 2013. Vaccine **32**(37), 4713–4725 (2014)

[12] Hie, B., Zhong, E.D., Berger, B., Bryson, B.: Learning the language of viral evolution and escape. Science **371**(6526), 284–288 (2021)

[13] Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K.P., Gal, Y., Marks, D.S.: Disease variant prediction with deep generative models of evolutionary data. Nature (2021)

[14] Thadani, N.N., Gurev, S., Notin, P., Youssef, N., Rollins, N.J., Sander, C., Gal,

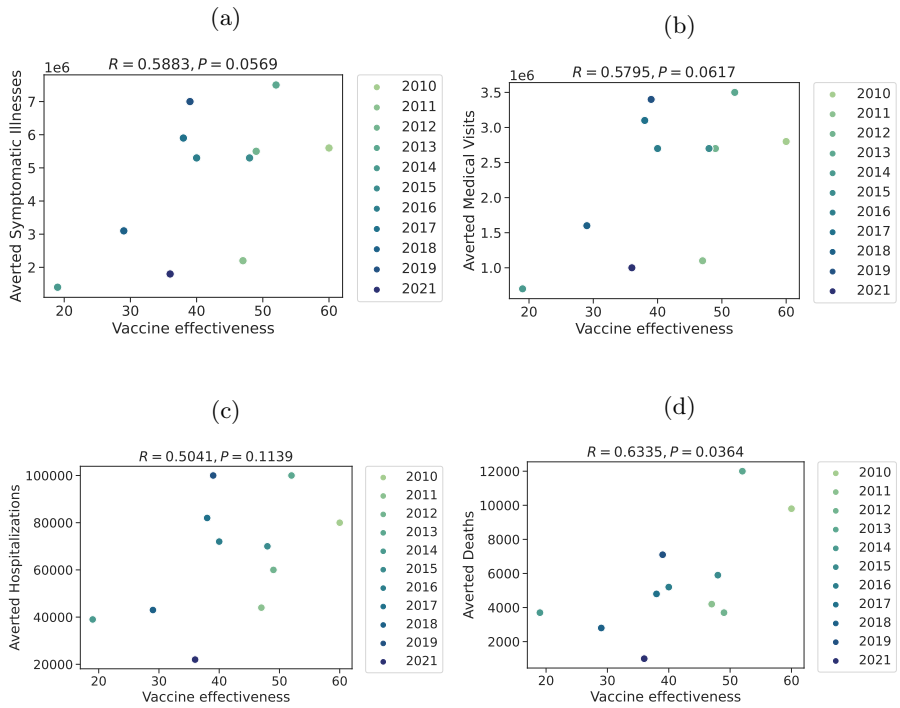

Fig. 9: **The vaccine effectiveness is correlated with disease burden prevention.** The vaccine effectiveness and the estimated number of influenza illnesses (**a**), medical visits (**b**), hospitalizations (**c**), and deaths (**d**) averted by vaccination. Pearson correlations (R) and corresponding P-values (P) are illustrated in the text.

Y., Marks, D.S.: Learning from pre-pandemic data to forecast viral escape. Nature (2023)

[15] Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S.F., Pyle, J.D., Yurkovetskiy, L., Bosso, M., Park, D.J., Babadi, M., MacInnis, B.L., *et al.*: Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. Science **376**(6599), 1327–1332 (2022)

[16] Han, W., Chen, N., Xu, X., Sahil, A., Zhou, J., Li, Z., Zhong, H., Gao, E., Zhang, R., Wang, Y., *et al.*: Predicting the antigenic evolution of sars-cov-2 with deep learning. Nature Communications **14**(1), 3478 (2023)

[17] Zeller, M.A., Gauger, P.C., Arendsee, Z.W., Souza, C.K., Vincent, A.L., Anderson, T.K.: Machine learning prediction and experimental validation of antigenic drift in h3 influenza a viruses in swine. MSphere **6**(2), 10–1128 (2021)

[18] Neher, R.A., Bedford, T., Daniels, R.S., Russell, C.A., Shraiman, B.I.: Prediction,

dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proceedings of the National Academy of Sciences **113**(12), 1701–1709 (2016)

[19] Yin, R., Thwin, N.N., Zhuang, P., Lin, Z., Kwoh, C.K.: Iav-cnn: a 2d convolutional neural network model to predict antigenic variants of influenza a virus. IEEE/ACM Transactions on Computational Biology and Bioinformatics **19**(6), 3497–3506 (2021)

[20] Xia, Y.-L., Li, W., Li, Y., Ji, X.-L., Fu, Y.-X., Liu, S.-Q., et al.: A deep learning approach for predicting antigenic variation of influenza a h3n2. Computational and mathematical methods in medicine **2021** (2021)

[21] Kostova, D., Reed, C., Finelli, L., Cheng, P.-Y., Gargiullo, P.M., Shay, D.K., Singleton, J.A., Meltzer, M.I., Lu, P.-j., Bresee, J.S.: Influenza illness and hospitalizations averted by influenza vaccination in the united states, 2005–2011. PloS one **8**(6), 66312 (2013)

[22] Reed, C., Chaves, S.S., Daily Kirley, P., Emerson, R., Aragon, D., Hancock, E.B., Butler, L., Baumbach, J., Hollick, G., Bennett, N.M., *et al.*: Estimating influenza disease burden from population-based surveillance data in the united states. PloS one **10**(3), 0118369 (2015)

[23] Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A.: Msa transformer. In: International Conference on Machine Learning, pp. 8844–8856 (2021). PMLR

[24] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)

[25] Gong, L.I., Suchard, M.A., Bloom, J.D.: Stability-mediated epistasis constrains the evolution of an influenza protein. Elife **2**, 00631 (2013)

[26] Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T., Yeo, W., *et al.*: Gisaid's role in pandemic response. China CDC weekly **3**(49), 1049 (2021)

[27] Ferruz, N., Schmidt, S., Höcker, B.: Protgpt2 is a deep unsupervised language model for protein design. Nature communications **13**(1), 4348 (2022)

[28] Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

[29] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[30] Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences **89**(22), 10915–10919 (1992)

[31] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J.: 1d convolutional neural networks and applications: A survey. Mechanical systems and signal processing **151**, 107398 (2021)

[32] Doud, M.B., Lee, J.M., Bloom, J.D.: How single mutations affect viral escape from broad and narrow antibodies to h1 influenza hemagglutinin. Nature communications **9**(1), 1386 (2018)

[33] Lina, B., Georges, A., Burtseva, E., Nunes, M.C., Andrew, M.K., McNeil, S.A., Ruiz-Palacios, G.M., Feng, L., Kyncl, J., Vanhems, P., *et al.*: Complicated hospitalization due to influenza: results from the global hospital influenza network for the 2017–2018 season. BMC infectious diseases **20**(1), 1–14 (2020)

[34] Jorquera, P.A., Mishin, V.P., Chesnokov, A., Nguyen, H.T., Mann, B., Garten, R., Barnes, J., Hodges, E., De La Cruz, J., Xu, X., *et al.*: Insights into the antigenic advancement of influenza a (h3n2) viruses, 2011–2018. Scientific Reports **9**(1), 2676 (2019)

[35] Han, A.X., Jong, S.P., Russell, C.A.: Co-evolution of immunity and seasonal influenza viruses. Nature Reviews Microbiology, 1–13 (2023)

[36] Steinegger, M., Söding, J.: Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**(11), 1026–1028 (2017)