Forecasting H1N1 Influenza Pandemic and Seasonal Evolution

Aarushi Mehrotra [1 2 3]   **Noor Youssef** [1]   **Sarah Gurev** [1 2 †]   **Debora Marks** [1 †]

Abstract

The rapid accumulation of mutations in influenza strains challenges vaccine efficacy by enabling immune escape. Current influenza vaccine strategies rely on annual updates that incorporate recently circulating strains, with strain selection guided by experimental neutralization assays and global surveillance data. However, these methods are resource-intensive and have limited predictive power, often resulting in mismatches between the strains included in the vaccine and the strains dominant at the time of vaccine rollout — ultimately reducing vaccine effectiveness. Understanding how specific mutations alter viral fitness and antigenicity could improve our ability to forecast emerging escape variants and design more effective vaccines. To address this, we adapt EVE, a deep generative model, and retrain on influenza hemagglutinin (HA) sequences to model the evolution of H1N1. By training on HA sequences sampled prior to the emergence of a given clade, our model consistently predicts the most frequent future mutations across all eight major H1N1 clades since the 2009 pandemic—without requiring prior knowledge of clade identity. This work highlights the potential of deep learning models to forecast influenza evolution and support proactive vaccine design.

1. Introduction

Influenza A represents one of the most significant causes of respiratory infections worldwide, with seasonal epidemics that infect one billion people and cause hundreds of thousands of deaths each year(WHO, 2025a). While the influenza vaccine is the primary line of defense and greatly reduces the burden of disease, vaccines for seasonal influenza have the lowest and most variable performance of any vaccine licensed for use in the United States, despite decades of investment(Perofsky & Nelson, 2020). Their effectiveness is challenged by rapid evolution and continuous antigenic shift and drift, which enables the virus to evade host immune responses, necessitating the annual reformulation of influenza vaccines to match circulating strains. Yet, vaccine strains (including influenza A(H1N1), A(H3N2), and B components) must be chosen more than six months in advance to allow time for production. The World Health Organization (WHO) recommendations are primarily based on surveillance sequencing and antigenic characterization of a limited number of globally circulating strains, but these have limited predictive power–traditional serological assays are resource-intensive, making it challenging to measure neutralization across thousands of circulating strains(Shu & McCauley, 2017). Over a decade, only three influenza seasons demonstrated more than 70% antigenic match between the selected vaccine strain and the dominant strains six months later(Choi et al., 2024). For example, the vaccine strain for the 2019-2020 flu season was a clade 6B.1A.1 strain, an emerging clade in February 2019 when WHO convened to formulate recommendations, which mostly disappeared by the start of November flu season as clade 6B.1A.5a took over (Fig. 1A, 1B). Such mismatches can significantly reduce vaccine effectiveness, emphasizing the need for improved predictive models of viral escape.

Ideally, at the time of vaccine selection we would be able to rank all circulating strains by their potential for dominance in the next influenza season and their antigenic distance from prior infections and vaccinations. More high-throughput experimental approaches are being developed that can measure neutralizing titers against dozens of strains simultaneously(Loes et al., 2024; Kikawa et al., 2025), although testing all circulating strains or potential future mutants remains intractable. Computational methods have also emerged to aid vaccine strain selection, however most depend heavily on large-scale antigenic data or current strain prevalence data, limiting their ability to forecast emerging variants(Agor & Özaltın, 2018; Gao et al., 2024; Lou et al., 2024; Hayati et al., 2020; Shi et al., 2024).
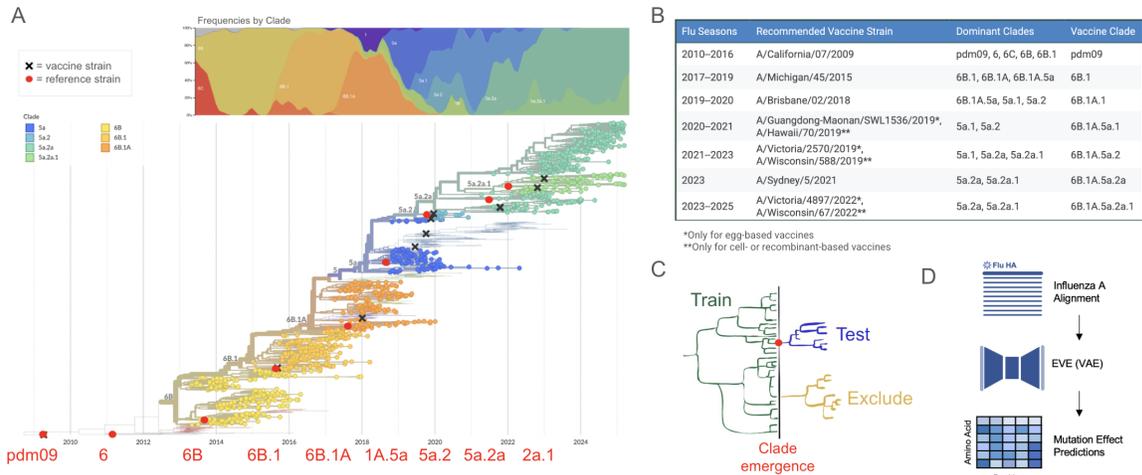
Figure 1. Data and modeling overview. **A**. Data curation strategy across 8 H1N1 clades since 2009. Tree from NextStrain(Hadfield et al., 2018; Sagulenko et al., 2018) using data from GISAID(Shu & McCauley, 2017). **B**. Recommendations for influenza vaccine composition from the World Health Organization(WHO, 2025b) (southern and northern hemisphere influenza seasons combined). **C**. Train and test split strategy used for all 8 clades. **D**. Pipeline from sequence alignment of training dataset to EVE model to mutation effect prediction.

Alternatively, information encoded in the evolutionary history of influenza can be used to make zero-shot predictions about future influenza evolution and key host adaptive or antigenic mutations. Annual surveillance efforts have resulted in hundreds of thousands of sequenced strains that are ideal for training evolutionary sequence models for quantifying both fitness and viral escape(Shu & McCauley, 2017). This sequencing data has previously been used for forecasting based on lineage growth, whose predictive ability suffers from the average 3-month lag time of influenza sequence deposition, but can instead be used to learn the underlying patterns of influenza function and structural constraints(Huddleston & Bedford, 2024). EVE– a variational autoencoder that models the fitness of protein mutations given a multiple sequence alignment–successfully forecasts the functional effects of mutations for SARS-CoV-2 Spike protein when trained on pre-pandemic coronavirus sequences and aids in evaluation of vaccine strains(Thadani et al., 2023; Youssef et al., 2025). In this study, we investigate whether EVE is sufficient to predict influenza A evolution, particularly of the H1N1 2009 pandemic.

The 2009 "swine flu" pandemic was a novel spillover of H1N1 from pigs, distinct from prior 1918 and 1977 H1N1 pandemics(Saunders-Hastings & Krewski, 2016). Between May and December 2009 alone, up to 2 million lives were lost. The 2009 strain then displaced previous H1N1 strains to begin producing seasonal outbreaks. At the start, mutations primarily resulted in adaptation to humans, for instance increasing binding to human-like $\alpha 2,6$-linked sialic acids, increasing replication in the respiratory tract, or elevating droplet transmission(Otte et al., 2016). The H1N1 compo-

nent of seasonal flu vaccines remained unchanged from the 2009 pdm09 strain until 2017, as circulating viruses maintained similar sera binding profiles for six years until antigenically distinct variants emerged(WHO, 2025b). Clade 6B.1 viruses obtained mutations in the Sa antigenic site that caused detectable differences in human post-vaccination sera, and the H1N1 vaccine component was consequently updated to A/Michigan/45/2015. Since then, antigenic evolution has dominated, resulting in five vaccine strain updates since 2019 (Fig. 1B).

We trace the evolutionary trajectory of the H1 hemagglutinin (HA) protein from the 2009 pandemic strain of Influenza A H1N1 (A/California/07/2009, or pdm09) through eight sequential, phylogenetically nested clades: 6, 6B, 6B.1, 6B.1A, 6B.1A.5a, 6B.1A.5a.2, 6B.1A.5a.2a, and 6B.1A.5a.2a.1(WIC, 2024). We demonstrate consistent success in using EVE to predict evolution–whether host adaptive or antigenic–at each of eight timepoints from 2009 to 2025, while training only on sequencing data available prior to that clade's emergence. This demonstrates the potential of using evolutionary models of historical influenza sequences to aid in vaccine strain selection.

2. Methods

A historical challenge in evaluating influenza models has been the absence of well-defined baseline reference sequences from which key mutations emerged. To address this, we perform phylogenetic analysis of H1N1 clades using the annual and interim reports sent from the Worldwide Influenza Centre at the Francis Crick Institute to the WHO to

inform vaccine composition(WIC, 2024). Tracking the evolution of H1N1 from 2009 using these reports, we defined each emerging clade's characteristic mutations, identified its most basal strain, and noted which years it was dominant. We used this basal strain as the reference sequence for modeling each clade, reverting any unique adaptations (e.g., egg-passaged Q240R) to ensure that the reference only contained the clade's characteristic mutations from the A/(H1N1)pdm09 sequence and the preceding clade. In total, we examined the following nested clade lineages: pdm09, 6, 6B, 6B.1, 6B.1A, 6B.1A.5a, 6B.1A.5a.2, 6B.1A.5a.2a, and 6B.1A.5a.2a.1. Supplemental Fig 1 summarizes the clades, reference strains, and training and test time alignments used.

We assembled a training dataset of all full-length influenza A hemagglutinin protein sequences from the Global Initiative on Sharing All Influenza Data (GISAID) database(Shu & McCauley, 2017), comprising 365,500 full-length sequences submitted before January 1, 2025. The training dataset for each clade-specific model was then built by filtering the full Influenza A dataset to include only sequences collected before that clade's emergence, then deduplicating and aligning them to the identified clade reference. Each model was evaluated on a corresponding test dataset containing all human sequences collected during the 2-4 year period that clade was dominant and annotated as belonging to that clade or its derivative subclades. For instance, sequences labeled 6B.1 or 6B.1A were included in the test set for the clade 6B model. We assigned clade labels based on our identified characteristic mutations, requiring only the new mutations specific to the child clade rather than cumulative acquisition of all ancestral mutations from pdm09. The test dataset mutations were further filtered to remove any mutation seen less than three times in order to ensure the mutations were actually present in the population. Reversion mutations back to a previous clade were also removed.

We estimate the fitness effect of mutations using EVE(Frazer et al., 2021), a deep variational autoencoder trained on influenza sequences in our pre-clade emergence datasets. EVE learns constraints underpinning structure and function for a given protein family and considers higher-order dependencies across positions. We compare with other alignment-based models with different epistatic assumptions: EVcouplings (pairwise couplings; (Hopf et al., 2017)) and a PSSM (site independent; (Hopf et al., 2017)), as well as EVEscape which combines fitness predictions learned from EVE with biophysical and structural features relevant to immune escape (Thadani et al., 2023).

Additionally, we benchmarked predictions against experimental data from Liu et al. 2024, which conducted deep mutational scanning (DMS) assays to quantify relative fitness scores across all mutations to the more variable globular head domain (HA1 subunit), which includes the receptor binding site and other key epitopes (Liu et al., 2024). For each clade, we selected the DMS dataset performed in the most closely related HA background: either pdm09 for early clades or A/Wisconsin/588/2019 for clade 5a.2 onwards. These DMS scores allowed comparison of computational predictions to experimentally derived fitness phenotypes, but do not capture immune escape.

3. Results

Extensive surveillance sequencing allows us to test EVE's ability to forecast frequent mutations in a retrospective study. We first focus on clade 6B.1A.5a.2a (hereafter refered to as 5a.2a) as it is the most widespread current H1N1 lineage. We trained EVE on only pre-2021 sequences and evaluated whether its top-scoring predictions corresponded to the mutations that emerged within the years the clade was dominant, in this case 2021 to 2025. To enforce biologically relevant mutational constraints, we restricted our analysis to the 4,442 possible single-nucleotide-accessible substitutions from the clade's reference sequence since multi-nucleotide mutations are very rare (<4% of the test dataset). The test dataset was also subsetted to HA1 for DMS comparison.

EVE identifies the most frequent mutations that occurred in clade 5a.2a (Fig. 2A). All models were more likely to label higher frequency mutations as escape than lower frequency ones, but EVE was consistently better than the baseline models, and the Wisconsin DMS at predicting the mutations actually observed in the outbreak at every frequency threshold, though the PSSM was surprisingly effective given its site independent assumption. For example, out of the three high-frequency HA1 mutations that emerged to define the subsequent Clade 5a.2a.1, only EVE and EVcouplings correctly classify all of them as escape (Supplemental Fig. 10). Other clade models do similarly well if not better, as five out of eight clade models reach 100% prediction of the highest frequency mutations (see Supplemental Figures 3-11 for more details on each clade).

Since EVE is able to assign high scores to high frequency clade mutations, we see that as a clade evolves, more of EVE's escape predictions are observed (Fig. 2B). At the clade's emergence in 2021, none of any model's top mutations were present in the population—an important sanity check confirming test set mutations are novel rather than trivial memorization of mutations present at training time. Over the next four years, as the clade accumulated mutations to escape multiple seasons of influenza vaccinations, predicted escape mutations began to appear. By the start of 2025, 93% of EVE's top 100 predictions had been observed at least three times, while the DMS was only able to recover 52% of the outbreak mutations—but both have predicted power, far outperforming the random baseline (in gray), which reflects the background rate of seen mutations
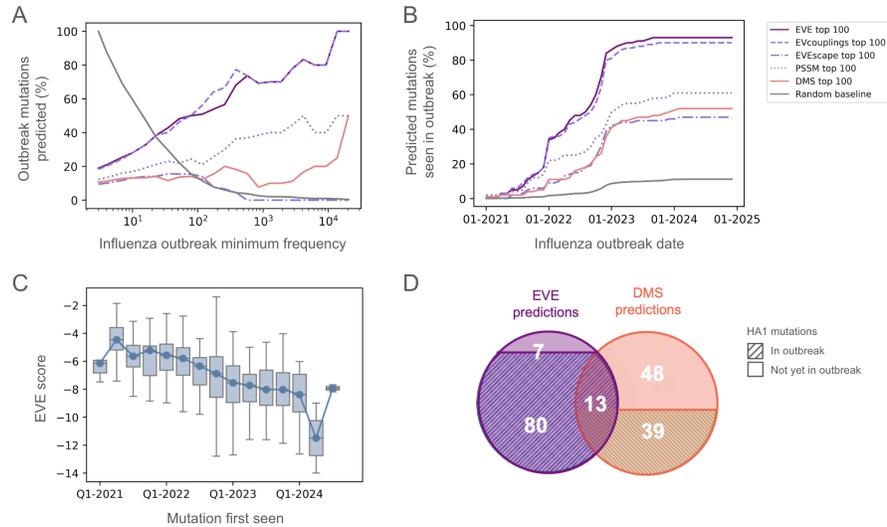
Figure 2. EVE predicts future outbreak HA1 mutations in the dominant H1N1 clade 6B.1A.5a.2a more accurately than deep mutational scans. We classify the top 100 mutations as model or DMS-predicted likely outbreak mutations. **A**. EVE and baselines are more likely to predict mutations that are seen in higher frequency mutations in the clade. **B**. By 2025, 93% of EVE's predicted escape mutations were observed, outperforming every other model's predictions. **C**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **D**. EVE and DMS predict distinct sets of mutations, though the small overlapping set all appear in the clade.

relative to all possible single-nucleotide HA1 mutations. Notably, higher EVE scores corresponded to mutations that were seen earlier in the outbreak (Fig. 2C), suggesting that more fit or immune-evasive substitutions are preferentially identified by EVE. Median EVE scores decline over time as more mutations—often with weaker selection advantages—accumulate, consistent with expectations of epistatic saturation and immunological niche filling over the course of clade evolution, with only two new mutations emerging by the second half of 2024. This is consistent with the plateau of predicted mutations seen after 2023 (Fig. 2B).

We directly compare EVE and DMS predictions on their ability to identify outbreak-relevant mutations (Fig. 2D). Out of the top 100 mutations from each method, only 13 are shared, highlighting their divergent scoring criteria. However, EVE's predictions are more reflective of natural viral evolution: 80 out of the 87 other top scorers have already been observed in the 5a.2a outbreak, compared to less than half that from DMS. Many DMS-predicted mutations are unseen in circulating sequences and are likely to remain unseen as new mutation rates drop (Fig. 2B and 2C). This underscores a limitation of even up-to-date, *in vitro* fitness-only mutagenesis scans in forecasting real-world evolution.

To evaluate performance across all clade models, we quantified the proportion of clade mutations observed ≥ 100 times that fall within the model's top 100 predicted escape mutations (Fig. 3A). This is a stringent benchmark that directly quantifies how well the model recapitulates known antigenic evolution—EVE's top 100 scores predict 52% of HA1 mu-

tations seen more than 100 times in Clade 5a.2a, compared to 23% by the PSSM model, 14% by EVEscape, and 12% by the Wisconsin DMS. The DMS consistently performs comparably to the random baseline as 65 out of 497 (13%) total mutations in Clade 5a.2a are observed more than 100 times, and for many clades predict almost no correct mutations. The outlier is Clade 5a.2 which has a single HA1 mutation observed more than 100 times, K226M (K209M in H1 numbering), which only the DMS predicts. Although frequent in 2020, it died out as Clade 5a.2a took over.

Additionally, EVE assigns higher scores to residues in the antigenically-relevant head than the conserved stem domain, and especially to clade-defining mutations (Fig. 3B-D). Top-ranked residues disproportionately fall within the canonically defined epitope regions (Sa, Sb, Ca, Cb, receptor-binding site), and high-scoring residues often include clade-defining mutations that mark key antigenic transitions. As an example, we highlight the S179N (S162N in H1 numbering) mutation that defines Clade 6B.1 and was key to finally moving towards a new vaccine recommendation after six years of the pdm09 strain vaccine. Lateral-patch-binding pdm09 antibodies could not neutralize this new strain because of this new glycosylation site in the Sa epitope(Guthmiller et al., 2021). EVE predicted this mutation as escape (within its top 100) in all three previous clade models: its escape score is in the top 2/3rd of the pdm09 model's escape predictions and in the top 1/3rd of Clade 6 and 6B models' escape predictions. EVE predicted the first major immunogenic H1N1 mutation at the very start of the H1N1 pandemic, six years before its dominance in 2015.
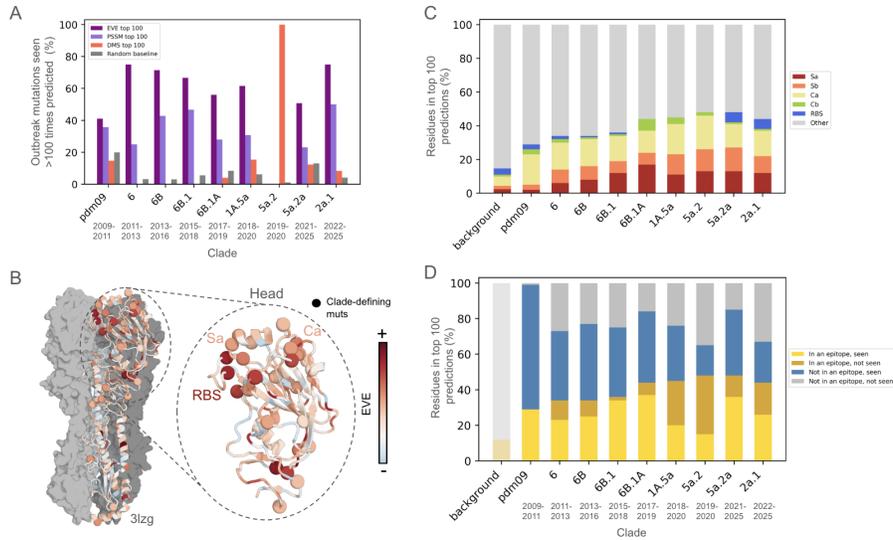
Figure 3. EVE prioritizes antigenically relevant, high frequency mutations across clades. **A**. Across clade models, a substantial fraction of high-frequency ($\geq$100x) clade mutations are ranked in the top 100 by EVE, outperforming DMS, PSSM, and random baselines. 5a.2 has a single seen mutation. **B**. EVE scores (site-level maximum) of pdm09 model mapped onto a representative HA structure (PDB: 3tzg) highlight high-scoring regions, especially the Receptor binding site (RBS) and antigenic sites Sa and Ca, within the head domain. Spheres indicate clade-defining mutations across all 9 clades. **C**. EVE-predicted escape residues are consistently enriched in known head epitopes compared to a background expectation. **D**. EVE-predicted escape residues in PDB epitopes are more likely to be observed.

4. Discussion

Our results demonstrate that unsupervised evolutionary models like EVE can prospectively predict key mutations in influenza A virus HA protein, even when trained only on historical sequence data. Across nine major H1N1 clades that have emerged since the 2009 pandemic, we show that the mutations prioritized by EVE before a clade's emergence are highly enriched for those that later become prevalent, even more so than deep mutational scanning assays. Importantly, we observe that EVE predictions are not only temporally accurate, but also structurally plausible. High-scoring mutations cluster in the HA head domain, especially at known antigenic sites, while mutations in conserved stalk helices are deprioritized. This aligns with known structural and immunological constraints on HA evolution(Wu & Wilson, 2020). Furthermore, we find that EVE is predictive of not only high-frequency mutations within a clade but also of the defining mutations of successor clades, highlighting the model's capacity to anticipate lineage-defining events.

Compared to existing models for influenza evolution—many of which rely on large-scale serological datasets or real-time tracking of strain frequency—our approach offers a complementary strategy. By leveraging only evolutionary sequence data, EVE enables forecasting of future mutations without requiring extensive lab-based assays or surveillance data that may lag behind real-time evolution. This is particularly advantageous in under-sampled regions or for forecasting

long-term evolutionary trajectories, such as in the selection of candidate vaccine strains. Notably, unlike prior work for SARS-CoV-2, we find the EVE is a better predictor of future evolution than EVEscape, likely because flu evolutionary sequences already contain constraints relevant to immune escape, and so do not benefit from explicit inclusion of escape-relevant terms (Thadani et al., 2023).

Looking forward, we are working on an epistasis model to combine single mutation scores for an interpretable strain-wide escape score to better forecast entire vaccine strains rather than individual mutations. We plan on extending our framework to H3 since historically, the H3N2 vaccine has been even less effective than the H1N1 vaccine due to rapid antigenic evolution(McLean & Belongia, 2021). We also plan to benchmark more computational methods, including modifications to EVEscape for influenza, as well as experimental scans measuring antibody escape. Overall, this work supports the hypothesis that deep evolutionary models, when trained on the right sequence context, can generalize to forecast viral escape—offering a scalable, data-efficient foundation for the future of vaccine strain design.

Acknowledgments

References

Agor, J. K. and Özaltın, O. Y. Models for predicting the evolution of influenza to inform vaccine strain selection. *Human vaccines & immunotherapeutics*, 14(3):678–683, 2018.

Choi, Y. J., Song, J. Y., Wie, S.-H., Choi, W. S., Lee, J., Lee, J.-S., Kim, Y. K., Kim, S. W., Lee, S. H., Park, K.-H., et al. Real-world effectiveness of influenza vaccine over a decade during the 2011–2021 seasons—implications of vaccine mismatch. *Vaccine*, 42(26):126381, 2024.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021.

Gao, C., Wen, F., Guan, M., Hatuwal, B., Li, L., Praena, B., Tang, C. Y., Zhang, J., Luo, F., Xie, H., et al. Maivess: streamlined selection of antigenically matched, high-yield viruses for seasonal influenza vaccine production. *Nature Communications*, 15(1):1128, 2024.

Guthmiller, J. J., Han, J., Li, L., Freyn, A. W., Liu, S. T. H., Stovicek, O., Stamper, C. T., Dugan, H. L., Tepora, M. E., Utset, H. A., Bitar, D. J., Hamel, N. J., Changrob, S., Zheng, N.-Y., Huang, M., Krammer, F., Nachbagauer, R., Palese, P., Ward, A. B., and Wilson, P. C. First exposure to the pandemic h1n1 virus induced broadly neutralizing antibodies targeting hemagglutinin head epitopes. *Science Translational Medicine*, 13(596):eabg4535, 2021. doi: 10.1126/scitranslmed. abg4535. URL https://www.science.org/doi/abs/10.1126/scitranslmed.abg4535.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

Hayati, M., Biller, P., and Colijn, C. Predicting the short-term success of human influenza virus variants with machine learning. *Proceedings of the Royal Society B*, 287 (1924):20200319, 2020.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P., Ingraham, J. B., Toth-Petroczy, A., Brock, K., Riesselman, A. J., Palmedo, P., et al. The evcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584, 2019.

Huddleston, J. and Bedford, T. Timely vaccine strain selection and genomic surveillance improves evolutionary forecast accuracy of seasonal influenza a/h3n2. *medRxiv*, 2024.

Kikawa, C., Loes, A. N., Huddleston, J., Figgins, M. D., Steinberg, P., Griffiths, T., Drapeau, E. M., Peck, H., Barr, I. G., Englund, J. A., et al. High-throughput neutralization measurements correlate strongly with evolutionary success of human influenza strains. *bioRxiv*, pp. 2025–03, 2025.

Liu, T., Reiser, W. K., Tan, T. J., Lv, H., Rivera-Cardona, J., Heimburger, K., Wu, N. C., and Brooke, C. B. Natural variation in neuraminidase activity influences the evolutionary potential of the seasonal h1n1 lineage hemagglutinin. *Virus Evolution*, 10(1):veae046, 2024.

Loes, A. N., Tarabi, R. A. L., Huddleston, J., Touyon, L., Wong, S. S., Cheng, S. M., Leung, N. H., Hannon, W. W., Bedford, T., Cobey, S., et al. High-throughput sequencing-based neutralization assay reveals how repeated vaccinations impact titers to recent human h1n1 influenza strains. *Journal of Virology*, 98(10):e00689–24, 2024.

Lou, J., Liang, W., Cao, L., Hu, I., Zhao, S., Chen, Z., Chan, R. W. Y., Cheung, P. P. H., Zheng, H., Liu, C., et al. Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations. *Nature communications*, 15(1):2546, 2024.

McLean, H. Q. and Belongia, E. A. Influenza vaccine effectiveness: new insights and challenges. *Cold Spring Harbor Perspectives in Medicine*, 11(6):a038315, 2021.

Otte, A., Marriott, A. C., Dreier, C., Dove, B., Mooren, K., Klingen, T. R., Sauter, M., Thompson, K.-A., Bennett, A., Klingel, K., et al. Evolution of 2009 h1n1 influenza viruses during the pandemic correlates with increased viral pathogenicity and transmissibility in the ferret model. *Scientific reports*, 6(1):28583, 2016.

Perofsky, A. C. and Nelson, M. I. The challenges of vaccine strain selection. *Elife*, 9:e62955, 2020.

Sagulenko, P., Puller, V., and Neher, R. A. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

Saunders-Hastings, P. R. and Krewski, D. Reviewing the history of pandemic influenza: understanding patterns of emergence and transmission. *Pathogens*, 5(4):66, 2016.

Shi, W., Wohlwend, J., Wu, M., and Barzilay, R. Vaxseer: Selecting influenza vaccine strains through evolutionary and antigenicity models. 2024.

Shu, Y. and McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Ritter, D., Sander, C., Gal, Y., and Marks, D. S. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, 2023.

WHO. Influenza (seasonal). *World Health Organization*, 2025a. URL https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).

WHO. Recommendations for influenza vaccine composition. *World Health Organization*, 2025b. URL https://www.who.int/teams/global-influenza-programme/vaccines/who-recommendations.

WIC. Worldwide influenza centre: Annual and interim reports. *Worldwide Influenza Centre, The Francis Crick Institute*, 2024. URL https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports.

Wu, N. C. and Wilson, I. A. Influenza hemagglutinin structures and antibody recognition. *Cold Spring Harbor perspectives in medicine*, 10(8):a038778, 2020.

Youssef, N., Gurev, S., Ghantous, F., Brock, K. P., Jaimes, J. A., Thadani, N. N., Dauphin, A., Sherman, A. C., Yurkovetskiy, L., Soto, D., Estanboulieh, R., Kotzen, B., Notin, P., Kollasch, A. W., Cohen, A. A., Dross, S. E., Erasmus, J., Fuller, D. H., Bjorkman, P. J., Lemieux, J. E., Luban, J., Seaman, M. S., and Marks, D. S. Computationally designed proteins mimic antibody immune evasion in viral evolution. *Immunity*, 2025. ISSN 1074-7613. doi: https://doi.org/10.1016/j.immuni.2025.04.015.

A. Supplemental Methods

Observed influenza sequences in GISAID reflect evolution under selection constraints for functional and infectious viruses. Generative sequence models express the probability that a sequence x would be generated by this process as $\mathbf{P}(x|\theta)$, where the parameters θ capture the constraints describing functional variants. A generative model trained on a multiple sequence alignment of observed viral protein variants can then be used to estimate the relative plausibility of a given mutant sequence as compared to wild-type by using the log ratio of sequence likelihoods as a heuristic. For EVE (Frazer et al., 2021), since an exact computation of the log likelihood of a sequence is intractable, we approximate it with the Evidence Lower Bound (ELBO) loss used to optimize the VAE:

$$E_{EVE}(x) = -\log \frac{\mathbf{P}(x|\theta)}{\mathbf{P}(w|\theta)} \sim \text{ELBO}(w) - \text{ELBO}(x)$$

To contextualize EVE's predictive performance, we compared against several sequence-based baselines, all computed from the same clade-specific multiple sequence alignments (MSAs) used for EVE.

Position-Specific Scoring Matrices (PSSM) were constructed by computing the log-odds enrichment of each amino acid at each site. Mutation scores were defined as the log-ratio between the wild-type and mutant residue frequencies at that position, normalized by pseudocounts to avoid zero divisions. PSSM provides a simple, interpretable baseline that captures conservation at the individual residue level without modeling interactions between sites.

EVcouplings (Hopf et al., 2019) models protein mutational constraints using pairwise statistical couplings inferred from evolutionary covariation within an MSA, scoring each mutation by the change in statistical energy ΔE between the mutant and wild-type sequence.

EVEscape (Thadani et al., 2023) builds upon EVE's fitness scores with structural features (e.g., antibody accessibility) and physicochemical dissimilarity from the reference strain to derive an escape likelihood for each mutation. We did not change any model components for EVcouplings or EVEscape, we simply retrained on our influenza MSAs.

B. Supplemental Figures

Clade	Characteristic Mutations	Years Dominant	Reference Strain	GISAID EPI ID	Num Training Sequences	Neff Training Sequences	Num Test Sequences	Num Test Mutations
pdm09	root	2009, 2010	A/California/07/2009	EPI273609	7733	2097	11836	740
6	P100S, D114N, S202T, S220T, I338V, E391K, S468N	2011, 2012	A/St._Petersburg/27/2011	EPI319527	15701	2965	1253	178
6B	K180Q, A273T, K300E, E516K	2013, 2014, 2015	A/South_Africa/3626/2013	EPI466626	22886	3965	6495	404
6B.1	A13T, S101N, S179N, I233T	2015, 2016, 2017	A/Michigan/45/2015	EPI662594	32707	5361	10860	462
6B.1A	S91R, S181T, I312V	2017, 2018	A/Paris/1447/2017	EPI1142011	46144	6958	11042	521
6B.1A.5a	N146D, S200P, S202I, N277D	2018, 2019	A/Norway/3433/2018	EPI1328929	54587	7933	7947	367
6B.1A.5a.2	L8M, K147N, N173K, L178I, V267A, E523D	2019, 2020	A/Wisconsin/588/2019	EPI1661758	63911	8834	2820	133
6B.1A.5a.2a	K71Q, A203T, Q206E, E241A, R276K, K325R	2021, 2022, 2023, 2024	A/India/Pun-NIV323546/2021	EPI1916988	78361	10016	53848	935
6B.1A.5a.2a.1	P154S, K159R, N277E, T294A, E373D, S468H	2022, 2023, 2024	A/Norway/25089/2022	EPI2149322	82203	10362	19345	555

Figure 1. Summary information for each clade, including defining mutations, basal reference strain, and training/testing alignments summaries. Mutation numbering is relative to the length of the full HA sequence from 1-566.
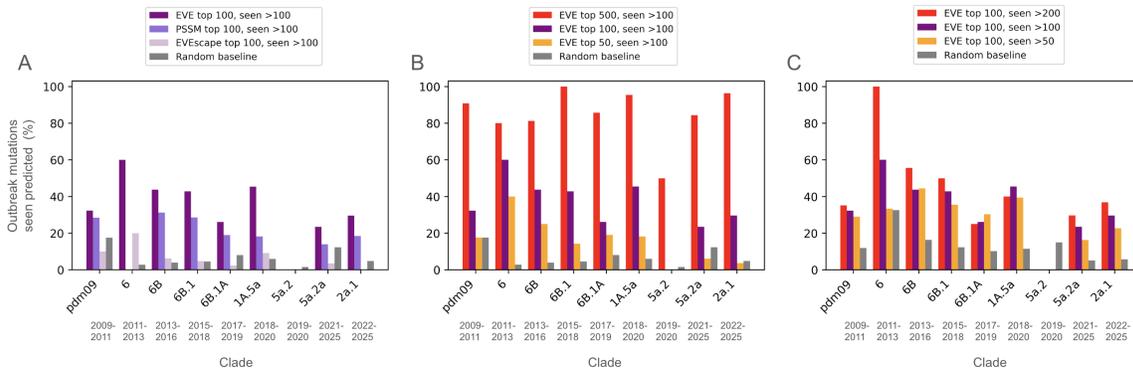
Figure 2. EVE summary statistics on full hemagglutinin, not subsetted to HA1 for DMS comparison as in Figure 3. Note that EVE top 100, seen $\geq$100 is consistent in purple across subfigures. **A**. Percentage of mutations seen $\geq$100x that are in the top 100 escape predictions across baselines. **B**. Percentage of mutations seen $\geq$100x that are in EVE's top 50, 100, and 200 escape predictions. **C**. Percent of mutations seen $\geq$50x, $\geq$100x, and $\geq$200x that are in EVE's top 100 escape predictions.
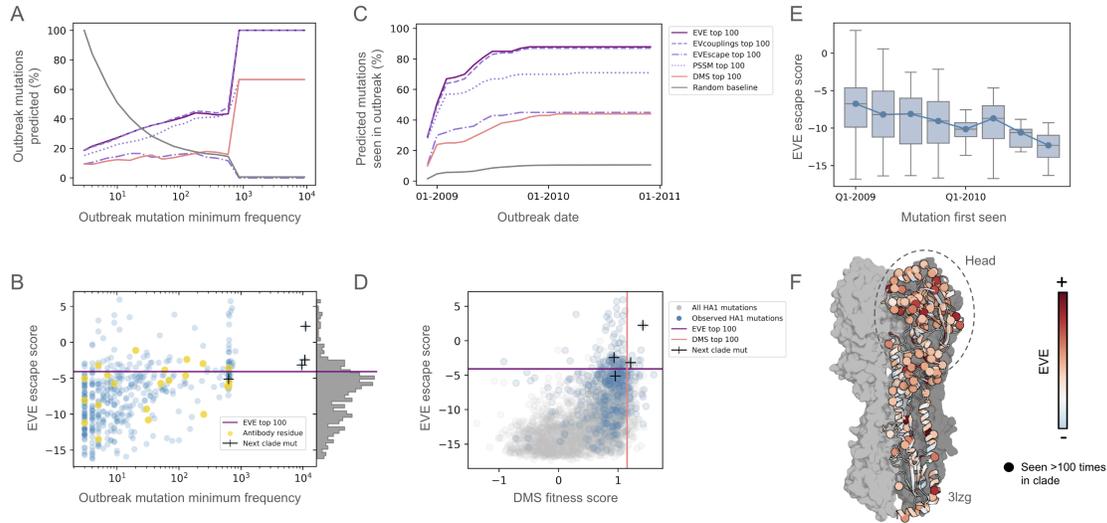
Figure 3. EVE accurately forecasts mutations seen after the 2009 H1N1 pandemic during the years after A/California/07/2009 emerged but before the first clade was defined. We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of the pdm09 model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2009 to 2011.

We note an anomaly in our pdm09 model, which starts out with 50% of its predicted escape mutations seen immediately on 2009-01-01 in the test dataset. We hypothesize that this is due to backfilled Collection Date annotations in GISAID—old sequences misentered as collected on 2009-01-01 although they were sequenced and submitted years later. Nine hundred sequences are listed as collected on January 1st, 2009 in GISAID, or equivalently 21% of the test mutations seen on January 1st itself, which is not likely to be true. Especially since none of our other models have this issue, we believe it is unique to an early time period where real-time sequence deposition was not yet normalized. We have updated the pdm09 model to stop training before 2008-12-31 so we can evaluate it on a test set starting from 2008-12-31 instead of 2009-01-01 like our other models.
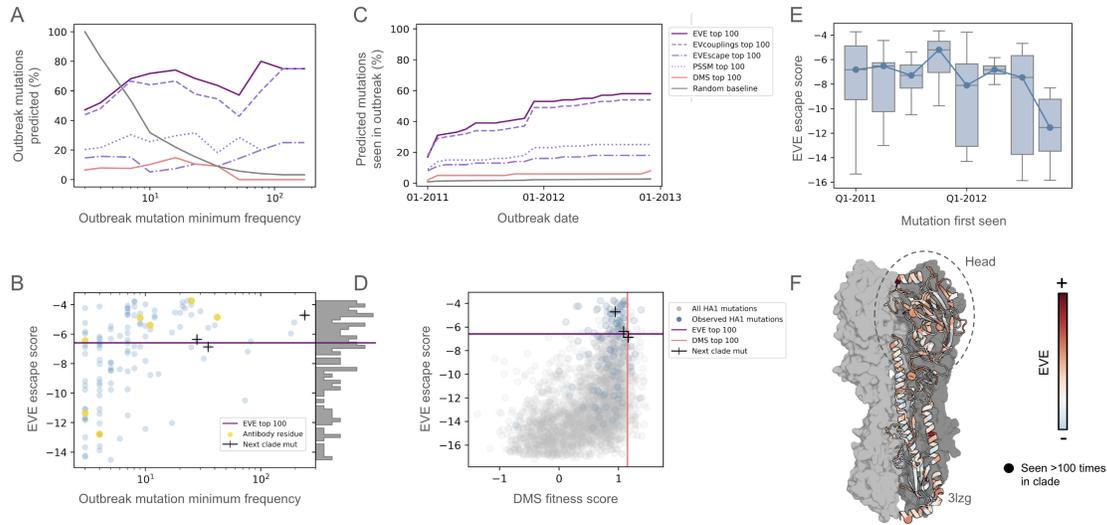
Figure 4. EVE accurately forecasts mutations seen in Clade 6 during the years it was dominant (2011-2013). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6 model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2011 to 2013.
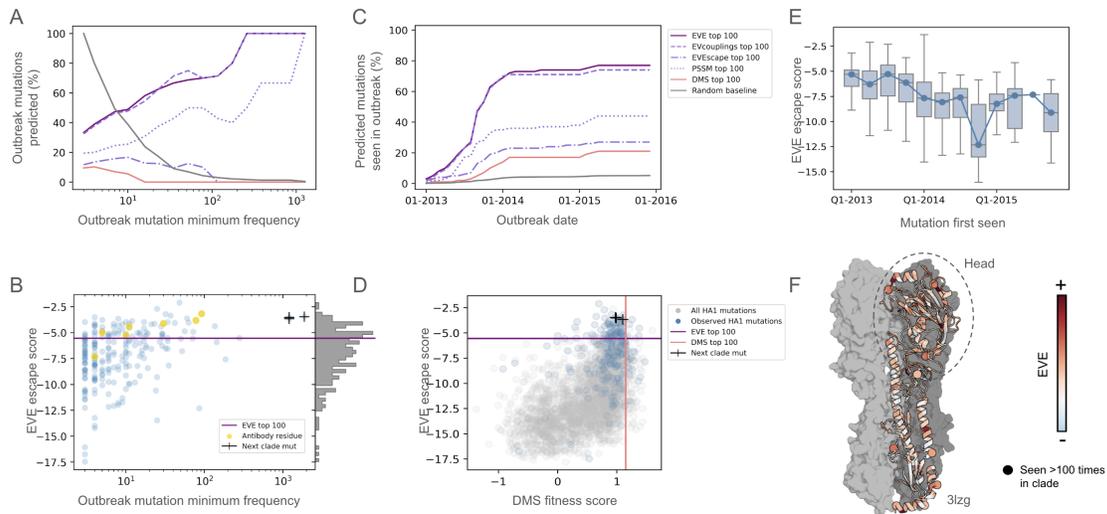
Figure 5. EVE accurately forecasts mutations seen in Clade 6B during the years it was dominant (2013-2016). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2013 to 2016.

Figure 6. EVE accurately forecasts mutations seen in Clade 6B.1 during the years it was dominant (2015-2018). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B.1 model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2015 to 2018.
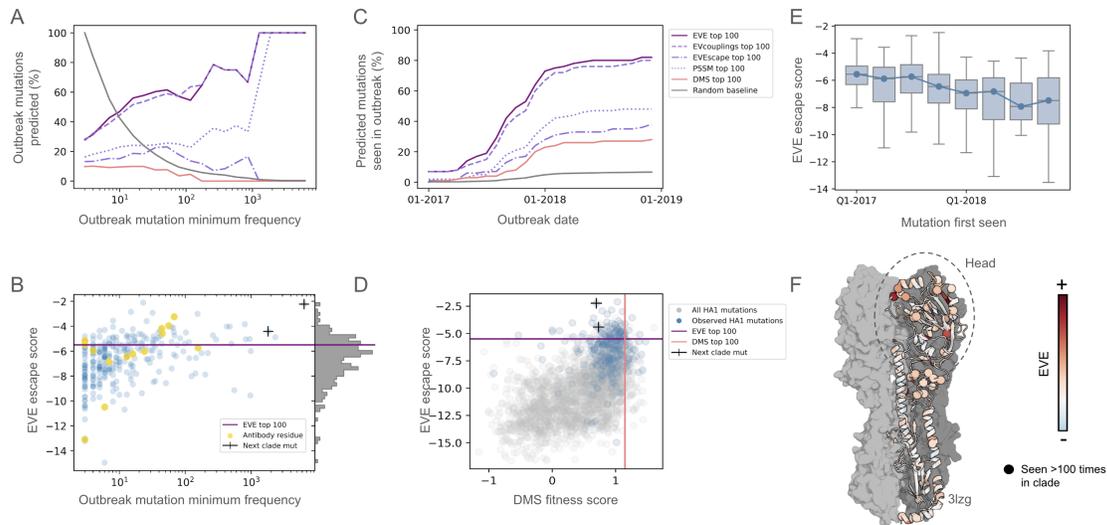
Figure 7. EVE accurately forecasts mutations seen in Clade 6B.1A during the years it was dominant (2017-2019). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B.1A model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2017 to 2019.
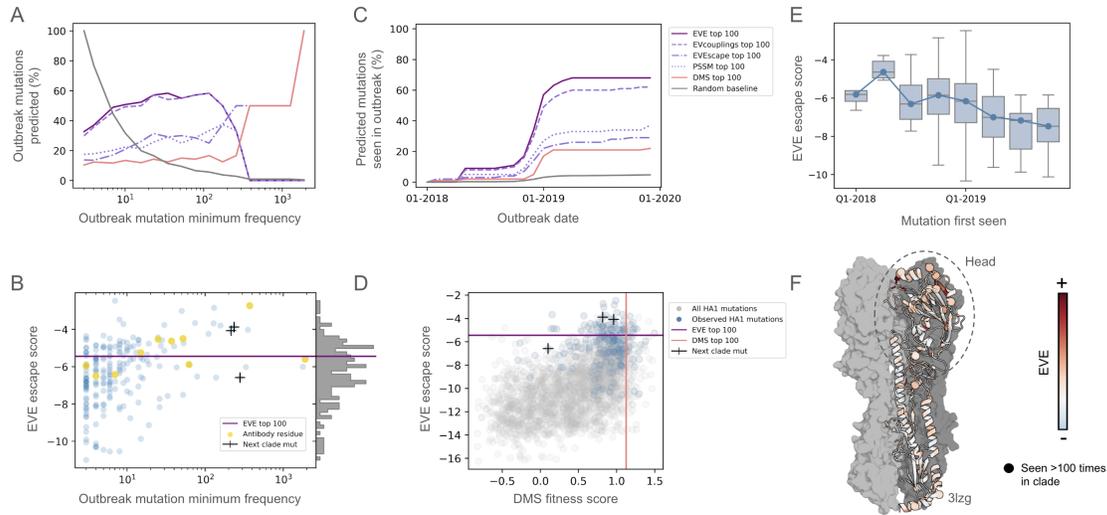
Figure 8. EVE accurately forecasts mutations seen in Clade 6B.1A.5a during the years it was dominant (2018-2020). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B.1A.5a model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2018 to 2020.
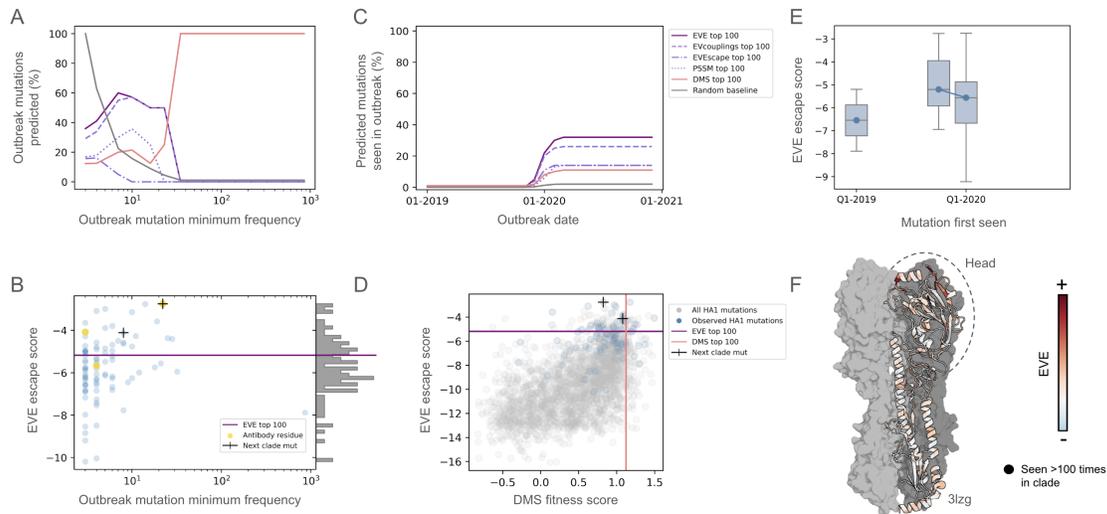
Figure 9. EVE accurately forecasts mutations seen in Clade 6B.1A.5a.2 during the years it was dominant (2019-2021). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B.1A.5a.2 model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2019 to 2021.

Figure 10. EVE accurately forecasts mutations seen in Clade 6B.1A.5a.2a during the years it was dominant (2021-2025). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B.1A.5a.2a model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2021 to 2025.
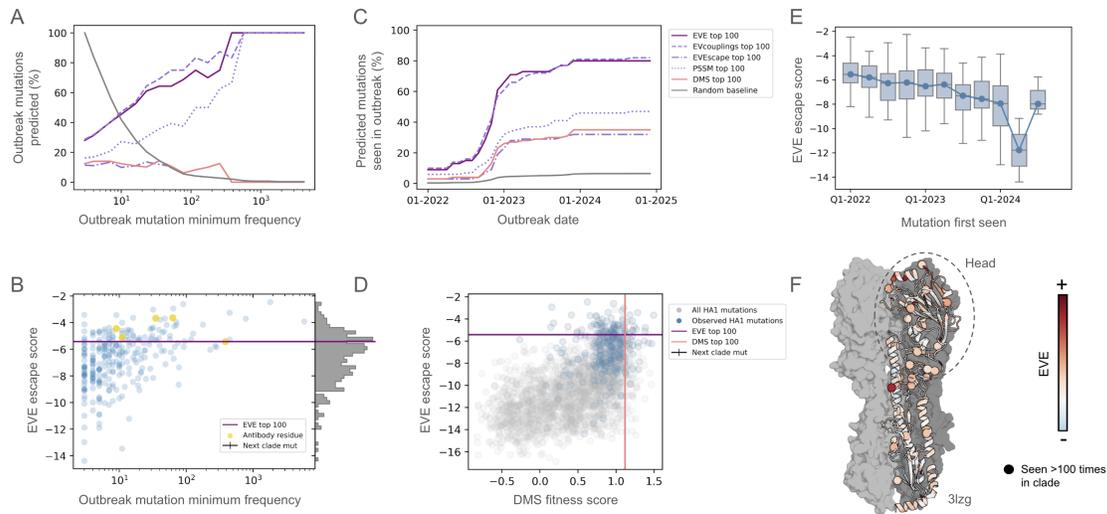
Figure 11. EVE accurately forecasts mutations seen in Clade 6B.1A.5a.2a.1 during the years it was dominant (2022-2025). We classify the top 100 mutations as model-predicted likely outbreak mutations. **A**. Percent of mutations seen in the clade that were predicted by EVE as a function of their minimum observed frequency. **B**. Distribution of mutations seen in the clade and their EVE scores. The top ten residues with the most number of antibodies bound in the PDB are colored in gold. **C**. Cumulative fraction of EVE-predicted mutations observed in the clade over time. **D**. Direct comparison of EVE vs DMS predicted escape mutations across all possible HA1 mutations and HA1 mutations seen in clade. **E**. Higher scoring EVE mutations are first seen earlier in the clade outbreak. **F**. EVE scores (site-level maximum) of Clade 6B.1A.5a.2a.1 model mapped onto a representative Hemagglutinin structure (PDB ID 3tzg). Spheres indicate mutations observed >100 times between 2022 to 2025.