Real-time Forecasting of Influenza Evolution

Aarushi Mehrotra[†]
Massachusetts Institute of Technology
aarushim@mit.edu

Navami Jain[†]
Harvard University
navami_jain@g.harvard.edu

Sarah Gurev*
Massachusetts Institute of Technology
sgurev@mit.edu

Noor Youssef*
Harvard Medical School
noor_youssef@hms.harvard.edu

Debora Marks*
Harvard Medical School
debora_marks@hms.harvard.edu

Abstract

Influenza A subtypes H1N1 and H3N2 are endemic in humans, with their high mutation rates requiring annual vaccine updates. However, vaccine strain selection currently relies on early neutralization assays and global surveillance data, resulting in mismatches between vaccine strains and circulating strains and ultimately reduce vaccine effectiveness. Accurately quantifying the impact of mutations on viral fitness and antigenicity could improve forecasting of emerging stains and enable more effective vaccine design. To address this, we track the phylogeny of H1N1 and H3N2 over the last 15 years to create forecasting benchmark datasets and evaluate state-of-the-art computational models for predicting their evolution. We find that computational models consistently outperformed experimental approaches at predicting new single mutations, with notable differences in performance across the strains. Finally, we study a period of high H1N1 incidence to explore how models can transfer learned evolutionary constraints across influenza subtypes. This work highlights the potential of deep learning models to forecast influenza evolution and support proactive vaccine design.

1 Introduction

Seasonal Influenza A viruses are responsible for epidemics infecting an estimated one billion people and causing hundreds of thousands of deaths each year[1], yet have the lowest and most variable effectiveness of any vaccine licensed for use in the US[2]. This variability arises from the virus's rapid evolution, which enables immune evasion and necessitates annual reformulation to match circulating strains. Current vaccine recommendations from the World Health Organization are based on surveillance sequencing and basic neutralization assays of a limited number of circulating strains. Yet, strain mismatch is common: over the past decade, only three influenza seasons demonstrated more than 70% antigenic match between the selected vaccine strain and the dominant strains six months later[3]. Such mismatches can significantly reduce vaccine effectiveness, emphasizing the need for improved predictive models of viral evolution.

Traditional serological assays [4–6], are resource-intensive and cannot feasibly measure neutralization across the vast diversity of circulating strains[7]. While computational methods offer scalability, many depend on large-scale antigenic data or current strain prevalence data—both of which have limited forecasting power[8–12]. An alternative approach is to leverage information encoded in the

[*] Senior authorship.
[†] Contributed equally.

Machine Learning in Structural Biology (2025).

evolutionary history of influenza (Fig. 1A,B), using the hundreds of thousands of sequenced strains from surveillance efforts[13–18].

Here, we systematically evaluate an array of computational models holding state-of-the-art performance on mutation effect prediction benchmarks[17, 19] on their ability to predict influenza A evolution, particularly for H1N1 and H3N2 clades from 2009 onward (Fig. 1A,B,C). The contrasting immunological landscapes between pandemic human-adaptation-driven H1N1 evolution[20, 21] and seasonal population-immunity-driven H3N2 evolution provide an ideal test of the model's generalizability. We evaluate 1) alignment-based models, which infer functional constraints from multiple sequence alignments (MSAs) of homologous proteins and have shown success in modeling viral evolution, immune escape, and vaccine effectiveness particularly for SARS-CoV-2 [14–16, 18, 22, 23]; and 2) PLMs, which are trained on a large corpora of protein sequences, offering an alignment-free approach which can, in principle, generalize across families[24–29], but have shown limited performance for viruses [17].

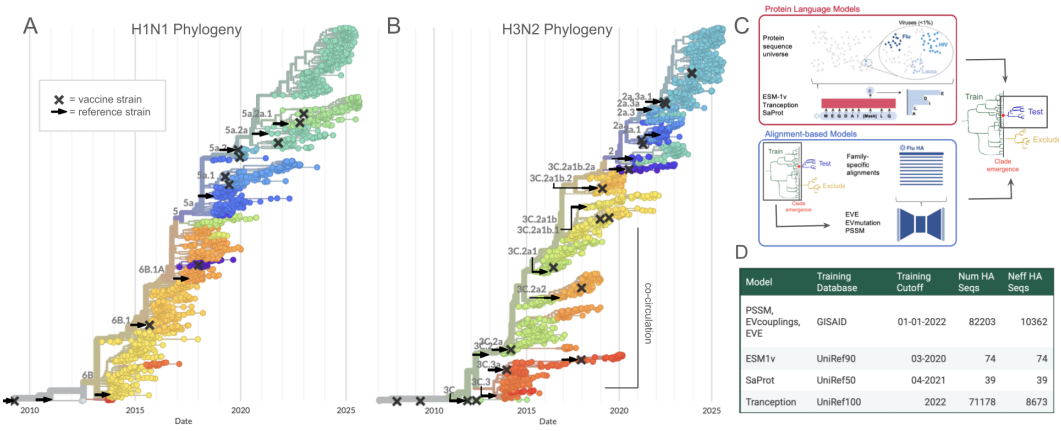


Figure 1: Modeling evolution of H1N1 and H3N2 clades with current computational models. **A-B.** Nested pandemic H1N1 phylogeny vs co-circulating seasonal H3N2 phylogeny since 2009, including the 9 H1N1 and 15 H3N2 clades we focus on modeling. Trees from NextStrain[30, 31] using data from GISAID[13]. **C.** Two classes of computational models, protein language models and alignment-based models, are evaluated. **D.** Representation of influenza in model training datasets.

Finally, we use the evolutionary model EVE to explore the subtype-level specificity of learned fitness constraints. We model the evolution of H1N1 following an antigenic shift event in 1989, and investigate how training on non-H1N1 subtypes changes prediction performance. Understanding how transferrable variant effect models are will shape strategies to model emerging subtypes with high risk of human spillover.

2 Results

We test three alignment-based models (PSSM[14], EVmutation[14], and EVE[15]); two PLMs (ESM-1v[24] and Tranception[25]); one structure-aware PLM (SaProt[32]); and four hybrid approaches (TranceptEVE[33], Tranception with MSA retrieval[25], VESPA[34], and SaProt-EVE[17]) (Fig. 1D). Here, we focus on the top performing models in each category and the most recent H1 and H3 clades (see S1,S2,S3,S4,S5,S6, for all). We focus on single nucleotide mutations since multi-nucleotide mutations are very rarely observed in flu clades post-2009 (<4% of test datasets).

We evaluate model performance by tracking how many mutations in each model's top 100 predictions are observed throughout a clade outbreak, defining precision as the percentage of predictions seen by the end of outbreak (Fig.S1,S2). Importantly, at the emergence of each clade, none of any model's top mutations were present in the population—confirming that test set mutations are novel rather than trivial memorization of mutations present at training time. However, we care not only that a model's predictions are seen but that it is able to predict the most frequently occurring mutations in a clade outbreak; we define a recall metric as the number of top 100 most frequent mutations that

were correctly predicted by the model. Crucially, all three top models were more likely to predict a mutation the more commonly it occurred (Fig. 2, Fig.S3,S4).

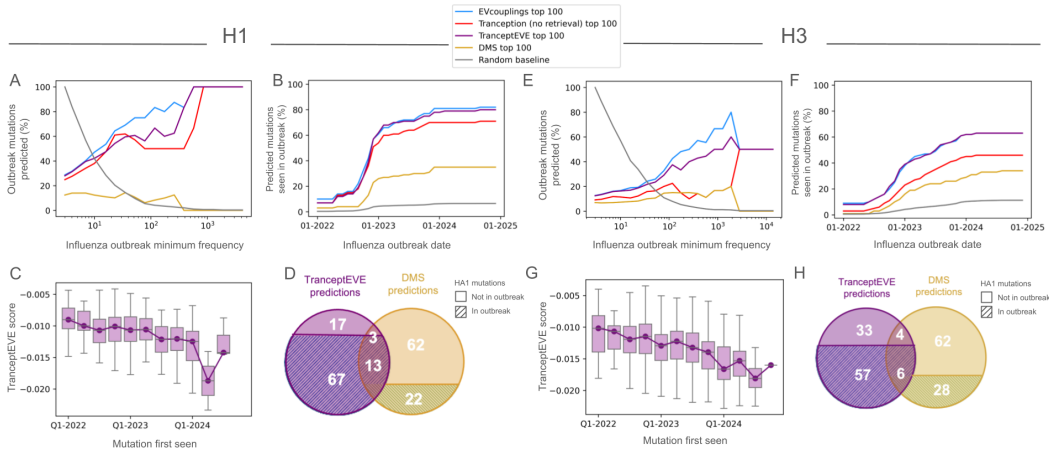
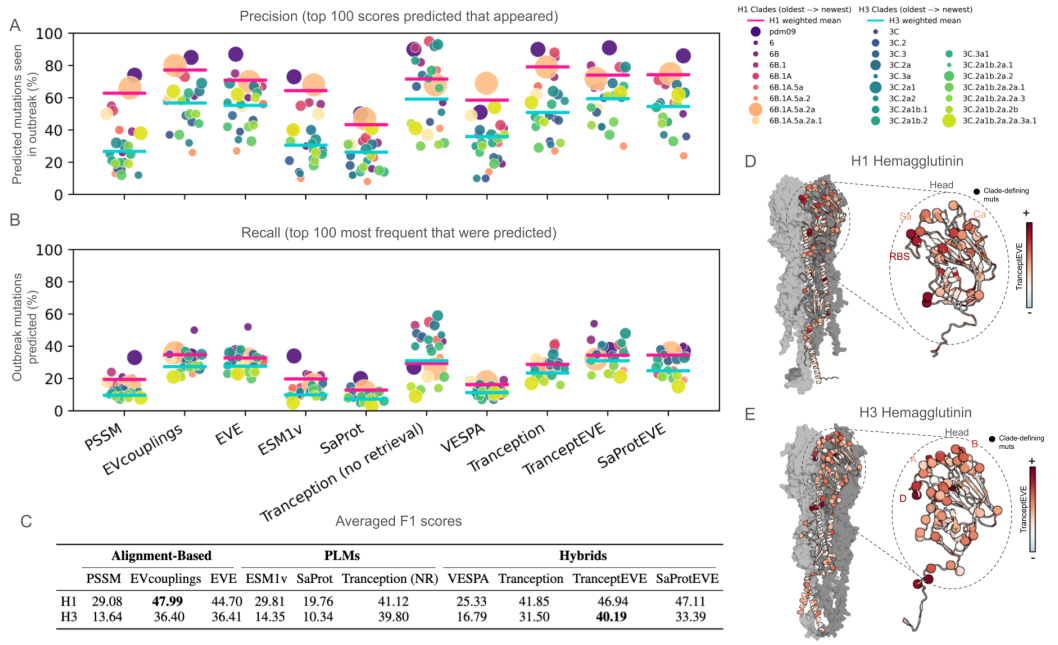


Figure 2: Computational models predict future outbreak HA1 mutations in the clades 6B.1A.5a.2a.1 (H1N1) and 3C.2a1b.2a.2a.3a.1 (H3N2) more accurately than deep mutational scans. A,E. Models are more likely to predict higher frequency mutations. **B,F.** By 2025, 80% of TranceptEVE's and EVcoupling's H1 predicted mutations and 60% of H3 predicted mutations were observed, outperforming every other model. **C,G.** Higher scoring TranceptEVE mutations are first seen earlier in the clade outbreak. **D,H.** TranceptEVE and DMS predict distinct sets of mutations.

This was not the case for the fitness DMS from H1N1 (A/California/7/2009) [35] and H3N2 (A/Massachusetts/18/2022) [36]. While the DMS had some predictive power, recovering <40% of mutations, it was unable to differentiate the most frequent mutations. This is reflected in how few predictions overlap between the DMS and TranceptEVE (16% for H1, 10% for H3, Fig. 2D,H). This suggests that while viral replication (H1N1 DMS[35]) and cell entry (H3N2 DMS[36]) may be good proxies for viral fitness, they are not as reflective of the mutation selection pressures in observed natural viral evolution as computational approaches are, which can recover double the outbreak mutations. This underscores a limitation of even up-to-date, *in vitro* fitness-only mutagenesis scans in forecasting real-world evolution. We also assess how different thresholds for top mutations affect precision and recall (Fig.S7,S8), with thresholds of 50, 100, and 200 giving comparable thresholds across the models.

TranceptEVE predicted at least 60% of the mutations seen for 7 of the 9 H1N1 clades(Fig.S1) and for 10 of the 15 H3N2 clades (Fig.S2). Notably, higher TranceptEVE scores corresponded to mutations that were seen earlier in the outbreak (Fig. 2C,G), suggesting that more fit or immune-evasive substitutions are preferentially identified by TranceptEVE. Median TranceptEVE scores decline over time as more mutations—often with weaker selection advantages—accumulate, consistent with expectations of mutation saturation over the course of clade evolution, with few mutations emerging in the later half of the clade's period of dominance. This is consistent with the plateau of predicted mutations seen after 2024 (Fig. 2B,F).

Across all models tested, we observe that larger clades have higher precision and that performance on H1N1 is stronger than H3N2 across all clades (Fig. 3A,B), despite slightly higher H3 representation in training sets. Among the protein language models tested in this paper, Tranception was the best performing (Fig. 3A,B,C), while SaProt and ESM1v did worse than the simplest alignment-based model, the position-specific scoring matrix—consistent with previous viral benchmarking studies [17]. This is unsurprising given that the model contains many more hemagglutinin sequences as a result of using Uniref100 as its training set (Fig. 1D), which has been shown to improve variant effect prediction for viruses [17]. Importantly however, we also see a drop-off in Tranception's performance that corresponds with the cutoff date of sequences in its training set (Fig. 1D,Fig.S1,S2). EVE and EVcouplings are the highest performing alignment-based models, performing on par with Tranception (Fig. 3C,S1,S2,S3,S4).



	Alignment-Based			PLMs			Hybrids			
	PSSM	EVcouplings	EVE	ESM1v	SaProt	Tranception (NR)	VESPA	Tranception	TranceptEVE	SaProtEVE
H1	29.08	**47.99**	44.70	29.81	19.76	41.12	25.33	41.85	46.94	47.11
H3	13.64	36.40	36.41	14.35	10.34	39.80	16.79	31.50	**40.19**	33.39

Figure 3: Models predict antigenically relevant, high frequency mutations in H1N1 and H3N2. **A.** Model precision and recall **B.** across all clades in H1N1 and H3N2. Clade datapoint size is proportional to number of total mutations. Clade datapoint color lightens from oldest to newest date of emergence. **C.** Averaged F1 scores weighted by clade size across all models. EVcouplings and TranceptEVE are the top performing. **D,E.** TranceptEVE scores (site-level maximum) of the first H1 and H3 models (clade pdm09 and 3C) mapped onto their AlphaFold HA structure highlight high-scoring regions, especially known epitopes within the head domain. Spheres indicate clade-defining mutations across all 9 or 15 clades.

Across models, higher scores are assigned to residues in the antigenically-relevant head rather than the conserved stem domain (Fig. 3D,E). For H1N1 and H3N2, top-ranked residues disproportionately fall within the canonically defined epitope regions and high-scoring residues often include clade-defining mutations that mark key antigenic transitions (Fig.S5,S6)). As an example, we highlight the S179N (S162N in H3 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[37]. EVE–without any training leakage–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 (Fig.S1). EVE predicted the first major immunogenic H1N1 mutation at the very start of the H1N1 pandemic, six years before its dominance in 2015.

Finally, we use the evolutionary model EVE to show that H1N1 evolution can be predicted during an earlier antigenic shift event in 1989 while relying on more than 10-fold fewer sequences (Table S5,Fig. S1,S2). Specifically, it can capture the most frequently occurring mutations while relying on sequences deposited prior to 1989 (Fig. 4A). Further, almost all the top predicted mutations from the alignment-based model EVE are seen over this two-decade period (Fig. 4B). All EVE models outperformed a DMS fitness assay performed on the 1989 H1N1 Siena strain strain[38].

A priority public health question is how evolutionary information from established influenza subtypes can be used to predict evolution in new emerging subtypes. To begin to answer this question, we examined how well non-H1N1 influenza evolution can model H1N1. Interestingly, excluding H1N1 sequences from the evolutionary model (Fig. 4A) led to a drop in the recall of the most frequent mutations. Moreover, exposing the model to future mutations seen in non-H1N1 sequences did not improve recall ability, in spite of the training set increasing by 100-fold. While this could suggest that distant influenza subtypes may not necessarily improve evolutionary prediction tasks, we also

see that each model is able to capture a distinct subset of the mutations frequently seen over the 20-year period. Interestingly, each method captured mutations occurring in antigenic regions of the HA protein ((Fig. 4C)), yet the overall performance of each is consistent. Understanding how to best leverage sequence data across influenza subtypes remains is an exciting future direction.

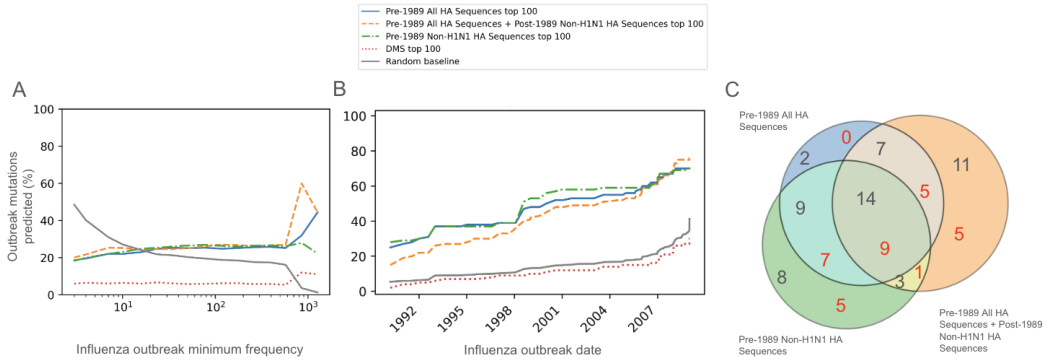


Figure 4: Models can recapitulate H1N1 evolution with very limited historical sequence data. **A.** Evolutionary models are able to predict higher frequency mutations better than DMS assays. This capacity dropped for models excluding historical H1N1 sequences. **B.** In the 20 years following 1989, close to 80% of the top predicted mutations were observed. **C.** The overlap of frequent mutations (seen > 100 times) that were predicted by each model. Gray represents the total number of mutations in each category, while red represents the total number in known antigenic regions.

3 Discussion

Our results demonstrate that the highest-performing models that can predict key mutations in the IAV HA protein are trained on historical evolutionary sequence data from UniRef100, GISAID, or both. However, PLMs suffered a decrease in performance for later clades where earlier HA strains represented in training sets become less relevant for variant effect prediction.

We show that the mutations prioritized by models before a clade's emergence are highly enriched for those that later become prevalent—even more so than deep mutational scanning assays—and align with known structural constraints on HA evolution[39]. Furthermore, we find that top performing models such as TranceptEVE are 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 (Fig.S5,S6).

While our work was shown to be applied to both pandemic strain and seasonal strains, each has distinct immunological pressures. A seasonal strain will have some immunological memory, and mutations that enable antibody escape, with greater differences in exposure histories and more complicated predictions over time. In contrast, a pandemic strain will initially accumulate mutations to adapt to humans, followed by drift to escape antibodies. Looking forward, we hope to incorporate information about existing population immunity in our mutation effect size calculations, building off EVEscape [40]. We are also working on an epistasis model to combine single mutation scores for better forecasting of entire vaccine strains. Finally, our results suggest subtype specificity of influenza sequences used in training shapes the performance of predictive models like EVE. The additon of cross-subtype sequences did not improve net recall or precision, but may still provide orthogonal information on how a strain could mutate. Whether more advanced architectures such as protein language models can better leverage cross-subtype information remains an open question.

Overall, this work offers a complementary strategy to existing models of influenza evolution, many of which require real-time surveillance data or large-scale experimental datasets. By leveraging only evolutionary sequence data, computational models enable forecasting of long-term evolutionary trajectories, offering a scalable data-efficient foundation for the future of vaccine strain selection.

Acknowledgements

The authors thank Tomas Lio Grudny, Pablo Cárdenas, and members of the Marks lab, as well as NextStrain and GISAID. This work was supported by the Coalition for Epidemic Preparedness Innovations (CEPI). The LaTeX template is heavily borrowed from LoG 2022.

References

[1] WHO. Influenza (seasonal). *World Health Organization*, 2025. URL https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). 1

[2] Amanda C Perofsky and Martha I Nelson. The challenges of vaccine strain selection. *Elife*, 9: e62955, 2020. 1

[3] Yu Jung Choi, Joon Young Song, Seong-Heon Wie, Won Suk Choi, Jacob Lee, Jin-Soo Lee, Young Keun Kim, Shin Woo Kim, Sun Hee Lee, Kyung-Hwa Park, 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. 1

[4] Caroline Kikawa, John Huddleston, Andrea N Loes, Sam A Turner, Jover Lee, Ian G Barr, Benjamin J Cowling, Janet A Englund, Alexander L Greninger, Ruth Harvey, et al. Near real-time data on the human neutralizing antibody landscape to influenza virus to inform vaccine-strain selection in september 2025. *bioRxiv*, pages 2025–09, 2025. 1

[5] Andrea N Loes, Rosario Araceli L Tarabi, John Huddleston, Lisa Touyon, Sook San Wong, Samuel MS Cheng, Nancy HL Leung, William W Hannon, Trevor Bedford, Sarah Cobey, 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.

[6] Caroline Kikawa, Andrea N Loes, John Huddleston, Marlin D Figgins, Philippa Steinberg, Tachianna Griffiths, Elizabeth M Drapeau, Heidi Peck, Ian G Barr, Janet A Englund, et al. High-throughput neutralization measurements correlate strongly with evolutionary success of human influenza strains. *bioRxiv*, pages 2025–03, 2025. 1

[7] 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. 1, 9

[8] Joseph K Agor and Osman Y Özaltın. Models for predicting the evolution of influenza to inform vaccine strain selection. *Human vaccines & immunotherapeutics*, 14(3):678–683, 2018. 1

[9] Cheng Gao, Feng Wen, Minhui Guan, Bijaya Hatuwal, Lei Li, Beatriz Praena, Cynthia Y Tang, Jieze Zhang, Feng Luo, Hang Xie, et al. Maivess: streamlined selection of antigenically matched, high-yield viruses for seasonal influenza vaccine production. *Nature Communications*, 15(1):1128, 2024.

[10] Jingzhi Lou, Weiwen Liang, Lirong Cao, Inchi Hu, Shi Zhao, Zigui Chen, Renee Wan Yi Chan, Peter Pak Hang Cheung, Hong Zheng, Caiqi Liu, et al. Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations. *Nature communications*, 15(1):2546, 2024.

[11] Maryam Hayati, Priscila Biller, and Caroline Colijn. Predicting the short-term success of human influenza virus variants with machine learning. *Proceedings of the Royal Society B*, 287(1924): 20200319, 2020.

[12] Wenxian Shi, Jeremy Wohlwend, Menghua Wu, and Regina Barzilay. Vaxseer: Selecting influenza vaccine strains through evolutionary and antigenicity models. 2024. 1

[13] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017. 2, 9

[14] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017. 2, 10, 11

[15] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021. 2, 10, 11, 12

[16] Nicole N Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J Rollins, Daniel Ritter, Chris Sander, Yarin Gal, and Debora S Marks. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, 2023. 2, 10

[17] Sarah Gurev, Noor Youssef, Navami Jain, and Debora S. Marks. Variant effect prediction with reliability estimation across priority viruses. *bioRxiv*, 2025. doi: 10.1101/2025.08.04.668549. URL https://www.biorxiv.org/content/early/2025/08/15/2025.08.04.668549. 2, 3, 11

[18] Noor Youssef, Sarah Gurev, Fadi Ghantous, Kelly P. Brock, Javier A. Jaimes, Nicole N. Thadani, Ann Dauphin, Amy C. Sherman, Leonid Yurkovetskiy, Daria Soto, Ralph Estanboulieh, Ben Kotzen, Pascal Notin, Aaron W. Kollasch, Alexander A. Cohen, Sandra E. Dross, Jesse Erasmus, Deborah H. Fuller, Pamela J. Bjorkman, Jacob E. Lemieux, Jeremy Luban, Michael S. Seaman, and Debora S. Marks. 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. 2

[19] Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, et al. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023. 2, 10, 11

[20] Patrick R Saunders-Hastings and Daniel Krewski. Reviewing the history of pandemic influenza: understanding patterns of emergence and transmission. *Pathogens*, 5(4):66, 2016. 2, 9

[21] Anna Otte, Anthony C Marriott, Carola Dreier, Brian Dove, Kyra Mooren, Thorsten R Klingen, Martina Sauter, Katy-Anne Thompson, Allan Bennett, Karin Klingel, 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. 2, 9

[22] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, October 2018. 2

[23] E. Laine, Y. Karami, and A. Carbone. Gemme: A simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.*, 36(11):1332, 2019. 2

[24] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021. 2, 10, 11

[25] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022. 2, 10, 11

[26] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

[27] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023.

[28] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. 10

[29] Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024. 2

[30] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018. 2

[31] Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018. 2

[32] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pages 2023–10, 2023. 2, 10, 11

[33] Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora S Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. December 2022. 2, 11

[34] Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, 141(10):1629–1647, 2022. 2, 11

[35] Tongyu Liu, William K Reiser, Timothy JC Tan, Huibin Lv, Joel Rivera-Cardona, Kyle Heimburger, Nicholas C Wu, and Christopher B Brooke. Natural variation in neuraminidase activity influences the evolutionary potential of the seasonal h1n1 lineage hemagglutinin. *Virus Evolution*, 10(1):veae046, 2024. 3

[36] Timothy C. Yu, Caroline Kikawa, Bernadeta Dadonaite, Andrea N. Loes, Janet A. Englund, and Jesse D. Bloom. Pleiotropic mutational effects on function and stability constrain the antigenic evolution of influenza hemagglutinin. *bioRxiv*, 2025. 3

[37] Jenna J. Guthmiller, Julianna Han, Lei Li, Alec W. Freyn, Sean T. H. Liu, Olivia Stovicek, Christopher T. Stamper, Haley L. Dugan, Micah E. Tepora, Henry A. Utset, Dalia J. Bitar, Natalie J. Hamel, Siriruk Changrob, Nai-Ying Zheng, Min Huang, Florian Krammer, Raffael Nachbagauer, Peter Palese, Andrew B. Ward, and Patrick C. Wilson. 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. 4

[38] Daniel P Maurer, Mya Vu, and Aaron G Schmidt. Antigenic drift expands viral escape pathways from imprinted host humoral immunity. *bioRxiv*, pages 2024–03, 2024. 4

[39] Nicholas C Wu and Ian A Wilson. Influenza hemagglutinin structures and antibody recognition. *Cold Spring Harbor perspectives in medicine*, 10(8):a038778, 2020. 5

[40] Nicole N. Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J. Rollins, Daniel Ritter, Chris Sander, Yarin Gal, and Debora S. Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023. 5

[41] WHO. Recommendations for influenza vaccine composition. *World Health Organization*, 2025. URL https://www.who.int/teams/global-influenza-programme/vaccines/who-recommendations. 9, 23

[42] James D Allen and Ted M Ross. H3n2 influenza viruses in humans: Viral mechanisms, evolution, and evaluation. *Human vaccines & immunotherapeutics*, 14(8):1840–1847, 2018. 9

[43] Barbara J Jester, Timothy M Uyeki, and Daniel B Jernigan. Fifty years of influenza a (h3n2) following the pandemic of 1968. *American journal of public health*, 110(5):669–676, 2020. 9

[44] Amanda C Perofsky, John Huddleston, Chelsea L Hansen, John R Barnes, Thomas Rowe, Xiyan Xu, Rebecca Kondor, David E Wentworth, Nicola Lewis, Lynne Whittaker, et al. Antigenic drift and subtype interference shape a (h3n2) epidemic dynamics in the united states. *Elife*, 13: RP91849, 2024. 9

[45] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pages 2022–02, 2022. 10

[46] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021. 11

[47] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992. 11

A Supplemental Methods

A.1 H1N1 Pandemic and Seasonal Evolution

The 2009 "swine flu" pandemic was a novel spillover of H1N1 from pigs, distinct from prior 1918 and 1977 H1N1 pandemics. Between these pandemics, the influenza virus evolved more slowly under antigenic drift [20].

Following the 2009 pandemic, up to 2 million lives were lost between May and December 2009 along. 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 α2,6-linked sialic acids, increasing replication in the respiratory tract, or elevating droplet transmission[21]. The H1N1 component 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[41]. 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.

A.2 H3N2 Seasonal Evolution

In 1968, a new H3N2 influenza strain (A/Hong Kong/1/1968 [HK/68]) first appeared, rapidly causing a global epidemic through 1972 that resulted in over one million fatalities worldwide [42, 43]. There is no evidence of H3N2 viruses circulating in humans prior to the epidemic. Since then, seasonal H3N2 has continually circulated in the human population, resulting in multiple epidemics and significant morbidity and mortality, more so than either H1N1 or influenza B. H3N2 viruses have also undergone antigenic change at higher rates than H1N1 viruses, and have lower vaccine effectiveness. Major antigenic transitions have since occurred, such as A/Sydney/5/1997-like strains and A/Fujian/411/2002-like strains. Following the 2009 H1N1 pandemic, H3N2-dominant seasons remained more frequent than H1N1 seasons, but H3N2 viruses showed reduced dominance compared to pre-2009 levels [44]. This decline in H3N2 predominance appears linked to increased genetic and antigenic diversification of H3N2, with multiple lineages with similar fitness co-circulating in each season, including vaccine strain updates most years. Antigenically distinct lineages 3C.2a and 3C.3a co-ciruclated beginning in 2012, and have since undergone further diversification.

A.3 Defining clades of H1N1 and H3N2

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 and H3N2 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[7]. Tracking the evolution of H1N1 and H3N2 from 2009 using these reports, we defined each emerging clade's characteristic mutations, noted which years it was dominant, and identified its basal, or earliest, strain. We used the basal strain as reference sequences 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. We examined the following 9 nested clade lineages for H1N1: pdm09, 6, 6B, 6B.1, 6B.1A, 6B.1A.5a, 6B.1A.5a.2, 6B.1A.5a.2a, and 6B.1A.5a.2a.1. We examine the following 15 clade lineages for H3N2: 3C, 3C.2, 3C.3, 3C.2a, 3C.3a, 3C.2a1, 3C.3a1, 3C.2a2, 3C.2a1b.1, 3C.2a1b.2, 3C.2a1b.2a.2, 3C.2a1b.2a.2a.1, 3C.2a1b.2a.2a.3, 3C.2a1b.2a.2b, 3C.2a1b.2a.2a.3a.1.

A.3.1 Assembling per-clade model training sets

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[13], comprising 365,500 full-length sequences submitted before January 1, 2025.

Alignment-based models rely on generating a multiple sequence alignment, or MSA. To therefore generate an MSA per clade, we filtered the full Influenza A dataset to include only sequences collected before that clade's emergence, then deduplicated and aligned them to the identified clade reference.

A.3.2 Assembling per-clade model test sets

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. To ensure this assumption would not mischaracterize strains, we verified the majority of strains assigned to a given clade contained at least 80% of the clade's ancestral mutations.

A.4 Alignment-based models

A.4.1 PSSM

Position-specific scoring matrix (PSSM) models assume each position in the protein evolves independently and assigns a prediction score for each mutation dependent on its frequency in the alignment. We used the site-wise maximum entropy model as implemented in[14].

A.4.2 EVmutation

To predict the effects of mutations that explicitly captures pairwise residue dependencies between positions, we used EVmutation as implemented in[14].

A.4.3 EVE

To predict the effects of mutations capturing high-order dependencies between positions, we used EVE, a Bayesian VAE model architecture, as implemented in[15]. We use single EVE models, rather than an ensemble of independent models as was reported in[16]. Note, that we use the negative of the evolutionary index reported by the model.

A.5 Protein language models

A.5.1 Tranception

Tranception [25] combines an autoregressive protein language model with inference-time retrieval from a MSA. We used Tranception Large (700M parameters) trained on UniRef100 using only the autoregressive inference without MSA retrieval as implemented in ProteinGym [19].

A.5.2 ESM-1v

ESM-1v [24] has a Transformer encoder architecture similar to BERT [Devlin et al., 2019] and was trained with a Masked-Language Modeling (MLM) objective on UniRef90. We use the implementation presented in ProteinGym [19] to handle sequences that are longer than the model context window (i.e., 1023 amino acids).

A.5.3 SaProt

SaProt [32] introduces a structure-aware vocabulary, into protein language modeling by training on Foldseek [45] 3Di tokens which represent the local geometric conformation information of each residue relative to its spatial neighbors. These 3Di tokens are combined with typical amino acid residue tokens as input to the SaProt model, which utilizes an ESM-2 Transformer architecture [28] but expands the embedding layer to encompasses 441 structurally-aware tokens instead of the original 20 amino acid residue tokens. We use both SaProt-650M-AF2, trained on approximately 40 million AF2 sequences/structures (from UniRef50) which notably excludes all viral proteins, and SaProt-650M-PDB, which continuously pre-trains the SaProt-650M-AF2 model on the PDB.

For structure inputs to Foldseek calculation, we fold monomeric forms of each DMS or WHO viral protein with AlphaFold3, where structures had not been folded previously in ProteinGym.

A.6 Hybrid models

A.6.1 SaProt-EVE

SaProt-EVE[17] combines the alignment-based, family-specific models with protein language models [17]. In this case, they use EVE and SaProt-PDB. Missing single mutation scores for EVE are first imputed based on the mean. Then, scores per model are standard scaled and shifted to be positive, and combined by taking the geometric mean.

A.6.2 VESPA

VESPA [34] combines the embeddings from ProtT5 [46] with a per-residue conservation prediction and supplements with BLOSUM substitution scores[47]. ProtT5 uses a T5 architecture which uses an encoder and decoder and was first trained on BFD and then finetuned on UniRef50.

A.6.3 Tranception with MSA retrieval

Tranception [25] combines an autoregressive protein language model with inference-time retrieval from a MSA. We used Tranception Large (700M parameters) trained on UniRef100 as implemented in ProteinGym. For retrieval, we use the same MSA (with the bit score and database chosen via the confidence metrics) as in alignment-based methods section. The MSA is not directly trained on, but retrieval inference uses the empirical distribution of amino acids observed across sequences in the MSA retrieved set of homologous sequences calculated via pseudocounts with Laplace smoothing[19]. Sequences are re-weighted as in[14]. The final log likelihood is a weighted average from autoregressive inference and the MSA log prior. The optimal aggregation coefficient was found to be 0.6 by grid search on a subset of DMSs.

A.6.4 TranceptEVE

TranceptEVE[33] is a hybrid method that combines Tranception[25] with an MSA-based EVE log prior[15]. The same EVE log prior is used to score all sequences of interest. Unlike Tranception, the aggregation coefficients used for both the EVE log prior and the MSA log prior are dependent on the depth of the retrieved MSA for a given protein family. If the protein family of interest has no or very few homologs, the autoregressive transformer is relied on, while if the MSA is deeper, the MSA and EVE log priors are weighted higher.

A.7 Mutation effect scoring

Generative models learn from the distribution of protein sequences collected as a result of billions of evolutionary experiments to capture the biochemical and structural constraints governing functional proteins. These models are trained to learn the distribution of natural, functional sequences.

For a given protein x composed of residues $(x_1, x_2, ..., x_L)$ the relative fitness of mutated protein compared to its wild-type can be calculated in the following ways depending on the modeling objective.

The fitness of a mutant sequence x^{mutant} is calculated as:

$$\log \frac{P(x^{\text{mutant}})}{P(x^{\text{wildtype}})}$$

For Tranception[25], an autoregressive model, the likelihood of x factorizes via the chain rule and is calculated as:

$$P(x) = \frac{1}{2}\Big[\prod_{i=1}^{L} P(x_i|x_{<i}) + \prod_{i=1}^{L} P(x_i|x_{>i})\Big]$$

In the masked language model setting, for ESM-1v [24] and SaProt [32], we use the masked marginal scoring function instead:

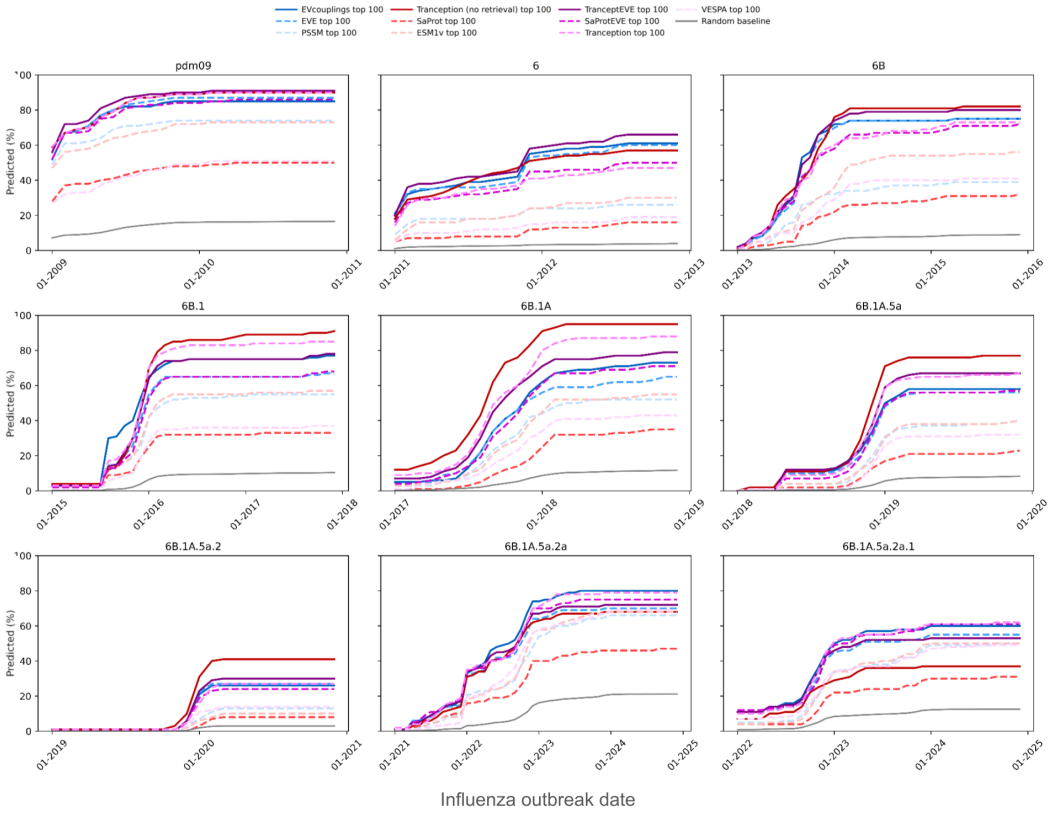
$$\sum_{i \in M} \log \frac{P(x_i = x_i^{\text{mutant}} \mid x_{-M})}{P(x_i = x_i^{\text{wildtype}} \mid x_{-M})}$$

where x_{-M} is the sequence x with masked residues at all mutated position M. Since we only consider single amino acid substitutions in this work, M contains only a single position.

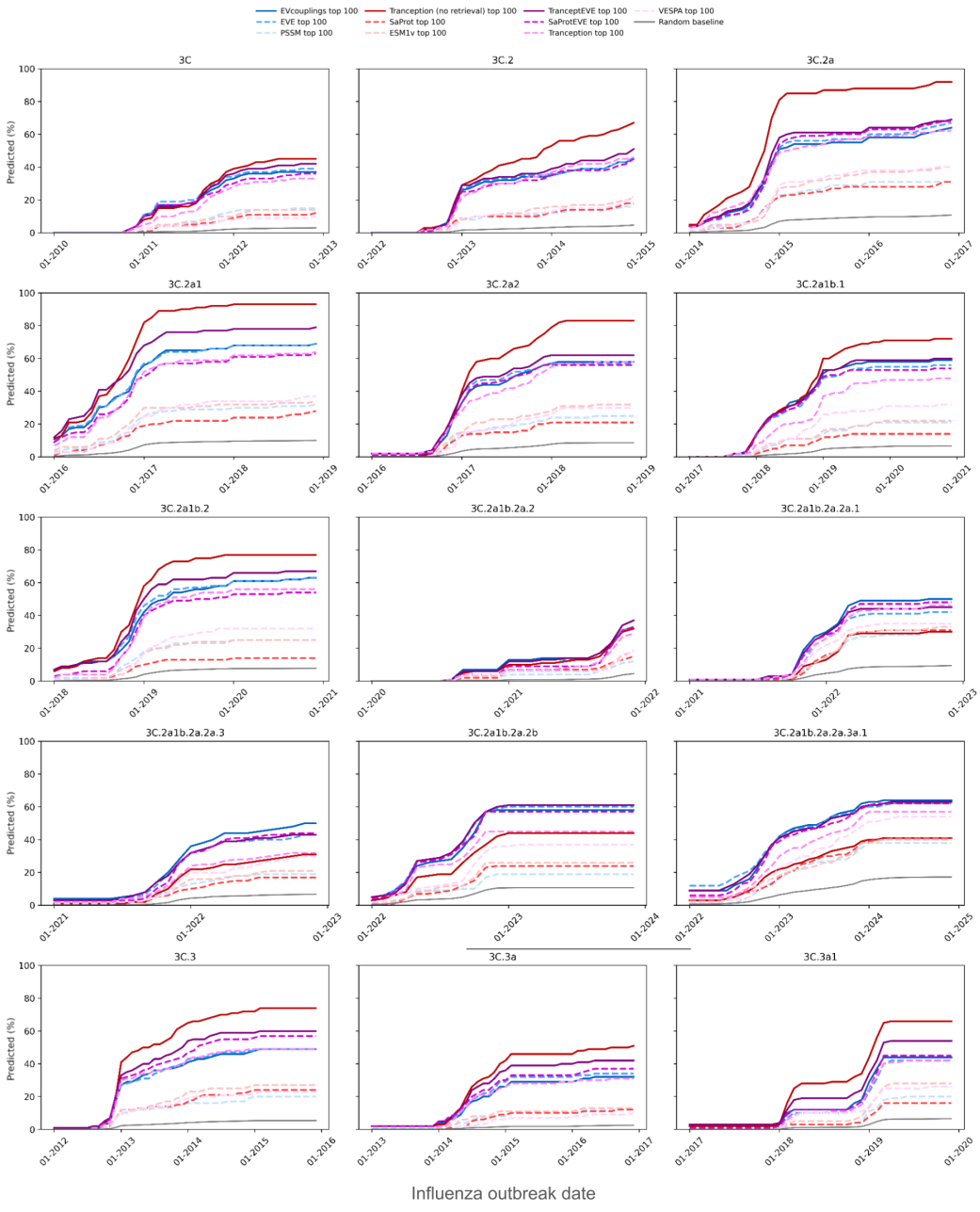
For a VAE, as in EVE[15], where the exact computation of log likelihood of a sequence is intractable, we approximate it with the Evidence Lower Bound (ELBO) used to optimize the VAE:

$$\log \frac{P(x^{\text{mutant}})}{P(x^{\text{wildtype}})} \approx ELBO(x^{\text{mutant}}) - ELBO(x^{\text{wildtype}})$$

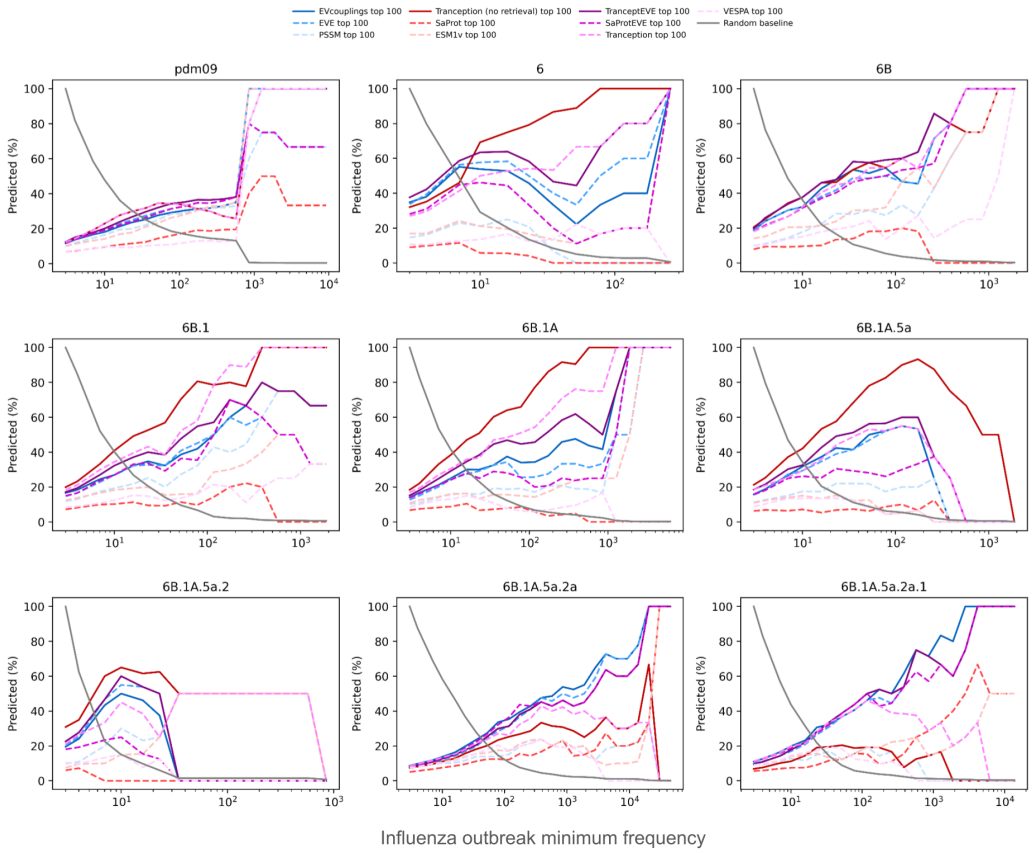
B Supplementary Figures



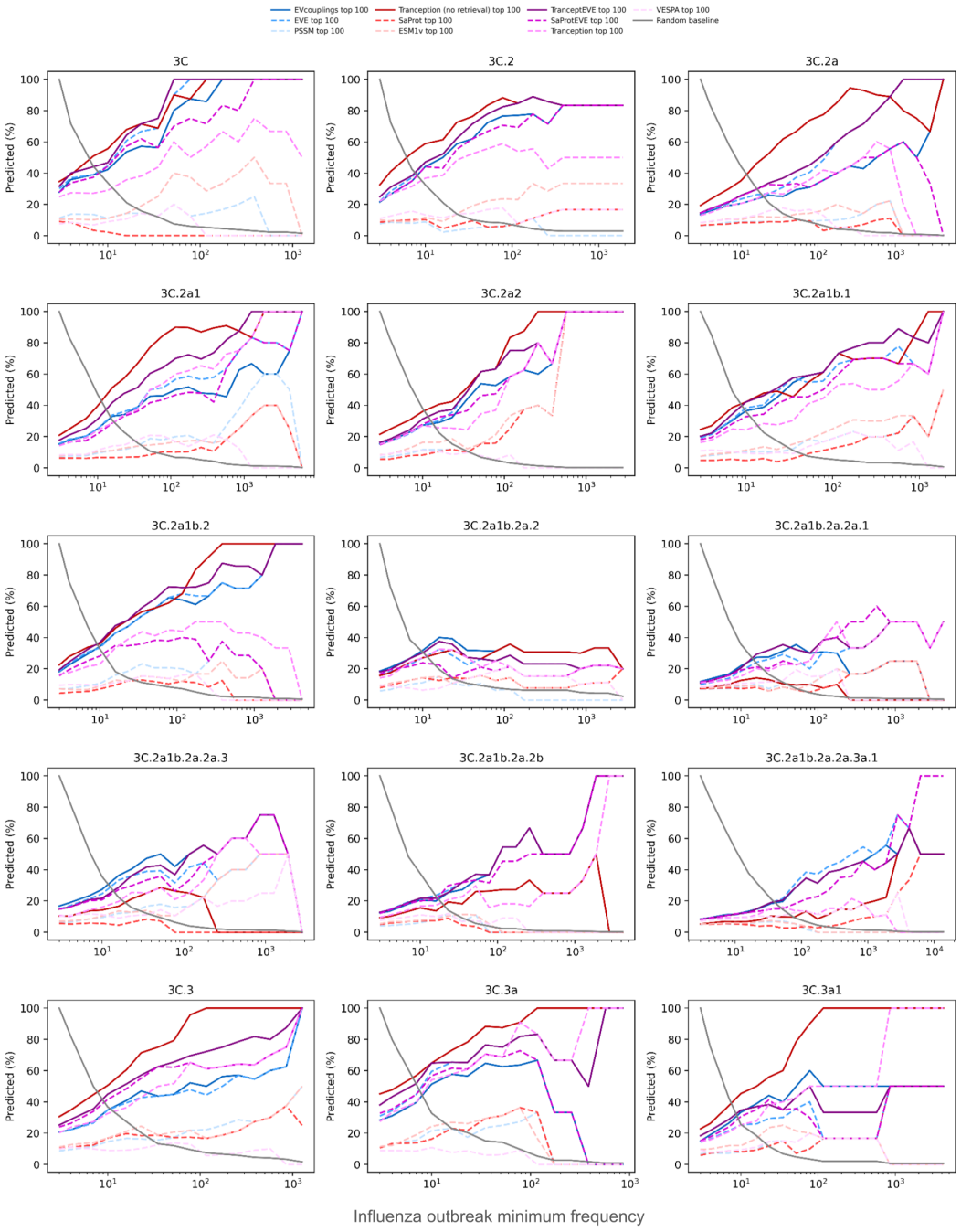
Supplementary Figure S1: Cumulative percentage of model-predicted mutations observed in each H1N1 clade over time.



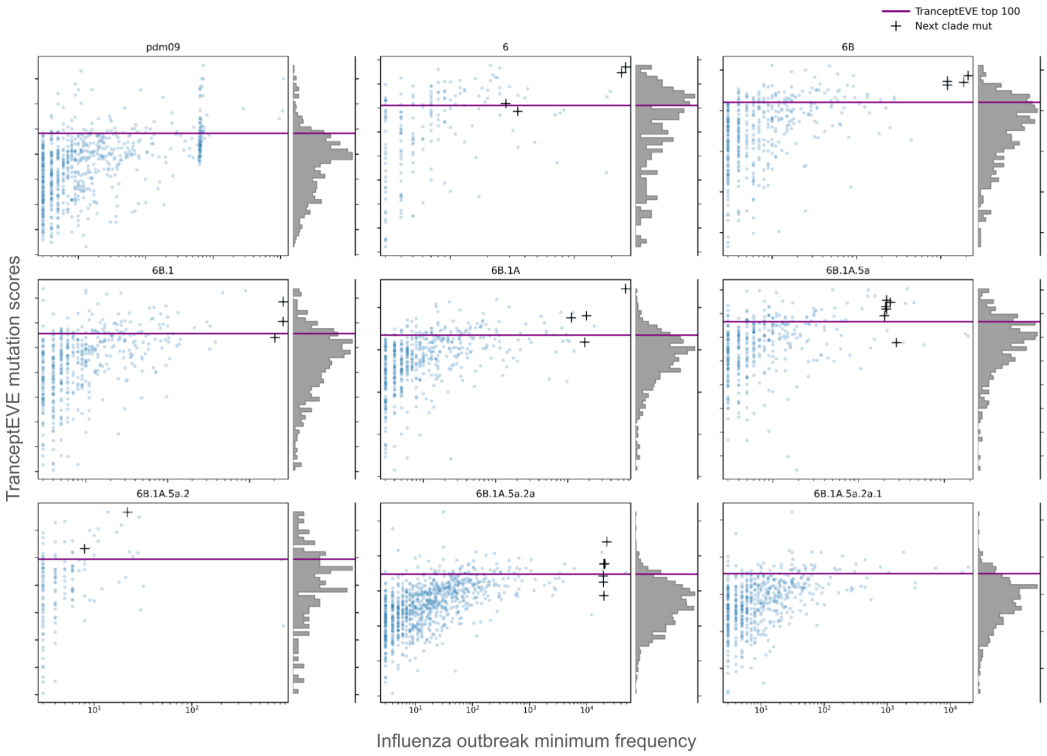
Supplementary Figure S2: Cumulative percentage of model-predicted mutations observed in each H3N2 clade over time.



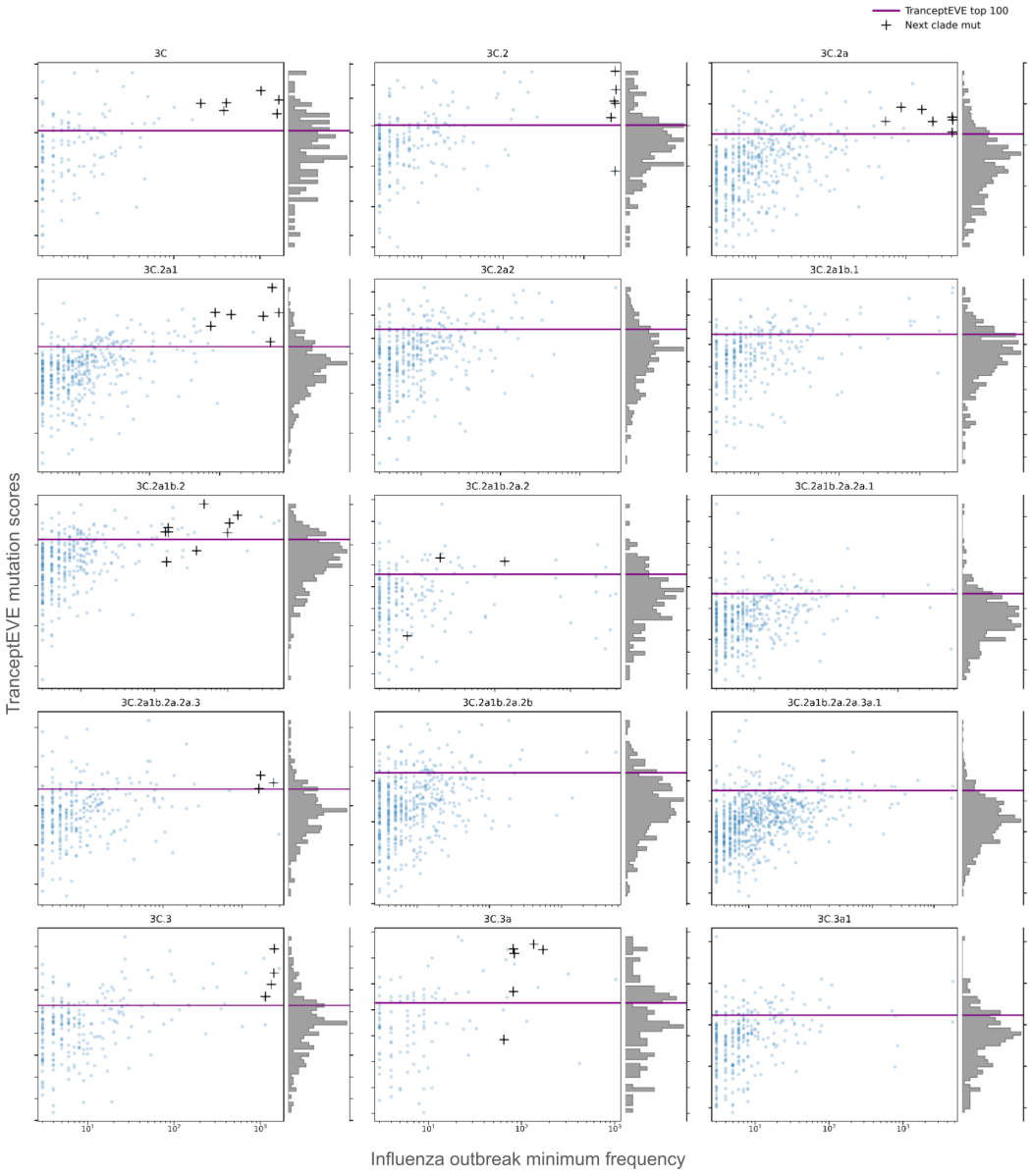
Supplementary Figure S3: Percent of highest frequency mutations in H1N1 clades that were predicted by models.



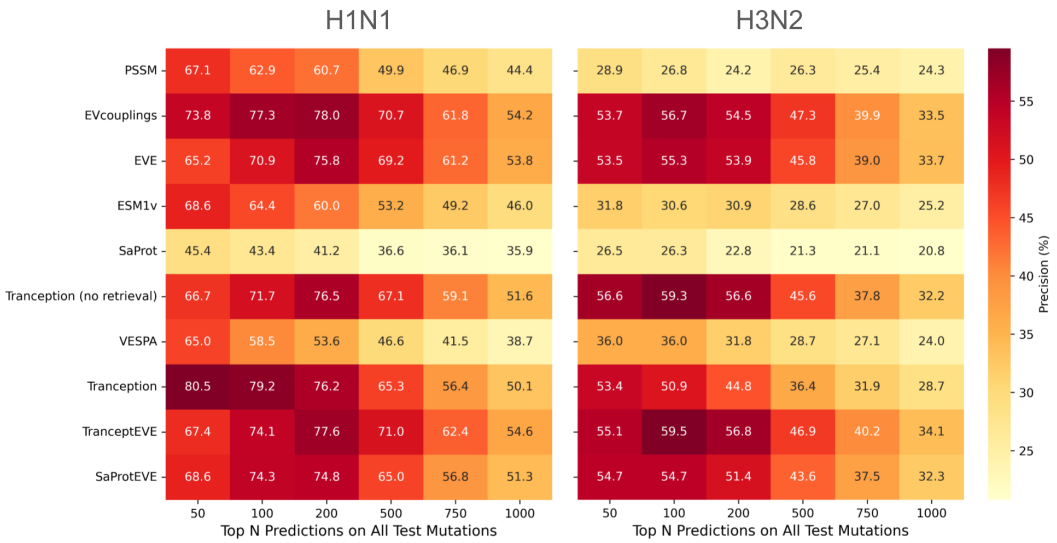
Supplementary Figure S4: Percent of highest frequency mutations in H3N2 clades that were predicted by models.



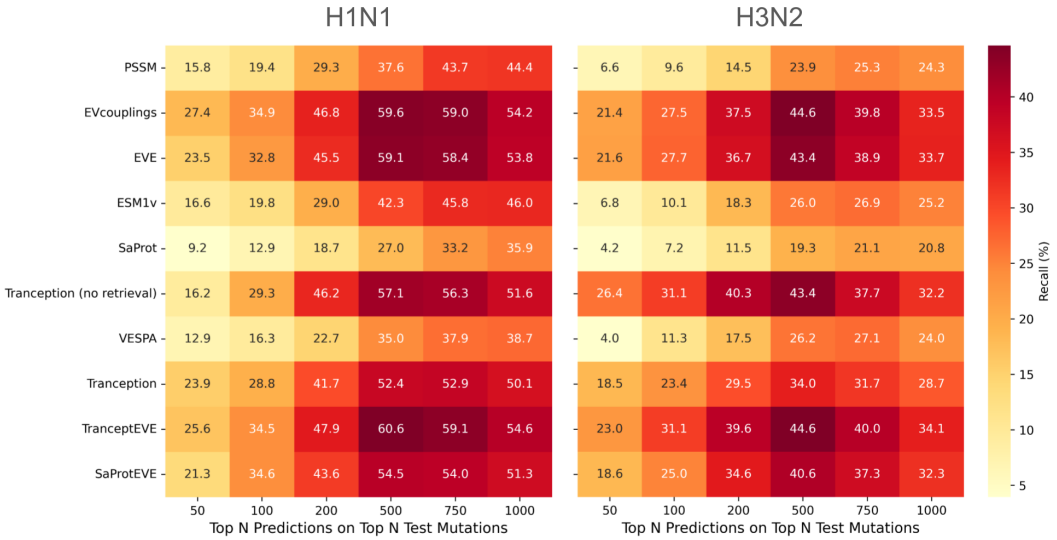
Supplementary Figure S5: Distribution of mutations seen in H1N1 clades and their TranceptEVE scores. High-scoring residues are often defining mutations for the subsequent clade (marked with a + sign). Prediction threshold (top 100) marked in purple.



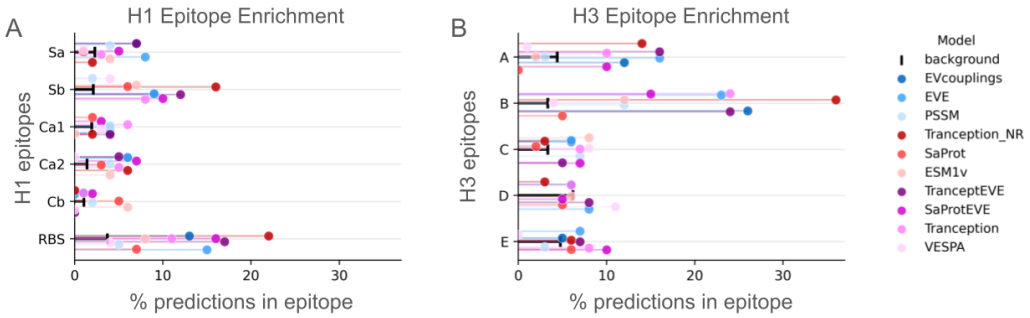
Supplementary Figure S6: Distribution of mutations seen in H3N2 clades and their TranceptEVE scores. High-scoring residues are often defining mutations for the subsequent clade (marked with a + sign). Prediction threshold (top 100) marked in purple.



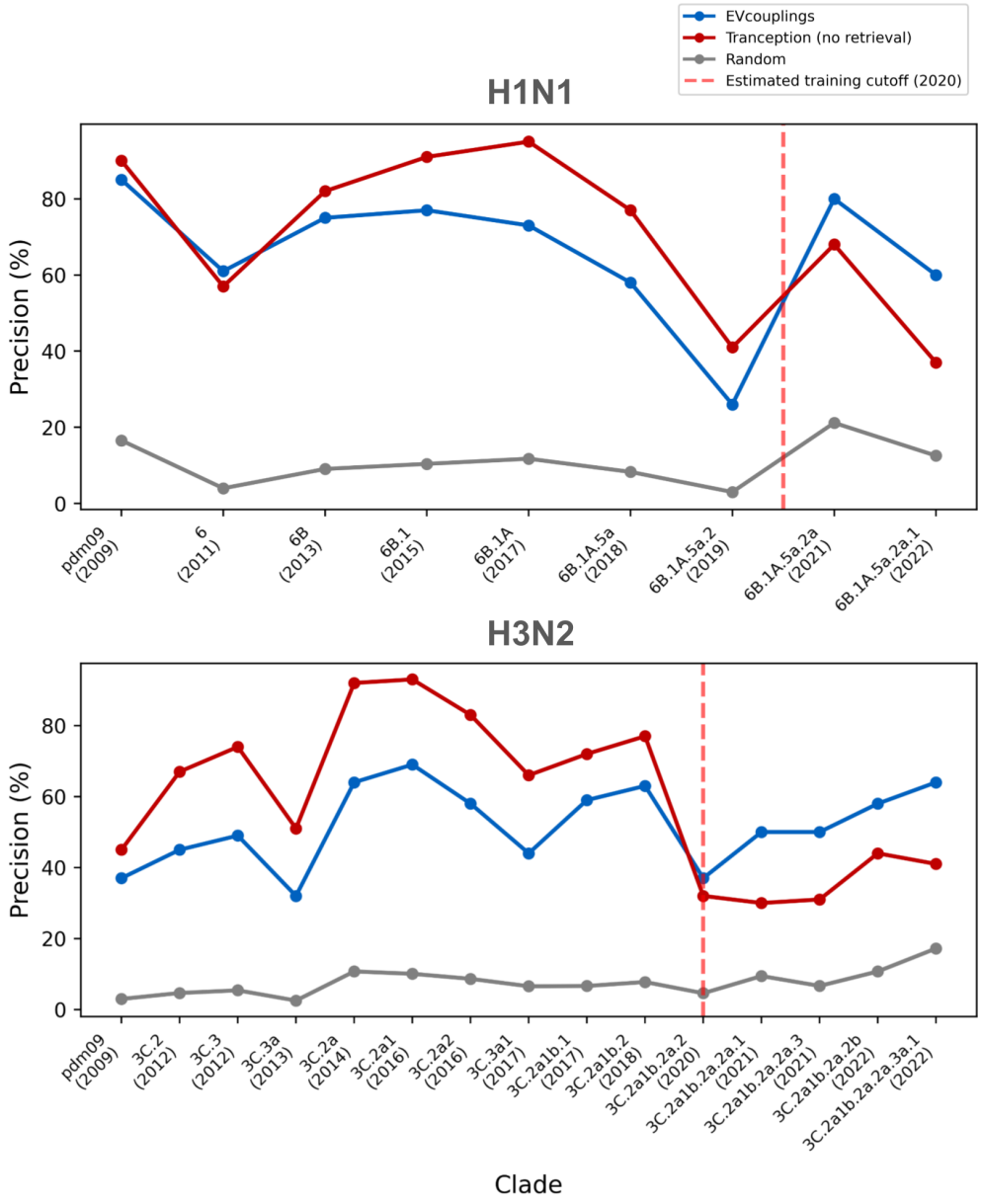
Supplementary Figure S7: Precision varies based on the threshold used for selecting top mutations. A threshold value of 100 was selected.



Supplementary Figure S8: Recall varies based on the threshold used for selecting top mutations. A threshold value of 100 was selected.



Supplementary Figure S9: TranceptEVE-predicted escape residues are consistently enriched in known head epitopes compared to a background expectation. The background distribution is based on epitope size.



Supplementary Figure S10: Performance of the PLM Tranception drops for later clades starting in 2020, which are likely to contain mutations not seen in the model's training set. Alignment-based methods such as EVCouplings overtake Tranception.

C Supplementary Tables

Clade	Characteristic Mutations	Years Dominant	Reference Strain	GISAID EPI ID	Num Training Sequences	Neff Training Sequences	Num Test Sequences	Num Unique Test Muts
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

Supplementary Table S1: Summary information for each H1N1 clade, including defining mutations, years dominant, basal reference strain, and training/testing alignments summaries. Mutation numbering is relative to the length of the full HA sequence from 1–566. Vaccine strains in bold.

Clade	Characteristic Mutations	Years Dominant	Reference Strain	GISAID EPI ID	Num Training Sequences	Neff Training Sequences	Num Test Sequences	Num Unique Test Muts
	root	2009	A/VICTORIA/208/2009	EPI272062				
3C	S61N, T64I, A214S, V239I, N328S	2010, 2011, 2012	A/Hong_Kong/3969/2011	EPI331093	12541	2524	1908	135
3C.2	Q49R, N161S, N294K, D505N	2012, 2013, 2014	A/BRISBANE/1/2013	EPI526663	19208	3461	3657	214
3C.3	Q49R, T144A, R158G, N161S, N294K, I156K	2012, 2013, 2014, 2015	A/Samara/73/2013	EPI460558	19206	3454	4974	253
3C.2a	L19I, N160S, F175Y, K176T, N241D, Q327H	2014, 2015, 2016	**A/Hong_Kong/4801/2014**	EPI539576	27345	4584	11759	490
3C.3a	A154S, F175S, N241D, K342R	2013, 2014, 2015, 2016	**A/Switzerland/9715293/2013**	EPI543763	22922	4014	1989	119
3C.2a1	N137K, N187K, I422V, G500E	2016, 2017, 2018	A/Oman/2585/2016	EPI769531	39185	6228	11439	465
3C.2a2	T147K, R158K, R277Q	2016, 2017, 2018	A/Norway/4465/2016	EPI868819	39216	6305	5688	396
3C.3a1	Y9C, L19I, S107N, N160K, F209S, I494M, D505N	2017, 2018, 2019	**A/Kansas/14/2017**	EPI1146345	46342	6920	6448	306
3C.2a1b.1	E78G, K108R, T144A, T151K, R158G, H327Q	2017, 2018, 2019, 2020	A/LaRioja/2202/2018	EPI1256086	46342	6920	5688	308
3C.2a1b.2	E78G, K108R, T147K, R158G, V545I, H327Q, A122V	2018, 2019, 2020	A/Norway/3275/2018	EPI1328875	54933	7912	6751	362
3C.2a1b.2a.2	K99E, Y110N, I538M, F209S, Y211F, Y9N, F175N, K176I, L180Q, G202D, D206N	2020, 2021, 2022	A/Bangladesh/4005/2020	EPI1838303	74018	9397	6554	217
3C.2a1b.2a.2a.1	H172S, D69G, D120G, K292R	2021, 2022	A/Slovenia/8720/2022	EPI2047241	78521	9813	15620	442
3C.2a1b.2a.2a.3	H172S, D69N, N112S, I208F, N394S	2021, 2022	A/Norway/24873/2021	EPI1958358	78557	9821	8370	321
3C.2a1b.2a.2b	E66K, F95V, I156K, F541S	2022, 2023	A/Georgia/02/2022	EPI2176976	82304	10262	12543	503
3C.2a1b.2a.2a.3a.1	T3A, E66K, I156K, I239V	2022, 2023, 2024	**A/Massachusetts/18/2022**	EPI2096148	82302	10219	27743	802

Supplementary Table S2: Summary information for each H3N2 clade, including defining mutations, years dominant, basal reference strain, and training/testing alignments summaries. Mutation numbering is relative to the length of the full HA sequence from 1–566. Vaccine strains in bold.

Flu Seasons	Recommended Vaccine Strain	Vaccine Clade	Dominant Clades
2010–2016	A/California/07/2009	pdm09	pdm09, 6, 6C, 6B, 6B.1
2017–2019	A/Michigan/45/2015	6B.1	6B.1, 6B.1A, 6B.1A.5a
2019–2020	A/Brisbane/02/2018	6B.1A.1	6B.1A.5a, 5a.1, 5a.2
2020–2021	A/Guangdong-Maonan/SWL1536/2019*, A/Hawaii/70/2019**	6B.1A.5a.1	5a.1, 5a.2
2021–2023	A/Victoria/2570/2019*, A/Wisconsin/588/2019**	6B.1A.5a.2	5a.1, 5a.2a, 5a.2a.1
2023	A/Sydney/5/2021	6B.1A.5a.2a	5a.2a, 5a.2a.1
2023–2025	A/Victoria/4897/2022*, A/Wisconsin/67/2022**	6B.1A.5a.2a.1	5a.2a, 5a.2a.1

Supplementary Table S3: Recommendations for influenza vaccine composition, H1N1 component, from the World Health Organization [41] (southern and northern hemisphere influenza seasons combined).

Flu Seasons	Recommended Vaccine Strain	Vaccine Clade	Dominant Clades
2008—2010	A/Brisbane/10/2007		
2010—2012	A/Perth/16/2009		
2012—2014	A/Victoria/361/2011	3C	3C.3, 3C.2, 3C.2a
2014—2015	A/Texas/50/2012	3C.1	3C.3, 3C.2a, 3C.3a, 3C.2
2015—2016	A/Switzerland/9715293/2013	3C.3a	3C.2a
2016—2018	A/Hong Kong/4801/2014	3C.2a	3C.2a, 3C.2a1, 3C.2a2, 3C.2a3, 3C.2a1b.1
2018—2019	A/Singapore/INFIMH-16-0019/2016	3C.2a1	3C.2a2, 3C.2a1b.1, 3C.2a1b.1b
2019	A/Switzerland/8060/2017	3C.2a2	3C.2a1b.1, 3C.2a1b.1b, 3C.2a1b.2, 3C.2a1b.2b
2019—2020	A/Kansas/14/2017	3C.3a1	3C.3a1, 3C.2a1b.1, 3C.2a1b.2a, 3C.2a1b.2b
2020	A/South Australia/34/2019	3C.2a1b.2	3C.2a1b.1a, 3C.2a1b.1b, **1**, **2**
2020—2021	A/Hong Kong/2671/2019*, A/Hong Kong/45/2019**	3C.2a1b.1b	3C.2a1b.1a, **1a**, **2**, **2a**, **2a.3**
2021—2022	A/Cambodia/e0826360/2020	3C.2a1b.2a.**1a**	**2a**, **2a.1**, **2a.3**
2022—2024	A/Darwin/9/2021*, A/Darwin/6/2021**	3C.2a1b.2a.**2a**	**2b**, **2a.3a.1**
2024—2025	A/Thailand/8/2022*, A/Massachusetts/18/2022**	3C.2a1b.2a.**2a.3a.1**	**2a.3a.1**

Supplementary Table S4: Recommendations for influenza vaccine composition, H3N2 component, from the World Health Organization [41] (southern and northern hemisphere influenza seasons combined).

Model	Reference Strain	Num Training Sequences	Num Unique Test Muts
Pre-1989 All HA Sequences	A/Siena/10/1989	733	882
Pre-1989 All HA Sequences + Post-1989 Non-H1N1 HA Sequences	A/Siena/10/1989	78498	882
Pre-1989 Non-H1N1 HA Sequences	A/Siena/10/1989	542	882

Supplementary Table S5: Summary information of models trained to predict evolution of the H1N1 A/Siena/10/1989 strain.