Evaluating H5N1 Vaccine Durability using Computationally-Designed Proteins

Anonymous Authors[1]

Abstract

The ongoing outbreaks of highly pathogenic avian influenza H5N1 viruses, particularly genotype D1.1 in clade 2.3.4.4b, and associated spillover events pose a concerning pandemic threat to humans. While immunity from seasonal flu exposure may confer limited cross-protection, H5N1 viruses are antigenically distinct and can evade this immunity. Moreover, while H5N1 vaccine stockpiles exist, they were developed against older clades and are a high mutational distance from currently circulating strains. Thus, they would confer only partial protection. We must ensure a new H5N1 vaccine provides sufficient protection against standing and future antigenic diversity likely to develop in a human population. To this end, we computationally design a panel of H5N1 proteins, VaxVal, to evaluate a candidate vaccine's durability and breadth. We show that deep learning models trained on historical Influenza hemagglutinin sequences can forecast close to 80% of the mutations that have occurred in clade 2.3.4.4b. Using these models, we successfully design 22 hemagglutinin variants, each carrying 2-4 mutations, that reflect antigenic changes across the protein. We show that constructs can easily escape protection by D1.1 vaccination and escape known broadly neutralizing monoclonal antibodies, sometimes close to 10-fold more than the wildtype. In forecasting immune escape, our pipeline can guide the design of broadly protective, long-lasting vaccines.

1. Introduction

A major challenge in mitigating viral outbreaks is anticipating how viral pathogens will evolve under immune pressure. Vaccines and other antiviral therapeutics are typically evaluated against currently circulating or historically observed strains, yet rapidly evolving RNA viruses can acquire mutations that reduce vaccine-induced protection. This challenge is particularly acute for highly pathogenic avian influenza viruses such as H5N1, which continue to circulate globally across animal reservoirs and pose an ongoing zoonotic and pandemic threat (Kok et al., 2025). Although current vaccine evaluation frameworks assess breadth against known viral diversity, they provide limited ability to prospectively evaluate protection against variants that have not yet emerged.

Influenza hemagglutinin (HA) protein represents an especially challenging target for vaccine design because immune pressure continuously drives antigenic evolution (Choi et al., 2024). Mutations within HA antigenic sites can substantially alter antibody recognition while preserving viral fitness, enabling immune escape and reducing vaccine effectiveness (Dadonaite et al., 2024). Traditional experimental approaches for assessing escape potential, including serial viral passaging or deep mutational scanning, are resource intensive and often limited to specific antibodies and viral backgrounds (Kikawa et al., 2025a; Loes et al., 2024; Kikawa et al., 2025b). A scalable framework capable of prospectively modeling plausible future HA evolution and directly evaluating vaccine robustness against these trajectories would substantially strengthen pandemic preparedness efforts.

Recent advances in computational evolutionary modeling have created new opportunities to forecast viral evolution directly from sequence data. Alignment-based generative models and protein language models (PLMs) trained on natural protein sequences can capture evolutionary constraints governing viral fitness and immune escape (Mehrotra et al., 2025). These models have demonstrated the ability to identify mutations enriched during viral evolution and predict functional consequences of sequence variation across diverse viral families (Gurev et al., 2025). Here we focus on H5N1 Avian influenza and to show how computationally predicted evolutionary trajectories can be translated into practical experimental tools for vaccine evaluation (Fig. 1a).

We first show that both alignment-based models and protein language models can successfully capture observed mutations in avian influenza HA, and to a much higher de-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

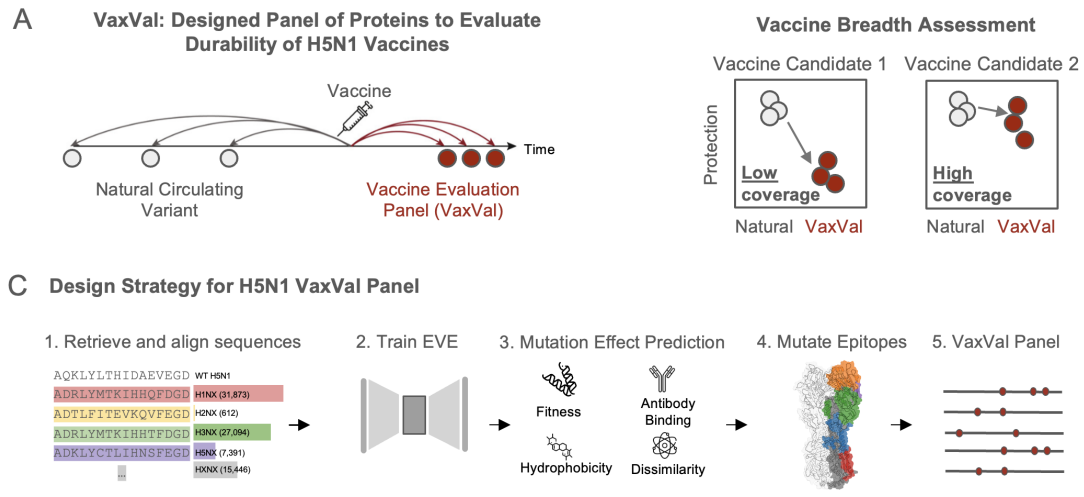

Figure 1. **Designed panel of H5 hemagglutinin proteins for vaccine evaluation.** A) Current vaccine evaluation approaches assess protection against circulating or past variants, which poorly predict efficacy against future viral evolution. To overcome this limitation, we developed VaxVal, a computationally designed panel of hemagglutinin variants that serves as a proxy for likely future evolutionary trajectories, enabling proactive vaccine evaluation. B) VaxVal panel design for H5N1 genotype D1.1. We trained the deep learning model EVE on historical hemagglutinin sequences predating the emergence of the D1.1 strain. We selected candidate mutations by integrating multiple constraints: high predicted fitness scores, known antibody binding, hydrophobicity, and dissimilarity between the mutant and wildtype amino acid. The final panel comprises 22 variants representing plausible future mutations.

gree than experimental deep mutational scans. We then use these models to computationally design VaxVal, a panel of 22 H5 hemagglutinin variants that include evolutionarily plausible combinations of mutations (Fig. 1b). Using this VaxVal panel, we evaluated the breadth of antibody responses elicited by an H5N1 vaccine candidate (D1.1). We identify multiple designed variants exhibiting reduced recognition by vaccine-induced antibody responses, demonstrating that computationally generated proteins can reveal potential escape pathways before their emergence in circulating viruses. By making this experimentally validated HA panel publicly available, we establish a scalable platform for prospective vaccine stress-testing that complements traditional surveillance and antigenic characterization approaches. More broadly, this framework provides a path toward evaluating vaccine resilience against future viral evolution rather than solely against past and present diversity.

2. Results

2.1. Computational models forecast H5N1 evolution

The extent to which computational models can be used to predict future variants for early vaccine evaluation, depends on their ability to predict how a virus will evolve. We first assessed whether computational models can predict the evolution of a recent viral clade. We used descendants of the recent D1.1 (A/American Wigeon/South Carolina/22-000345-001/2021) strain (emergence in November 2021) as the test set. We evaluated six models: (i) the alignment-based model EVE (Frazer et al., 2021); (ii) Ensemble ESM-1v, a sequence-only protein language model (PLM) (Meier et al., 2021); (iii) Tranception, a sequence-only PLM trained on UniRef100 (Notin et al., 2022); (iv) two versions of SaProt, a structure-aware PLM trained on either AlphaFoldDB (SaProt-AF) or further refined using structures from the Protein Data Bank (SaProt-PDB) (Su et al., 2023); and (v) Tranception with MSA retrieval (Notin et al., 2022) and (vi) VESPA (Marquet et al., 2022).

First, we assessed whether models could recall the most frequently observed mutations (Fig. 2a). Because the vast majority of available influenza sequences are derived from the H1 and H3 subtypes (Fig. 1a), we evaluated EVE models trained on datasets with different subtype specificities. We found that EVE robustly recalled the most frequently occurring mutations regardless of the training dataset used, though the prediction set did not completely overlap (S1). EVE correctly identified all mutations observed more than 100 times in the test set. In contrast, protein language models (PLMs) showed substantially more variable performance (Fig. 2a). Among PLMs, Tranception outperformed SaProt-AF, SaProt-PDB, ESM-1v, and the hybrid model VESPA. Incorporating H5N1-specific multiple sequence alignment retrieval into Tranception further improved performance, yielding recall comparable to EVE.

Next, we assessed the extent to which models could recall mutations that emerged over time. Our highest-performing models, EVE and Tranception with retrieval, captured approximately 80% of all mutations observed during the sub-

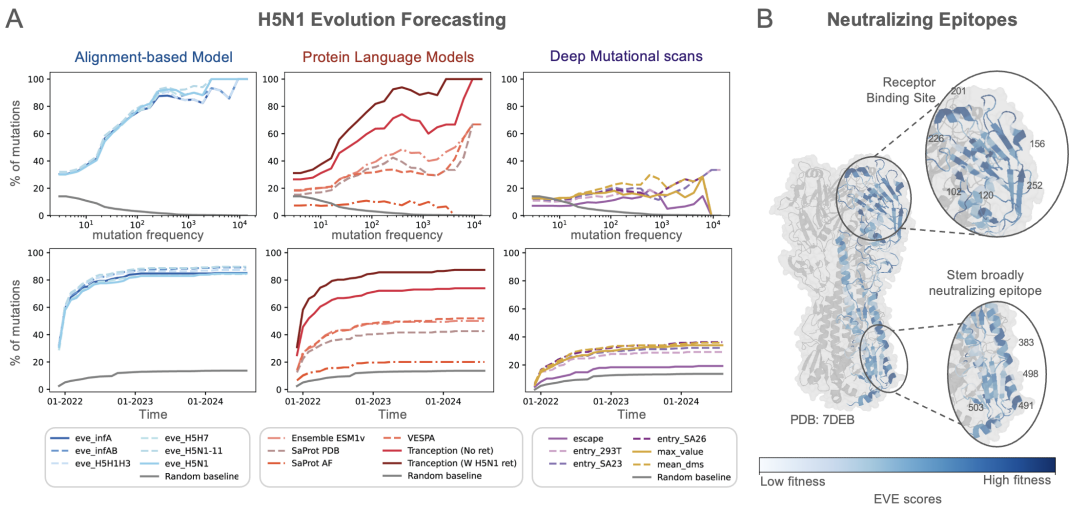

Figure 2. **Computational models forecast H5N1 hemagglutinin evolution.** A) Performance of alignment-based models, protein language models, and deep mutational scans (DMS) in predicting the most frequently observed mutations (top) and recalling mutations that arose following the emergence of the H5N1 D1.1 lineage (bottom). The alignment-based model EVE and the hybrid model Tranception with retrieval (which integrates a protein language model with alignment-derived evolutionary information) achieve the highest performance across both tasks. DMS experiments measured distinct phenotypic components of viral fitness (e.g., escape from sera, entry into different human cell lines). DMS scores were also aggregated per mutation using either a max or mean operator across all measured phenotypes. B) EVE scores are enriched in antigenically relevant regions proximal to the receptor-binding domain, as well as within a stem epitope targeted by broadly neutralizing antibodies.

sequent two years of D1.1 evolution. We note that perfect recall is not expected, as some mutations may arise under weak or near-neutral selection.

In contrast to computational models, deep mutational scanning (DMS) experiments measuring individual phenotypes recalled fewer than 40% of mutations observed during the evolution of the D1.1 lineage. The DMS datasets quantified four phenotypic readouts: (i) escape from polyclonal sera and (ii–iv) viral entry across three distinct cell lines. Each assay captured only a subset of observed evolutionary variation, suggesting that no single phenotype is sufficient to explain viral evolutionary dynamics. Because viral fitness likely reflects a composite of multiple selective pressures, we next evaluated whether aggregating measurements across assays improved predictive performance. Specifically, we computed per-mutation maximum and mean scores across all phenotypic readouts. However, neither aggregation strategy substantially improved recall (Fig. 2a).

Given these results, we selected EVE trained on all Influenza A hemagglutinin sequences for VaxVal panel design. As shown in Fig. 2b, EVE scores highlight known antigenically relevant regions across the protein, particularly epitopes proximal to the receptor-binding domain and within the conserved stalk domain.

2.2. VaxVal Panel Identifies Escape from D1.1 Vaccination

In order to design a comprehensive protein evaluation panel for vaccine protection, we must identify prominent antigenic regions on the hemagglutinin protein. To do this, we identified all known antibody-binding footprints (defined as residues within 5 Å of antibody atoms) for influenza A hemagglutinin. We then spatially clustered all antibody-contacting residues into five distinct epitopes (Fig. 3a). Three epitopes were located in the HA head region (E1–E3) and two in the stem region (E4–E5). We used EVE, trained on all influenza A hemagglutinin sequences, to generate all possible double, triple, and quadruple mutants within each epitope, sampled sequentially across sites. Mutants were ranked using a composite score incorporating predicted mutational fitness, hydrophobicity and charge dissimilarity relative to wild-type residues, and solvent accessibility as a proxy for antibody exposure. Designs containing mutations within 10 Å of one another were excluded to avoid spatially clustered substitutions. A final set of 22 VaxVal constructs was selected based on this ranking (Fig. 3b).

We used the VaxVal panel to evaluate the breadth of protection elicited by D1.1 immunization. Three rabbits were immunized with D1.1, receiving a primary immunization followed by a booster dose after two weeks; sera were collected one week after boosting. Results from an initial subset of 11 VaxVal constructs are shown in Fig. 3c, along

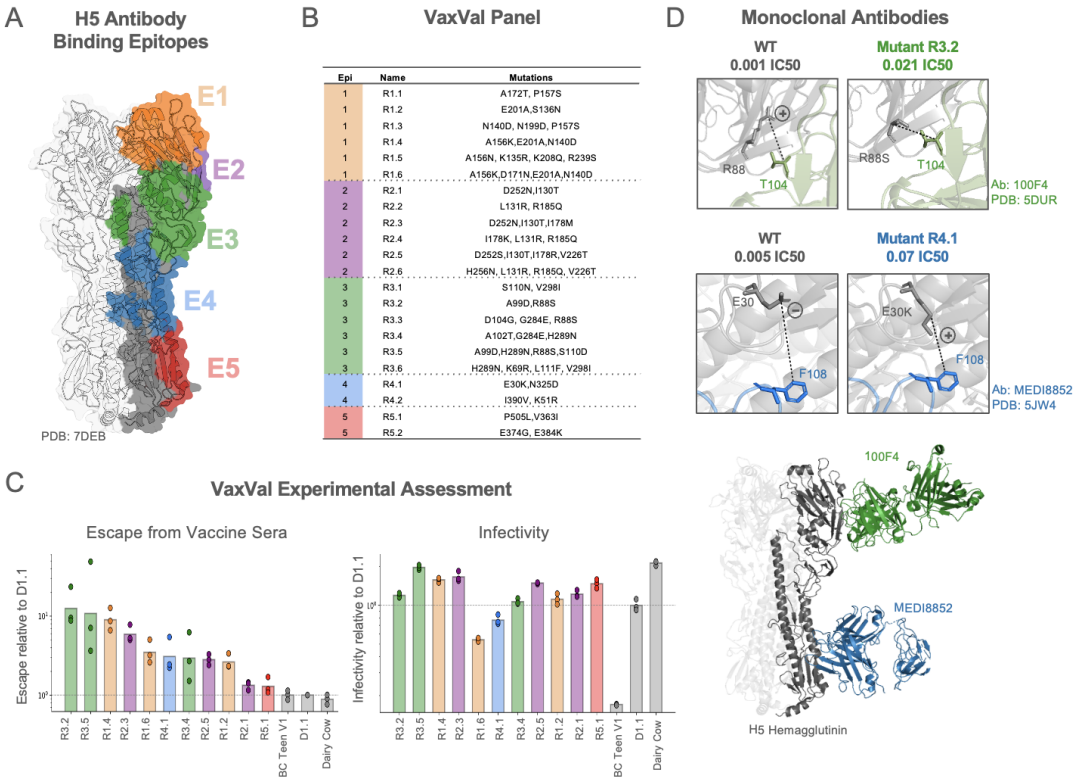

Figure 3. **VaxVal panel identifies antibody escape mutations in H5 hemagglutinin.** A) Defining antibody-binding epitopes across H5 hemagglutinin (HA). Antibody contact residues (5A) were identified from structures in the PDB and grouped by spatial clustering into five distinct epitopes. B) Design of the VaxVal panel. Twenty-two HA variants were generated, each incorporating mutations within a one of the five defined epitopes. C) VaxVal variants exhibit escape from vaccine-induced polyclonal sera. All variants show reduced neutralization by sera from immunized rabbits relative to the wild-type, vaccine-matched D1.1 strain (top). Variants with the greatest escape predominantly include mutations in epitopes E1, E2 and E3. Many high-escape variants retain infectivity, indicating preservation of functional viral entry (bottom). D) Constructs escape broadly neutralizing H5 monoclonal antibodies more than WT, with mutations disrupting native binding the antibodies' known HA epitopes.

with two naturally circulating variants: A/dairy cow/New Mexico/A240920343-90/2024 (GISAID ID 19091701; hereafter "Dairy Cow") and a low-frequency D1.1 variant identified from an infected adolescent ("BC Teen V1"). Both naturally occurring variants exhibited escape levels comparable to the D1.1 wild-type strain, despite differing substantially in sequence space: the BC Teen V1 variant is separated from D1.1 by three mutations, whereas the Dairy Cow variant differs by ten mutations.

The VaxVal panel revealed substantial resistance to neutralization by D1.1-induced sera, even though constructs contained only 2–4 mutations. In particular, designed variants R3.2, R3.5, R1.4, and R2.3 exhibited approximately 10-fold reductions in neutralization relative to wild-type D1.1. The most resistant constructs predominantly carried mutations in the HA head domain, consistent with its known immunodominance relative to the more conserved stalk region (Zost et al., 2019). Despite their escape from vaccine-induced sera, all constructs retained infectivity, with many showing infectivity levels comparable to or exceeding that of wild-type D1.1 (Fig. 3c, full results in S2).

To further characterize the effects of mutations in the VaxVal variants, we measured neutralization using two broadly neutralizing monoclonal antibodies (mAbs), MEDI8852 and 100F4, targeting the stem and head of the protein, respectively (Fig. 3a). In Fig. 3d, we highlight three high-escape variants, each originating from a distinct epitope group, and show how mutations alter binding to their corresponding cognate mAbs. Together, these results demonstrate VaxVal's ability to both generate plausible evolutionary variants and systematically assess potential immune escape across antigenically distinct regions of the protein. While our panel is designed around H5N1, we anticipate that the workflow can be scaled across pandemic threat viruses.

2.3. Impact Statement

Our workflow has the potential to improve vaccine efficacy and prepare for pandemic threats. In doing so, it can steer healthcare towards being more preventative than reactive.

References

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.

Dadonaite, B., Ahn, J. J., Ort, J. T., Yu, J., Furey, C., Dosey, A., Hannon, W. W., Vincent Baker, A. L., Webby, R. J., King, N. P., Liu, Y., Hensley, S. E., Peacock, T. P., Moncla, L. H., and Bloom, J. D. Deep mutational scanning of H5 hemagglutinin to inform influenza virus surveillance. *PLOS Biology*, 22(11):e3002916, November 2024. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002916. URL https://dx.plos.org/10.1371/journal.pbio.3002916.

Dana, J. M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., and Velankar, S. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research*, 47(D1):D482–D489, 2019. doi: 10.1093/nar/gky1114. URL https://academic.oup.com/nar/article/47/D1/D482/5184711.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42 (D1):D1140–D1146, 2014. doi: 10.1093/nar/gkt1043.

Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004. doi: 10.1093/nar/gkh340. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC390337/.

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.

Gurev, S., Youssef, N., Jain, N., and Marks, D. S. Variant effect prediction with reliability estimation across priority viruses. *bioRxiv*, pp. 2025.08.04.668549, January 2025. doi: 10.1101/2025.08.04.668549. URL http://biorxiv.org/content/early/2025/08/15/2025.08.04.668549.abstract.

Kikawa, C., Huddleston, J., Loes, A. N., Turner, S. A., Lee, J., Barr, I. G., Cowling, B. J., Englund, J. A., Greninger, A. L., Harvey, R., et al. Near real-time data on the human neutralizing antibody landscape to influenza virus to inform vaccine-strain selection in september 2025. *bioRxiv*, pp. 2025–09, 2025a.

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, 2025b.

Kok, A., Wilks, S. H., Tureli, S., James, S. L., Bestebroer, T. M., Burke, D. F., Funk, M., Van Der Vliet, S., Spronken, M. I., Rijnink, W. F., Pattinson, D. J., De Meulder, D., Rosu, M. E., Lexmond, P., Van Den Brand, J. M. A., Herfst, S., Smith, D. J., Fouchier, R. A. M., and Richard, M. A vaccine central in A(H5) influenza antigenic space confers broad immunity. *Nature*, 647(8091):1005–1013, November 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09626-3. URL https://www.nature.com/articles/s41586-025-09626-3.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

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.

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., and Rost, B. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, 141(10):1629–1647, 2022.

Mehrotra, A., Jain, N., Gurev, S., Youssef, N., and Marks, D. Real-time forecasting of influenza evolution. In *Machine Learning in Structural Biology Workshop*, 2025. URL https://www.mlsb.io/papers_2025/112.pdf. Poster presented at MLSB 2025.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A. N., Marks, D., and Gal, Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.

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, 2023a.

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, October 2023b. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06617-0. URL https://www.nature.com/articles/s41586-023-06617-0.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Gilchrist, C. L., Söding, J., and Steinegger, M. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022.

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*, 58(6):1411–1421.e6, June 2025. ISSN 10747613. doi: 10.1016/j.immuni.2025.04.015. URL https://linkinghub.elsevier.com/retrieve/pii/S1074761325001785.

Zost, S. J., Wu, N. C., Hensley, S. E., and Wilson, I. A. Immunodominance and antigenic variation of influenza virus hemagglutinin: Implications for design of universal vaccine immunogens. *The Journal of Infectious Diseases*, 219(Supplement$_1$) : $S38 - -S45, 042019. ISSN0022 -$ 1899. *doi* : . URL https://doi.org/10.1093/infdis/jiy696.

A. Methods

A.1. Alignment-based models

We assessed the performance of 3 alignment-based models: a position-specific scoring matrix, potts model, and variational autoencoder.

A.1.1. EVE

To predict the effects of mutations capturing high-order dependencies between positions, we used EVE, a Bayesian Variational Autoencoder, as implemented in (Frazer et al., 2021). The architecture consists of a symmetric encoder and decoder architecture, each with 3 layers with 2,000-1,000-300 and 300-1,000-2,000 units respectively, as well as a 50-dimensional latent space. As generative models, VAEs can learn a complex distribution of the high-dimensional data on which they are trained, in our case, sequences from a specific protein family. We use single EVE models, rather than an ensemble of independent models as was reported in (Thadani et al., 2023a). Note, that we use the negative of the evolutionary index reported by the model.

A.2. Protein language models

A.2.1. TRANCEPTION

Tranception (Notin et al., 2022) 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 (Notin et al., 2023). Tranception averages the log ratios from both left-to-right and right-to-left scoring.

A.2.2. ESM-1V

ESM-1v (Meier et al., 2021) 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 (Notin et al., 2023) to handle sequences that are longer than the model context window (i.e., 1023 amino acids), and ensemble across 5 models.

A.2.3. SAPROT

SaProt (Su et al., 2023) introduces a structure-aware vocabulary, into protein language modeling by training on Foldseek (van Kempen et al., 2022) 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 (Lin et al., 2023) but expands the embedding layer to encompasses 441 structurally-aware tokens instead of the original 20 amino acid residue tokens. We use (1) SaProt-650M-AF2, trained on approximately 40 million AF2 sequences/structures (from UniRef50) which notably explicitly excludes all viral proteins though implicitly included hundreds of thousands of prophages; and (2)SaProt-650M-PDB, which continuously pre-trains the SaProt-650M-AF2 model on the PDB.

For structure inputs to Foldseek calculation, we folded monomeric forms of H5N1 D1.1 Hemagglutinin with AlphaFold3.

A.3. 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 ProGen2(Notin et al., 2022), an autoregressive model, the likelihood of x factorizes via the chain rule and is calculated as:

$$P(x) = \prod_{i=1}^{L} P(x_i|x_{<i})$$

For Tranception, an autoregressive model with bidirectional scoring:

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

In the masked language model setting, for ESM-1v (Meier et al., 2021) and SaProt (Su et al., 2023), 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(Frazer et al., 2021), 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:

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

The ELBO term itself is estimated via Monte Carlo sampling, using 20k samples from the approximate posterior distribution. These approximations have been shown to provide strong results in practice(Frazer et al., 2021). Note that this is the negative of the evolutionary index score outputted by the EVE model.

A.4. Forecasting D1.1 Evolution

To assess model performance in forecasting D1.1, we used the Global Initiative on Sharing All Influenza Data (GISAID) to identify all mutations arising from D1.1 Briefly, we took all hemagglutinin sequences deposited after November 2021 (the emergence date of D1.1 WT) and filtered for H5N1 sequences with a mutation distance upto 5 from the WT and used MUSCLE (Edgar, 2004) to generate an alignment, from which we calculated the frequency of all mutations.

A.5. Building Alignments for EVE

We used GISAID to build alignments of Influenza hemagglutinin sequences prior to November of 2021. Sequences were deduplicated and aligned to the D1.1 WT.

A.6. Deep mutational scanning assays

All deep mutational scanning assays were taken from (Dadonaite et al., 2024) which briefly, uses a lentivirus-based DMS assay to measure phenotypic effects of all single mutations on the D1.1 strain, which include cell entry, sialic acid receptor preference, HA stability, and serum neutralization (using sera from vaccinated mice or sera from vaccinated or infected ferrets).

A.7. Designing VaxVal Panel

To design the VaxVal panel, we used the Structural Antibody Database (SAbDab) (Dunbar et al., 2014) to scrape all PDBs with an antibody bound to an Influenza A hemagglutinin protein, giving a total of 210 structures. The antibody footprints of hemagglutinin were considered those within 5 Angstroms of an antibody residue, and were mapped onto the 7DEB PDB

structure of an H5 protein using SIFTS (Dana et al., 2019). Positions were filtered on the condition of interacting with atleast 3 unique antibodies. The final set of sites were converted into 3D coordinates and clustered into 5 distinct epitopes by agglomerative clustering (Pedregosa et al., 2011).

To design constructs per epitope, the top scoring mutations were selected by the following percentiles– 3% for epitope 4, 1% for epitope 1, 1% for epitope 3, 5% for epitope 5, and 1.5% for epitope 2. From this, all sets of double, triple, and quadruple mutations were generated as a plausible construct. Because the head of hemagglutinin is more antigenic, proteins with mutations in the head were prioritized–giving 6 proteins mutated in Epitope 1, 6 mutated in Epitope 2, 6 mutated in Epitope 3, 2 mutated in Epitope 4, and 2 mutated in Epitope 5 (Fig. 3a). Constructs were prioritized based on a composite score that sums the minimum scores assigned to any mutation in the construct. The scores correspond to 4 metrics: percentile in the epitope's fitness distribution, distance to any other mutating residue, surface accessibility, and residue dissimilarity. The latter two are taken from the EVEscape framework developed by (Thadani et al., 2023b). Constructs with the highest scores were selected.

A.8. Recombinant DNA sequence plasmid design

All recombinant DNA work follows a previous protocol as outlined by (Youssef et al., 2025).

A.9. Infectivity and neutralization assays

HIV pseudotyped H5N1 influenza virus-like particles were made by transfecting HEK-293T cells with plasmids encoding the luciferase reporter, the HIV gag-pol, the cow derived H5 hemagglutinin and the cow derived N1 neuraminidase, in a 5:4:1:0.25 ratio (total 2,250 ng of DNA) and infectivity was performed on HEK-293T cells using a previously described protocol (Youssef et al., 2025).

A.10. Rabbit immunization

Rabbit immunization experiments for D1.1 was conducted by Genscript (https://www.genscript.com). 3 New Zealand white rabbits were used and given a 200 microgram mRNA vaccine followed by a booster shot two weeks later. Sera was collected following one week of the booster shot. Rabbit sera neutralization and monoclonal antibody neutralization also followed the procedure in (Youssef et al., 2025).

B. Supplementary Figures

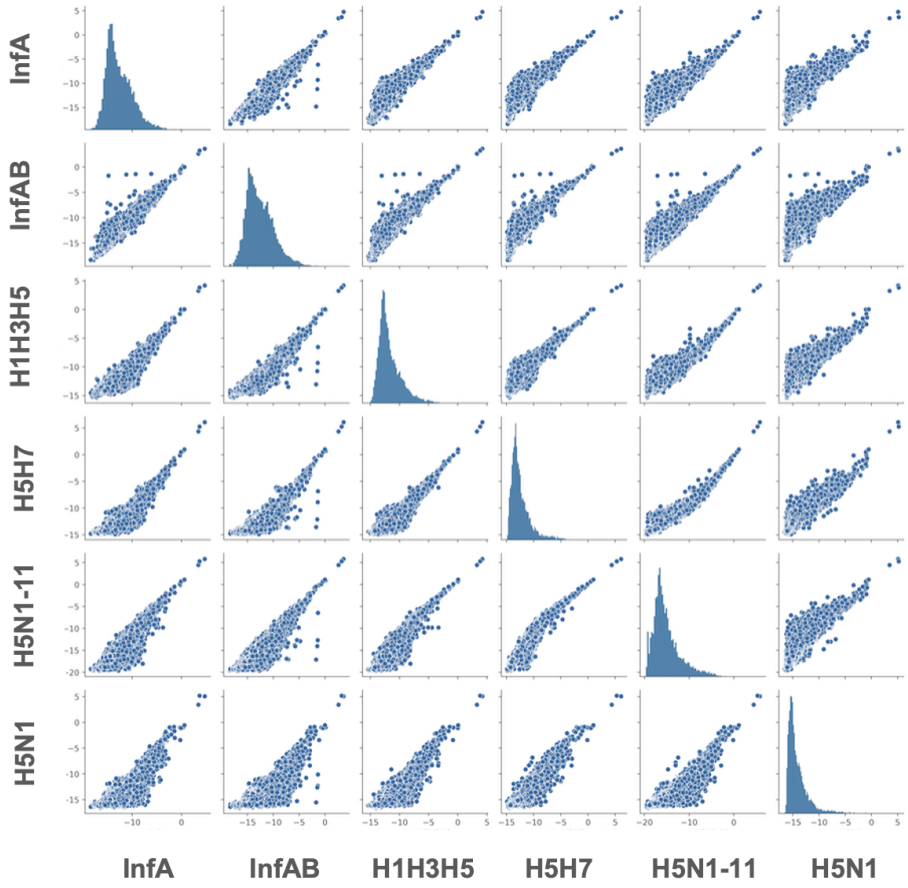

Figure 1. Correlation of EVE scores across different alignments

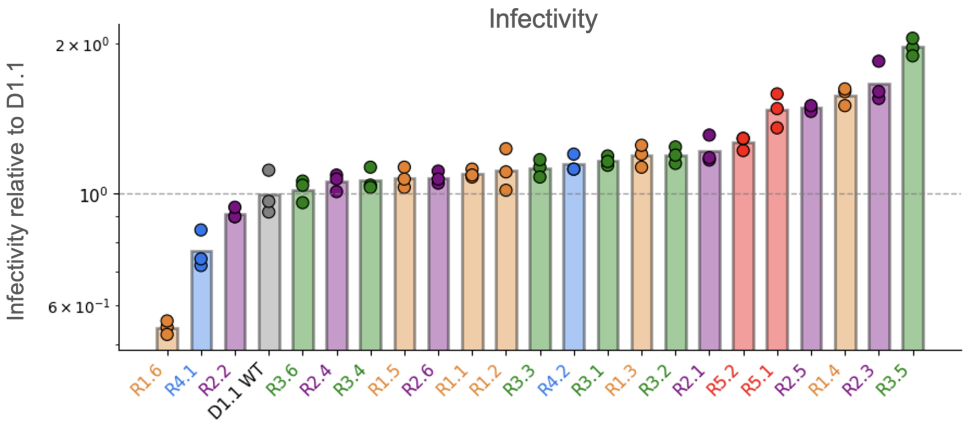

Figure 2. Infectivity results from all 22 VaxVal Constructs