EVALUATING VARIANT EFFECT PREDICTION ACROSS VIRUSES

Sarah Gurev[12*†], **Noor Youssef**[13*†], **Navami Jain**[14], **Aarushi Mehrotra**[1], **Sarrah Rose Mikhail Leung**[1], **Abigail Jackson**[15] , **Debora Marks**[13†]

ABSTRACT

Viruses are a major threat to global health due to their rapid evolution, extensive diversity, and frequent cross-species transmission. Although advances in machine learning and the expanding availability of sequence and structural data have accelerated large-scale mutation effect prediction, viral proteins, and particularly fast-evolving antigenic proteins, pose unique biological and data-related challenges that may limit model performance. We introduce EVEREST, a curated dataset for evaluating model performance on (i) forecasting real-world viral evolution (31 clades across 4 viruses) and (ii) concordance with lab-based deep mutational scanning assays (45 proteins, >340,000 variants). Using EVEREST, we show that state-of-the-art protein language models trained across the protein universe substantially under-perform on viral proteins relative to alignment-based models trained on homologous proteins. This under-performance persists even in low-sequence regimes, as is the case during a novel viral outbreak. We develop calibrated reliability metrics to quantify confidence in model predictions where no evaluation datasets exist. For more than half of the WHO-prioritized pandemic-threat viruses, current models fail to produce reliable predictions, highlighting the urgent need for more data or new modeling approaches. Together, these findings reveal key factors driving model under-performance and provide actionable recommendations for improving viral mutation effect prediction in preparation for current and future outbreaks.

INTRODUCTION

Viral pathogens pose a persistent and growing threat to global health. Their rapid evolution and high mutation rates often undermine the durability of vaccines and therapeutics, while increasing human–animal interactions heighten the risk of zoonotic spillovers. In response to these concerns, the World Health Organization (WHO) recently identified 40 priority and prototype RNA viruses across 16 families with high pandemic potential and limited medical countermeasures[1]. Accurately forecasting how these viruses evolve could transform surveillance and therapeutic design by identifying immune or treatment-evading strains early, from the first time they are sequenced. In the case of SARS-CoV-2, computational approaches were able to flag Variants of Concern months before their designation by the WHO[2,3]. While both experimental and computational tools exist to predict the functional impact of mutations, experimentally testing large numbers of variants remains a significant challenge and laboratory settings capture specific phenotypes which often do not fully recapitulate real-world selective pressures. Despite these limitations, experimental assays have proven valuable for predicting the evolution of key SARS-CoV-2 mutations ahead of their emergence during the COVID-19 pandemic[4,5].

Computational approaches offer a scalable alternative that learn fitness constraints directly from evolutionary sequences. Some models estimate mutation effects from large-scale sequencing of a specific viral population, through clade-growth, mutation-counting, and other phylogenetic approaches[6–11]. While valuable, these methods are primarily retroactive, quantifying the fitness impact of mutations after they have emerged during viral evolution. The challenge for many viruses, including the WHO priority pathogens, is the ability to forecast the fitness effects of as-yet-unseen mutations, and before extensive surveillance sequencing is available. Other computational approaches, commonly used in protein design[12–16], overcome this challenge by learning fitness constraints from broader evolutionary sequences. These approaches differ by their input data, training objective, model scale and performance on different tasks[17]. For example, alignment-based variant effect predictors (VEPs) learn from multiple sequence alignments (MSAs) of homologous proteins[18–20]; Protein language models (PLMs) learn from large sets of sequences from across the tree of life[21–26]; Inverse folding models are trained to predict an amino acid sequence given a three dimensional structure[27,28];

*Equal contribution
†Correspondence: sgurev@mit.edu, noor_youssef@hms.harvard.edu, debbie@hms.harvard.edu
[1]Systems Biology, Harvard Medical School; [2]FutureHouse; [3]Broad Institute; [4]Bioinformatics and Integrative Genomics, Harvard Medical School; [5]Harvard-MIT Health Sciences and Technology

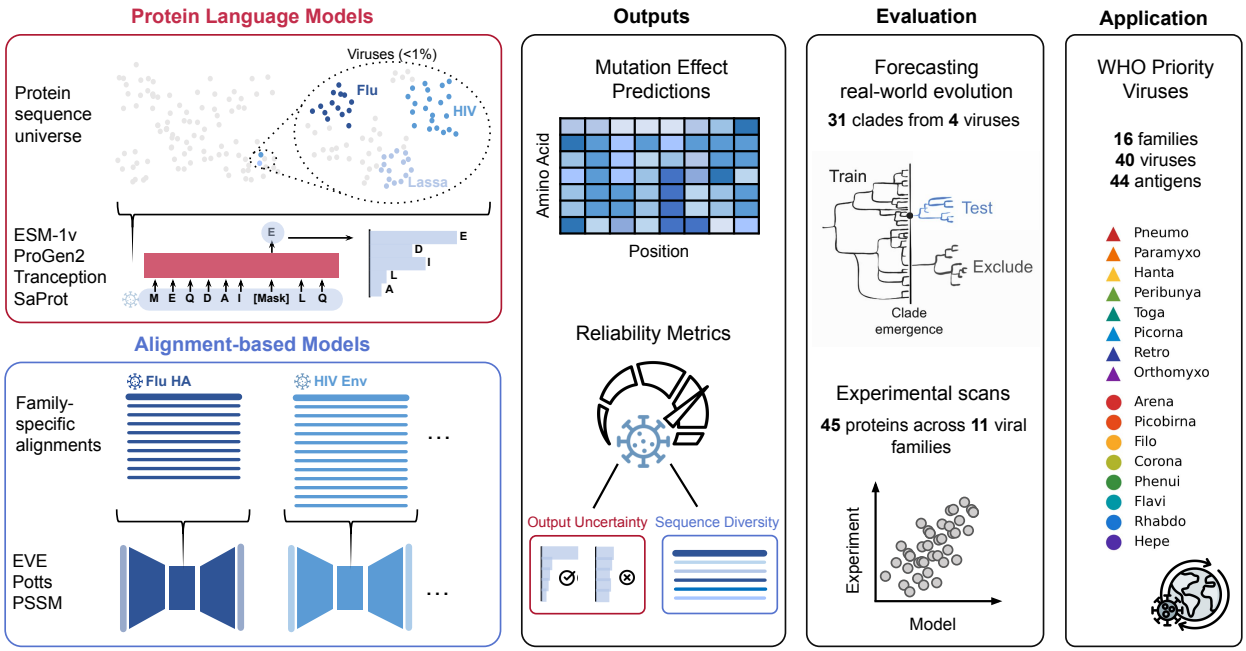


Figure 1: Comprehensive evaluation of variant effect prediction models for viral evolution. We evaluate the performance of four Protein Language Models (PLMs: ESM1v, ProGen2, Tranception, and SaProt), three alignment-based models (PSSM, Potts model, and EVE) and four hybrid models (VESPA, Tranception with MSA, TranceptEVE, and SaProt-EVE) using EVEREST, a curated dataset for viral tasks. This assessment includes large-scale experimental deep mutational scans (45 DMS across 11 viral families) and real-world viral evolution forecasting (31 clades across 4 viruses; SARS-CoV-2, HIV, and influenza H3N2 and H1N1). We calibrate per-protein reliability metrics to estimate model uncertainty in the absence of experimental data and apply this to quantify model reliability across 40 WHO-prioritized viruses.

and tree-based approaches look for conservation across subtrees of a protein family[29]. Here, we focus on the most commonly used approaches for modeling viral evolution: alignment-based VEPs, PLMs, and hybrid models (Fig. 1).

Alignment-based models have shown success in modeling viral evolution, immune escape, and vaccine effectiveness, especially for highly-sequenced viruses like HIV and SARS-CoV-2[2,3,18–20,30–36]. Given a query protein, related proteins are retrieved from sequence databases and are aligned to form a MSA. This set of sequences is assumed to include positive examples enriched for the same properties as the query. Therefore, a density model of these homologous, "evolutionary" sequences can be used as proxy for fitness constraints. The generalizability of alignment-based VEPs to high-risk viral families is currently unknown and may be limited, as many of the high-risk viruses currently lack diverse sequence data. Moreover, these VEPs cannot accommodate insertions and deletions—common features of viral evolution[37]—due to their reliance on fixed-length sequence representations.

PLMs offer an alignment-free alternative for modeling protein mutation effects. Trained on large corpora of sequences, these models can handle sequences of arbitrary length and, in theory, generalize across protein families[21–26]. PLMs span a broad design space, ranging from single-sequence models to architectures that incorporate MSA information during inference or training[22,38–40], as well as hybrid or structure-aware models[41–43]. Single-sequence PLMs, the most common class, are trained on millions of natural proteins across diverse families to learn masked amino-acid prediction tasks. Although PLMs now achieve state-of-the-art performance on many general protein benchmarks[17], their effectiveness on viral proteins has yet to be systematically evaluated, despite their increasing use in this domain[2,44–46].

Unique features of viral evolution challenge extrapolation of best modeling practices from evaluations on non-viral tasks[17]. As obligate intracellular parasites, viruses evolve under strong diversifying selection, particularly in antigenic regions. As such, viral proteins (especially those in RNA viruses) mutate at rates far exceeding those of cellular organisms and exhibit high structural and functional plasticity[47,48] (Fig. S1A). Consequently, generalized substitution matrices such as BLOSUM often fail to capture the patterns of mutations in viral proteins[2,49]. These unique viral

features raise a key question: can existing models *reliably* forecast viral mutation effects and evolutionary trajectories—especially during the early stages of an outbreak when experimental data is limited or not yet available?

To address this gap, we developed EVEREST (Evaluating Viral variant Effect prediction with Reliability ESTimation), a comprehensive benchmark tailored to viruses. EVEREST includes evaluation on two key tasks: forecasting real-world viral evolution and predicting variant effects from viral deep mutational scans (DMS). Real-world forecasting provides the strongest test of model performance but evaluation on this task is only possible for a small number of extensively sequenced viruses, such as SARS-CoV-2, HIV, and seasonal influenza (H1N1 and H3N2). To enable broader evaluation across diverse viruses, we include evaluations across 45 DMS datasets spanning 11 viral families. Using EVEREST, we evaluate 11 state-of-the-art models, including alignment-based VEPs, PLMs and hybrid models, and introduce calibrated metrics that quantify model reliability.

Across the 40 WHO-prioritized viruses, we find that current state-of-the-art models frequently fail to generate reliable predictions for surface glycoproteins—the primary antigenic determinants (hereafter, "antigens") critical to vaccine design and immune resistance—substantially limiting their immediate utility for pandemic preparedness. This systematic assessment reveals the underlying reasons existing models underperform on viral proteins and provides actionable guidelines for improving future architectures and training strategies. Together, these insights lay the foundation for developing models capable of early and accurate forecasting of viral evolution, a critical component of global pandemic preparedness.

Results and Discussion

EVEREST experimental and real-world forecasting benchmarks

The EVEREST benchmark includes two evaluation datasets, measuring two complementary tasks: forecasting real-world viral evolution (forecasting benchmark) and predicting variant effects measured by viral deep mutational scanning assays (DMS benchmark; Fig. 1). For the forecasting benchmark, we identified 31 clades across four well-characterized viral protein antigens—SARS-CoV-2 spike, HIV envelope, and influenza H3N2 and H1N1 hemagglutinins—and assessed models on their ability to predict emerging mutations and prioritize highly frequent variants during evolution (Fig. S1). For alignment-based models, we implemented chronological train–test splits, restricting training data to sequences available prior to each clade's emergence. In contrast, we note that since PLMs cannot be retrained at this scale, there is unavoidable data leakage (where mutations seen in the test set are included in training), resulting in inflated performance in all but the most recent clades. To contextualize forecasting performance, we compare models to the predictive ability of high-throughput fitness DMS measurements[5,50–53], while acknowledging that not all DMS assays are intended to serve as forecasting tools.

For the DMS benchmark, we follow well-established practices for evaluating the ability of general protein design models to predict mutation effects and design functional proteins[17,18]. We curated a viral DMS benchmark that includes 45 viral proteins from 11 viral families, and more than 340,000 mutations—vastly broadening viral diversity compared to our forecasting benchmark[4,50,54–78]. Measured experimental phenotypes include general fitness properties (e.g., infectivity, replication, and cell entry) and specific traits (e.g., stability and expression; Fig. S1, Table S1). We exclude antibody escape assays due to the limited number of antibodies or sera tested for most viruses, which can result in apparent false positives where a model correctly predicts escape but the DMS assay does not capture it.

Using EVEREST, we evaluate three alignment-based methods (PSSM[18], Potts model[18], and EVE[19]); three PLMs (ESM-1v[21], Tranception[22], and ProGen2[24]); one structure-aware PLM (SaProt[41]); and four hybrid approaches (TranceptEVE[38], Tranception with MSA retrieval[22], VESPA[79], and SaProt-EVE). These models were chosen for their diverse architectures, training datasets, and high performance on prior general benchmarks (Table S2).

Protein Language Models underperform on viral proteins

Protein language models (PLMs), deep learning models trained on large-scale protein sequence datasets, have achieved state-of-the-art performance on mutation effect prediction in general protein benchmarks[17]. As a result, prior studies have assumed that PLMs would represent a paradigm shift over alignment-based models for viral evolution prediction[44]. However, despite their success in other domains, we find that single-sequence PLMs underperform on viral proteins compared to proteins from other taxa, with most PLMs having less accurate predictions than even the simplest site-independent alignment-based model (Fig. 2A, dotted lines).

A likely driver of PLM underperformance on viral proteins is the relatively poor representation of viruses in PLM training datasets (<1% of training; Fig. 2B-C). PLMs are typically trained on clustered sequence databases, such

as UniRef90 or UniRef50, which retain only one representative per sequence cluster with ≥90% or ≥50% identity, respectively. While clustering reduces redundancy and improves performance for well-sequenced protein families[21], it disproportionately filters out already rare viral sequences. While non-viral proteins maintain a ratio of 1 effective sequence (clustered at 90% identity) for every 8 raw (unclustered) sequences, most viruses exhibit a drastic reduction to 1:550 (Fig. 2B). This disproportionate downsampling is likely due to over-sequencing of narrow parts of the viral sequence landscape and the orders of magnitude higher mutation rates of viruses[47], leading to related sequences which more quickly evolve past the limits of homology detection (Fig. S1A). There are two notable exceptions to this downsampling trend: HIV and phages (Fig. 2B). HIV, with its extreme genetic variability and large sequencing efforts[80], has sufficient sequence diversity to allow for structural contact recovery from coevolutionary analysis, a property rarely observed in other viral antigens (Fig. S10, S5). Phages, as mostly double-stranded DNA viruses, have lower mutation rates than most eukaryote-infecting RNA viruses (Fig. S1A), resulting in higher number of identifiable homologs.

Two modeling choices tend to mitigate the poor representation of viral proteins in training datasets: (i) using larger models that may extract more informative features from sparsely represented proteins, and (ii) training on more comprehensive datasets with greater viral diversity. We find that increasing model size consistently improves accuracy on viral mutation effect prediction, in contrast to the performance plateau observed for non-viral proteins (Fig. 2D, Fig. S2)[23,25,81,82]. Further supporting this, we find that ESM-2 models trained on the same dataset but spanning 8 million to 15 billion parameters show steady performance gains on viral proteins, as well as on non-viral proteins that are also poorly represented in training, with little to no improvement for well-represented proteins (Fig. S3).

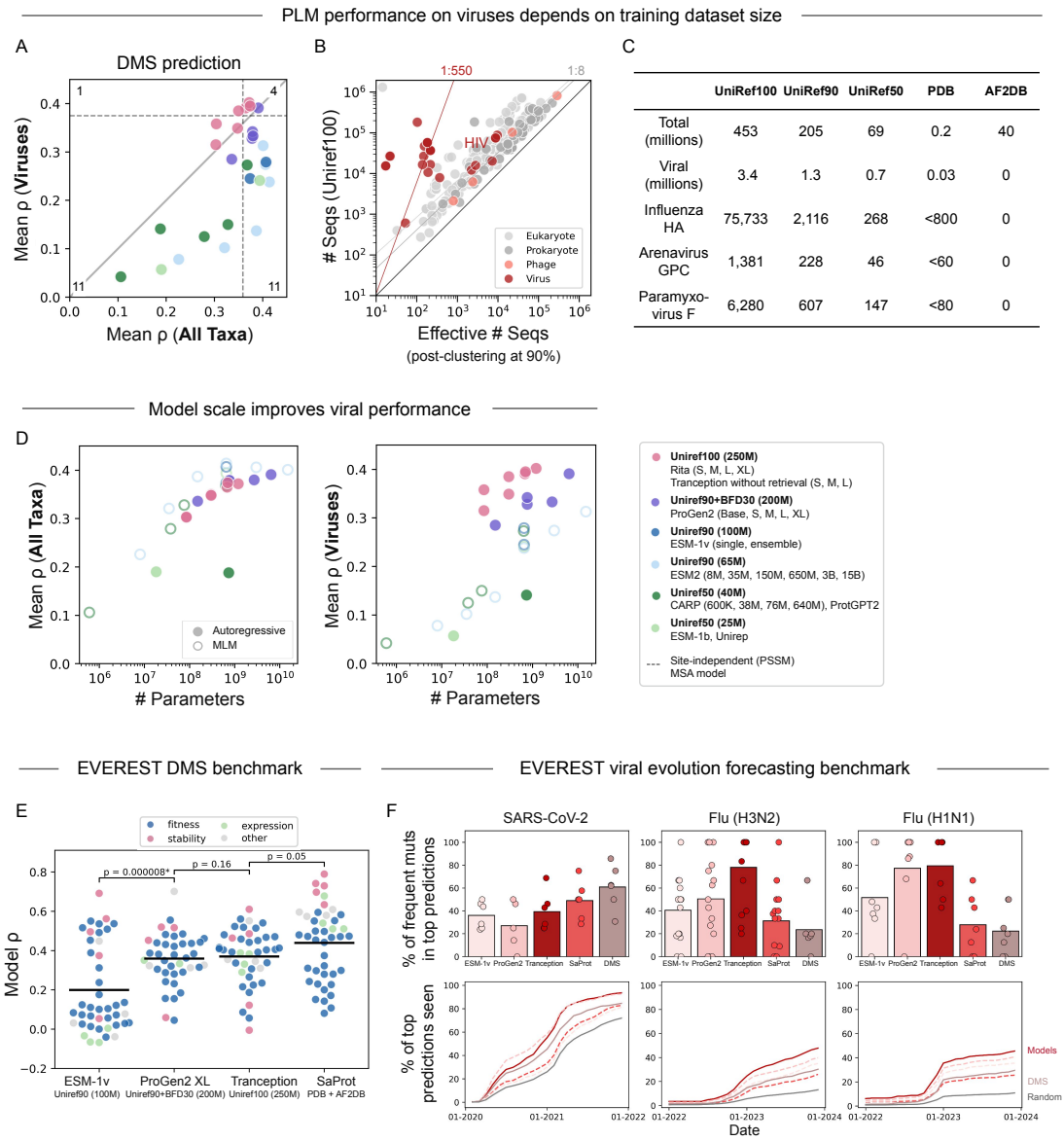
Training dataset choice also affects performance in viral compared to non-viral proteins. PLMs trained on more comprehensive datasets (e.g., unclustered UniRef100) outperform those trained on clustered datasets (e.g., UniRef90, UniRef50) only for eukaryote-infecting viruses (Fig. 2A, S2, Table S3). For other taxa—including phages, which are not disproportionately downsampled in clustered databases (Fig. 2B)—models trained on UniRef90 or UniRef50 achieve superior performance compared with UniRef100-trained models (Fig. S2). The inclusion of metagenomic datasets (e.g., BFD30, BFD90, or OMG) provides minimal benefit for viral mutation effect prediction (Fig. S4). For example, ProGen2 trained on UniRef90+BFD30 performs similarly to the same model trained on UniRef90+BFD90 (Fig. S5B, S4), and both underperform compared to models trained on UniRef100, even when controlling for model size (Fig. 2D). Likewise, metagenomic sequences do not improve performance for SaProt-OMG relative to SaProt, despite the former being a substantially larger model (Fig. S5C, S4). Together, these results suggest that larger, unclustered datasets yield superior PLM performance on viral proteins and that metagenomic sequences do not compensate for the loss of natural evolutionary diversity introduced by clustering.

A confounding factor when interpreting training dataset-size effects is that models trained on the largest datasets (e.g., Tranception and ProGen2) employ an autoregressive objective, predicting each residue conditioned on all previous residues, rather than masked language modeling (MLM), which predicts only a small fraction of randomly masked residues (Fig. 2D, Table S3). It remains unclear whether performance gains arise primarily from increased dataset size or from the more challenging autoregressive objective, which introduces longer contiguous spans of uncertainty and may promote better generalization. Disentangling these contributions represents an important direction for future investigation.

We also examined other features that may contribute to performance. While most PLMs use input windows of 1,024 tokens, the effective context length is typically closer to half the training length[83,84], which is smaller than the length of most viral antigens. Nonetheless, context length alone is unlikely to explain the observed performance disparities. ProGen2 Base and ProGen2 Medium share identical architectures and training data but differ only in context length (2,048 vs. 1,024 tokens); the longer-context model performs slightly worse on viral proteins (Spearman's $\rho = 0.328$ vs. 0.342; Table S3)[24]. We also find that ensembling across multiple random seeds provides modest improvements across taxa, including viruses (Table S3).

Taken together, these observations suggest that larger PLMs trained on larger, unclustered datasets (e.g,. Tranception-L and ProGen2-XL) are better suited for viral applications in zero-shot settings than PLMs trained on smaller datasets with an MLM objective (e.g., ESM-2[44–46]).

Tranception outperforms other PLMs on EVEREST Using our EVEREST viral forecasting and DMS benchmarks, we further evaluate four PLMs trained on different databases and with different training objectives: ESM-1v (ensemble) trained on UniRef90 (100M sequences), ProGen2-XL trained on UniRef90 + BFD30 (200M), Tranception-L trained on UniRef100 (250M) and the structure-aware PLM SaProt trained on the PDB + AF2DB (40M structures). All models share transformer-based architectures: both SaProt and ESM-1v use a BERT-style MLM objective, whereas Tranception and ProGen2 are autoregressive models[21,22,24,41].



	UniRef100	UniRef90	UniRef50	PDB	AF2DB
Total (millions)	453	205	69	0.2	40
Viral (millions)	3.4	1.3	0.7	0.03	0
Influenza HA	75,733	2,116	268	<800	0
Arenavirus GPC	1,381	228	46	<60	0
Paramyxo-virus F	6,280	607	147	<80	0

Figure 2: Evaluation of protein language models for viral evolution and mutation-effect prediction.
A. Protein language models (PLMs) show reduced performance on viral proteins compared to other taxa. Performance is measured as the average Spearman correlation with experimental deep mutational scanning (DMS). Results include 218 DMS assays across all taxa versus < 20 assays from eukaryote-infecting viral proteins from prior benchmark[17]. Dotted line indicates the average performance of alignment-based PSSM model, without any of the viral-specific alignment optimization discussed in the Alignment-based models section. See Table S3 for details.
B. Sequence clustering disproportionately reduces viral diversity. While non-viral proteins maintain a ratio of 1 effective sequence (clustered at 90% identity) for every 8 raw (unclustered) sequences, most viruses exhibit a drastic reduction to 1:550.
C. Viruses are underrepresented in PLM training, making up $<1\%$ of sequence databases.
D. Unlike performance on other taxa (left) which saturates with model scale, viral performance (right) increases with both model size and training dataset volume. Each point denotes an individual PLM, colored by training dataset and objective (autoregressive vs. masked language modeling).
E. PLM performance on EVEREST DMS benchmark, measured as Spearman correlation between model predictions and experimental DMS scores. Other PLMs show significant improvement over ESM-1v ($P < 0.01$; paired t-test).
F. PLM performance on EVEREST viral forecasting benchmark. Top: percentage of highly frequent mutations captured by model predictions across 6 SARS-CoV-2 clades, 9 H1N1 clades and 15 H3N2 clades. Bottom: percentage of model-predicted mutations subsequently observed during evolution. Time-course analyses shown for representative clades—SARS-CoV-2 Wuhan (wild-type vaccine strain) and current flu vaccine strains for H1N1 (6B.1A.5a.2a.1), and H3N2 (3C.2a1b.2a2a.3a.1). Model predictions correspond to top 250 ranked mutations (see Fig. S7,S8 for alternate thresholds). Frequent mutations seen $>$1k times for flu and $>$10k times for SARS-CoV-2.

ProGen2-XL, Tranception, and SaProt had comparable performance across the 45 DMS assays, significantly outperforming ESM-1v (Fig. 2E, $P < 0.0001$; paired t-test). ESM-1v only outperformed ProGen2 and Tranception on proteins with high numbers of homologous sequences in Uniref90 (i.e., HIV and phage proteins; Fig. 2B, S5A,F). SaProt's strong performance on DMS assays is primarily driven by experiments measuring stability, expression, and other properties where structural context may be especially informative (Fig. 2E, Fig. S5A). The comparable performance of ProGen2-XL and Tranception is likely due to ProGen2 having almost 10 times as many parameters as Tranception, which may be compensating for its more limited training dataset. When controlling for model size, Tranception consistently outperforms ProGen2 (Fig. 2D-E, S5A,E).

In contrast to its strong performance on DMS datasets, SaProt performs substantially worse on real-world viral forecasting compared to other PLMs (Fig. 2F, 4F). Tranception was the top-performing model, both in identifying the most frequently occurring mutations (top panels, Fig. 2F, 4F, S6A) and in predicting emerging mutations (bottom panels). PLMs generally outperform DMS experiments for forecasting, with the exception of SARS-CoV-2 for which the available DMS assays measured only mutations that were already highly prevalent during the pandemic.

PLM performance on real-world evolution forecasting is complicated by data leakage (when test data is included in training). Because these models are trained on large datasets, many of the sequences used in our forecasting tests are likely present in their training data (Table S3). Consistent with this, Tranception's accuracy declines for more recent clades, which are less likely to contain mutations seen during training (Fig. S6B). Nevertheless, in the most recent influenza clades (representing current H1N1 and H3N2 vaccine strains) where leakage is minimal, Tranception and ProGen2-XL outperformed the DMS experiment at forecasting (Fig. 2F, bottom).

Future directions for PLM improvements for viruses Overall, our comprehensive comparison of modeling approach uncovers the drivers of successful viral modeling, and suggests that the low performance of existing PLMs on viral proteins is primarily driven by the poor representation of viral sequences in training data. While architectural improvements and increasing model size may help, these gains are ultimately constrained by the lack of viral training data. Increasing the breadth and diversity of viral sequences—particularly from underrepresented lineages—will be essential to build models that perform robustly across the viral protein universe. Models may benefit from training not only on unclustered datasets but also focusing on viral-specific databases including GISAID, Pathoplexus[85,86] and even the metagenome-derived VIRE[87], as well as relevant subsets of Uniref or larger databases like BFD, MGnify, and Logan[88–90]. Although if the goal is not single-virus specific, care must be taken during sampling to avoid over-training on any one highly sequenced virus.

Incorporating structural information into PLMs may help mitigate the challenges posed by the sparse and highly divergent nature of viral sequence data. Structural features can evolve up to ten times slower than amino acid sequences[91] and some proteins, including viral proteins, retain structural similarity despite sharing less than 10% sequence identity[92] (Fig. S9A), enabling more reliable detection of distant homologs and improved protein function annotation[93–95]. The benefits of structure-aware training is evident by SaProt's higher performance on almost half of the EVEREST DMS benchmarks (including some fitness DMSs for relatively well represented viruses in the PDB) compared to sequence-only models (Fig. 2C,E, S5A). The inclusion of even a small number of viral structures found in the PDB yielded consistent improvement boost over SaProt-AF2 (which trained only on the AlphaFold2 database with no viral structures; Fig. S5C,D).

We find some evidence that structural training data may have a greater impact on performance than sequence data alone. For instance, SaProt outperforms ESM-1v despite being trained on orders of magnitude fewer amino acid data points (Fig. 2B, S9C). Second, the availability of additional training points yielded a larger improvement in performance for SaProt compared to Tranception (Fig. S9D). Additionally, SaProt's performance is better predicted by the combined number of sequence and structure training points than by either alone (Fig. S9B). The growing availability of viral-specific structural databases[94,96–98], holds promise for improving viral modeling as structural information may provide local structural context or the ability to link relevant sequences with remote structural homology.

Alignment-based Models have high and consistent performance on Viruses

Alignment relevance improves performance. Rather than learn from across the protein universe (and therefore primarily non-viral sequences), alignment-based models are trained on a set of homologous sequences and can therefore learn relevant evolutionary constraints for specific protein (Fig. 3A)[2,3,18–20,30–36]. By focusing on homologous sequences, alignment-based models avoid pitfalls of PLM training datasets and perform equally well on viruses as other taxa (Fig. 3B), with even the simplest alignment-based model (site-independent PSSM) outperforming most PLMs for viruses (Fig. 2A).

A common challenge in using alignment-based methods is identifying which set of sequences will yield optimal model results in the absence of any evaluation data. Prior strategies for alignment selection often include selecting MSAs that recover structural contacts[18]. However, most viral MSAs—except for a few notable cases such as HIV—fail to recapitulate structural contacts from couplings analysis (Fig. S10). In the absence of structural ground truth, a common heuristic is to instead maximize "alignment depth" (measured as the ratio of the effective sequence number to protein length, Neff/L), while ensuring high positional coverage[18]. Neff represents the effective number of non-redundant sequences computed by down-weighting highly similar sequences. Alignment depth can be categorized as low (Neff/L<1), medium (1<Neff/L<100), or high (Neff/L>100)[17]. To investigate the impact of alignment selection on model performance, for each protein in our EVEREST DMS benchmark we generated 18 MSAs by querying three sequence databases (UniRef90, UniRef100, and UniRef100+MGnify+BFD) and six depths for homology search (Fig. 3A).

Contrary to common practice, we find that deeper alignments do not necessarily improve performance (Fig. 3C-D). For example, while model performance for influenza A virus (IAV) polymerase improved with alignment depth (Fig. 3C, Top), deeper alignments yielded worse performance for Coxsackie B3 protease (Fig. 3C, Bottom). In fact, for most proteins in our EVEREST DMS benchmark, maximizing alignment depth reduced performance, particularly for proteins with the most sequences retrieved (Fig. 3D). We hypothesize that beyond a minimum alignment depth, incorporating additional sequences—especially from distantly related organisms—may introduce noise or conflicting constraints, as these sequences are subject to distinct selective pressures.

To test this, we examined whether performance correlated with "alignment relevance", quantified as the proportion of MSA sequences sharing >90% identity with the target viral protein, given Neff/L >1 . Alignment relevance is a stronger predictor of model performance than either alignment depth or the fraction of viral sequences in the alignment (Fig. 3E; S11A-B), though maximizing relevance also increases the fraction of viral sequences in the alignment (Fig. S11F). For both the coxsackievirus protease and IAV polymerase, selecting the alignment that maximized relevance improved Potts model performance (Fig. S11C,E).

Selecting alignments based on the relevance criteria significantly improved Potts model performance across the EVEREST DMS benchmark (from an average Spearman of 0.32 to 0.39; $P < 0.001$, paired t-test). Performance improvements were more pronounced for predicting fitness assays, where functional differences in divergent sequences seen in deeper alignments may be misleading, as opposed to stability or expression prediction where the learned phenotype may be more consistent across proteins from diverse species (Fig. S11G). Given these results, we propose a practical heuristic for alignment selection in viral proteins: (1) select alignments with at least medium depth (Neff/L >1), and (2) among these, choose the alignment with the highest fraction of sequences sharing >90% identity to the target. For proteins with no alignment reaching Neff/L >1 (only one protein in our dataset), we instead select the alignment that maximizes alignment depth.

Few of the alignments selected using relevance criteria required metagenomic sequences found in BFD or MGnify (Fig. S11D). This is likely because alignment relevance selects for closely related sequences while disfavoring the inclusion of highly divergent proteins found in BFD+MGnify[99]. In some cases, these larger databases also caused memory issues during model training (Table S4), as BFD+MGnify is almost 9 times the size of Uniref100[89,100,101]. These results suggest that, for eukaryote-infecting viruses, expanding alignments to include metagenomic sequencing may rarely provide additional benefit, though this may change for more viral-focused datasets, or when focusing on non-fitness phenotypes.

EVE outperforms other alignment-based models on EVEREST. Using the alignment relevance selection strategy, we train three alignment-based VEPs with increasing epistatic complexity: a position-specific scoring matrix (PSSM; site-independent)[18], Potts model (an epistatic pairwise maximum entropy model)[18], and EVE (a fully connected variational autoencoder)[19].

We found that EVE consistently outperformed both PSSM and Potts models across nearly all EVEREST DMS datasets, and had a higher average Spearman correlation (ρ = 0.44 vs. 0.38 for both PSSM and Potts model; $P < 0.0001$, paired t-test; Fig. 3F, S13, Table S5). This is consistent with prior reports that capturing higher-order epistasis, using a variational autoencoder, improves prediction accuracy[2,19,102]. Surprisingly given the frequent use of pairwise epistatic alignment-based models for viral fitness prediction[20,30–36], the Potts model and site-specific PSSM had comparable overall DMS performance for viruses, in contrast to other taxa (Table S2). This may be because pair-wise terms (inferred by the Potts model) are incorrectly learned in the limited diversity regime (consistent with poor structure reconstruction from coupling analyses for viral proteins; Fig. 2B, S10) and using alignment relevance advantages site-independent terms (by learning site-specific frequencies from a set of sequences with high similarity to the query). Supporting this, we find that PSSMs are more likely to outperform Potts models when the alignment is composed of more relevant sequences to the query (Fig. S12C).

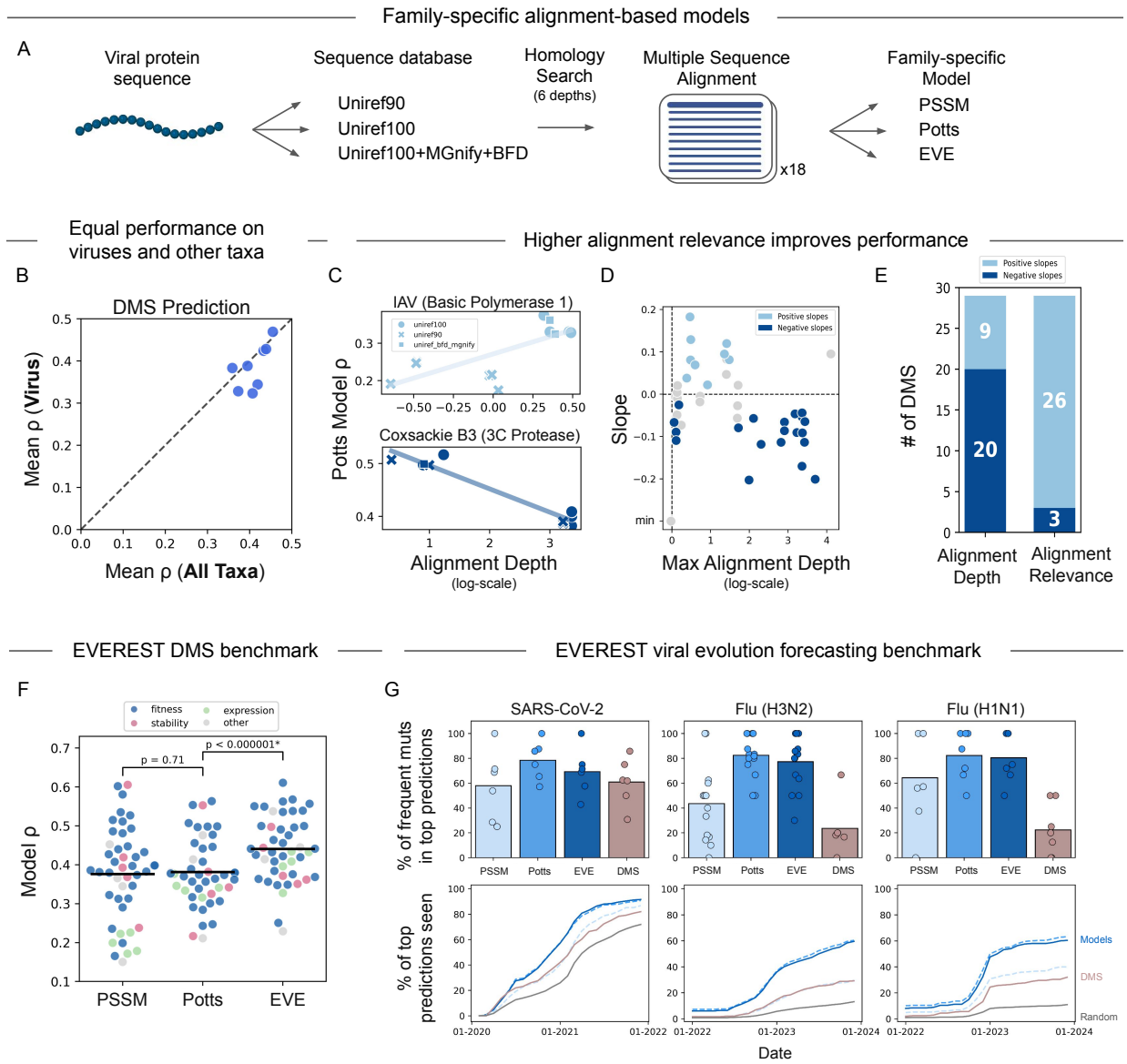


Figure 3: Evaluation of alignment-based models for viral evolution and mutation-effect prediction.
A. Overview of MSA generation for alignment-based models.
B. Unlike protein language models, alignment-based models have similar performance for viral proteins compared to proteins from other taxa. Performance is measured as Spearman correlation with experimental deep mutational scanning (DMS) assays from prior benchmark[17].
C. Increasing alignment depth (Neff/L) improves Potts model[18] performance for influenza polymerase, but decreases performance for coxsackievirus protease. Performance measured as Spearman correlation between model prediction and DMS.
D. Alignment depth is a poor predictor of Potts model performance. Colors indicate whether performance increased (light blue), decreased (dark blue), or was unchanged (gray; Spearman range < 0.1) with increasing alignment depth.
E. Alignment relevance, measured as the fraction of sequences within 90% identity to the query sequence, is a better predictor of model performance than alignment depth. Reported is the number of viral proteins with positive slopes between Potts model performance and alignment-selection metric for DMSs where Spearman range is >0.1.
F. Alignment-based VEP performance on EVEREST DMS benchmark. EVE[19] is a significant improvement over Potts and PSSM models ($P < 0.001$, paired t-test). Alignments selected using alignment relevance criteria.
G. Alignment-based VEP performance on EVEREST viral forecasting benchmark. Top: percentage of highly frequent mutations captured by model predictions across 6 SARS-CoV-2 clades, 9 H1N1 clades and 15 H3N2 clades. Bottom: percentage of model-predicted mutations subsequently observed during evolution. Time-course analyses shown for representative clades—SARS-CoV-2 Wuhan (wild-type vaccine strain) and current flu vaccine strains for H1N1 (6B.1A.5a.2a.1), and H3N2 (3C.2a1b.2a2a.3a.1). Model predictions correspond to top 250 ranked mutations (see Fig. S7,S8 for alternate thresholds). Frequent mutations seen $>$1k times for flu and $>$10k times for SARS2.

Despite the comparable performance of PSSM and Potts models on the DMS benchmark ($P > 0.5$; paired t-test), we find that Potts models significantly outperform PSSM on forecasting viral evolution, with comparable accuracy to EVE models (Fig. 3G, 4F, S6A; $P < 0.0001$; paired t-tests). Differences in performance on the EVEREST DMS dataset and the viral forecasting dataset highlight that models that are better able to recapitulate specific phenotypes measured in laboratory setting are not necessarily better at capturing the full set of fitness constraints acting on evolving viruses, for which sequences from prior evolution may be particularly informative.

COMPARING AND COMBINING ALIGNMENT-BASED MODELS AND PLMS

Alignment-based model EVE outperforms PLMs at predicting viral mutation effects We next compare the top-performing alignment-based variant effect predictor (EVE) and PLMs (Tranception and SaProt). Across the EVEREST DMS benchmark, EVE significantly outperformed Tranception ($P < 0.01$, paired t-test) and showed comparable performance to SaProt ($P > 0.05$, paired t-test; Fig. 4A). SaProt's advantage was primarily driven by superior performance on stability assays, where it outperformed both EVE and Tranception ($P < 0.01$ and $P < 0.05$, respectively, paired t-tests). In contrast, for DMSs assessing general fitness, which better reflect selective pressures in naturally evolving strains, EVE outperformed both PLM models ($P < 0.001$, paired t-test; Fig. 4A, Table S5).

In contrast to the view of PLMs being better suited in low-sequence regimes[44], we find that EVE's relative advantage over PLMs was particularly pronounced for proteins with few homologous sequences, as is likely in a new viral outbreak scenario (Fig. 4B, S14E). EVE's success over PLMs must come not because the alignment has access to more data–as this comparison is with the Uniref100-trained Tranception model–but because of explicit access to the few related sequences that may become diluted by unrelated sequence families in a PLM. While both Tranception and SaProt showed an improvement in performance with the number of relevant training points (Fig. S9C-D), EVE's performance was consistent across different number of related sequences (Fig. S9D, S12A, S14E)—for example, while EVE had high performance for well-sequenced viruses, like HIV and Flu antigens ($\rho > 0.4$), they were far from the best performing proteins, which are primarily non-antigens (Fig. S12A, S14C-D).

The contrasting observations that PLMs benefit from training on larger datasets with more remote homologs while alignment-based models perform better with more focused alignments, likely stems from the fact that alignment-based models (besides redundancy-weighting) treat all sequences identically regardless of sequence and functional similarity to the target. In contrast, PLMs, by learning across the protein universe, must by necessity learn to reason about sequences with varying degrees of relevance.

The superior performance of EVE on fitness DMS assays would suggest that it is better suited for predicting real-world viral evolution. While Tranception and EVE have similar performance in forecasting ($P > 0.5$; paired t-tests; Fig. 4F), this is likely due to data leakage where sequences in the test set are seen during Tranception training. Consistent with this, the apparent advantage of Tranception over EVE is most pronounced for clades that emerged earlier (and were included in training), but diminishes for the most recently circulating clades, which were not part of the training data (Fig. S6B). Moreover, while SaProt excels at predicting stability, only a subset of stable proteins satisfy other necessary fitness constraints, such as receptor binding, conformational changes, or adaptive selection pressures captured by evolutionary sequences. This difference is evident when comparing the distribution of top scores for SaProt and EVE—many mutations receive equivalently high SaProt scores, while only a small fraction achieve the highest EVE scores (Fig. S18). High-frequency mutations that emerge during evolution tend to require both high EVE and high SaProt scores, but EVE's more selective scores makes it a better discriminator of evolutionary trajectories.

Hybrid models improve performance over alignment-based or PLMs individually Several strategies have been proposed to combine alignment-based models with PLMs. One approach is to fine-tune PLMs using sets of unlabeled homologous sequences ("evotuning")[103]. Another is to construct hybrid models that combine mutation scores between both approaches. We examined multiple hybrid strategies, including Tranception with site-independent MSA retrieval[22], TranceptEVE (Tranception-EVE hybrid)[38], a newly constructed SaProt-EVE hybrid, and VESPA[79], which integrates PLM embeddings with conservation and BLOSUM substitution scores. Ideally, a hybrid model should perform at least as well as its best component and, where possible, surpass it through synergistic integration.

While evotuning did improve PLM performance (SaProt $\rho = 0.28$ vs. best evotuning $\rho = 0.34$ for Lassa DMS; Methods), it did not reach the performance of even a site-independent alignment-based model ($\rho = 0.51$), indicating that simple fine-tuning is not always sufficient to surpass alignment-based approaches. These observations are preliminary, and further evaluation at scale is warranted.

Alternatively, we find that hybridizing alignment-based models with PLMs (e.g., SaProt-EVE and TranceptEVE) yields superior performance relative to either modeling approach alone (Fig. 4C-D, S14A,S15B-C). In fact, SaProt-

EVE and TranceptEVE significantly outperformed all other models—including alignment-based models, PLMs, and alternative hybrids—on the EVEREST DMS benchmark ($P < 0.005$, paired t-test; Fig. 4D, S15D). However, the benefits of hybrid approaches for viral forecasting were more modest (Fig. 4E-F), even while plagued with the same data leakage concerns as PLMs which should inflate performance. Overall, they represent a robust default strategy when experimental validation is unavailable, multiple phenotypes are of interest, or a single, reliable model is preferred over multi-model evaluation.

CALIBRATING RELIABILITY ESTIMATES

We have thus far identified conditions under which different modeling approaches perform best. What remains unclear is how to know *a priori*, in the absence of experimental data, how reliable model predictions are for a specific viral protein. Reliability estimation has been useful in other biological domains, most notably for AlphaFold2[105], but such estimates have not yet accompanied variant effect predictions. To address this, we calibrate reliability metrics on our EVEREST DMS benchmark tailored to each modeling approach.

For alignment-based VEPs, we find that the number of effective sequences within 90% identity (Neff @ 90% ID) is the best predictor of model reliability (Fig. 5A, S12A). For most PLMs, we find that pseudo-perplexity (PPPL), a measure of model uncertainty given a specific input[25,106], correlates with model performance across proteins (Fig. 5B,S16A). We scaled PPPL to a value between 0 and 1, where higher values reflect greater confidence. This metric was a strong predictor of SaProt performance ($\rho = 0.75$), outperforming the number of homologous training sequences and structures ($\rho = 0.68$; Fig. S9). For some models other metrics were more predictive of performance—for example, Tranception performance was better explained by number of relevant training data points (Fig 9D).

For each of the MSA and PLM reliability metrics, we identify stringent thresholds for high versus low model reliability based on maximizing sensitivity and specificity (Youden's J) for predicting high model performance (defined as Spearman $\rho > 0.4$). We find that for EVE, a MSA reliability metric of 30 effective sequences within 90% identity effectively separates high versus low performance ($P < 0.01$; t-test). For SaProt, a scaled PPPL of 0.88 (interpretable as an average of ≈3 amino acid choices per residue) similarly separates high performance ($P < 0.0001$; t-test). Importantly, proteins exceeding the high confidence threshold consistently have $\rho > 0.4$, indicating that these cutoffs are conservative estimates.

The MSA and PLM reliability scores also predicted which of the two modeling approaches would perform best for a given viral protein (Fig. S14B). Notably, across fitness assays, non-antigens have consistently higher MSA reliability scores (likely as non-antigens are under less antigenic pressure and therefore have more homologs within 90% identity; Fig. 5A, S14C-D). However, many of the highest PLM reliability scores are for common antigens in the PDB, despite potential for the greater mutability of antigens being interpreted as greater model uncertainty (Fig. 5B, S14C-D).

When applied to the EVEREST DMS dataset, hybrid models with multiple reliability metrics can work in tandem–SaProt-EVE only underperforms (Spearman $\rho < 0.4$) when both reliability scores fall below their respective thresholds. Conversely, all viral DMS experiments with both reliability scores above threshold exhibit correlations exceeding 0.5 (Fig. 5C), demonstrating that high-confidence predictions across both modeling approaches provide synergistic estimates of variant effects representative of both the model's internal confidence in its predictions and the richness of evolutionary information supporting those predictions.

Future directions While EVE and SaProt reliability thresholding significantly separate DMSs by performance, the MSA reliability metric is less accurate than the PLM metric. PPPL is an intrinsic measure of a model's predictive uncertainty, and is a property of the PLM predictions itself rather than the input data (which is the case for EVE). We find that an equivalent "pseudo-perplexity" measure for EVE does not distinguish high versus low performance, this is likely due to EVE's consistent performance in the EVEREST DMS benchmark so there is less signal to separate high versus low performance (Fig. S16C). Future work evaluating this and other intrinsic alignment-based model reliability metrics is warranted.

EVALUATING 40 WHO PRIORITY VIRAL PATHOGENS

The WHO recently prioritized 40 RNA viruses (with 44 antigens as some viruses have multiple antigenic proteins) for research and countermeasure development, based on their potential for transmission, virulence, and the current lack of effective medical interventions[1]. This list includes both extensively studied pathogens—such as influenza, HIV, and SARS-CoV-2—and less characterized viruses like Lassa, Nipah, and Lujo viruses[107]. Leveraging our calibrated reliability metrics, we aimed to identify the most effective modeling approaches for antigenic proteins from each WHO-prioritized virus, particularly those where scarce experimental data limits direct model evaluation. We assessed

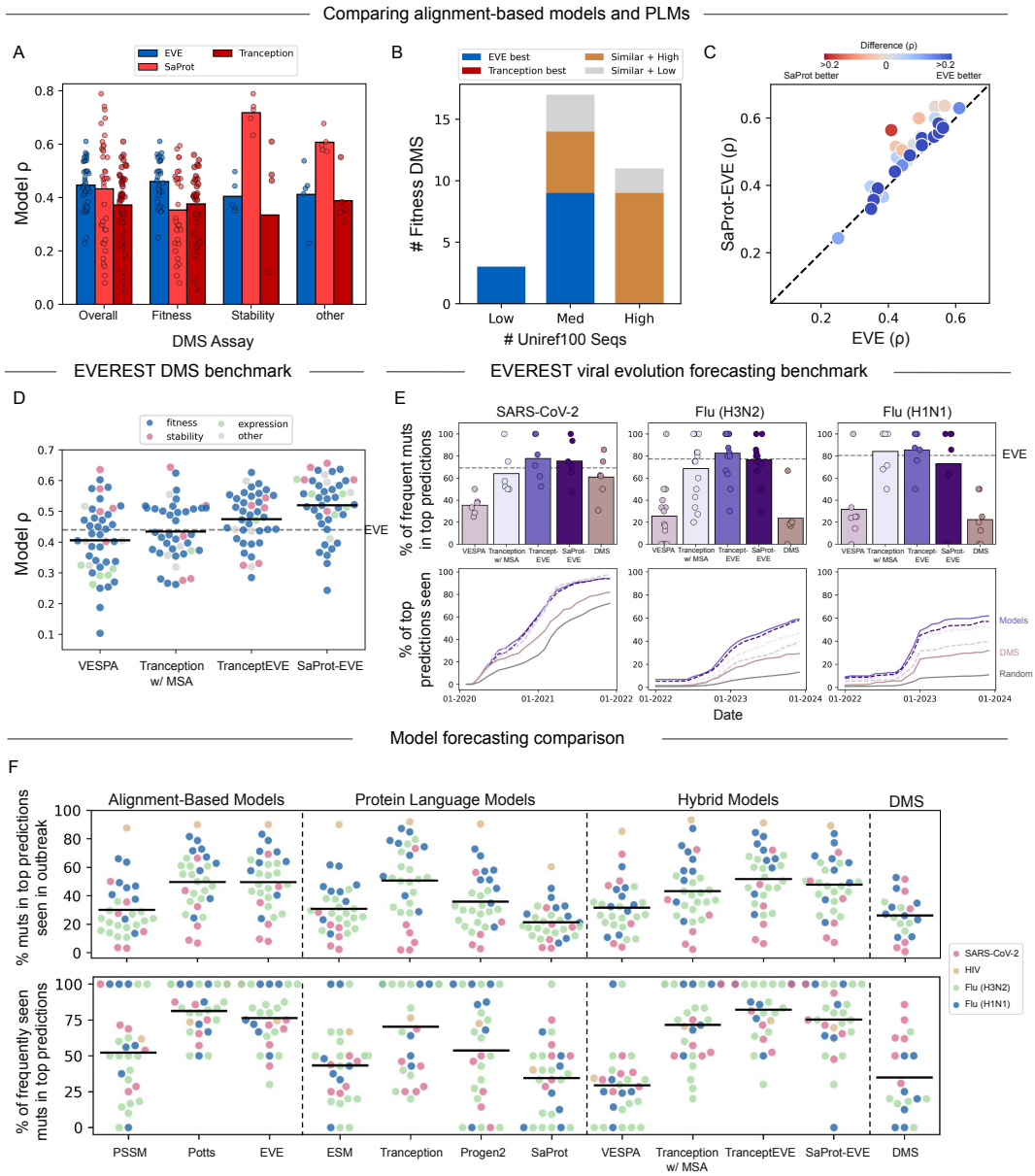


Figure 4: Combining PLM and alignment-based models improves performance.
A. Alignment-based models and PLMs perform best on different DMS phenotypes.
B. Relative performance of EVE and Tranception at different sequence-depth-regimes: Low (<3000), Medium (3000-30,000), High (>30,000). Colors indicate the best performing model (ρ difference > 0.1 between EVE and Tranception), or if the models had similar and high performance (ρ difference < 0.1, and $\rho > 0.4$) or similar and low performance (ρ difference < 0.1, and $\rho < 0.4$). This analysis considered fitness DMSs. Number of sequences determined by MMseqs search on UniRef100[104]
C. SaProt-EVE performs as well as EVE and exceeds EVE when SaProt is a better model.
D. Hybrid models performance on EVEREST DMS benchmark. SaProt-EVE and TranceptEVE have improved performance relative to either model alone (EVE, dotted line), as well as other hybrid models ($P < 0.005$, paired t-test).
E. Hybrid model performance on EVEREST viral forecasting benchmark. Top: percentage of highly frequent mutations captured by model predictions across 6 SARS-CoV-2 clades, 9 H1N1 clades and 15 H3N2 clades. Bottom: percentage of model-predicted mutations subsequently observed during evolution. Time-course analyses shown for representative clades—SARS-CoV-2 Wuhan (wild-type vaccine strain) and current flu vaccine strains for H1N1 (6B.1A.5a.2a.1), and H3N2 (3C.2a1b.2a2a.3a.1). Model predictions correspond to top 250 ranked mutations (see Fig. S7,S8) for alternative thresholds). Frequent mutations seen >1k times for flu and >10k times for SARS-CoV-2.
F. Percentage of model-predicted mutations subsequently observed during evolution (top) and percentage of highly frequent mutations captured by model predictions (bottom) across 6 SARS-CoV-2 clades, 9 H1N1 clades, 15 H3N2 clades and 1 HIV clade.

whether current models provide sufficient accuracy or whether improved models and additional data are needed to enhance predictions.

Using our calibrated metrics, we find that existing state-of-the-art models will likely fail to reliably predict the functional impacts of mutations in a large fraction of these antigens (Fig. 5E; Table S6). Of the 44 WHO viral antigens, only 21 have sufficient sequence diversity to be accurately modeled using EVE. Only 7 of the 44 antigens—mostly influenza hemagglutinins—demonstrated high PLM confidence for SaProt, the PLM with greatest average reliability (Fig. 5D). EVE's low reliability may seem surprising given its strong performance on DMS benchmarks, but these benchmarks likely overestimate model generalizability due to their focus on well-characterized viruses (Fig. 17C). Applying reliability estimates to WHO-prioritized viruses therefore provides a more balanced assessment of current model limitations.

These findings highlight that additional sequences and structures are required for more accurate modeling of pandemic risk pathogens. Many under-sequenced viral families and structural classes cluster together with low reliability scores (Fig. 5E, S17A-B). For instance, most Class II fusion accompanying proteins show low reliability and have some of the lowest numbers of related sequences in UniRef (both pre- and post-clustering), reflecting biases in available training data and structural coverage. Given these limitations, aggregating scores from the best-performing models is expected to mitigate some of these issues. Moreover, even "unreliable" predictions from understudied viruses may still prove useful for prioritizing mutations or identifying conserved regions in the absence of experimental data. We provide model predictions for all 44 antigens across the 40 prioritized viruses, alongside estimated model reliability scores, as a community resource.

We anticipate these predictions can be used in multiple ways, given that most of these viruses lack extensive experimental characterization. For therapeutic design, it is essential to identify evolutionarily conserved regions less likely to mutate in the future. For example, for the Zika envelope protein dimer, a class II fusion protein, model predictions reflect the known evolutionary constrained fusion loop and variable glycan loop[109] (Fig. 5F). For newly emerging viruses, top mutation predictions could also be incorporated into therapeutic designs. As an example of the predictive power of these models, we highlight the H1N1 "swine flu" pandemic—a novel spillover from pigs distinct from prior flu pandemics—and the predictability of the S179N mutation that necessitated the first vaccine strain update[110]. EVE, without any training leakage, predicted this mutation in the top scores of all three clade models prior to the update. Evolutionary sequence models therefore identified this major immunogenic change at the very start of the H1N1 pandemic, six years before its eventual dominance.

BIOSECURITY ASSESSMENT

Biosecurity requires a careful balancing of risks and benefits, which are inherently coupled. Improvements in predictive modeling can substantially benefit public health, while simultaneously increasing the potential for misuse. The same capabilities that enable rational vaccine and therapeutic design could, in principle, be repurposed to engineer more virulent or immune-evasive pathogens.

To limit the potential for misuse of the EVEREST dataset, we deliberately assess only a narrow and constrained component of the broader viral design problem relevant to biosecurity. Our analyses are limited to models that predict the effects of single amino-acid substitutions in individual viral proteins. We exclude nucleotide-level effects, combinations of mutations, insertions and deletions, and full viral genome contexts. We further restrict our benchmarks to functions relevant to viral surveillance and therapeutic design, rather than explicitly harm-enhancing phenotypes such as increased pathogenicity or virulence, and we exclude viral antibody-escape deep mutational scanning datasets. In addition, we focus on functions learnable from naturally occurring evolutionary sequences and structures, rather than those relevant under misuse-oriented conditions (e.g., non-physiological temperature or pH ranges).

To enable pandemic preparedness, we show for the first time that current mutation effect predictors are unreliable for more than half of pandemic-threat viruses. While this hinders the immediate biosecurity risk of these models generalizing to a wide range of viruses with potential for harm, it also highlights that limits of current models. Nevertheless, viral performance has improved over time (Fig. S16B), and is likely continue to improve with better models, more viral sequences and newly available structure data[94,96–98], and finetuning on specific viruses[46,111–113]. In this context, EVEREST provides not only a framework for evaluating models relevant to pandemic preparedness, but also a mechanism for monitoring the emergence of dual-use biosecurity risks as predictive capabilities advance.

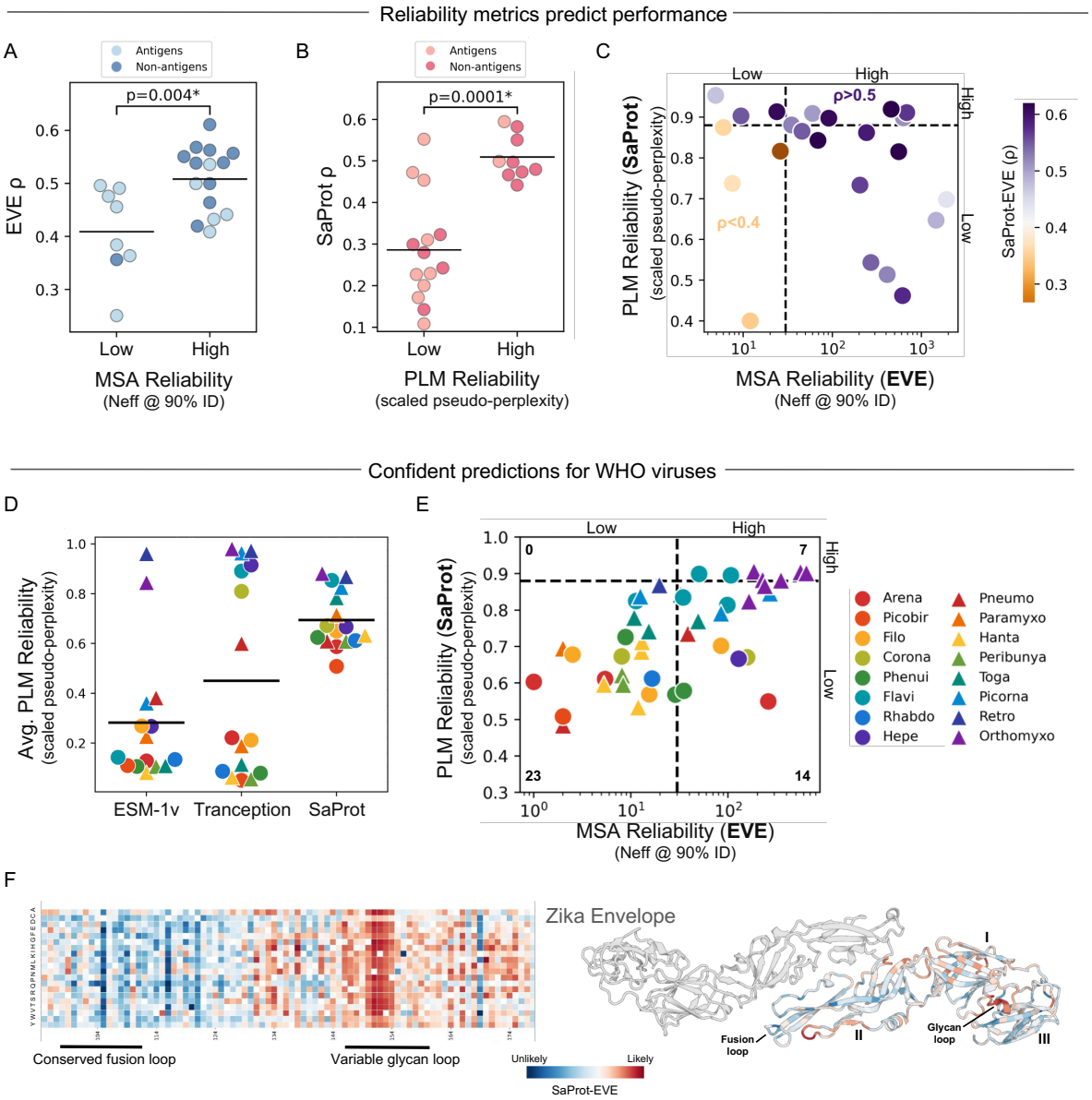


Figure 5: Quantifying model reliability across WHO priority viruses.
A. EVE models with high MSA reliability scores have significantly better performance than those with low reliability. MSA reliability is measured as the number of effective sequences (Neff) with greater than 90% identity to the query (Neff @ 90% ID).
B. SaProt models with high PLM reliability scores have significantly better performance than those with low reliability. PLM reliability is measured as a scaled pseudo-perplexity.
C. Hybrid model SaProt-EVE has correlation below 0.4 only where both reliability metrics for EVE and SaProt are low and correlation greater than 0.5 when both reliability metrics are high.
D. Inferring PLM reliability for 44 viral antigens from 40 WHO-prioritized viruses. Structure-aware PLM, SaProt, has on average higher reliability than sequence-only PLMs, ESM-1v and Tranception. Plotted are the average reliability score for a given viral family.
E. EVE and SaProt reliability scores for all WHO viral antigens. More than half of WHO antigens have both low MSA and PLM confidence. Based on our viral benchmarks, antigens with high model confidence are expected to yield Spearman correlations greater than 0.4 with DMS experiments.
F. SaProt-EVE predictions for priority virus Zika envelope, showing *in silico* mutational scan and site-average fitness scores mapped on the structure (AlphaFold3[108] structure and PDB 5ire). Fusion loop (98-107) and glycan 150 loop (147-161).

Conclusion

In this study, we present the first virus-specific, large-scale comparison of variant effect predictors, evaluating alignment-based models, protein language models (PLMs), and hybrid approaches across 45 deep mutational scanning experiments and 31 real-world forecasting datasets using our EVEREST benchmark. We also introduce calibrated reliability metrics that quantify model confidence in the absence of validation data—a scenario typical of emerging viral outbreaks. Applying these metrics reveals that only approximately half of the 40 WHO priority RNA viruses can currently be predicted with high reliability, underscoring fundamental limitations in existing methods and data for many under-studied pandemic threats.

Alignment-based models trained on family-specific viral sequences frequently outperform PLMs trained on large, general protein corpora. Notably, this advantage persists even in low-sequence-depth regimes characteristic of novel outbreaks, where "single-sequence" PLMs were previously assumed to yield superior performance. These results suggest that widely used PLMs such as ESM-2 are often ill-suited for viral mutation effect prediction, despite their growing use in this domain[44–46].

The relative underperformance of PLMs on viruses likely reflects their limited representation in training data, which is disproportionately reduced in clustered sequence datasets (e.g., UniRef90 and UniRef50). Whereas PLMs benefit from scale and diversity across distant protein families in training, alignment-based approaches benefit from training on more evolutionarily relevant sequences with higher identity to the query. This contrast highlights a fundamental modeling difference: alignment-based methods effectively treat all homologous sequences as equally informative (after redundancy weighting), while PLMs implicitly attempt to reason over relevance—a task made difficult when viral sequences are sparse or underrepresented.

These findings suggest several priorities for future research. First, PLM performance on viral proteins may be improved through training on larger, unclustered datasets, viral-specific pretraining or fine-tuning, domain adaptation, or the integration of additional data modalities[46,111–113]. Structure-aware modeling, in particular, may help compensate for sparse viral sequence data by leveraging more conserved structural constraints. Similar strategies may be necessary for RNA language models applied to viruses, which likely suffer from analogous dataset limitations[114]. Second, alignment-based models could be strengthened by weighting sequences according to evolutionary proximity to the query, rather than relying solely on clustering heuristics, and by developing intrinsic, model-based reliability estimates. Third, alternative approaches also offer promising ways of improving viral mutation effect predictions. For instance, GEMME—a phylogeny-based model incorporating mutation frequency, evolutionary accessibility, and subtree conservation—achieves performance comparable to the best models in our benchmarks (fitness DMS prediction, $\rho = 0.5$; Table S5). Fourth, evaluation strategies should extend beyond single-mutation concordance, incorporating multi-mutation genotypes and strain-level effects that better reflect real viral evolution.

Many of the guidelines for future model improvements outlined here for viruses are likely to generalize to other proteins with limited sequence diversity. Yet, for many pandemic-risk viruses, with limited existing data, methodological advances alone are unlikely to suffice. Increased sequencing efforts, not only on the WHO priority viral species themselves, but on related viruses in their natural reservoirs, will be essential for improving modeling and enabling pandemic readiness. Notably, SARS-CoV-2 modeling success stems not from the millions of highly similar COVID-19 pandemic sequences, but from approximately 3,000 diverse pre-pandemic coronavirus sequences sampled primarily from bats and pigs, which provided the essential evolutionary and functional template for rapid pandemic response[2].

Data and Code Availability

Data, results, and code are available at https://github.com/debbiemarkslab/priority-viruses for 45 viral DMSs, 44 WHO-priority viral antigens, and viral forecasting of 31 clades. Code for model training and for mutation effect scoring is available through ProteinGym[17] at https://github.com/OATML-Markslab/ProteinGym. DMS data used for model evaluation collected and standardized from papers listed in Table S1[4,50,54–78]. Sequence data analyzed for forecasting from GISAID and LANL[85,115].

Acknowledgments

The authors thank Aaron Kollasch and other members of the Marks lab for valuable discussions. Special thanks to the teams of experimentalists who developed and performed the viral DMS assays this work is built on. If you are using these assays in your work, please cite the corresponding papers in Table S1. This work was supported by the Coalition for Epidemic Preparedness Innovations (CEPI), including Danny Scarponi, and biosecurity analysis completed alongside experts from CEPI, particularly Andrew Hebbeler (Biosecurity Director). S.G. was supported

by funding from the FutureHouse AI-for-Science Postdoctoral Fellowship; work at FutureHouse is supported by the generosity of Eric and Wendy Schmidt. We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, as well as LANL, on which some of this research is based.

REFERENCES

[1] WHO. Pathogens prioritization: a scientific framework for epidemic and pandemic research preparedness. *World Health Organization: Geneva, Switzerland*, 2024.

[2] 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.

[3] 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.

[4] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *cell*, 182(5):1295–1310, 2020.

[5] Bernadeta Dadonaite, Jack Brown, Teagan E McMahon, Ariana G Farrell, Marlin D Figgins, Daniel Asarnow, Cameron Stewart, Jimin Lee, Jenni Logue, Trevor Bedford, et al. Spike deep mutational scanning helps predict success of sars-cov-2 clades. *Nature*, 631(8021):617–626, 2024.

[6] Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, 2014.

[7] Brian Lee, Ahmed Abdul Quadeer, Muhammad Saqib Sohail, et al. Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *Nature Communications*, 16(1):441, 2025.

[8] Noémie Lefrancq, Loréna Duret, Valérie Bouchez, et al. Learning the fitness dynamics of pathogens from phylogenies. *Nature*, pages 1–8, 2025.

[9] Muhammad Saqib Sohail, Raymond HY Louie, Matthew R McKay, and John P Barton. Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature biotechnology*, 39(4):472–479, 2021.

[10] Jesse D Bloom and Richard A Neher. Fitness effects of mutations to sars-cov-2 proteins. *Virus Evolution*, 9(2): vead055, 2023.

[11] Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *Elife*, 3:e03568, 2014.

[12] Jeffrey A Ruffolo, Stephen Nayfach, Joseph Gallagher, Aadyot Bhatnagar, Joel Beazer, Riffat Hussain, Jordan Russ, Jennifer Yip, Emily Hill, Martin Pacesa, et al. Design of highly functional genome editors by modelling crispr–cas sequences. *Nature*, 645(8080):518–525, 2025.

[13] Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. *Nature biotechnology*, 42(2):216–228, 2024.

[14] William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

[15] Xinran Lian, Nikša Praljak, Subu K Subramanian, Sarah Wasinger, Rama Ranganathan, and Andrew L Ferguson. Deep-learning-based design of synthetic orthologs of sh3 signaling domains. *Cell Systems*, 15(8):725–737, 2024.

[16] Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1):2403, April 2021.

[17] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

[18] 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.

[19] 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.

[20] Jaclyn K Mann, John P Barton, Andrew L Ferguson, Saleha Omarjee, Bruce D Walker, Arup Chakraborty, and Thumbi Ndung'u. The fitness landscape of hiv-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS computational biology*, 10(8):e1003776, 2014.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[28] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR, 2022.

[29] 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.

[30] Gregory R Hart and Andrew L Ferguson. Empirical fitness models for hepatitis c virus immunogen design. *Physical Biology*, 12(6):066006, 2015.

[31] Andrew L Ferguson, Jaclyn K Mann, Saleha Omarjee, Thumbi Ndung'u, Bruce D Walker, and Arup K Chakraborty. Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[32] Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt. Epistatic models predict mutable sites in sars-cov-2 proteins and epitopes. *Proceedings of the National Academy of Sciences*, 119(4): e2113118119, 2022.

[33] Raymond HY Louie, Kevin J Kaczorowski, John P Barton, Arup K Chakraborty, and Matthew R McKay. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proceedings of the National Academy of Sciences*, 115(4):E564–E573, 2018.

[34] John P Barton, Nilu Goonetilleke, Thomas C Butler, Bruce D Walker, Andrew J McMichael, and Arup K Chakraborty. Relative rate and location of intra-host hiv evolution to evade cellular immunity are predictable. *Nature communications*, 7(1):11660, 2016.

[35] Ahmed A Quadeer, Raymond HY Louie, and Matthew R McKay. Identifying immunologically-vulnerable regions of the hcv e2 glycoprotein and broadly neutralizing antibodies that target them. *Nature communications*, 10(1):2073, 2019.

[36] Ahmed A Quadeer, John P Barton, Arup K Chakraborty, and Matthew R McKay. Deconvolving mutational patterns of poliovirus outbreaks reveals its intrinsic fitness landscape. *Nature communications*, 11(1):377, 2020.

[37] Pedro Miguel Carneiro Jeronimo, Cleber Furtado Aksenen, Igor Oliveira Duarte, Roberto D Lins, and Fabio Miyajima. Evolutionary deletions within the sars-cov-2 genome as signature trends for virus fitness and adaptation. *Journal of Virology*, 98(1):e01404–23, 2024.

[38] 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. *bioRxiv*, December 2022.

[39] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[40] Ruben Weitzman, Peter Mørch Groth, Lood Van Niekerk, Aoi Otani, Yarin Gal, Debora Marks, and Pascal Notin. Protriever: End-to-end differentiable protein homology search for fitness prediction. *arXiv preprint arXiv:2506.08954*, 2025.

[41] 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.

[42] Yang Tan, Bingxin Zhou, Lirong Zheng, Guisheng Fan, and Liang Hong. Semantical and geometrical protein encoding toward enhanced bioactivity and thermostability. *Elife*, 13:RP98033, 2025.

[43] Steffanie Paul, Aaron Kollasch, Pascal Notin, and Debora Marks. Combining structure and sequence for superior fitness prediction. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[44] Kieran D Lamb, Joseph Hughes, Spyros Lytras, Francesca Young, Orges Koci, James Herzig, Simon C Lovell, Joe Grove, Ke Yuan, and David L Robertson. From a single sequence to evolutionary trajectories: protein language models capture the evolutionary potential of sars-cov-2 protein sequences. *bioRxiv*, pages 2024–07, 2024.

[45] Brent Allman, Luiz Vieira, Daniel J Diaz, and Claus O Wilke. A systematic evaluation of the language-of-viral-escape model using multiple machine learning frameworks. *bioRxiv*, pages 2024–09, 2024.

[46] Brian L Hie, Kevin K Yang, and Peter S Kim. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell systems*, 13(4):274–285, 2022.

[47] Bodo Linz, Helen M Windsor, John J McGraw, Lori M Hansen, John P Gajewski, Lynn P Tomsho, Caylie M Hake, Jay V Solnick, Stephan C Schuster, and Barry J Marshall. A mutation burst during the acute phase of helicobacter pylori infection in humans and rhesus macaques. *Nature communications*, 5(1):4165, 2014.

[48] Nobuhiko Tokuriki, Christopher J Oldfield, Vladimir N Uversky, Igor N Berezovsky, and Dan S Tawfik. Do viral proteins possess unique biophysical features? *Trends in biochemical sciences*, 34(2):53–59, 2009.

[49] 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.

[50] Bernadeta Dadonaite, Katharine HD Crawford, Caelan E Radford, Ariana G Farrell, C Yu Timothy, William W Hannon, Panpan Zhou, Raiees Andrabi, Dennis R Burton, Lihong Liu, et al. A pseudovirus system enables deep mutational scanning of the full sars-cov-2 spike. *Cell*, 186(6):1263–1278, 2023.

[51] Frances C Welsh, Rachel T Eguia, Juhye M Lee, Hugh K Haddox, Jared Galloway, Nguyen Van Vinh Chau, Andrea N Loes, John Huddleston, C Yu Timothy, Mai Quynh Le, et al. Age-dependent heterogeneity in the antigenic effects of mutations to influenza hemagglutinin. *Cell Host & Microbe*, 32(8):1397–1411, 2024.

[52] C Yu Timothy, 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.

[53] 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.

[54] Sam Sinai, Nina Jain, George M Church, and Eric D Kelsic. Generative aav capsid diversification by latent interpolation. *bioRxiv*, pages 2021–04, 2021.

[55] Florian Mattenberger, Victor Latorre, Omer Tirosh, Adi Stern, and Ron Geller. Globally defining the effects of mutations in a picornavirus capsid. *Elife*, 10:e64256, 2021.

[56] Beatriz Álvarez-Rodríguez, Sebastian Velandia-Álvarez, Christina Toft, and Ron Geller. Mapping the mutational landscape of a full viral proteome reveals distinct profiles of mutation tolerability. *bioRxiv*, pages 2024–03, 2024.

[57] Amporn Suphatrakul, Pratsaneeyaporn Posiri, Nittaya Srisuk, Rapirat Nantachokchawapan, Suppachoke Onnome, Juthathip Mongkolsapaya, and Bunpote Siridechadilok. Functional analysis of flavivirus replicase by deep mutational scanning of dengue ns5. *bioRxiv*, pages 2023–03, 2023.

[58] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

[59] William Bakhache, Walker Orr, Lauren McCormick, and Patrick T Dolan. Uncovering structural plasticity of enterovirus a through deep insertional and deletional scanning. *Research Square*, 2024.

[60] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of hiv envelope. *Elife*, 7:e34420, 2018.

[61] Maria Duenas-Decamp, Li Jiang, Daniel Bolon, and Paul R Clapham. Saturation mutagenesis of the hiv-1 envelope cd4 binding loop reveals residues controlling distinct trimer conformations. *PLoS pathogens*, 12(11): e1005988, 2016.

[62] Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to hiv's envelope protein on viral replication in cell culture. *PLoS pathogens*, 12(12):e1006114, 2016.

[63] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6):155, 2016.

[64] Nicholas C Wu, Arthur P Young, Laith Q Al-Mawsawi, C Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Robert G Chin, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports*, 4(1):4942, 2014.

[65] Juhye M Lee, John Huddleston, Michael B Doud, Kathryn A Hooper, Nicholas C Wu, Trevor Bedford, and Jesse D Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35):E8276–E8285, 2018.

[66] Bernadeta Dadonaite, Jenny J Ahn, Jordan T Ort, Jin Yu, Colleen Furey, Annie Dosey, William W Hannon, Amy L Vincent Baker, Richard J Webby, Neil P King, et al. Deep mutational scanning of h5 hemagglutinin to inform influenza virus surveillance. *PLoS biology*, 22(11):e3002916, 2024.

[67] Michael B Doud, Orr Ashenberg, and Jesse D Bloom. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molecular biology and evolution*, 32(11):2944–2960, 2015.

[68] Li Jiang, Ping Liu, Claudia Bank, Nicholas Renzette, Kristina Prachanronarong, Lutfu S Yilmaz, Daniel R Caffrey, Konstantin B Zeldovich, Celia A Schiffer, Timothy F Kowalik, et al. A balance between inhibitor binding and substrate processing confers influenza drug resistance. *Journal of molecular biology*, 428(3):538–553, 2016.

[69] Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS genetics*, 11(7):e1005310, 2015.

[70] YQ Shirleen Soh, Louise H Moncla, Rachel Eguia, Trevor Bedford, and Jesse D Bloom. Comprehensive mapping of adaptation of the avian influenza polymerase protein pb2 to humans. *Elife*, 8:e45079, 2019.

[71] Yuan Li, Sarah Arcos, Kimberly R Sabsay, Aartjan JW Te Velthuis, and Adam S Lauring. Deep mutational scanning reveals the functional constraints and evolutionary potential of the influenza a virus pb1 protein. *Journal of virology*, 97(11):e01329–23, 2023.

[72] Nicholas C Wu, Yushen Du, Shuai Le, Arthur P Young, Tian-Hao Zhang, Yuanyuan Wang, Jian Zhou, Janice M Yoshizawa, Ling Dong, Xinmin Li, et al. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza a virus m segment. *BMC genomics*, 17:1–15, 2016.

[73] Tyler Starr. Deep mutational scanning of sars-related cov rbds. `https://github.com/tstarrlab/SARSr-CoV-RBD_DMS`, 2024.

[74] Ashley L Taylor and Tyler N Starr. Deep mutational scans of xbb. 1.5 and bq. 1.1 reveal ongoing epistatic drift during sars-cov-2 evolution. *PLoS Pathogens*, 19(12):e1011901, 2023.

[75] Julia M Flynn, Neha Samant, Gily Schneider-Nachum, David T Barkan, Nese Kurt Yilmaz, Celia A Schiffer, Stephanie A Moquin, Dustin Dovala, and Daniel NA Bolon. Comprehensive fitness landscape of sars-cov-2 mpro reveals insights into viral resistance mechanisms. *Elife*, 11:e77433, 2022.

[76] Xinyu Wu, Margareta Go, Julie V Nguyen, Nathan W Kuchel, Bernadine GC Lu, Kathleen Zeglinski, Kym N Lowes, Dale J Calleja, Jeffrey P Mitchell, Guillaume Lessene, et al. Mutational profiling of sars-cov-2 papain-like protease reveals requirements for function, structure, and drug escape. *Nature Communications*, 15(1): 6219, 2024.

[77] Brendan B Larsen, Teagan McMahon, Jack T Brown, Zhaoqian Wang, Caelan E Radford, James E Crowe, David Veesler, and Jesse D Bloom. Functional and antigenic landscape of the nipah virus receptor-binding protein. *Cell*, 188(9):2480–2494, 2025.

[78] Caleb R Carr, Katharine HD Crawford, Michael Murphy, Jared G Galloway, Hugh K Haddox, Frederick A Matsen, Kristian G Andersen, Neil P King, and Jesse D Bloom. Deep mutational scanning reveals functional constraints and antibody-escape potential of lassa virus glycoprotein complex. *Immunity*, 57(9):2061–2076, 2024.

[79] 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.

[80] L Buonaguro, ML Tornesello, and FM Buonaguro. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *Journal of virology*, 81(19):10209–10219, 2007.

[81] Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

[82] Eli Weinstein, Alan Amin, Jonathan Frazer, and Debora Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness. *Advances in neural information processing systems*, 35:5484–5497, 2022.

[83] Luiz C Vieira, Morgan L Handojo, and Claus O Wilke. Scaling down for efficiency: Medium-sized protein language models perform well at transfer learning on realistic datasets. *bioRxiv*, pages 2024–11, 2025.

[84] Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of llms fall short? *arXiv preprint arXiv:2410.18745*, 2024.

[85] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael Tc Lee, Winston Yeo, Gisaid Core Curation Team, and Sebastian Maurer-Stroh. GISAID's role in pandemic response. *China CDC Wkly*, 3(49):1049–1051, December 2021.

[86] Elena Dalla Vecchia. Pathoplexus: towards fair and transparent sequence sharing. *The Lancet Microbe*, 5(12), 2024.

[87] Suguru Nishijima, Anthony Fullam, Thomas SB Schmidt, Michael Kuhn, and Peer Bork. Vire: a metagenome-derived, planetary-scale virome resource with environmental context. *Nucleic Acids Research*, page gkaf1225, 2025.

[88] M Steinegger and J Söding. Clustering huge protein sequence sets in linear time. nat commun 9: 2542, 2018.

[89] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A Gurbich, Varsha Kale, Anton Korobeynikov, Shriya Raj, Alexander B Rogers, Ekaterina Sakharova, Santiago Sanchez, Darren J Wilkinson, and Robert D Finn. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 12 2022. doi: 10.1093/nar/gkac1080.

[90] Rayan Chikhi, Brice Raffestin, Anton Korobeynikov, Robert Edgar, and Artem Babaian. Logan: planetary-scale genome assembly surveys life's diversity. *bioRxiv*, pages 2024–07, 2024.

[91] Kristoffer Illergård, David H Ardell, and Arne Elofsson. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

[92] Anne Lopes, Paulo Tavares, Marie-Agnès Petit, Raphaël Guérois, and Sophie Zinn-Justin. Automated classification of tailed bacteriophages according to their neck organization. *BMC genomics*, 15:1–17, 2014.

[93] Jonathon Mifsud, Spyros Lytras, Michael R Oliver, Kamilla Toon, Vincenzo A Costa, Edward C Holmes, and Joe Grove. Mapping glycoprotein structure reveals flaviviridae evolutionary history. *Nature*, 633(8030):695–703, 2024.

[94] Jason Nomburg, Erin E Doherty, Nathan Price, Daniel Bellieny-Rabelo, Yong K Zhu, and Jennifer A Doudna. Birth of protein folds and functions in the virome. *Nature*, 633(8030):710–717, 2024.

[95] Timothy K Soh, Sofia Ognibene, Saskia Sanders, Robin Schäper, Benedikt B Kaufer, and Jens B Bosse. A proteome-wide structural systems approach reveals insights into protein families of all human herpesviruses. *Nature Communications*, 15(1):10230, 2024.

[96] Ulad Litvin, Spyros Lytras, Alexander Jack, David L Robertson, Joe Grove, and Joseph Hughes. Viro3d: a comprehensive database of virus protein structure predictions. *bioRxiv*, pages 2024–12, 2024.

[97] Roni Odai, Michele Leemann, Tamim Al-Murad, Minhal Abdullah, Lena Shyrokova, Tanel Tenson, Vasili Hauryliuk, Janani Durairaj, Joana Pereira, and Gemma Catherine Atkinson. The viral alphafold database of monomers and homodimers reveals conserved protein folds in viruses of bacteria, archaea, and eukaryotes. *bioRxiv*, 2025. doi: 10.1101/2025.05.14.653371.

[98] Rachel Seongeun Kim, Eli Levy Karin, Milot Mirdita, Rayan Chikhi, and Martin Steinegger. Bfvd—a large repository of predicted viral protein structures. *Nucleic Acids Research*, 53(D1):D340–D347, 2025.

[99] 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.

[100] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods*, 16(7):603–606, 2019.

[101] Consortium UniProt. Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 49(D1): D480–D489, 2025.

[102] 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.

[103] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[104] Maria Hauser, Martin Steinegger, and Johannes Söding. Mmseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, 32(9):1323–1330, 2016.

[105] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[106] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.

[107] Julian W Tang, Tommy T Lam, Hassan Zaraket, W Ian Lipkin, Steven J Drews, Todd F Hatchette, Jean-Michel Heraud, Marion P Koopmans, Ashta Mary Abraham, Amal Baraket, et al. Global epidemiology of non-influenza rna respiratory viruses: data gaps and a growing need for surveillance. *The Lancet Infectious Diseases*, 17(10): e320–e326, 2017.

[108] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[109] Lianpan Dai, Jian Song, Xishan Lu, Yong-Qiang Deng, Abednego Moki Musyoki, Huijun Cheng, Yanfang Zhang, Yuan Yuan, Hao Song, Joel Haywood, et al. Structures of the zika virus envelope protein and its complex with a flavivirus broadly protective antibody. *Cell host & microbe*, 19(5):696–704, 2016.

[110] Jenna J Guthmiller, Julianna Han, Lei Li, Alec W Freyn, Sean TH Liu, Olivia Stovicek, Christopher T Stamper, Haley L Dugan, Micah E Tepora, Henry A Utset, et al. First exposure to the pandemic h1n1 virus induced broadly neutralizing antibodies targeting hemagglutinin head epitopes. *Science translational medicine*, 13(596): eabg4535, 2021.

[111] Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, Genotype to Phenotype Japan (G2P-Japan) Consortium, and Kei Sato. A protein language model for exploring viral fitness landscapes. *bioRxiv*, pages 2024–03, 2024.

[112] Spyros Lytras, Adam Strange, Jumpei Ito, and Kei Sato. Inferring context-specific site variation with evotuned protein language models. *bioRxiv*, pages 2025–02, 2025.

[113] Tianlai Chen, Madeleine Dumas, Rio Watson, Sophia Vincoff, Christina Peng, Lin Zhao, Lauren Hong, Sarah Pertsemlidis, Mayumi Shaepers-Cheu, Tian Zi Wang, et al. Pepmlm: target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling. *ArXiv*, pages arXiv–2310, 2024.

[114] Rohit Arora, Murphy Angelo, Christian Andrew Choe, Courtney Shearer, Aaron Kollasch, Fiona Qu, Ruben Weitzman, Artem Gazizov, Sarah Gurev, Erik Xie, et al. Rnagym: Large-scale benchmarks for rna fitness and structure prediction. *bioRxiv*, pages 2025–06, 2025.

[115] Cristian Apetrei, Beatrice Hahn, Andrew Rambaut, Steven Wolinsky, J Rodney Brister, B Keele, and C Faser. Hiv sequence compendium 2021. *Theoretical Biology and Biophysics Group*, 2021.

[116] 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`.

[117] Eric Lewitus, Yifan Li, Hongjun Bai, Phuc Pham, and Morgane Rolland. Hiv-1 gag, pol, and env diversified with limited adaptation since the 1980s. *Mbio*, 15(3):e01749–23, 2024.

[118] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[119] Mika O Salminen, BO Johansson, Anders Sönnerborg, Seyoum Ayehunie, Deanna Gotte, Pauli Leinikki, Donald S Burke, and FRANCINE E McCUTCHAN. Full-length sequence of an ethiopian human immunodeficiency virus type 1 (hiv-1) isolate of genetic subtype c. *AIDS research and human retroviruses*, 12(14):1329–1339, 1996.

[120] 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, December 2018.

[121] Victor M Markowitz, I-Min A Chen, Krishna Palaniappan, Ken Chu, Ernest Szeto, Yuri Grechkin, Anna Ratner, Biju Jacob, Jinghua Huang, Peter Williams, et al. Img: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research*, 40(D1):D115–D122, 2012.

[122] Andre Cornman, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, pages 2024–08, 2024.

[123] L Steven Johnson, Sean R Eddy, and Elon Portugaly. Hidden markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):431, August 2010.

[124] 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.

[125] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77, 2020.

[126] Thomas A Hopf, Anna G Green, Benjamin Schubert, Sophia Mersmann, Charlotta PI Schärfe, John B Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J Riesselman, Perry Palmedo, et al. The evcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584, 2019.

[127] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[128] N. Ferruz, S. Schmidt, and B. Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(4348), 2022. doi: 10.1038/s41467-022-32007-7.

[129] E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G.M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16:1315–1322, 2019. doi: 10.1038/s41592-019-0598-1.

[130] J Fernando Bazan and Robert J Fletterick. Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications. *Proceedings of the National Academy of Sciences*, 85 (21):7872–7876, 1988.

[131] Antonina Andreeva, Dave Howorth, John-Marc Chandonia, Steven E Brenner, Tim JP Hubbard, Cyrus Chothia, and Alexey G Murzin. Data growth and its impact on the scop database: new developments. *Nucleic acids research*, 2007.

[132] Lesley H Greene, Tony E Lewis, Sarah Addou, Alison Cuff, Tim Dallman, Mark Dibley, Oliver Redfern, Frances Pearl, Rekha Nambudiry, Adam Reid, et al. The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic acids research*, 2007.

[133] Khushboo Bafna, Christopher L Cioffi, Robert M Krug, and Gaetano T Montelione. Structural similarities between sars-cov2 3clpro and other viral proteases suggest potential lead molecules for developing broad spectrum antivirals. *Frontiers in Chemistry*, 10:948553, 2022.

[134] Wenyu Li, Srinivasa M Srinivasula, Jijie Chai, Pingwei Li, Jia-Wei Wu, ZhiJia Zhang, Emad S Alnemri, and Yigong Shi. Structural insights into the pro-apoptotic function of mitochondrial serine protease htra2/omi. *Nature structural biology*, 9(6):436–441, 2002.

[135] Thomas A Hopf, Charlotta PI Schärfe, João PGLM Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre MJJ Bonvin, and Debora S Marks. Sequence co-evolution gives 3d contacts and structures of protein complexes. *elife*, 3:e03430, 2014.

A METHODS

A.1 EVEREST DEEP MUTATION SCANNING BENCHMARK

We curated and standardized a collection of 45 viral deep mutational scanning (DMS) datasets from 25 publications[4,50,54–78], more than doubling the number of viral datasets included in previous general-purpose benchmark (ProteinGym[17]). The resulting EVEREST DMS benchmark spans viruses of major importance for vaccine development—including SARS-CoV-2 and other sarbecoviruses, seasonal and pandemic influenza viruses, and high-priority pathogens such as Lassa and Nipah—as well as viruses relevant to viral vector design (e.g., AAV).

We systematically searched for all available viral DMS experiments and included datasets measuring single–amino acid substitutions and phenotypes directly relevant to pandemic preparedness, such as protein expression, host receptor binding, viral replication, and infectivity. Datasets were excluded when the assayed phenotype was not directly comparable across viruses or was not aligned with the goals of this benchmark. Specifically, we excluded assays measuring drug inhibition or viral escape (from antibodies or polyclonal sera). While relevant for pandemic preparedness, viral escape datasets were excluded due to the limited number of antibodies or sera tested for most viruses, which can result in apparent false positives where a model correctly predicts escape but the DMS assay does not capture it.

There are some limitations in using DMSs as a benchmark. DMS phenotype single proteins under laboratory conditions that incompletely capture the full constraints a virus faces in the real world. DMSs also measure a narrow fitness landscape of largely single mutations. Yet, for comparing model performance broadly across diverse viral families, a new viral-specific DMS dataset remains our best available benchmark.

A.2 EVEREST VIRAL FORECASTING BENCHMARK

We created a new benchmark for evaluating the ability to predict viral evolution across multiple viruses and numerous phylogenetic clades: 9 influenza (H1N1), 15 influenza (H3N2), 1 HIV, and 6 SARS-CoV-2. Extensive surveillance sequencing allows us to test the ability of evolutionary sequence models to forecast frequent mutations that emerge in each clade in a retrospective study. We focus on single nucleotide mutations since multi-nucleotide mutations are very rare (i.e., $< 4\%$ of the test dataset for influenza, and $< 1\%$ for SARS-CoV-2). For alignment-based models, we perform strict time separation between training and testing datasets, while for PLMs we acknowledge and assess data leakage (test mutations included within PLM training data) that likely artificially inflate performance.

A.2.1 INFLUENZA H1N1 AND H3N2

Clade definitions 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[116]. Tracking the evolution of H1N1 and H3N2 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 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. We examined the following nested clade lineages for H1N1: pdm09, 6, 6B, 6B.1, 6B.1A, 6B.1A.5a, 6B.1A.5a.2a, and 6B.1A.5a.2a.1. We examine the following clade lineages for H3N2: 3C, 3C.2, 3C.2a, 3C.2a.1, 3C.2a2, 3C.2a.1b.1, 3C.2a.1b.2, 3C.2a.1b.2a.2, 3C.2a.1b.2a.2a.3, 3C.2a.1b.2a.2b, 3C.2a.1b.2a.2a.3a.1, 3C.3, 3C.3a, and 3C.3a.1.

Training Datasets 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[85], comprising 365,500 full-length sequences submitted before January 1, 2025. To generate an MSA per clade for alignment-based models, 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.

Testing Datasets Each model was evaluated on a corresponding test dataset containing all human sequences collected during the 2-4 year period for which 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. For forecasting analyses, we restricted to mutations observed at least 3 times, and classified mutations as "high-frequency" if they were observed more than 1,000 times.

A.2.2 HIV

Training Dataset A set of HIV-1 subtype-stratified multiple sequence alignments were sourced from Lewitus et al. (2024) [117], in which HIV sequences from Los Alamos National Laboratory (LANL) were filtered and stratified to create complete and independent sequence sets that capture the diversity of each subtype phylogeny [115]. All non-subtype C envelope (Env) protein alignments were concatenated and aligned to a reference sequence using MAFFT [118], creating an alignment of 3,452 effective sequences [118]. This reference sequence was sourced from the LANL database as one of the earliest published subtype C Env sequences which was recorded in the 1980s [119].

Testing Dataset All subtype C Env sequences available in the LANL HIV Sequence Database [115] were downloaded and aligned to the 1980 reference sequence [119] using MAFFT [118]. Analysis was stratified to all sequences in the years 1986-1995 to create a time period more comparable to other virus in the EVEREST forecasting benchmark, with an end date chosen for when the observed fraction of mutations began to plateau, although this is still several times longer than the other test periods choices. This test split is phylogenetic rather than temporal (i.e., training data even for alignment based models includes sequences after this testing period). For forecasting analyses, we restricted to mutations observed at least 100 times, and classified mutations as "high-frequency" if they were observed more than 10,000 times (due to increased sequencing).

A.2.3 SARS-CoV-2

Training Dataset We selected five SARS-CoV-2 variants which were ancestors to the majority of circulating strains during the COVID-19 pandemic: Wuhan, BA.1, BA.4, BA.5, XBB, and JN.1. For each variant, we constructed a training alignment consisting of (i) all homologous Spike protein sequences available prior to 2020 ($\sim$ 3,000 sequences), and (ii) all SARS-CoV-2 Spike sequences deposited in GISAID [85] prior to the emergence of that specific variant (Fig. S1D). Variant emergence dates were obtained from NextStrain [120].

Testing Dataset For each parent variant, we used NextStrain [120] to identify all descendant lineages. We then subset GISAID to include only sequences sampled after the variant's emergence and assigned to its descendant lineages. From these sequences, we extracted all single-nucleotide mutations observed within the two years following the variant's first appearance. For forecasting analyses, we restricted to mutations observed at least 100 times, and classified mutations as "high-frequency" if they were observed more than 10,000 times (due to increased sequencing).

A.3 Sequence and structure datasets

Both alignment-based and protein language models rely on sequence and structure databases either for retrieving homologous protein sequences or for PLM training.

UniRef100, is a database of a non-redundant protein sequence database containing all UniProt Knowledgebase records plus selected UniParc records [101]. UniRef90 and UniRef50, are 90% and 50% identity-clustered versions of UniRef100. The Big Fantastic Database (BFD) integrates sequences from UniProt (Swiss-Prot and TrEMBL) [101], Metaclust [88], and the Soil and Marine Eukaryotic Reference Catalog [100]. MGnify includes proteins from metagenomic assemblies [89].

AlphaFold databse (AF2DB) is comprised of 40 million AlphaFold predicted structures for UniRef50 sequences, which notably excludes all viral proteins. Open MetaGenomic (OMG) corpus (50% ID clustering) from combining IMG [121] and MGnify contains 200 million representative sequences from clusters with at least two members, a more 3-fold increase in sequence diversity compared to UniRef50 [122].

A.4 Alignment-based models

A.4.1 Generation of multiple sequence alignments

All alignment-based models rely on a method for generating a multiple sequence alignment on which they are trained. Multiple sequence alignments of the corresponding protein family were obtained using the method outlined in Hopf et al. (2017) [18]. This involved five search iterations of the profile HMM homology search tool JackHMMER [123] against the specified sequences database. We evaluated the impact of searching against three database with vastly different number of sequences: UniRef100, UniRef90, and UniRef100+BFD+MGnify.

We generated alignments with six different length-normalized bit scores: 0.5, 0.3, 0.1, 0.05, 0.03, and 0.01 bits per residue. A challenge during the process of fitting alignment-based models was the memory required to query large

sequence databases. Some models encountered memory limitations that prevented training on the largest alignments (Table S4), but these would mostly be excluded by the choice of Neff cutoff. This highlights the computational burden of alignment-based approaches, particularly when using comprehensive databases like BFD and MGnify, with 2.1 billion and 2.4 billion sequences, respectively.

The alignments were post-processed to exclude positions with more than 50% gaps and to exclude sequence fragments that align to less than 50% of the length of the target sequence. We slightly relax the constraint on minimum column coverage for sequences in the training MSAs (50% instead of the 70% in Frazer et al. (2021) [19]) as it led to superior fitness prediction performance in our hyperparameter tuning analyses for the different viruses modeled in our prior work[2]. Similarly, following prior work, we also cluster sequences at 99% sequence identity (referred to as theta) instead of 80% sequence identity[2,19,102] due to limited viral sequence diversity and an expectation that viral sequence are more likely to contain constraint information within smaller differences than are proteins from higher-order organisms. For the best performing model EVE, where clustering is used strictly for sampling, there is no difference in performance however between clustering differences (Fig. S12B). See Reliability Estimation below for more details on sequence clustering and alignment selection.

A.4.2 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 Hopf et al. (2017) [18].

A.4.3 Potts Model

To predict the effects of mutations that explicitly captures pairwise residue dependencies between positions, we used EVmutation as implemented in Hopf et al. (2017) [18]. In a Potts model, protein families are represented as probability distributions governed by two constraint types: individual position preferences for specific amino acids/nucleotides and pairwise coupling constraints between positions. This approach assumes that complex correlations in sequence alignments arise from simpler underlying positional couplings, which traditional metrics like mutual information cannot properly disentangle due to transitive correlations. The resulting distribution represents the maximum entropy solution—the least-structured distribution that still matches the observed single-site and pairwise frequencies in the alignment.

A.4.4 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) [19]. 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. (2023) [2]. Note, that we use the negative of the evolutionary index reported by the model.

A.4.5 Imputing missing data

For accurate comparisons across the same sets of mutations in DMSs, we impute missing model values (for positions with insufficient alignment coverage, using a lenient coverage threshold) for alignment-based models with the mean value across the target protein. All mutation scores are calculated by PLMs.

A.5 Protein language models

A.5.1 Tranception

Tranception[22] 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[17]. Tranception averages the log ratios from both left-to-right and right-to-left scoring.

A.5.2 ESM-1v

ESM-1v[21] 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[17] to handle sequences that are longer than the model context window (i.e., 1023 amino acids), and ensemble across 5 models.

A.5.3 ProGen2

ProGen2[24] is an autoregressive transformer with left-to-right causal masking. We primarily use ProGen2-XL trained (6.4B parameters) trained on both Uniref90 and BFD30.

A.5.4 SaProt

SaProt[41] introduces a structure-aware vocabulary, into protein language modeling by training on Foldseek[124] 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[25] 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; (2)SaProt-650M-PDB, which continuously pre-trains the SaProt-650M-AF2 model on the PDB; and (3) SaProt-1.3B-AFDB-OMG-NCBI with 40 million AF2 structures, and 200 million OMG (50% identity filtering) and 150 million NCBI (70% identity filtering) sequences trained with multimodal input integration.

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 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[22], 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[21] and SaProt[41], 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[19], 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[19]. Note that this is the negative of the evolutionary index score outputted by the EVE model.

A.7 Reliability Estimation

A.7.1 Alignment-based Reliability: Neff at 90% Identity

To account for uneven sampling of protein families, by both sequencing efforts and natural processes, alignment-based models reweight sequences in a multiple sequence alignment by a measure of their uniqueness. Each sequence in an alignment is assigned a weight $\pi_x = 1/S$, where S is the number of sequences in that alignment within a given Hamming distance. For non-viruses this cutoff is typically set to 0.2 (80% sequence identity)[17]. Following prior work[2,18], we use a cutoff of 0.01 (99% sequence identity) that works well for viral proteins, due to the relatively limited sequence diversity and an expectation that small difference in viral sequences will have comparatively large impacts on overall fitness constraints.

The effective number of non-redundant homologous sequences (Neff) is calculated using the sum of sequence weights as a measure of the sequence information present in the training alignment[18].

$$\text{Neff} = \sum_{i=1}^{N} \pi_{x_i}$$

where N is the number of sequences in the alignment. Neff has also been expressed as the exponential of the Shannon entropy averaged over all columns of the multiple sequence alignment.

Following Hopf et al. (2017) [18], we divide Neff by the length of the target sequence (L). This ratio provides a normalized measure of sequence diversity relative to the protein length, which helps assess whether there is sufficient evolutionary information to reliably model the constraints on each position in the sequence. Alignment depth can be split into Low (Neff/L $<$1), Medium (1 $<$Neff/L $<$100), and High (Neff/L $>$100) as in ProteinGym[17].

As our work demonstrates the greater importance of sequences within 90% identity of the target sequence, we define an alignment-based confidence metric as the effective number of sequences within 90% identity of the target:

$$\text{Neff @90\%ID} = \sum_{i=1}^{N} \pi_{x_i} \cdot \mathbb{1}[Dist(x_i, x_{\text{target}}) \geq 0.9]$$

Alignment selection criteria We distinguish between the metric used for confidence when comparing across viral proteins and the metric used for deciding on the optimal alignment within one viral protein (that have been made using different bitscores, databases, or other parameters). To understand the modeling choices to optimize alignment-based model performance for viruses, for each viral protein in our DMS benchmark, we generated MSAs using three reference databases: UniRef90, UniRef100, and UniRef100+BFD+MGnify, while systematically varying alignment depth at six sequence similarity thresholds during homology search (Fig. 3A). When deciding the optimal alignment for one viral protein, we demonstrate that the proportion of sequences in the alignment that are highly relevant to the target sequence (within 90% identity) is critical. We therefore choose the best alignment that has (1) Neff/L $>$1 and (2) the highest proportion of sequences within 90% identity. While this heuristic works for alignments created using a standard alignment-generating software like JackHMMER[123], taking this approach to the limit–maximizing the proportion of highly related sequences by selecting only sequences above the 90% identity threshold–is unlikely to be optimal. We choose 90% as the identity cutoff as the percentage in which the greatest fraction of DMSs across taxa (70%) showed positive correlation between alignment relatedness at that percentage and model performance

(Fig. S11H). We also note that while alignment relevance is a better predictor of performance than fraction of viral sequences in the alignment, increasing relevance does by necessity also increase the fraction viral (Fig. S11F).

A.7.2 PLM Reliability: Scaled Pseudo-Perplexity

In an encoder-only PLM like ESM or SaProt, an estimation of perplexity of a sequence x of length L can be obtained by calculating the exponential of the negative pseudo-log-likelihood, or pseudo-perplexity:

$$\text{PPPL}(x) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L} \ln p(x_i|x_{\{-i\}})\right),$$

where $x_{\{-i\}}$ is the set of all residues except x_i [106].

For SaProt, we calculate pseudo-perplexity only of the amino acid tokens, not structural tokens, and with the structural token unmasked when a particular amino acid token x_i is masked out and predicted by the model based on all other tokens $x_{\{-i\}}$. This matches the masking strategy used during training. Moreover, the structural token is left visible to reduce the model's emphasis on predicting it, as despite inaccuracies in foldseek tokens there is still valuable context provided[41].

For the autoregressive PLM Tranception, we directly calculate the perplexity of a sequence x using the exponentiated average log-likelihood conditioned on the preceding amino acid tokens when scored from both the left and right directions:

$$\text{PPL}(x) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L} \frac{\ln p(x_i|x_{<i}) + \ln p(x_i|x_{>i})}{2}\right)$$

.

We create a PLM reliability metric calculated based on the (pseudo-)perplexity of the wildtype protein as $\frac{20-\text{PPPL}(wildtype)}{19}$, scaling the result to a value between 0 and 1, where higher scores indicate greater confidence. This scaling reflects perplexity's interpretation as the effective number of amino acids the model is choosing between, where a perfect (pseudo-)perplexity of 1 represents complete certainty and a model completely uncertain between 20 amino acids would approach a pseudo-perplexity of 20. While we note a pseudo-perplexity of 1 represents complete certainty, in truth more than 1 amino acid should be functional at a given position and therefore there should be some uncertainty unless the model has overfit to the wildtype sequence. So, the confidence metric should rarely reach a score of 1, and a score of 0.9 would be expected and in fact desirable for high performing models. Note that negative values are technically possible with this metric, but for our purposes can be capped at 0.

We also experiment with using a "pseudo-perplexity" of the EVE Evidence Lower Bound (ELBO), a lower bound to the log-marginal likelihood of each amino acid at each position, as an alternate reliability score, though this is difficult to interpret given the overall high confidence and performance of the model.

A.7.3 Reliability Thresholds

We choose an optimal reliability threshold that maximizes the sensitivity and specificity (Youden's J) for the reliability measure predicting high DMS correlation (defined as Spearman $\rho > 0.4$ between model and DMS). To calibrate these reliability estimates, we use a subset of 23 "comparable" viral DMSs. We expect correlations with DMS to be more comparable when we focus on fitness assays (measuring cell entry, replication, or infectivity), proteins with length less than 1024 (as the context size of PLMs), and studies performed after 2015 (many earlier DMSs have low replicate correlations). Given the relatively low number of DMSs to calibrate against, often a range of reliability scores will have the same sensitivity and specificity, and we choose to round down within this range. While we use Spearman $\rho > 0.4$ as a means to choose reliability thresholds here for demonstration, we expect that others will use the EVEREST benchmark as a means to define their own thresholds based on desired performance levels.

A.8 Hybrid Models

A.8.1 VESPA

VESPA [79] combines the embeddings from ProtT5 [99] with a per-residue conservation prediction and supplements with BLOSUM substitution scores [49]. ProtT5 uses a T5 architecture which uses an encoder and decoder and was first trained on BFD and then finetuned on UniRef50.

A.8.2 TRANCEPTION WITH MSA RETRIEVAL

Tranception[22] 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[17]. Sequences are re-weighted as in Hopf et al. (2017) [18]. 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.8.3 TRANCEPTEVE

TranceptEVE[38] is a hybrid method that combines Tranception[22] with an MSA-based EVE log prior[19]. 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.8.4 SAPROT-EVE

For comparison, we make a hybrid of SaProt and EVE that combines the advantages of alignment-based, family-specific models with structure-aware protein language models. We impute the EVE scores with the mean value, standard scale each of the models, shift the scores positive, and then take the geometric mean. We select the geometric mean instead of an arithmetic mean as it is less sensitive to outliers.

A.8.5 EVOTUNING SAPROT

"Evotuning" is an alternative to training protein language models on new datasets, where pre-trained models are fine-tuned on a set of homologous, evolutionary sequences and generally expected to improve predictive capabilities for the targeted protein groups[103]. We evaluated several fine-tuning strategies for the SaProt backbone on a Lassa dataset of 3102 sequences, constructed from both Foldseek and JackHMMER (optimal alignment-based model selection) searches of the Lassa virus glycoprotein complex (GPC). In all cases, the "backbone" denotes the pretrained SaProt transformer encoder and the "task head" the downstream regression head used for DMS score prediction. We considered three fine-tuning modes:
- Mode Full: full model trainable.
- Mode Unfreeze-8: backbone frozen except the last 8 layers.
- Mode Unfreeze-16: backbone frozen except the last 16 layers.

We report Spearman correlation between model predictions and experimental DMS scores. For each run we give the best single-epoch value, and where noted we also report the mean over all epochs in that run.

As an initial hyperparameter search, we performed a learning-rate sweep for Mode Unfreeze-8 with learning rate ($lr = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$) for 60 epochs. For each setting, we evaluated on the Lassa DMS the following models:
- Mode Unfreeze-8, $lr = 10^{-3}$: average 0.19 (best 0.31)
- Mode Unfreeze-8, $lr = 10^{-4}$: average 0.19 (best 0.31)
- Mode Unfreeze-8, $lr = 10^{-5}$: average 0.30 (best 0.34)
- Mode Unfreeze-8, $lr = 10^{-6}$: average 0.32 (best 0.34)

We therefore fixed $lr = 10^{-6}$ for all subsequent SaProt experiments on Lassa-GPC. Under this setting (60 epochs), we evaluated on the Lassa DMS the following models:
- Mode Full: average 0.33 (best 0.34)
- Mode Unfreeze-16: average 0.32 (best 0.34)

To test whether span-level reconstruction improves SaProt's representations by creating a more difficult training objective[125], we augmented the standard token-level masked language modeling (MLM) objective with a SpanBERT-style span prediction loss. In addition to randomly masking individual residues, we (i) sample contiguous random spans of residues and mask their internal tokens, and (ii) apply a span-boundary objective that trains the span-boundary representations to reconstruct the entire masked span. We refer to this configuration as SaProt + SpanBERT-style loss. For these runs, we fine-tuned SaProt on Lassa-GPC using Mode Full with $lr = 10^{-6}$ for 60 epochs, achieving an average Spearman of 0.29 (best 0.32).

Dataset We created a Lassa GPC dataset with both amino acid and foldseek tokens based on combining the JackHMMer selected alignment (amino acids only) used throughout this paper with a Foldseek search (amino acids and structural tokens). For the JackHMMER alignment, our goal was to assign each raw JackHMMER hit a corresponding structural sequence suitable as input to SaProt. We first assembled a pool of usable structural sequences. For each JackHMMER accession, we checked whether it matched a representative accession in the BFVD or an exact UniProt to PDB mapping, and when such a match was available we extracted the associated 3Di sequence. In parallel, we retrieved the corresponding raw amino acid sequences from the UniProt API. We then expanded this pool by running Foldseek against PDB and AF2 (UniRef50) structure databases, which yielded additional pairs of raw sequences and 3Di sequences. For all non-PDB structures, we also downloaded per residue pLDDT scores and masked 3Di tokens at positions with confidence below 0.7.

Next, we used MMseqs2 to map the original set of raw JackHMMER hits, treated as queries, to this structural sequence database. For each query, we selected the most stringent query coverage threshold in the set 70%, 60%, 50%, 40%, 30%, 20%, 10% that retained at least one match, and within the surviving hits we chose the sequence with the highest sequence identity. If a PDB hit was present whose sequence identity was at least 80% of that top hit, we instead selected the PDB sequence. Otherwise, we retained the global top identity hit. For each JackHMMER hit, we then took the chosen structural sequence and performed a global pairwise alignment to the original raw query sequence. Along this alignment, we grafted the 3Di tokens onto the query by copying tokens at aligned residue–residue positions, skipping positions aligned to gaps in the query, and inserting "#" mask tokens at query insertions or at positions that had previously been masked as low confidence because of pLDDT.

A.9 ProteinGym

ProteinGym is a collection of benchmarks for comparing the ability of models to predict the effects of protein mutations[17]. We used this prior benchmark to compare to non-viral DMSs and to extrapolate the results of the models tested here to the > 50 alignment-based, protein language model, hybrid, and inverse folding models evaluated previously (but on a more limited set of viral assays). We use ProteinGym v1.1 containing 218 deep mutational scans[17]. Multiple sequence alignments used in ProteinGym were created for each assay by performing five search iterations of the profile HMM homology search tool JackHMMER[123] on the UniRef100 database of non-redundant proteins[101]. Alignments were created using 9 bit score thresholds, from 0.1 to 0.9, and selected using structural ground truth by the highest number of significant evolutionary couplings (ECs) from EVcouplings[126], detailed in Notin et al. (2022) [22].

B SUPPLEMENTARY TABLES

ID	Virus	Protein	Paper
Fitness			
LASSA_GP_Carr	Lassa	GP	Carr et al. (2024) [78]
SARS2_MRPO_Flynn	SARS-CoV-2	MRPO	Flynn et al. (2022) [75]
SARS2_BA1_SPIKE_Dadonaite	SARS-CoV-2 BA.1	SPIKE	Dadonaite et al. (2023) [50]
SARS2_DELTA_SPIKE_Dadonaite	SARS-CoV-2 Delta	SPIKE	Dadonaite et al. (2023) [50]
DENV_POLG_Suphatrakul	Dengue virus	POLG	Suphatrakul et al. (2023) [57]
IAV_H1_HA_Doud	Influenza A virus H1N1	H1	Doud and Bloom (2016) [63]
IAV_H1_HA_Wu	Influenza A virus H1N1	H1	Wu et al. (2014) [64]
IAV_PB2_Soh	Influenza A virus H1N1	PB2	Soh et al. (2019) [70]
IAV_H3_HA_Lee	Influenza A virus H3N2	H3	Lee et al. (2018) [65]
IAV_H5_HA_Dadonaite	Influenza A virus H5N1	H5	Dadonaite et al. (2024) [66]
IAV_H3_NP_Doud	Influenza A virus H1N1	NP	Doud et al. (2015) [67]
IAV_H1_NP_Doud	Influenza A virus H1N1	NP	Doud et al. (2015) [67]
IAV_PA_Wu	Influenza A virus H1N1	PA	Wu et al. (2015) [69]
IAV_RDRP_Li	Influenza A virus H1N1	RDRP	Li et al. (2023) [71]
NIPAH_F_Larsen	Nipah Henipavirus	F	Larsen et al. (2025) [77]
AAV2_CAPSD_Sinai	AAV2	CAPSD	Sinai et al. (2021) [54]
CVB3_2A_Alvarez	Coxsackievirus B3	2A	Álvarez-Rodríguez et al. (2024) [56]
CVB3_2B_Alvarez	Coxsackievirus B3	2B	Álvarez-Rodríguez et al. (2024) [56]
CVB3_2C_Alvarez	Coxsackievirus B3	2C	Álvarez-Rodríguez et al. (2024) [56]
CVB3_3A_Alvarez	Coxsackievirus B3	3A	Álvarez-Rodríguez et al. (2024) [56]
CVB3_3B_Alvarez	Coxsackievirus B3	3B	Álvarez-Rodríguez et al. (2024) [56]
CVB3_3C_Alvarez	Coxsackievirus B3	3C	Álvarez-Rodríguez et al. (2024) [56]
CVB3_3D_Alvarez	Coxsackievirus B3	3D	Álvarez-Rodríguez et al. (2024) [56]
CVB3_VP1_Alvarez	Coxsackievirus B3	VP1	Álvarez-Rodríguez et al. (2024) [56]
CVB3_VP3_Alvarez	Coxsackievirus B3	VP3	Álvarez-Rodríguez et al. (2024) [56]
EV_CAPSD_Bakhache	Enterovirus A	CAPSD	Bakhache et al. (2024) [59]
EV_REP_Bakhache	Enterovirus A	REP	Bakhache et al. (2024) [59]
CVB3_POLG_Mattenberger	Coxsackievirus B3	POLG	Mattenberger et al. (2021) [55]
HIV1_BF520_ENV_Haddox	HIV-1 BF520	ENV	Haddox et al. (2018) [60]
HIV1_HV1B9_ENV_DuenasDecamp	HIV-1 HV1B9	ENV	Duenas-Decamp et al. (2016) [61]
HIV1_BG505_ENV_Haddox	HIV-1 BG505	ENV	Haddox et al. (2018) [60]
Stability			
PESV_POLG_Tsuboyama	Porcine enteric sapovirus	POLG	Tsuboyama et al. (2023) [58]
BPP22_COAT_Tsuboyama	Salmonella phage P22	COAT	Tsuboyama et al. (2023) [58]
BPT7_RPOL_Tsuboyama	Enterobacteria phage T7	RPOL	Tsuboyama et al. (2023) [58]
LAMBDA_HCP_Tsuboyama	Enterobacteria phage lambda	HCP	Tsuboyama et al. (2023) [58]
BP434_RPC1_Tsuboyama	Enterobacteria phage 434	RPC1	Tsuboyama et al. (2023) [58]
Expression			
SARS2_PRD0038_RBD_Starr	Bat coronavirus PRD0038	RBD	Starr (2024) [73]
RmYN02_RBD_Starr	Bat coronavirus RmYN02	RBD	Starr (2024) [73]
RsYN04_RBD_Starr	Bat coronavirus RsYN04	RBD	Starr (2024) [73]
SARS2_XBB15_RBD_Taylor	SARS-CoV-2 XBB.1.5	RBD	Taylor and Starr (2023) [74]
SARS2_RBD_Starr_expression	SARS-CoV-2	RBD	Starr et al. (2020) [4]

Table S1 continued on next page

Table S1 continued from previous page

ID	Virus	Protein	Paper
Binding			
SARS2_RBD_Starr_binding	SARS-CoV-2	RBD	Starr et al. (2020) [4]
IAV_NA_Jiang	Influenza A virus H3N2	NA	Jiang et al. (2016) [68]
Activity			
SARS2_PLPRO_Wu_activity	SARS-CoV-2	PLPRO	Wu et al. (2024) [76]
Abundance			
SARS2_PLPRO_Wu_abundance	SARS-CoV-2	PLPRO	Wu et al. (2024) [76]

Supplementary Table 1: Deep Mutational Scanning studies of 45 viral proteins across 25 papers, with fitness, stability, expression, and other assays.

Model Type	Model	Avg Spearman	Viral Avg Spearman
Alignment-based model	EVE (ensemble)	0.439	0.428
	EVE (single)	0.433	0.424
	DeepSequence (ensemble)	0.419	0.344
	DeepSequence (single)	0.407	0.323
	Potts model	0.395	0.388
	Site-Independent	0.359	0.383
Hybrid - Alignment & PLM	TranceptEVE L	0.456	0.453
	TranceptEVE M	0.455	0.441
	TranceptEVE S	0.452	0.433
	Tranception L	0.434	0.432
	MSA Transformer (ensemble)	0.434	0.414
	Tranception M	0.427	0.415
	MSA Transformer (single)	0.421	0.390
	Tranception S	0.418	0.405
	Tranception M no retrieval	0.348	0.349
	Unirep evotuned	0.347	0.349
Hybrid - Structure & PLM	SaProt-AF2 (650M)	0.457	0.300
	ProtSSN (ensemble)	0.449	0.356
	ProtSSN (k=20 h=1280)	0.442	0.347
	ProtSSN (k=20 h=512)	0.441	0.359
Hybrid - other	VESPA	0.436	0.432
	VESPAl	0.394	0.392
Protein language model	ProGen2 XL	0.391	0.391
	ProGen2 L	0.380	0.333
	ProGen2 M	0.379	0.342
	ProGen2 Base	0.378	0.328
	Tranception L no retrieval	0.374	0.395
	ESM-1v (ensemble)	0.407	0.279
	ESM-1v (single)	0.374	0.258
	CARP (640M)	0.368	0.273
	RITA XL	0.372	0.402
	RITA L	0.365	0.391
	RITA M	0.350	0.385
	ProGen2 S	0.336	0.285
	CARP (76M)	0.328	0.150
	ESM-2 (3B)	0.406	0.274
	ESM-2 (15B)	0.401	0.313
	ESM-2 (150M)	0.387	0.137
	ESM-2 (650M)	0.414	0.238
	ESM-2 (35M)	0.321	0.102
Inverse folding model	ESM-IF1	0.422	0.374
	MIF	0.383	0.359
	ProteinMPNN	0.258	0.248
Tree-based model	GEMME	0.455	0.469

Supplementary Table 2: Model performance previously available in ProteinGym[17]. Note this analysis only covers half of the now curated 45 viral datasets and does not include SaProt-PDB, but can be used to contextualize our findings in terms of 44 models.

Model	Parameters	Data Size	Objective	Cutoff	Context	All Taxa (ρ)	Virus (ρ)	Reference
Uniref100								
RITA XL	1.2B	280M	AR	2022	1024	0.372	0.402	Hesslow et al. (2022) [23]
Tranception L	700M	250M	AR	2022	1024	0.374	0.395	Notin et al. (2022) [22]
RITA L	680M	280M	AR	2022	1024	0.365	0.391	Hesslow et al. (2022) [23]
RITA M	300M	280M	AR	2022	1024	0.350	0.385	Hesslow et al. (2022) [23]
RITA S	85M	280M	AR	2022	1024	0.304	0.358	Hesslow et al. (2022) [23]
Tranception M	300M	250M	AR	2022	1024	0.348	0.349	Notin et al. (2022) [22]
Tranception S	85M	250M	AR	2022	1024	0.303	0.315	Notin et al. (2022) [22]
Uniref90+BFD30								
ProGen2 XL	6.4B	230M	AR	2022	1024	0.391	0.391	Nijkamp et al. (2023) [24]
ProGen2 L	2.7B	230M	AR	2022	1024	0.380	0.333	Nijkamp et al. (2023) [24]
ProGen2 M	764M	230M	AR	2022	1024	0.379	0.342	Nijkamp et al. (2023) [24]
ProGen2 Base	764M	230M	AR	2022	2048	0.378	0.328	Nijkamp et al. (2023) [24]
ProGen2 S	151M	230M	AR	2022	1024	0.336	0.285	Nijkamp et al. (2023) [24]
Uniref90								
ESM-2 (650M)	650M	65M	MLM	2022	1024	0.414	0.238	Lin et al. (2023) [25]
ESM-1v (ens.)	650M	98M	MLM	2020	1024	0.407	0.279	Meier et al. (2021) [21]
ESM-2 (3B)	3B	65M	MLM	2022	1024	0.406	0.274	Lin et al. (2023) [25]
ESM-2 (15B)	15B	65M	MLM	2022	1024	0.401	0.313	Lin et al. (2023) [25]
ESM-2 (150M)	150M	65M	MLM	2022	1024	0.387	0.137	Lin et al. (2023) [25]
ESM-1v	650M	98M	MLM	2020	1024	0.374	0.245	Meier et al. (2021) [21]
ESM-2 (35M)	35M	65M	MLM	2022	1024	0.321	0.102	Lin et al. (2023) [25]
ESM-2 (8M)	8M	65M	MLM	2022	1024	0.226	0.078	Lin et al. (2023) [25]
Uniref50								
ESM-1b	650M	27M	MLM	2018	1024	0.394	0.241	Rives et al. (2021) [127]
CARP (640M)	640M	40M	MLM	2023	4096	0.368	0.273	Yang et al. (2024) [26]
CARP (76M)	76M	40M	MLM	2023	4096	0.328	0.150	Yang et al. (2024) [26]
CARP (38M)	38M	40M	MLM	2023	4096	0.279	0.125	Yang et al. (2024) [26]
ProtGPT2	738M	40M	AR	2022	1024	0.188	0.141	Ferruz et al. (2022) [128]
Unirep	18.2M	24M	AR	2018	280	0.190	0.057	Alley et al. (2019) [129]
CARP (600K)	600K	40M	MLM	2023	4096	0.106	0.042	Yang et al. (2024) [26]

Supplementary Table 3: Single-sequence Protein Language Model performance of models in ProteinGym[17] with model details collected from model references. No hybrid models or retrieval included. Training objectives are autoregressive (AR) and masked language model (MLM); all architectures are transformer-based besides CARP which is CNN-based and Unirep which is an RNN. Performance recorded as average Spearman across ProteinGym DMSs–this analysis only covers half of the now curated 45 viral datasets, but can be used to contextualize our findings. Data size in approximate number of sequences (note different databases have different sequence length distributions[99]). Date cutoffs and dataset splits influence training data size even when same database is used. Some model details were estimated from published papers when not explicitly stated by the authors. Ens.: Ensemble.

Bitscore	UniRef90	UniRef100	UniRef100+MGnify+BFD
0.5	57	57	48
0.4	57	56	37
0.3	55	51	28
0.05	54	51	21
0.04	53	45	12
0.03	44	37	10

Supplementary Table 4: Number of successful Potts models given the equal memory constraints at different bitscores across UniRef90, UniRef100, and UniRef100+MGnify+BFD datasets. Failed models with high numbers of sequences would mostly be excluded by selected Neff cutoff, as only a subset of bit scores/databases failed per viral DMS.

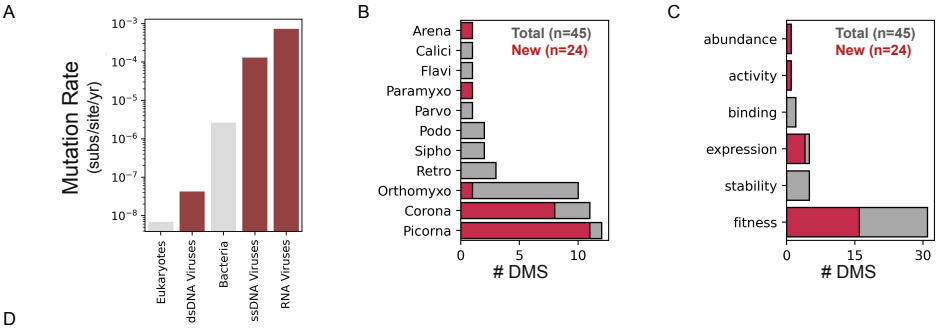
	Model	All Viral DMS	Fitness Only	Fitness+Antigens Only
Alignment	PSSM	0.38	0.41	0.37
	Potts model	0.38	0.39	0.36
	EVE	**0.44**	**0.46**	**0.43**
PLM	ESM-1v - Uniref90	0.19	0.19	0.31
	Progen - Uniref90 + BFD	0.36	0.35	0.33
	Tranception - Uniref100	0.37	**0.38**	0.35
	SaProt - PDB + AF2	**0.44**	0.35	**0.39**
Hybrid	VESPA	0.41	0.40	0.38
	Tranception with retrieval	0.43	0.44	0.42
	TranceptEVE	0.47	0.48	0.45
	SaProt-EVE	**0.52**	**0.50**	**0.48**
Tree-based	**GEMME**	**0.49**	**0.50**	**0.46**

Supplementary Table 5: Model Performance (average Spearman ρ) on EVEREST DMS Benchmark with 45 viral DMSs, or on the subset of scans with an encompassing fitness assay measuring infectivity, replication, or cell entry, or the subset of fitness assays that are antigens.

ID	Family	Structural Class	MSA confidence	PLM confidence
Junin_GPC	Arena	Class I Fusion	5	0.61
Lassa_GPC	Arena	Class I Fusion	259	0.55
Lujo_GPC	Arena	Class I Fusion	1	0.60
MERS_Spike	Corona	Class I Fusion	8	0.67
SARS2_Spike	Corona	Class I Fusion	157	0.67
Ebola_GP	Filo	Class I Fusion	85	0.70
Marburg_GP	Filo	Class I Fusion	15	0.57
SEBOV_GP	Filo	Class I Fusion	3	0.68
DENV_E	Flavi	Class II Fusion	107	0.90
TBE_E	Flavi	Class II Fusion	98	0.81
WNV_E	Flavi	Class II Fusion	11	0.82
YFV_E	Flavi	Class II Fusion	35	0.83
Zika_E	Flavi	Class II Fusion	50	0.90
HTNV_Gc	Hanta	Class II Fusion	13	0.69
HTNV_Gn	Hanta	Class II AP	12	0.53
SNV_Gc	Hanta	Class II Fusion	13	0.71
HEV3_orf2	Hepe	Capsid	128	0.67
IAV_H1	Orthomyxo	Class I Fusion	556	0.90
IAV_H10	Orthomyxo	Class I Fusion	164	0.82
IAV_H2	Orthomyxo	Class I Fusion	184	0.91
IAV_H3	Orthomyxo	Class I Fusion	349	0.88
IAV_H5	Orthomyxo	Class I Fusion	638	0.90
IAV_H6	Orthomyxo	Class I Fusion	226	0.88
IAV_H7	Orthomyxo	Class I Fusion	238	0.87
NiV_F	Paramyxo	Class I Fusion	2	0.69
NiV_G	Paramyxo	Non-fusion RBP	9	0.73
OROV_G1	Peribunya	Class II Fusion	8	0.62
OROV_G2	Peribunya	Class II AP	8	0.60
RVF_Gn	Phenui	Class II AP	29	0.57
SFTS_Gc	Phenui	Class II Fusion	9	0.73
SFTS_Gn	Phenui	Class II AP	35	0.58
hPBV_capsid	Picobirna	Capsid	2	0.51
hPBV_capsid	Picobirna	Capsid	2	0.51
EV-D68_VP1	Picorna	Capsid	85	0.79
EV71_VP1	Picorna	Capsid	269	0.85
polio_capsid	Picorna	Capsid	13	0.84
hMPV_F	Pneumo	Class I Fusion	39	0.74
hMPV_G	Pneumo	Non-fusion RBP	2	0.48
HIV1_Env	Retro	Class I Fusion	20	0.87
VSV_GP	Rhabdo	Class III Fusion	17	0.61
CHIKV_E1	Toga	Class II Fusion	32	0.84
CHIKV_E2	Toga	Class II AP	49	0.77
VEE_E1	Toga	Class II Fusion	11	0.78
VEE_E2	Toga	Class II AP	15	0.74

Supplementary Table 6: MSA (Neff @ 90% identity) and PLM (scaled pseudo-perplexity, SaProt) reliability scores for 44 antigens of 40 WHO priority and prototype RNA viruses. RBP: Receptor Binding Protein, AP: Accompanying Protein.

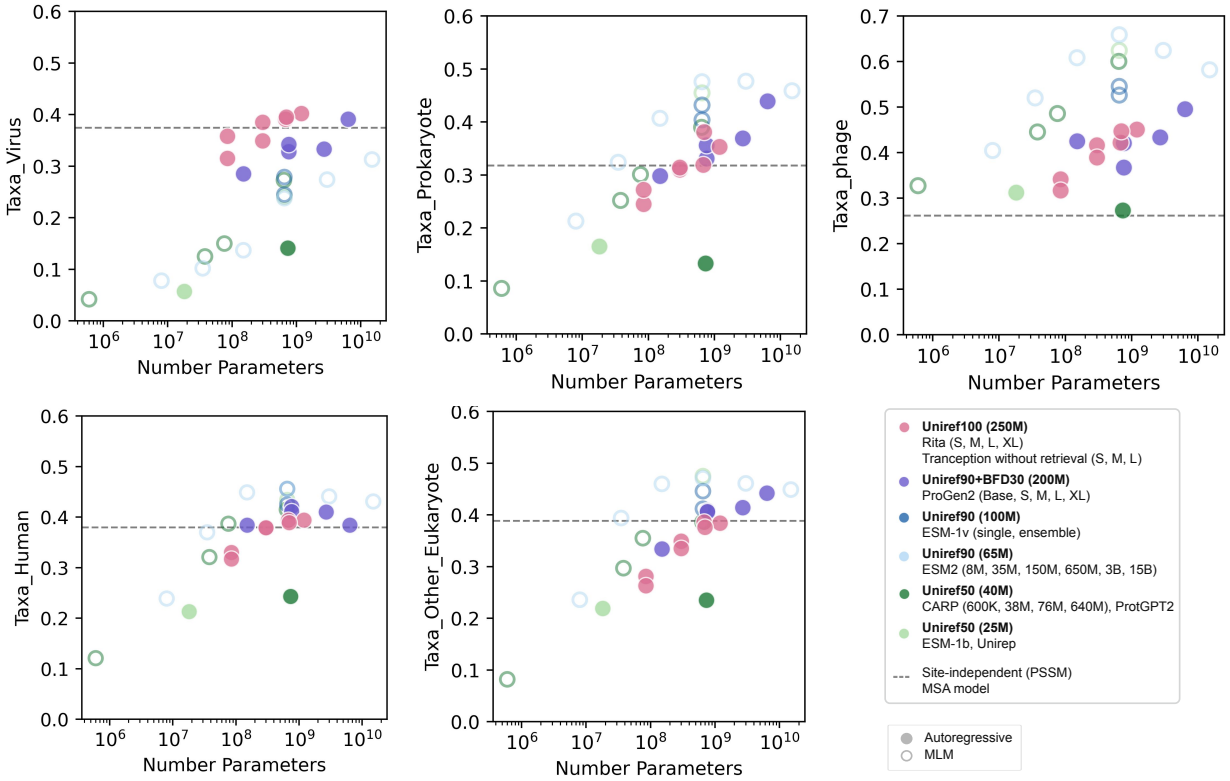
C SUPPLEMENTARY FIGURES



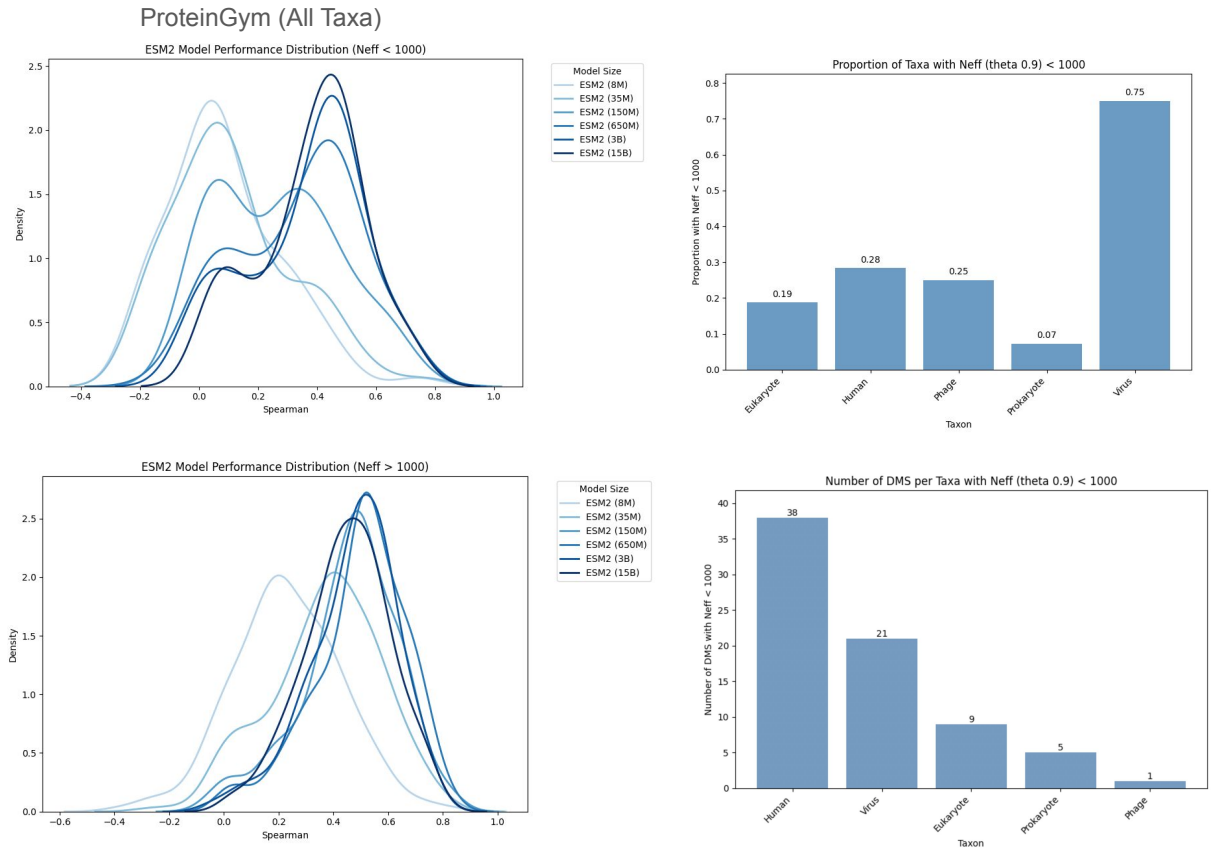
D

Virus	Clade	Emergence date	# training seqs	# test muts	DMS
SARS-CoV-2 (test mutations seen >= 100x)	Wuhan	2020-01-10	4577	181	Dadonite *et al* 2023. *Cell* (variant = Delta)
	BA.1	2021-06-03	6786	135	Dadonite *et al* 2023. *Cell* (variant = BA.1)
	BA.4	2021-11-30	9856	30	Dadonite *et al* 2024. *Nature* (variant = BA.2)
	BA.5	2022-01-07	10956	73	Dadonite *et al* 2024. *Nature* (variant = BA.2)
	XBB	2022-07-01	13881	167	Dadonite *et al* 2024. *Nature* (variant = XBB.1.5)
	JN.1	2023-06-24	16309	29	Dadonite *et al* 2024. *Nature* (variant = XBB.1.5)
H3N2 (test mutations seen >= 3x)	3C	2010-01-01	12589	133	–
	3C.2	2012-01-01	19256	211	–
	3C.3	2012-01-01	19254	244	–
	3C.3a	2013-01-01	22970	115	–
	3C.2a	2014-01-01	27393	484	–
	3C.2a1	2016-01-01	39233	455	–
	3C.2a2	2016-01-01	39264	396	–
	3C.3a1	2017-01-01	46390	298	–
	3C.2a1b.1	2017-01-01	46390	299	Welsh *et al* 2024. *Cell Host & Microbe* (A/Hong Kong/45/2019)
	3C.2a1b.2	2018-01-01	54981	350	Welsh *et al* 2024. *Cell Host & Microbe* (A/Hong Kong/45/2019)
	3C.2a1b.2a.2	2020-01-01	74066	208	Welsh *et al* 2024. *Cell Host & Microbe* (A/Hong Kong/45/2019)
	3C.2a1b.2a.2a.1	2021-01-01	78569	423	Welsh *et al* 2024. *Cell Host & Microbe* (A/Hong Kong/45/2019)
	3C.2a1b.2a.2a.3	2021-01-01	78605	302	Welsh *et al* 2024. *Cell Host & Microbe* (A/Hong Kong/45/2019)
	3C.2a1b.2a.2b	2022-01-01	82352	483	Welsh *et al* 2024. *Cell Host & Microbe* (A/Hong Kong/45/2019)
	3C.2a1b.2a.2a.3a.1	2022-01-01	82350	774	Yu *et al* 2025. BioRxiv (A/Massachusetts/18/2022)
H1N1 (test mutations seen >= 3x)	pmd09	2009-01-01	7785	739	Liu *et al* 2024. *Virus* (variant = pdm09 HA + pdm09 NA)
	6	2011-01-01	15749	178	Liu *et al* 2024. *Virus* (variant = pdm09 HA + pdm09 NA)
	6B	2013-01-01	22934	404	Liu *et al* 2024. *Virus* (variant = pdm09 HA + pdm09 NA)
	6B.1	2015-01-01	32755	462	Liu *et al* 2024. *Virus* (variant = pdm09 HA + pdm09 NA)
	6B.1A	2017-01-01	46192	521	Liu *et al* 2024. *Virus* (variant = pdm09 HA + pdm09 NA)
	6B.1A.5a	2018-01-01	54635	367	Liu *et al* 2024. *Virus* (variant = pdm09 HA + pdm09 NA)
	6B1A.5a.2	2019-01-01	63959	133	Liu *et al* 2024. *Virus* (variant = pdm09 HA + WI19 NA)
	6B.1A.5a.2a	2021-01-01	78409	934	Liu *et al* 2024. *Virus* (variant = pdm09 HA + WI19 NA)
	6B.1A.5a.2a.1	2022-01-01	82251	554	Liu *et al* 2024. *Virus* (variant = pdm09 HA + WI19 NA)
HIV (test mutations seen >= 3x)	Subtype C	–	71694	1542	–

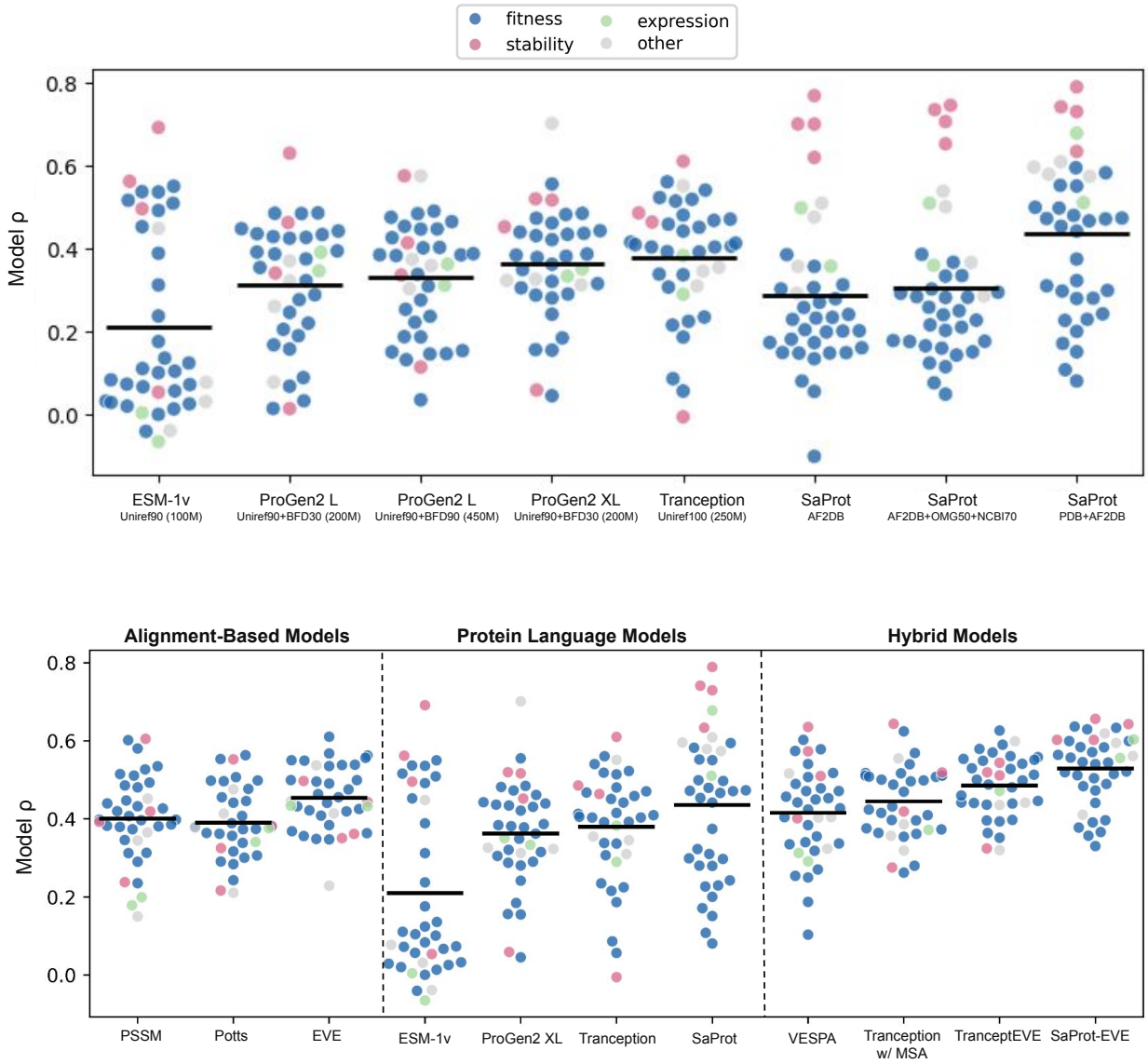
Supplementary Figure 1: Overview of EVEREST virus-specific benchmarking dataset. (A) Viruses, especially RNA viruses, require curated benchmarking because of unique features including have higher mutation rates than other organisms. Geometric average mutation rates (substitutions per site per year) from Linz et al.[47]. (B-C) EVEREST DMS benchmark comprised of 45 experimental scans, approximately doubling the number of viral assays included in prior benchmarks[17]. EVEREST covers 11 viral families (B) and diverse assay phenotypes (C). (D) EVEREST real-work forecasting benchmark, includes 31 clades from 4 virus: SARS-CoV-2, H3N2, H1N1, and HIV, and compares the forecasting ability of fitness DMSs (matched to the starting clade when possible)[5,50–53].



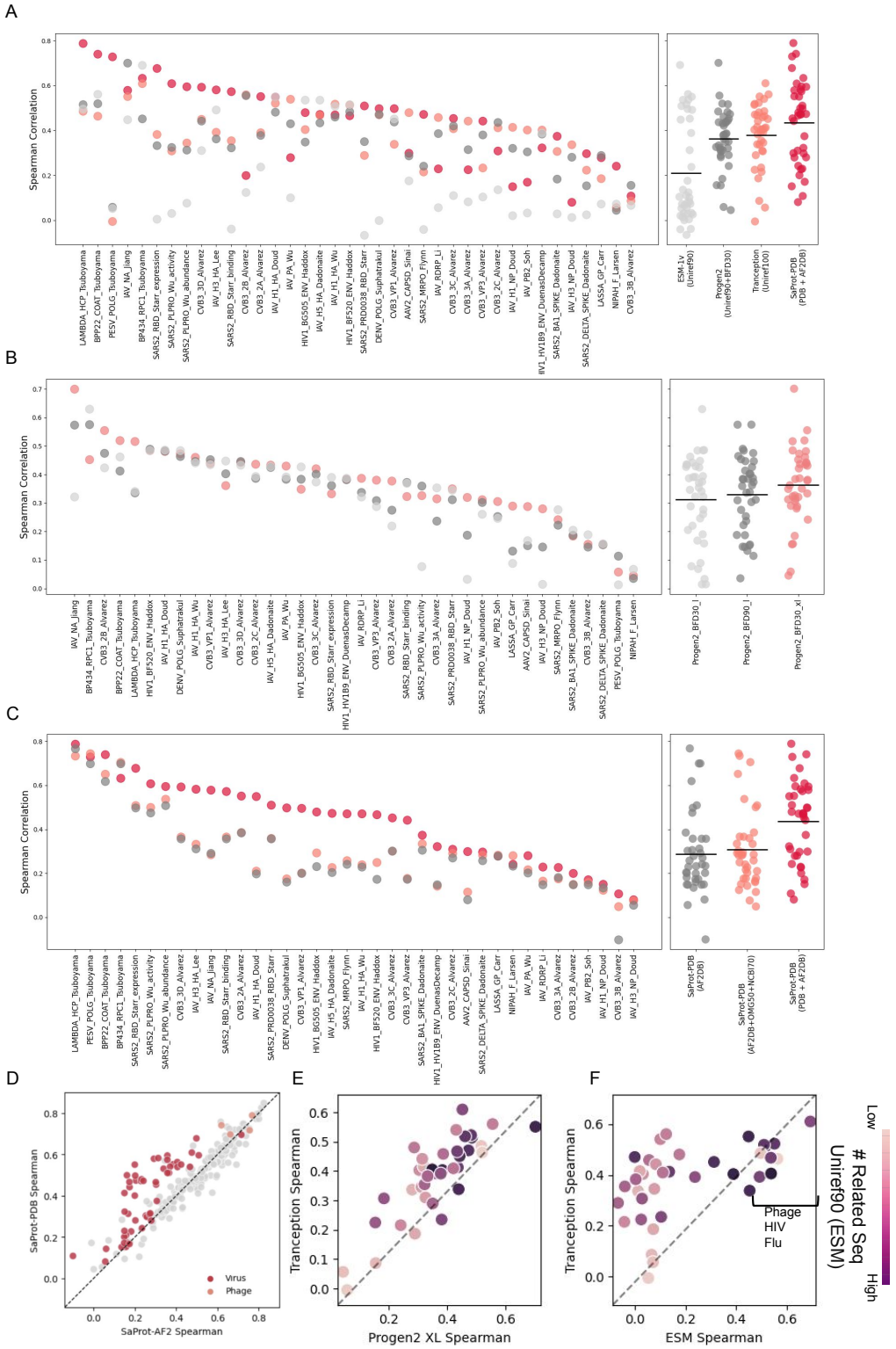
Supplementary Figure 2: Relationship between protein language model size and training dataset and model performance (average Spearman) across different taxa. Scaling number of PLM parameters leads to higher average Spearman correlation with viral proteins (top left). The best PLMs for viruses are autoregressive with larger training datasets, while masked language models generally have better performance on other taxa. Phages (also all stability assays) separated out from eukaryotic-infecting viruses due to differences in performance. Most PLMs outperform site-independent alignment-based model (dotted line) for all taxa except viruses.



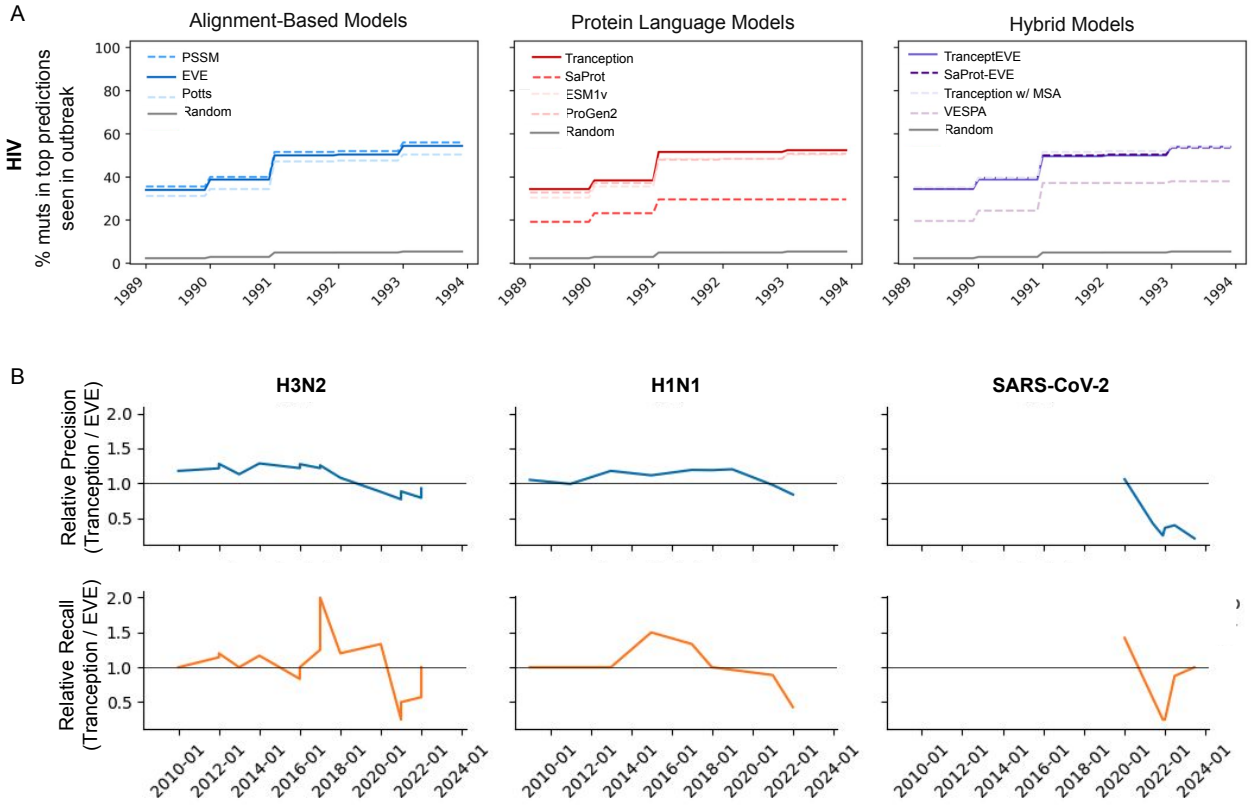
Supplementary Figure 3: The masked protein language model ESM2 benefits from increased number of parameters especially when there is a low number of related sequences in the PLM training data (calculated as number of effective sequences in a JackHMMER alignment of Uniref100 with clustering at theta 90%, since ESM2 is trained on Uniref90) when looking DMSs from all taxa in ProteinGym. This impacts the majority of viruses as 75% of the viruses in ProteinGym have Neff less than 1000, while between 7% and 28% of other taxa have such low effective number of sequences post-clustering (Neff; though the low Neff group has a substantial number of non-viruses, as in total there are more than 5 times the number of non-viral DMS as viral; bottom right). When Neff is greater than 1000, once the number of parameters is at least 150M there is little difference in performance as model size increases.



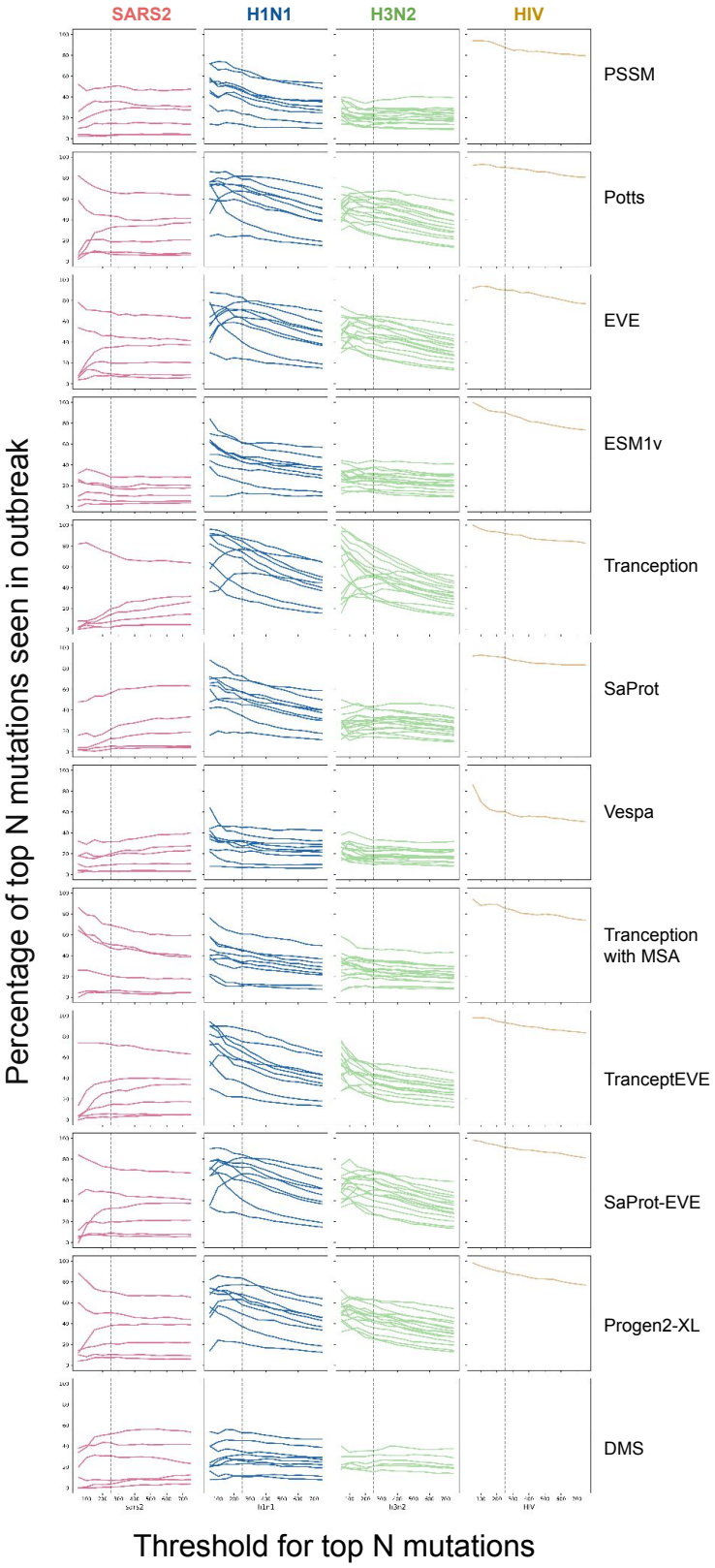
Supplementary Figure 4: Performance of alignment-based, protein language models, and hybrids on EVEREST DMS benchmark.



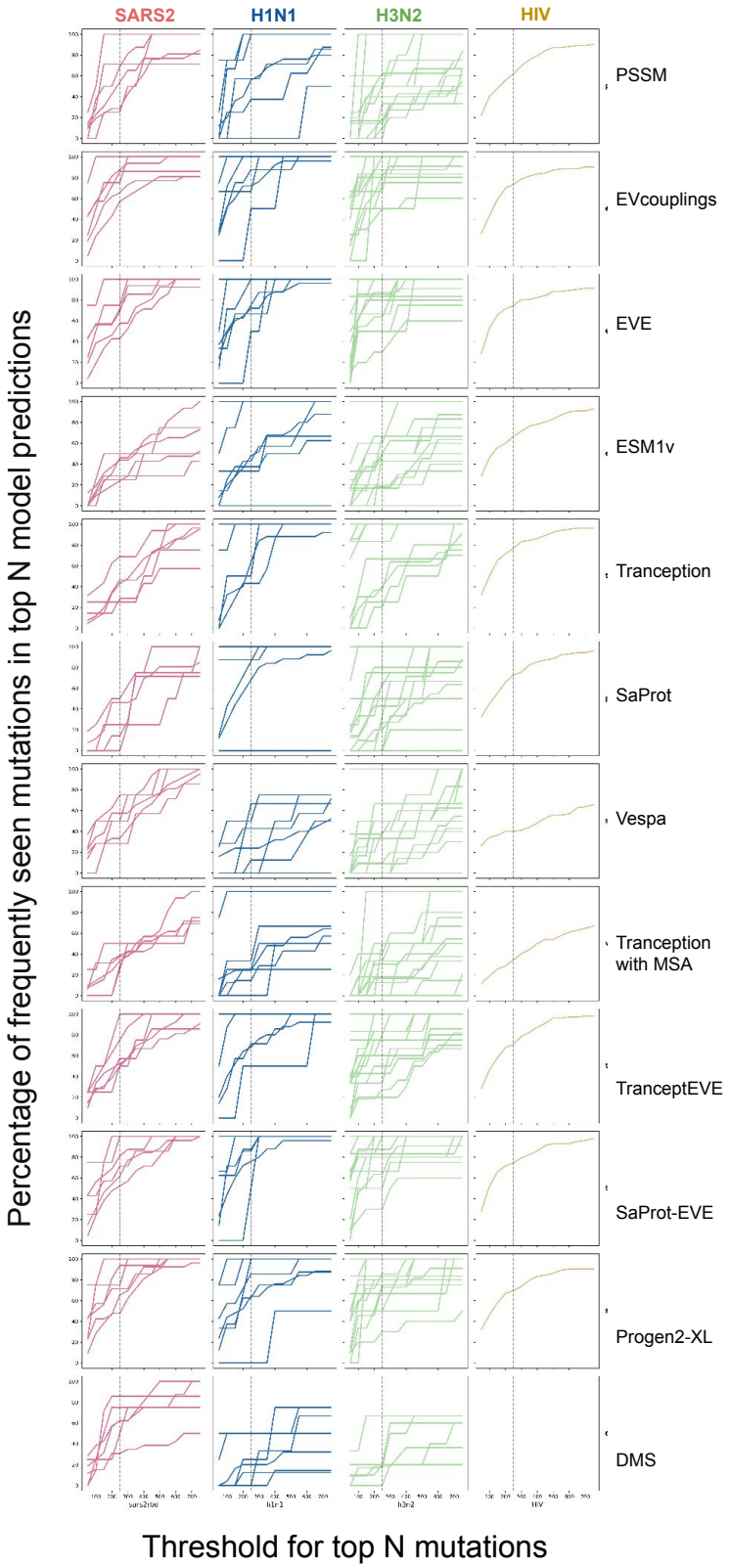
Supplementary Figure 5: Breakdown of protein language model performance on each protein in EVEREST DMS benchmark. (A) Performance of models shown in main: ESM-1v, ProGen2, Tranception and SaProt. (B) Performance of different ProGen2 models with different model size and training datasets. (C) Performance of different SaProt models with different training datasets. (D) Comparison of the Spearman rank correlation of SaProt-PDB to SaProt-AF2 for all DMSs from ProteinGym. SaProt-PDB outperforms SaProt-AF2 for most eukaryotic viral proteins, but not for phages. (E-F) Differences in performance between Tranception and ProGen2 (E) or ESM-1v (F) colors by number of related sequences in Uniref90 training dataset.



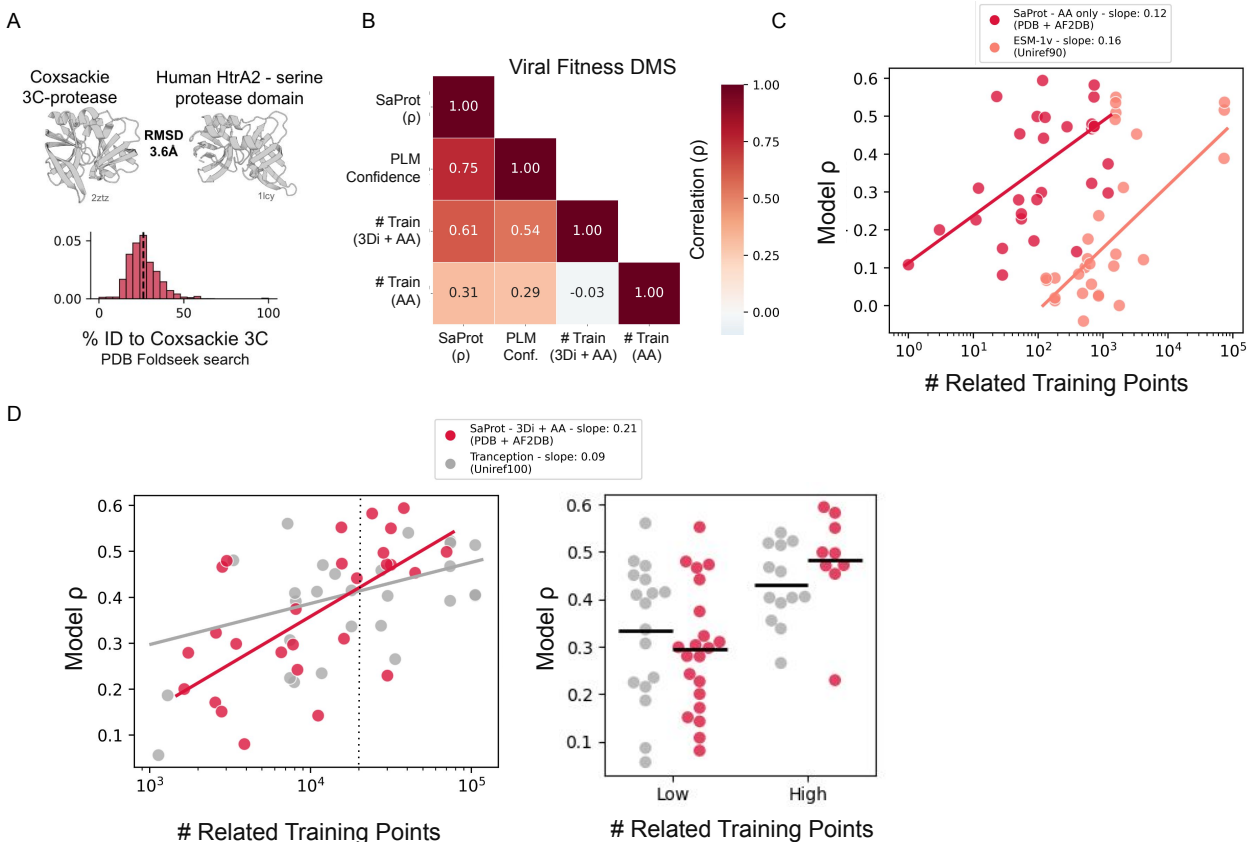
Supplementary Figure 6: Model performance on EVEREST real-world viral evolution forecasting benchmark. (A) Time-course analyses shown for HIV clade. Model predictions correspond to top 250 ranked mutations. Highly frequent mutations are those observed >1k times for influenza and >10k times for SARS-CoV-2 and HIV. (B) Relative performance of of Tranception versus EVE across time. Top panel is relative performance of model-predicted mutations subsequently occurring during evolution. Bottom panel is relative performance on model recalling frequent mutations.



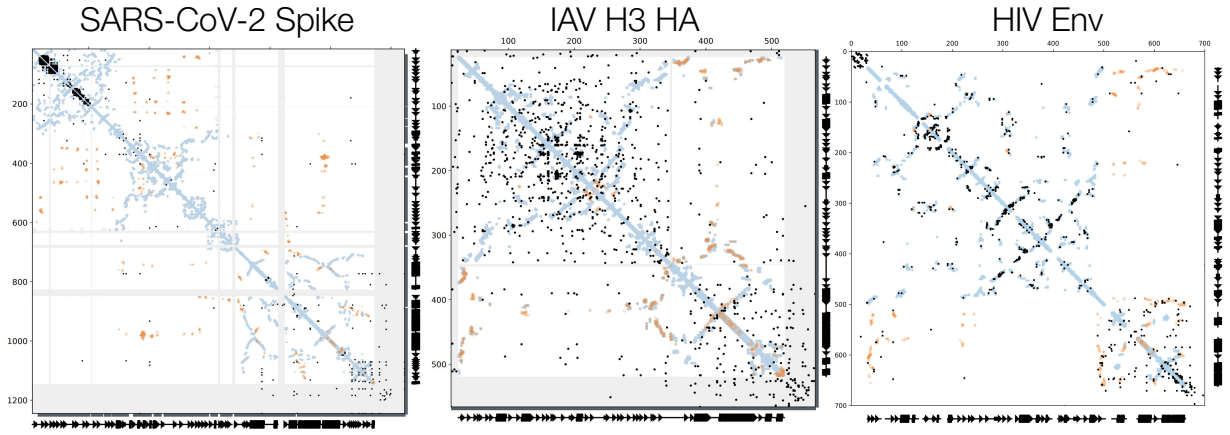
Supplementary Figure 7: Model performance at different top prediction cutoffs. Performance measured as the percentage of mutations seen during evolution. Each column is a virus with 6 SARS-CoV-2 clades, 9 H1N1 clades, 15 H3N2 clades and 1 HIV clade. Highly frequent mutations are those observed >1k times for influenza and >10k times for SARS-CoV-2 and HIV.



Supplementary Figure 8: Model performance at different top prediction cutoffs. Performance measured as percentage of highly frequent mutations captured by model predictions. Each column is a virus with 6 SARS-CoV-2 clades, 9 H1N1 clades, 15 H3N2 clades and 1 HIV clade. Highly frequent mutations are those observed >1k times for influenza and >10k times for SARS-CoV-2 and HIV.



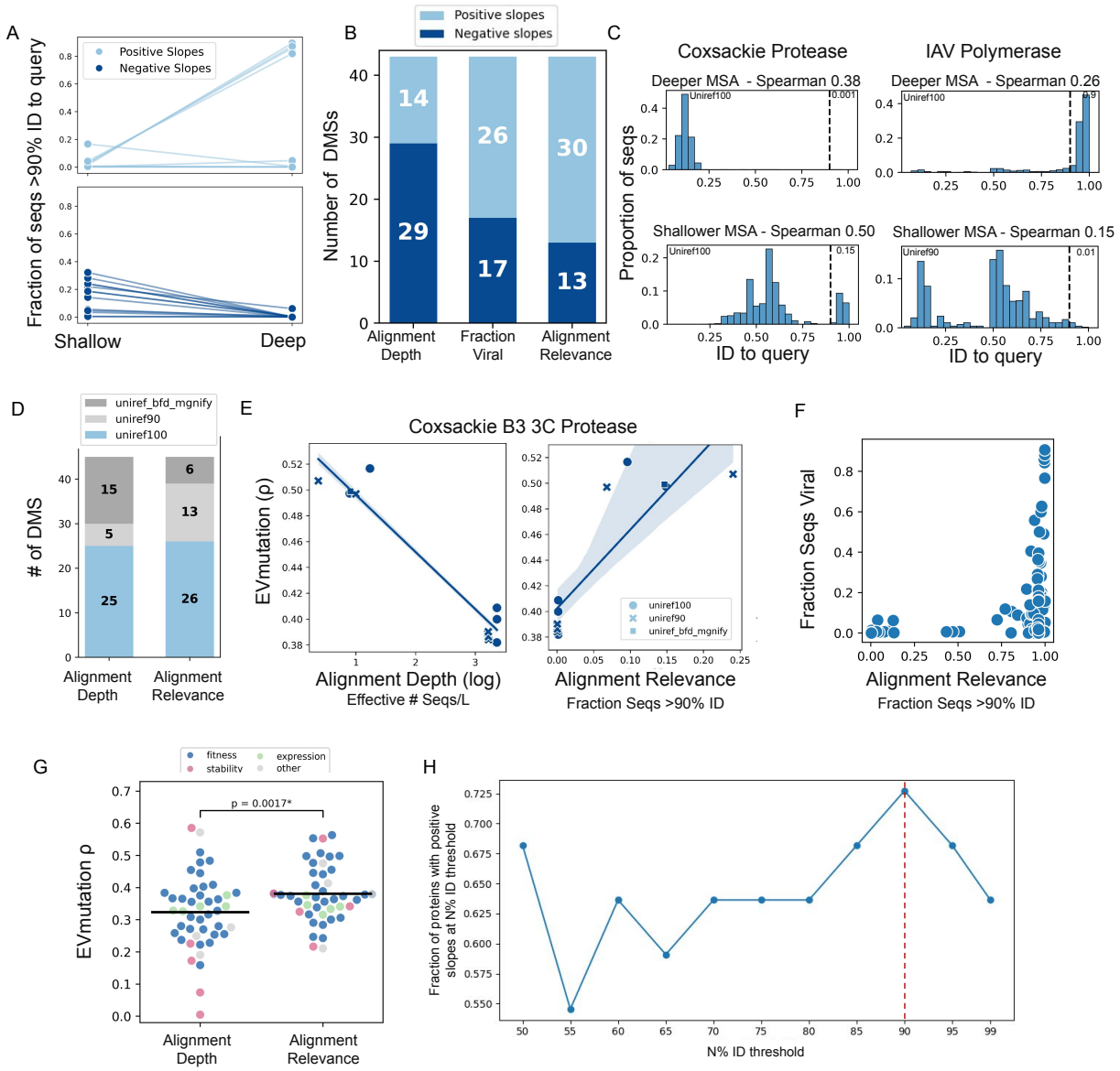
Supplementary Figure 9: (A) By incorporating structural information, many more homologous proteins can be retrieved than by sequence alone when using comparable approaches—between $10^3 - 10^5$ using Foldseek versus fewer than 10^3 using MMseqs2. For instance, a Foldseek search using coxsackievirus 3C protease retrieved thousands of structurally related proteins, most with less than 50% sequence identity. Among them was human HtrA2, a serine protease which shares structural similarity with cysteine proteases (such as the coxsackievirus 3C protease) across the tree of life[130] and membership in the same CATH and SCOP superfamilies[131,132] (Fig. S9). Whether this is an example of true homology (divergence from a common ancestor) or convergent evolution on a common biochemistry is a matter of debate[133]. Coxsackievirus 3C protease structure (2ztz chain A) and Human HtrA2 serine protease domain (1lcy, selected residues 1-210[134]) share a common fold with RMSD of 3.6Å. RMSD identified by CE algorithm for alignment, defined over 160 residues, robust for structures with low sequence similarity. Pairwise sequence identity of all sequences from the Foldseek (3Di + AA, n=2164) search of Coxsackievirus 3C in the PDB, dotted line shown for HtrA2. While such remote homologs are retrievable with more robust sequence homology methods like JackHMMER, their great distance in sequence space may prevent sequence-only PLMs from learning their relevance, and could benefit from structure tokens to increase this signal. These remote structural analogs therefore likely contribute to SaProt's improved generalization across divergent protein families given sufficient structure information. (B) Correlations across viral fitness DMS benchmark between SaProt performance (Spearman correlation with DMS), SaProt pseudo-perplexity-based reliability score, number of related sequences in training data identified via sequence homology search, and number or related sequences in training data identified via sequence and structural homology search. (C) PLM performance corresponds to number of relevant sequences in training datasets (via mmseqs search of amino acids (AA) only). SaProt (trained on PDB + AF2DB) learns from much fewer related sequences (by amino acid identity) than ESM-1v (trained on UniRef90) to achieve greater correlation with DMS, despite using an ESM architecture. (D) PLM performance heavily corresponds to number of relevant sequences and structures in training datasets (via Foldseek search using amino acids and 3Di tokens) (left). SaProt may have a performance boost or be more consistent when there are high numbers of related training points (though this is less common to have high numbers of relevant training points in the PDB+AF2DB). Swarmplot (right) shown when splitting number of relevant sequences for each appropriate training dataset based on whether greater than 20,000 (dotted line on left). Slopes calculated with robust linear model estimation using RANSAC. However, the extent to which errors arise in predicted viral structures resulting from low sequence diversity[105] remains to be investigated and may impact downstream performance, even when only local structural information is encoded. Moreover, it is also possible that learning from the PDB has been optimized such that its few data points are exceptionally useful for mutation effect prediction as more important sequences are generally selected for structural characterization.



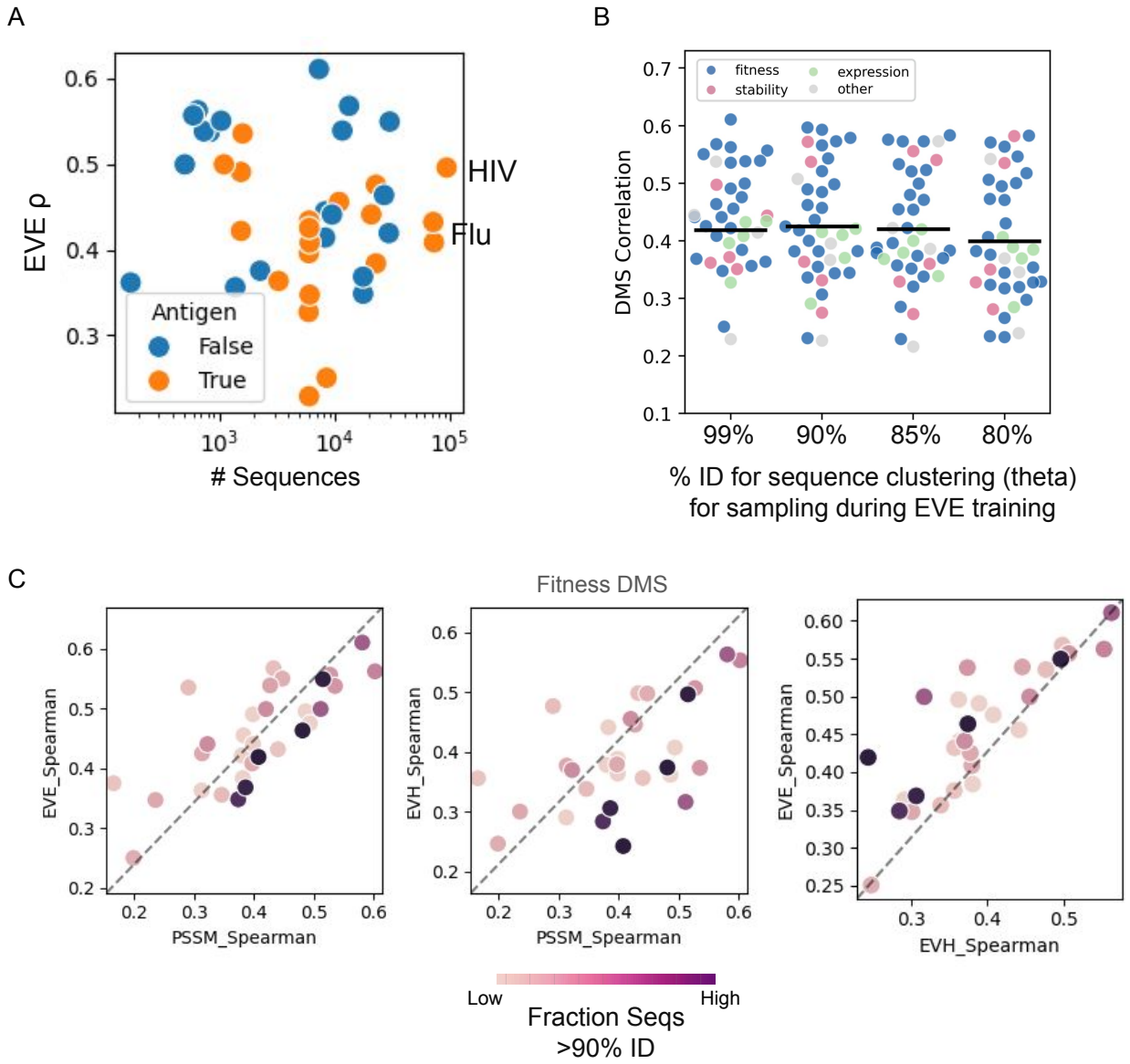
Evcouplings is unable to return correct structural contacts for most viruses due to low sequence diversity, even when mutation effect predictions are accurate

One of the few successful structure contact predictions

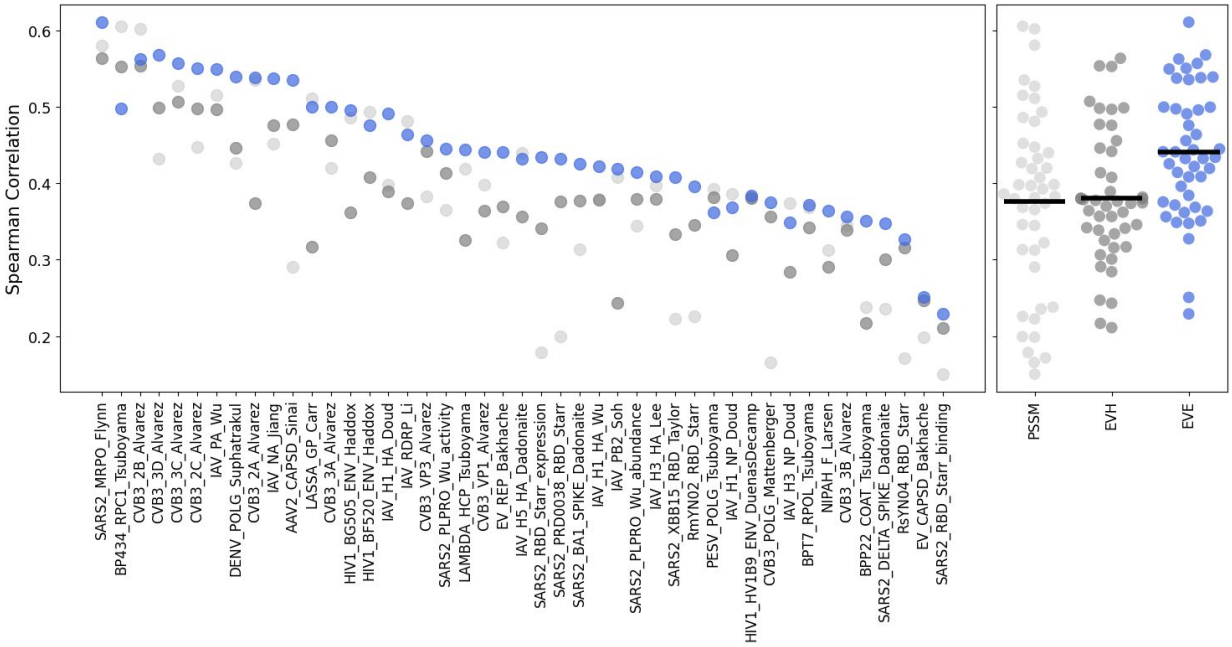
Supplementary Figure 10: Evolutionary couplings fail to recuperate structure contacts for most viral proteins. The ability to predict structural contacts is a strong predictor of model performance for other taxa. Contact maps from UniRef100 alignments with bit score 0.1 for SARS-CoV-2 Spike, Influenza H3 and HIV Envelope proteins. HIV Env is one of the few examples of viral proteins where evolutionary couplings successfully predict structure contacts. Blue are inter-molecular contacts and orange are intra-molecular contacts from PDB, black are top evolutionary couplings (predicted co-evolving residues pairs) from the Potts model[135].



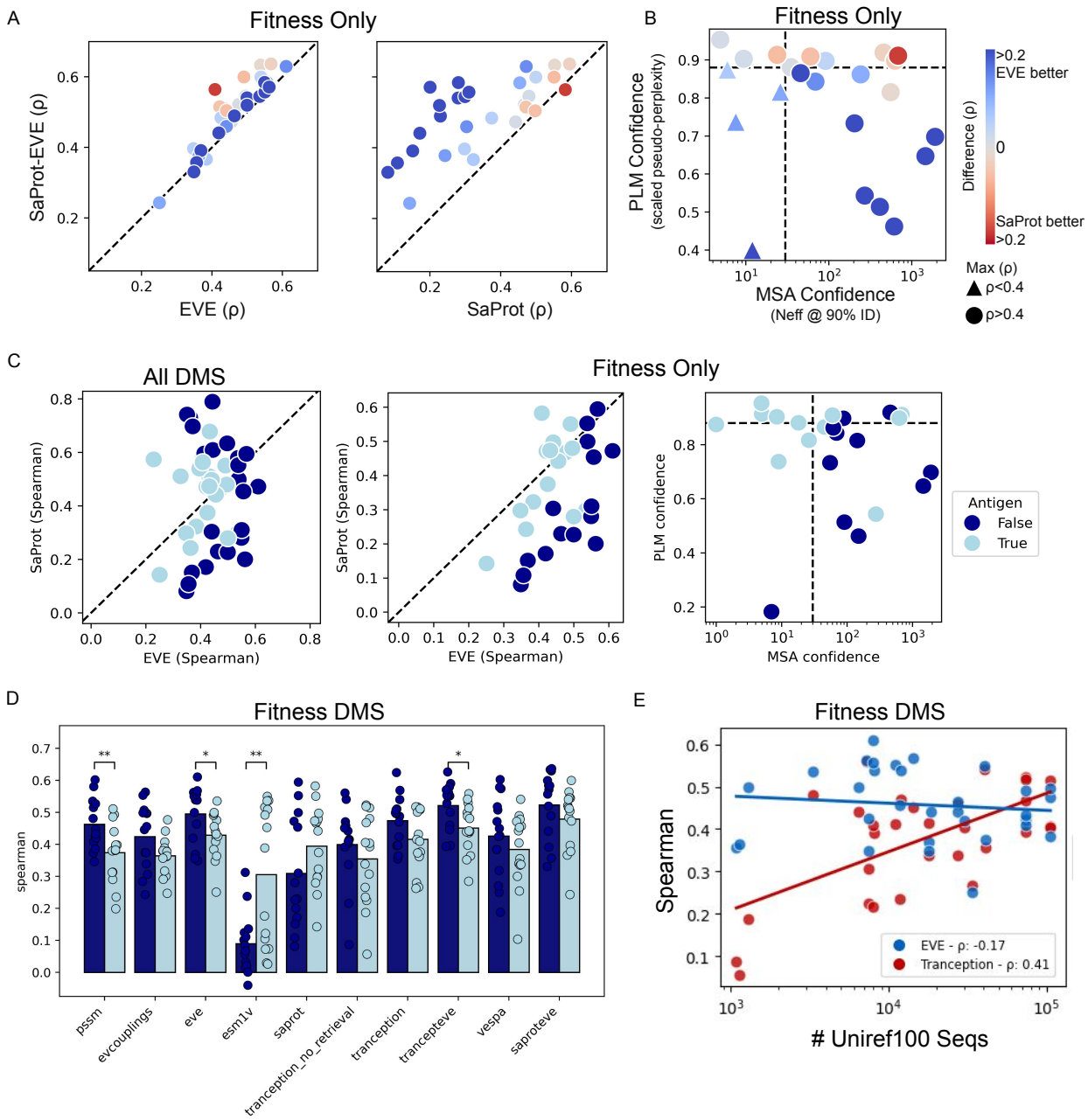
Supplementary Figure 11: Alignment relevance is a better predictor of model performance than commonly used alignment depth. Alignment relevance is measured as the fraction of sequences in the alignment within greater than 90% identity to the quety. Alignment depth is measured as the number of effective sequences over protein length. (A) Deeper alignments improve (or reduce) performance when a larger (or smaller) proportion of sequences are within 90% identity of the target sequence. Proteins that benefited from deeper alignments had a higher proportion of closely related sequences, while those that performed worse contained more distantly related sequences. (B) Alignment relevance is the best predictor of PSSM performance over alignment depth (Neff / L) and the fraction of viral sequences in the alignment. Reported are the number of DMS assays with positive (light blue) or negative (dark blue) slopes in performance as a function of different alignment metrics. Only proteins with significant differences in performance (Spearman difference between best and worst alignment >0.1) are reported. (C) Increasing alignment relevance increases Spearman correlation from 0.38 to 0.50 for the coxsackievirus protease (query). Using a deeper alignment, the majority of sequences have less than 25% identity to query (with less than 0.1% of sequences within 90% identity, dotted line). Using a shallower alignment, 15% of sequences are within 90% identity of the query. Conversely for IAV polymerase, the deeper alignment has a higher relevance and yields higher performance. (D) Fraction of selected alignments using each database for old and new alignment selection metric. (E) Using alignment relevance as a metric for alignment selection instead of depth is a better predictor of Spearman performance for coxsackievirus protease. (F) Relationship between alignment relevance and fraction of viral sequences in the alignment across Uniref100 alignments at different bitscores. (G) Improved model performance (Spearman correlation between Potts model prediction and DMS) across EVEREST DMS benchmark when using alignment relevance versus depth for alignment selection. (H) Model performance at different % ID threshold for defining alignment relevance across a subset of ProteinGym.



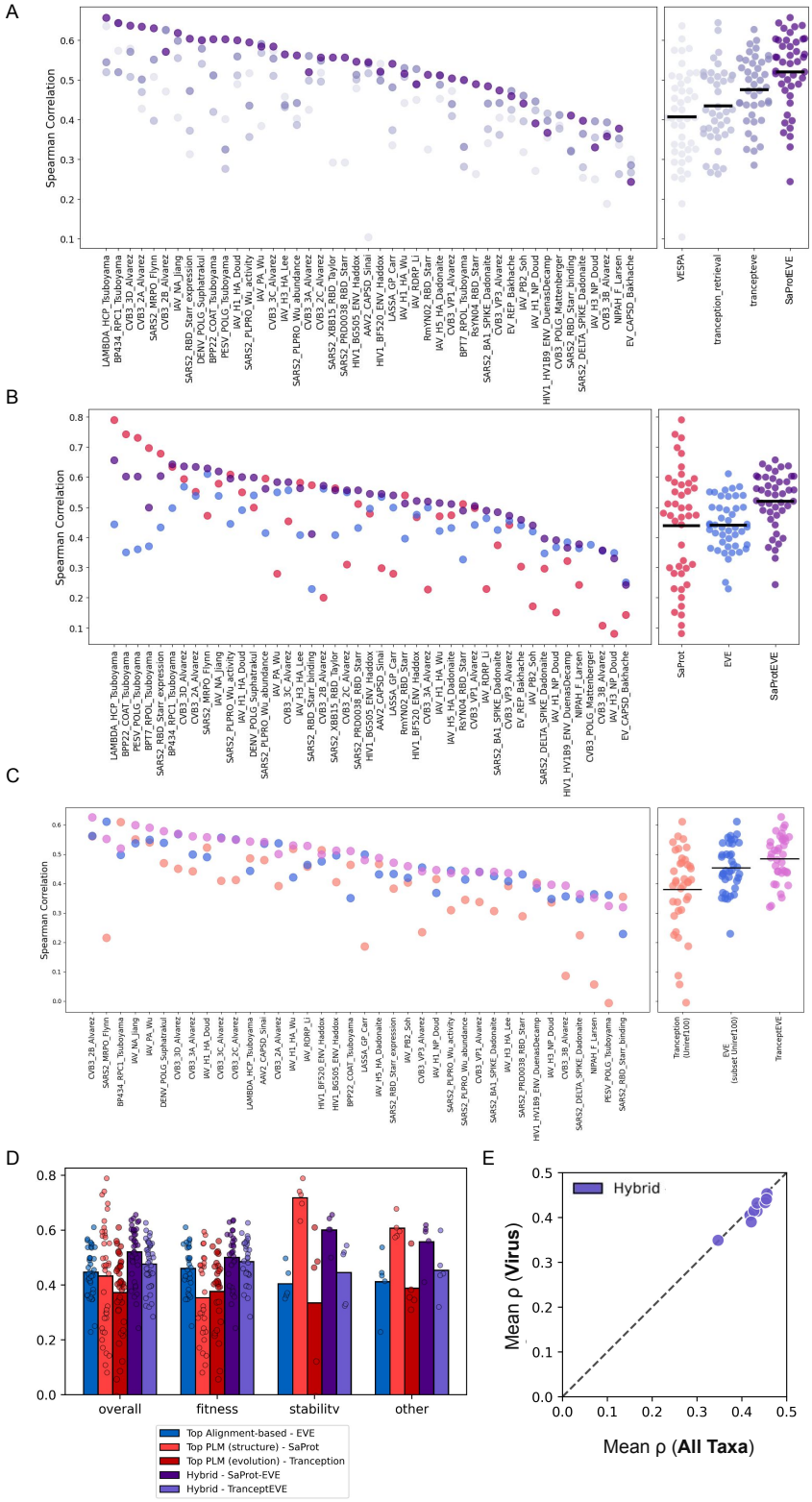
Supplementary Figure 12: (A) No relationship between raw number of sequences and EVE performance, with no clear performance boost for the most highly sequenced viruses (HIV and Flu) (B) Similar model performance (Spearman correlation between EVE model prediction and DMS) across EVEREST DMS benchmark when using different percentage sequence identity cutoffs (thetas) for clustering to determine sampling frequency during EVE training. (C) Relationship between alignment relevance and performance of different alignment-based models. PSSMs are more likely to outperform Potts models when the alignment is composed of more relevant sequences to the query.



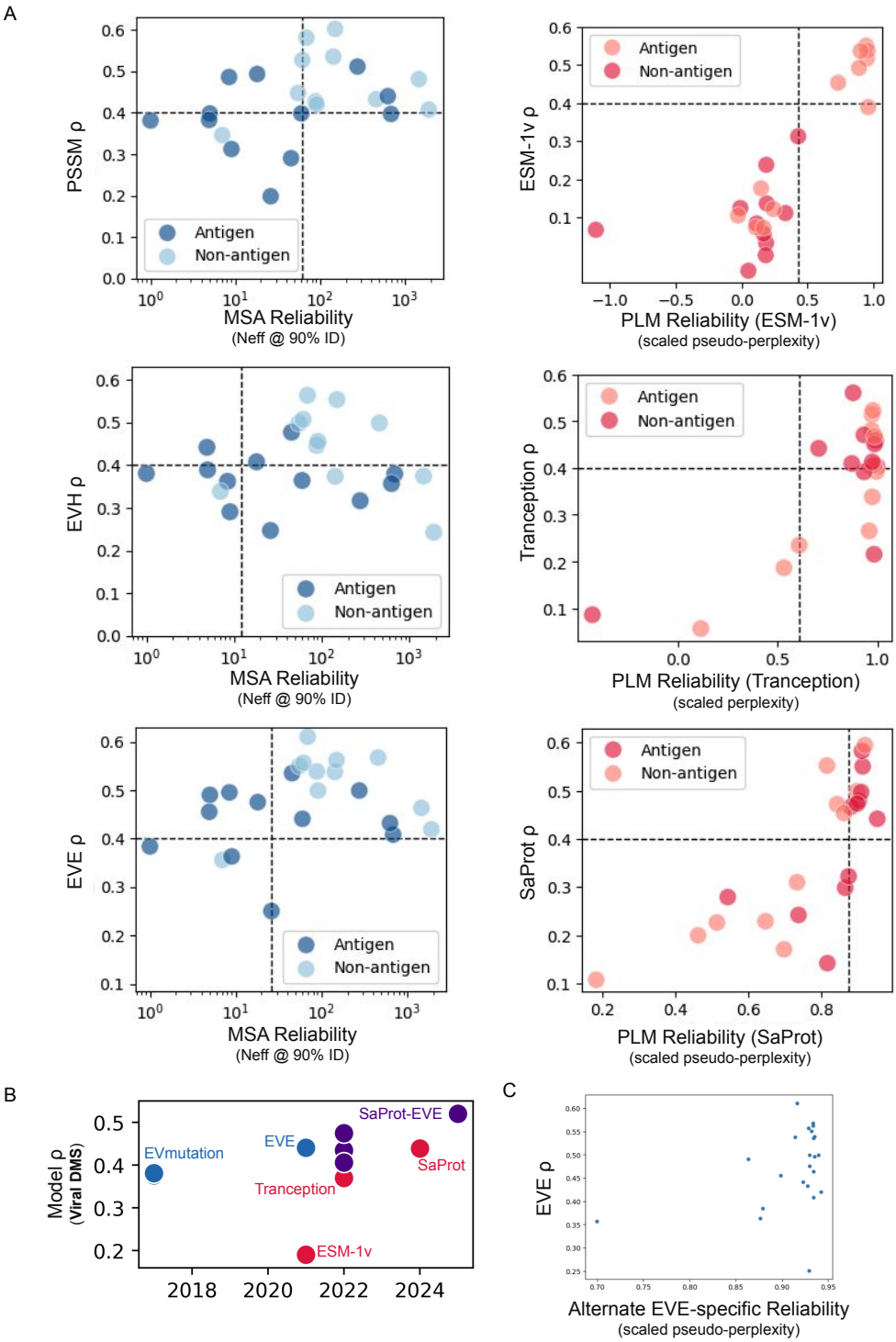
Supplementary Figure 13: Spearman rank correlation between alignment-based model score and DMS score across EVEREST DMS benchmark. EVE outperforms other alignment-based models.



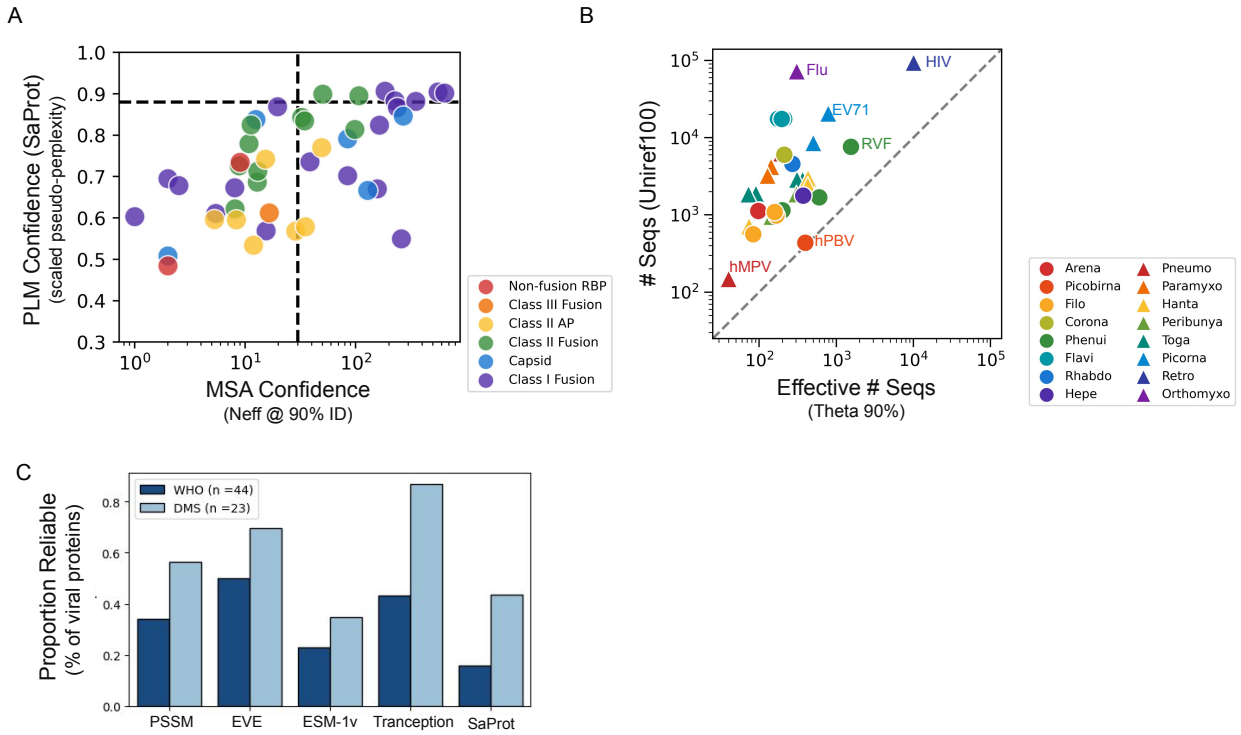
Supplementary Figure 14: (A) SaProt-EVE performance meets or exceeds EVE or SaProt correlations for all viral fitness DMSs. (B) Viral proteins in fitness DMS with high EVE reliability scores (Neff90) have better performance with EVE; those with higher SaProt reliability scores (scaled pseudo-perplexity) are better modeled with SaProt. Size depends on if max Spearman for DMS across EVE and SaProt is greater than 0.4. Lines shown at optimal thresholds that maximize sensitivity and specificity for whether confidence metric is predictive of model-DMS correlation being higher than 0.4 (high confidence lines). (C) Non-antigens have higher consistently higher EVE scores and MSA confidence values for a particular SaProt score range, whereas the highest SaProt scores are a mix of antigens and non-antigens. (D) Some models have different performances on fitness DMSs for antigens and non-antigens. Significance by paired t-test. (E) Tranception performance increases with more related sequences in Uniref100 by MMseqs search. The associated best EVE model performance shown with respect to the Uniref100 number of sequences, to inform model selection.



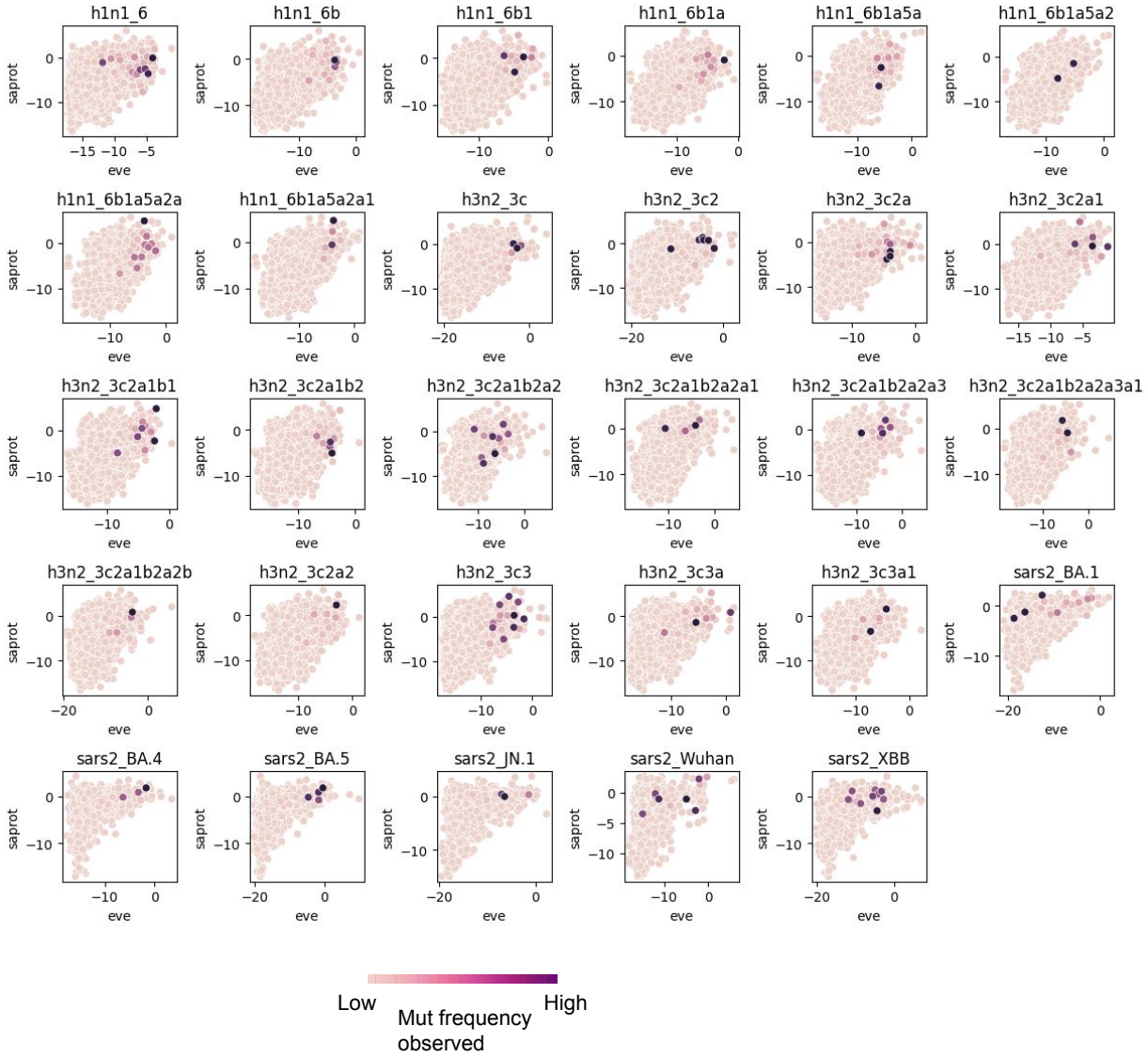
Supplementary Figure 15: (A-C) Breakdown of hybrid models performance on each protein in EVEREST DMS benchmark. (A) Performance of models shown in main: VESPA, Tranception with MSA, TranceptEVE, and SaProt-EVE. (B) SaProt, EVE and SaProt-EVE (C) Tranception, EVE and TranceptEVE. (D) Alignment-based models, PLMs, and hybrids perform best on different DMS phenotypes. (E) Unlike protein language models, hybrid models have similar performance for viral proteins compared to proteins from other taxa. Performance is measured as Spearman correlation with experimental deep mutational scanning (DMS) assays from prior benchmark[17].



Supplementary Figure 16: (A) Reliability scores predict performance for a given viral protein for PSSM, EVE, ESM-1v, and SaProt models, though not for Potts model or Tranception. For alignment-based models we use effective number of sequences at 90% sequence identity to the query and for PLMs we use scaled perplexity (autoregressive) or pseudo-perplexity (MLM). The optimal threshold that maximizes sensitivity and specificity for predicting high EVE performance (defined as Spearman $\rho > 0.4$) was 30 (dotted line). For comparability, we subset to the set of fitness assays, with length <1024 (context size of PLMs) and date post-2015. See methods for more details. (B) Average spearman on viral-specific DMS benchmarks across evaluated models have improved over time. (C) Correlation between EVE performance and an alternate attempt at a reliability metric using a "pseudo-perplexity" of the EVE Evidence Lower Bound (ELBO), a lower bound to the log-marginal likelihood of each amino acid at each position.



Supplementary Figure 17: A. PLM (SaProt) and MSA reliability scores for all 44 WHO priority or prototype RNA viral antigens. WHO priority pathogens colored by structural class. B. Number of related sequences in a JackHMMER search of Uniref100 and when clustered with theta 0.9 for each priority virus. HIV has the most sequences, including when looking at diverse effective number sequences at theta 0.9. C. Proportion of reliable viral proteins across viral DMS (subset of 23 comparable fitness assays) or 44 WHO priority or prototype RNA viral antigens. Models included where reliability estimates predict performance (see Fig. S16). DMS benchmarks overestimates generalizability of model performance due to focus on well-characterized viruses.



Supplementary Figure 18: Distribution of model predicted mutation scores for SaProt and EVE colored by mutation frequency.