VARIANT EFFECT PREDICTION WITH RELIABILITY ESTIMATION ACROSS PRIORITY VIRUSES

Sarah Gurev[*†]
Systems Biology
Harvard Medical School

Noor Youssef[*†]
Systems Biology
Harvard Medical School

Navami Jain
Systems Biology
Harvard Medical School

Debora S. Marks[†]
Systems Biology
Harvard Medical School

ABSTRACT

Viruses pose a significant threat to global health due to their rapid evolution, adaptability, and increasing potential for cross-species transmission. While advances in machine learning and the growing availability of sequence and structure data offer promise for large-scale mutation effect prediction, viruses present unique biological and informational constraints that may challenge these models. To quantify this, we introduce EVEREST—a framework for Evolutionary Variant Effect prediction with Reliability ESTimation—which assesses model performance on viral mutational fitness prediction using a curated benchmark of 45 viral deep mutational scanning datasets (over 340 thousand variants) and quantifies model reliability in the absence of experimental data. This large-scale evaluation revealed wide differences in prediction accuracy across models and viral families. Protein language models have reach state-of-the-art performance at many mutation effect tasks, yet their effectiveness for viruses has been unclear despite their increasing deployment. We find that protein language models, trained on diverse sequence corpora, underperform on viruses compared to alignment-based models trained on much smaller sets of homologous sequences. We calibrate reliability estimates against the DMS data, and then apply this framework across 40 WHO-prioritized pandemic-threat viruses (over 400 thousand variants across 16 viral families), discovering that current models fail to reliably predict mutations in over half of these viruses. Our findings uncover key factors leading to underperformance, offer actionable recommendations for improving viral mutation effect prediction, and provide an objective framework for analyzing dual-use biosecurity risk.

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[1]. Accurately forecasting how these viruses evolve could transform surveillance and therapeutic design by identifying immune or treatment-evading strains early. While both experimental and computational tools exist to predict the functional impact of mutations, many high-risk viruses cannot be studied safely in the laboratory, and experimentally testing the vast space of possible variants remains a significant challenge. Computational approaches offer a scalable alternative—but their performance in viral contexts has not yet been systematically evaluated.

Computational approaches for predicting mutational effects generally fall into two categories: alignment-based models and protein language models (PLMs; Fig. 1A, S1). Alignment-based models infer functional constraints from multiple sequence alignments (MSAs) of homologous proteins and have shown success in modeling viral evolution, immune escape, and vaccine effectiveness particularly for SARS-CoV-2[2–8]. However, their generalizability to high-risk viral families may be limited, as many lack the diverse sequence data required for accurate inference. Moreover, these models cannot accommodate insertions and deletions—common features of viral evolution[9]—due to their reliance on fixed-length sequence representations.

PLMs offer an alternative, alignment-free approach. Trained on large corpora of protein sequences, these models can, in principle, generalize across families and handle sequences of arbitrary length[10–15]. Some PLMs also leverage alignment information during inference or training[11,16–18] or integrate structural context using vocabularies of both sequence and structural tokens or as hybrids[19–21]. Although PLMs have achieved state-of-the-art performance on

[*]Equal contribution
[†]Correspondence: sgurev@mit.edu, noor_youssef@hms.harvard.edu, debbie@hms.harvard.edu

many general protein benchmarks[22], their effectiveness in viral applications—especially under conditions of limited sequence data—remains unclear despite their increasing use in this domain[23–25].

Unique features of viral proteins make it challenging to extrapolate best modeling practices from evaluations on non-viral tasks. Viral proteins mutate at rates far exceeding those of cellular organisms and exhibit distinct biophysical properties, such as low stability, loosely packed cores, and structural and functional plasticity[26,27] (Fig. S2). As obligate intracellular parasites, viruses evolve under strong diversifying selection, particularly in antigenic regions. Consequently, generalized substitution matrices such as BLOSUM often fail to capture the patterns of mutations in viral proteins[2,28]. 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 are limited or not yet available?

We present EVEREST (Evolutionary Variant Effect prediction with Reliability ESTimation), a framework for systematically evaluating single mutation effect prediction in viral proteins. We curated a virus-specific benchmark dataset and evaluated a diverse suite of models—three alignment-based methods (PSSM[5], EVmutation[5], and EVE[6]); two PLMs (ESM-1v[10] and Tranception[11]); one structure-aware PLM (SaProt[19]); and three hybrid approaches (Tran-

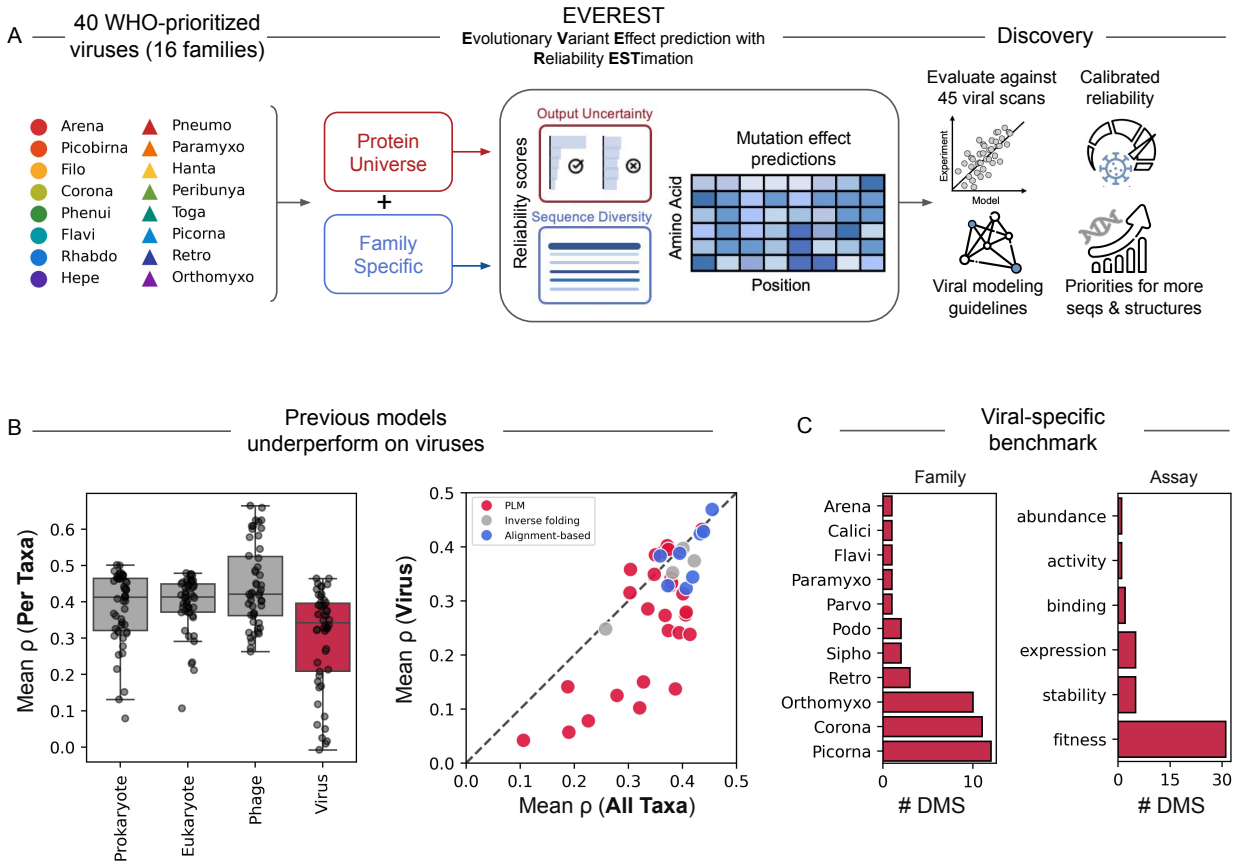


Figure 1: Current computational models underperform on viral proteins relative to non-viral proteins.
A. Overview of the EVEREST framework, which evaluates computational models—including alignment-based methods, protein language models (PLMs), and hybrid approaches—on viral mutation effect prediction. We apply this framework to evaluate model reliability across 44 WHO-prioritized viral antigens.
B. Computational models have lower performance for viral proteins compared to proteins from other taxa (left), driven primarily by reduced performance of PLMs on viral proteins (right). Performance is measured as Spearman correlation with experimental deep mutational scanning (DMS) assays. Results include 218 DMS assays across all taxa versus fewer than 20 assays from eukaryote-infecting viral proteins from a prior benchmark[22].
C. Viral-specific DMS evaluation dataset includes 45 viral proteins: over 340,000 variants across 11 viral families and multiple assay types. See Fig. S4 for details.

ceptEVE[16], Tranception with MSA retrieval[11], and VESPA[29]). These models were selected for diverse architectures and training strategies, as well as high performance on general benchmarks[22] (Table S1). Calibrating against our viral benchmark, we introduce per-protein model-specific reliability metrics quantifying model confidence in the absence of experimental data. This assessment helps uncover why models have historically underperformed on viruses and provides actionable guidelines for future improvement, enabling the development of models for early and accurate prediction of viral evolution essential for pandemic preparedness.

By quantifying model performance and reliability, EVEREST can also be used by public health organizations as an objective, transparent framework to identify models and pathogens with potential biosecurity risks. Applying EVEREST to the 44 WHO-prioritized viral antigens, we show that current state-of-the-art models often fail to provide high-confidence predictions, limiting their immediate utility for pandemic preparedness—and, conversely, their potential for misuse. However, as viral sequence and structure databases continue to expand and newer modeling approaches are adopted, we expect model capabilities to increase, including our abilities to predict diverse functions and mutation effects in whole-viral contexts. In collaboration with biosecurity experts, EVEREST can enable proactive governance and risk mitigation to accompany these advances.

Results and Discussion

Current computational models for mutation effect prediction perform markedly worse on eukaryote-infecting viral proteins than proteins from other taxa (Fig. 1B, S3, Table S1). This performance gap is primarily driven by PLMs, as alignment-based models maintain stable performance across viral and non-viral proteins and outperform PLMs on viral prediction tasks (Fig. 1B, S3).

To systematically compare modeling strategies in viral contexts, we curated a virus-specific benchmark of 45 deep mutational scanning (DMS) assays[30–56]. This doubles the number of viral assays available in prior general-purpose benchmarks[22], and includes more than 340 thousand variants tested across 11 viral families (Fig. 1C). This dataset covers "general fitness" assays measuring phenotypes such as infectivity, replication, and cell entry—broad indicators of viral function directly relevant to pandemic forecasting—as well as additional assays focused on specific properties like stability and expression, but excludes viral anitbody escape (Fig. S4, Table S2). By evaluating model performance across this diverse set of assays and viruses, we aim to identify modeling strategies that generalize across viral families and can be deployed rapidly in outbreak scenarios involving WHO-prioritized viruses.

Alignment-based Models

Alignment-based models rely on homologous sequence alignments to learn evolutionary constraints. Yet, a central challenge in using alignment-based methods is identifying *a priori* which set of sequences will yield optimal model results. Prior strategies for alignment selection often include selecting MSAs that recover structural contacts[5,22]. However, most viral MSAs—except for a few notable cases such as HIV—fail to recapitulate known structural contacts from couplings analysis (Fig. S5). 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[5,22]. Alignment depth can be categorized as low (Neff/L$<$1), medium (1$<$Neff/L$<$100), or high (Neff/L$>$100)[22].

For each viral protein in our benchmark, we generated MSAs using three reference databases: UniRef90, UniRef100, and UniRef100+BFD+MGnify, each with more than 450 million, 205 million, and 4.5 billion sequences, respectively. We systematically varied alignment depth at six sequence similarity thresholds (Fig. 2A). We then trained three alignment-based models with increasing epistatic complexity: a position-specific scoring matrix (PSSM; site-independent)[5], EVmutation (pairwise couplings)[5], and EVE (a fully connected latent variable model)[6].

Contrary to common practice, we found that while deeper alignments improved performance for some viruses (showing a positive correlation between alignment depth and agreement with DMS measurements), they greatly reduced performance for others (Fig. 2B). In fact, for most proteins in our viral benchmark, maximizing alignment depth reduced performance, particularly for proteins with the most sequences retrieved (Fig. 2C). We hypothesize that beyond a minimum alignment depth, incorporating additional sequences—especially from distantly related backgrounds—may introduce noise or conflicting constraints, as these sequences are subject to distinct selective pressures.

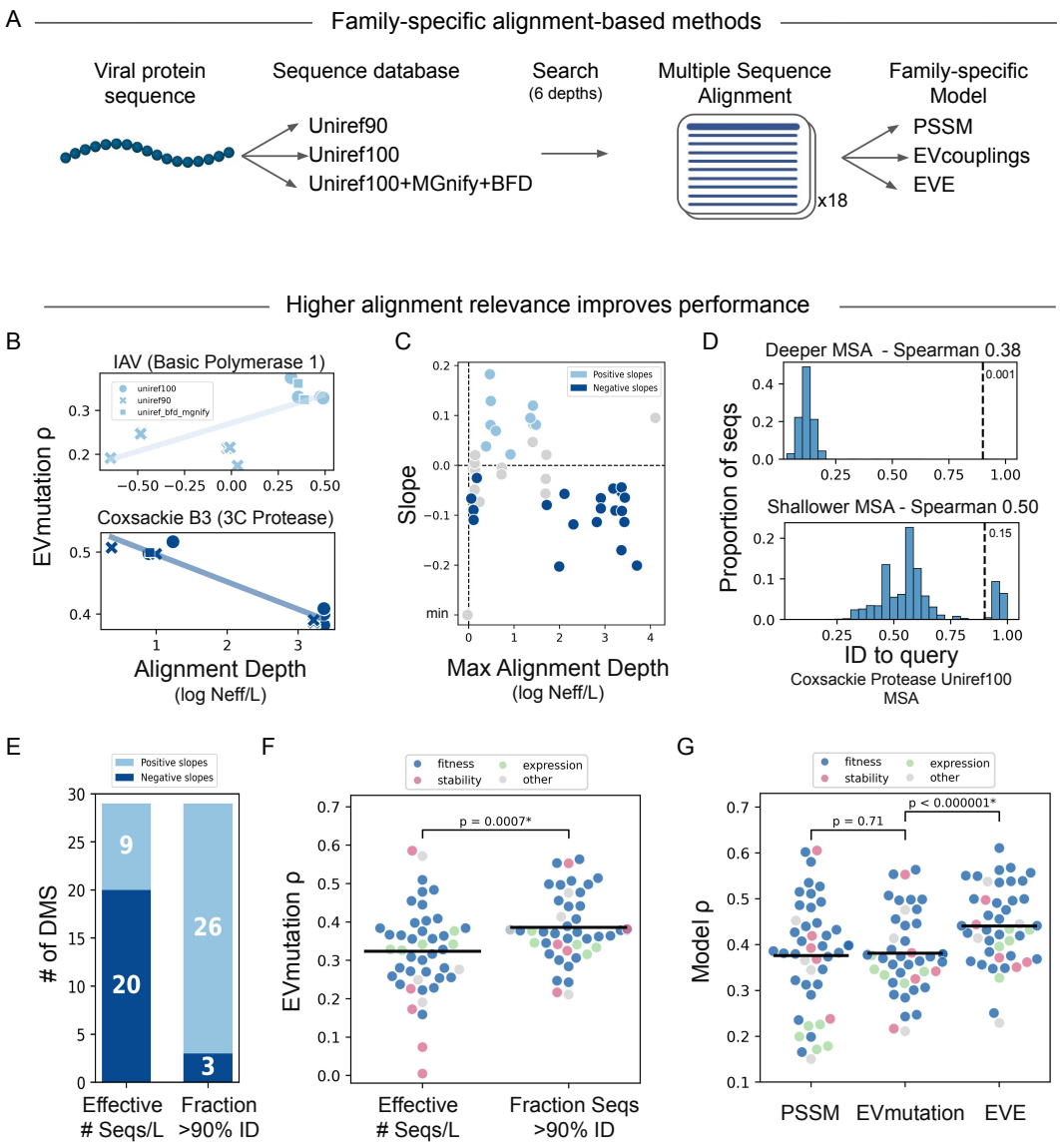


Figure 2: Balancing alignment depth and relevance improves model performance.
A. Overview of MSA generation and alignment-based models.
B. Increasing alignment depth (Neff/L) improves EVmutation[5] performance for influenza polymerase, but decreases performance for coxsackievirus protease. Performance measured as Spearman correlation between model prediction and DMS.
C. Alignment depth is a poor predictor of EVmutation performance. Colors indicate whether performance increased (blue), decreased (red), or was unchanged (gray; Spearman range < 0.05) with increasing alignment depth. Only one protein has the maximum Neff/L of all alignments categorized as "low" (less than 0 in log-scale).
D. Increasing alignment relevance (fraction of sequences >90% identity) increases Spearman correlation from 0.38 to 0.50 for the coxsackievirus protease (query). Deeper alignment contains mostly distant sequences (less than 25% identity), while a shallower but more relevant alignment contains 15% of sequences within 90% identity (dotted line).
E. Alignment relevance is a better predictor of model performance than alignment depth. Reported is the number of viral proteins with positive slopes between EVmutation performance and alignment-selection metric for DMSs where Spearman range is >0.05.
F. Improved model performance across viral benchmarks when selecting alignments based relevance versus depth. DMSs colored by assay type.
G. EVE[6] is a significant improvement over EVmutation ($P < 0.001$, paired t-test). Alignments selected using fraction 90% identity.

To test this, we examined whether performance correlated with "alignment relatedness", quantified as the proportion of MSA sequences sharing >90% identity with the target viral protein. In the case of the coxsackievirus protease, selecting the alignment that maximized alignment relatedness improved EVmutation performance from a Spearman correlation of 0.38 to 0.50 (Fig. 2D, S6). We found that this relatedness heuristic was a stronger predictor of model accuracy than alignment depth (Fig. 2E), and that selecting alignments based on this significantly improved EVmutation performance across the benchmark (from an average Spearman of 0.32 to 0.39; $P < 0.001$, paired t-test; Fig. 2F). The increased relevance of the alignments is especially helpful for predicting fitness assays, where functional differences in divergent sequences are especially important. Moreover, shallower related alignments were equally effective as deeper alignments for stability or expression prediction, where it might be expected that deeper alignments would be preferred as the learned phenotype is more consistent across the alignment. We therefore 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.

Most alignments selected with this heuristic used the UniRef100 sequence database—a non-redundant protein database from UniProt[57] (Fig. S6). This is likely because our selection method prioritizes alignments enriched in functionally relevant, closely related sequences while disfavoring the inclusion of overly divergent ones—common in larger metagenomics databases like BFD and MGnify, despite the prevalence of viruses (especially phages) in metagenomic sequencing[58,59]. Protein fragments are also more common in BFD, and their removal would also be incentivized by a focus on alignment relatedness[60]. In some cases, these larger databases also caused memory issues during model training (Table S3), as BFD+MGnify is almost 9 times the size of Uniref100[57,61,62]. 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.

We used this alignment selection strategy to train EVE models for each viral protein in the benchmark. We found that EVE consistently outperformed both PSSM and EVmutation across nearly all viral proteins, and had a higher average Spearman correlation ($\rho = 0.44$ vs. $\rho = 0.38$; Fig. 2G, S7, Table S4). This is consistent with prior reports that capturing higher-order epistasis, using a variational autoencoder, improves prediction accuracy[2,6,7]. Surprisingly, EVmutation and site-specific PSSM had comparable performance for viruses, in contrast to other taxa (Table S1). EVmutation's pairwise terms may be uninformative for mutation effect prediction due to insufficient sequence diversity, just as most viral multiple sequence alignments fail to capture meaningful structural couplings (Fig. S5).

Protein Language Models

PLMs trained on large-scale protein sequence datasets can in principle learn general constraints governing protein structure and function (Fig. 3A) and therefore hold promise for generalizing to underrepresented protein classes. However, we find that most single-sequence PLMs, including state-of-the-art ESM-2 models, perform worse on viral mutation effect prediction than even the simplest site-independent alignment-based model (PSSM; Fig. 3B, dotted lines). This underperformance suggests that current PLMs are not well-suited for modeling viral proteins.

One contributing factor is that PLMs are typically trained on clustered subsets of UniRef100, such as UniRef90 or UniRef50, which retain only one representative per sequence cluster with ≥90% or ≥50% identity, respectively. Clustering reduces redundancy and improves performance for non-viral proteins by limiting over-representation of well-sequenced families[10], but it disproportionately filters out viral sequences (Fig. 3C, S8). As a result, viral proteins—already underrepresented (<1% of training datasets)—are further diminished during clustering. Most viral proteins fall well below the expected relationship between pre- and post-clustering counts observed for non-viral proteins (Fig. 3C), likely due to their orders of magnitude higher mutation rates and strong and divergent selection pressures (Fig. S2). HIV is a notable exception, as its extreme genetic variability[63] allows the recovery of structural contacts through coevolutionary analysis, a property rarely observed in other viral proteins (Fig. S5). Phages also deviate from this trend, likely due to lower mutation rates (as mostly double-stranded DNA viruses) resulting in higher number of identifiable homologs (Fig. S2). Additionally, the superior performance of PLMs on phage datasets (Fig. 1D, S3) may partially reflect the nature of their DMS experiments, which focus on protein stability—a property that PLMs generally predict well[22].

Model architecture and training objective may also contribute to performance differences(Table S5). Most PLMs operate with a maximum input length of 1,024 tokens, which should accommodate most viral proteins with the exception of long polyproteins and coronavirus spikes. However, the effective context length is typically limited to half the training length[64,65], which limits applicability to most viral antigens. Yet it is unlikely length is the primary driver of performance differences: ProGen2 Base and ProGen2 Medium share identical architectures and training data, differing only in context length (2,048 vs. 1,024 tokens, respectively), yet the larger context model ProGen2 Base performs slightly

worse on viral proteins (Spearman's $\rho = 0.328$ vs. $\rho = 0.342$; Table S5)[13]. In addition, most PLMs are trained using a masked language modeling (MLM) objective that masks only 15% of residues. By contrast, autoregressive models predict every residue given its predecessors—posing a harder task with more numerous and contiguous unknowns that may encourage better generalization on underrepresented sequences like viral proteins. Current autoregressive models are also trained on significantly larger datasets, making it difficult to disentangle the effect of objective function from data volume (Fig. 3D), though it is less clear that the relative success of ProGen2 for viruses necessarily results from the inclusion of BFD30.

We identified two general factors that lead to improved performance on viral proteins: larger viral representation in training datasets and larger models. Training on more comprehensive datasets—such as unclustered UniRef100 with much more viral sequences—substantially improves model accuracy on viral proteins (Fig.3B, Table S5). This suggests that viral applications where models trained on smaller datasets (i.e., ESM-2 trained on Uniref90[23–25]) are employed would likely benefit in a zero-shot setting from switching to PLMs with unclustered training sets. Likewise, scaling up model size leads to increased accuracy on viral tasks, in contrast to the performance plateau often observed

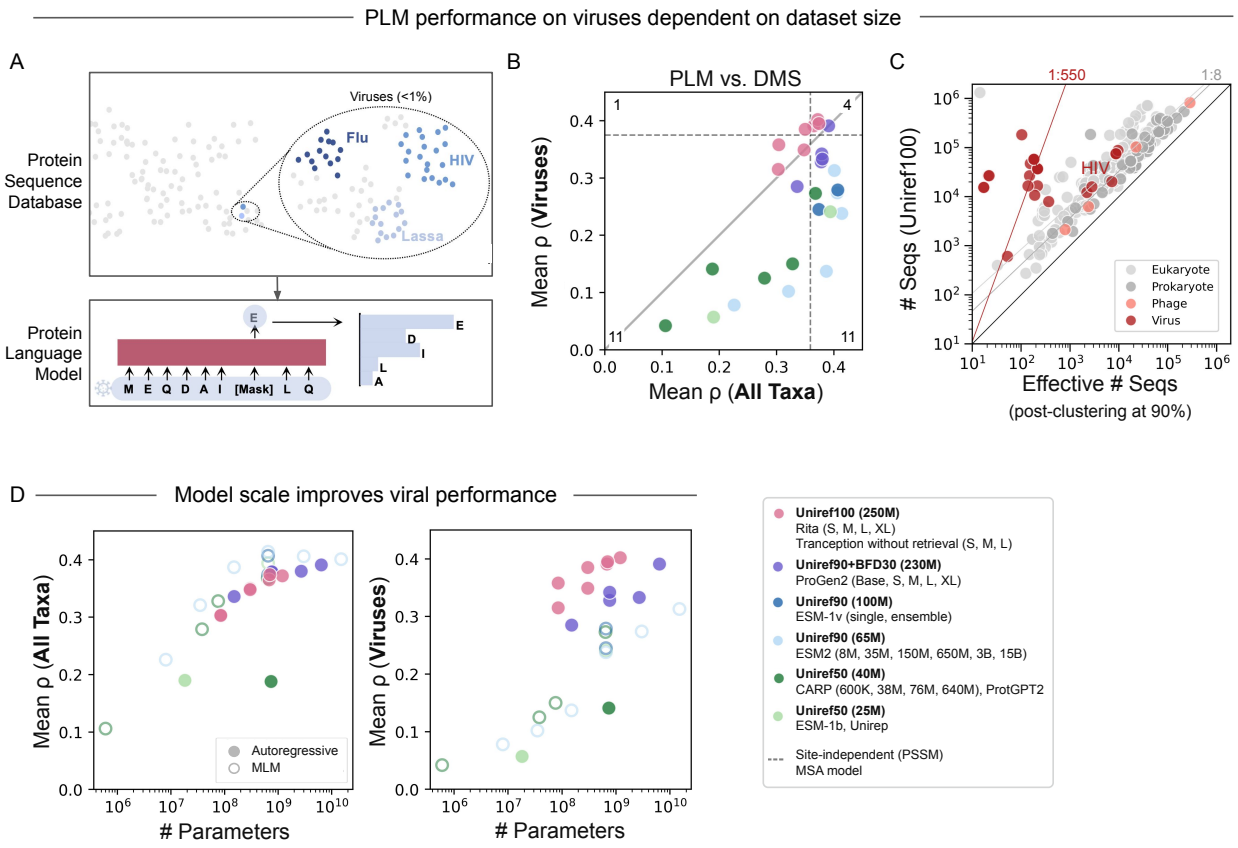


Figure 3: Sequence clustering disproportionately reduces PLM performance on viral proteins.
A. Overview of PLM training on large protein sequence database (commonly used Uniref100 or Uniref90 have <1% viral sequences), and scoring of mutations from a single sequence without family-specific information.
B. Unlike for general performance across all taxa, PLM accuracy on viral proteins is strongly dependent on the training dataset, shown by color (numbers in legend are approximate). Performance measured as average Spearman correlation across all 218 DMS datasets (x-axis) and the viral subset (y-axis) from a prior benchmark, ProteinGym[22]. Dotted line indicates the performance of a baseline PSSM model. See Table S5 for more details.
C. Number of effective sequences (clustered by 90% identity) compared to raw number of sequences in Uniref100. Prokaryotes and Eukaryotes follow a ratio of 1 effective sequence post-clustering: 8 raw sequences pre-clustering when fitting a regression line, while many viruses follow a ratio of 1:550, with the exception of phages and HIV.
D. Increasing model size leads to improved viral performance (right), even as aggregate performance on all taxa saturates (left). Each point represents a PLM colored by training dataset and training objective. Impact of training dataset confounded by exclusively autoregressive models training on Uniref100.

on non-viral proteins (Fig. 3D, Fig. S8)[12,14,66,67]. We hypothesize that larger models are better able to generalize from sparse viral examples by leveraging broader evolutionary patterns learned from the non-viral protein universe. Supporting this, ESM-2 models trained on the clustered Uniref90 dataset shows consistent gains with scale—from 8 million to 15 billion parameters—when considering viruses or other proteins with shallow alignments (fewer than 1,000 effective sequences at 90% identity; Fig. S9).

Our systematic comparative 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 and Pathoplexus[68,69] or relevant subsets of larger databases like BFD, MGnify, and Logan[62,70,71], although if the model goal is not-virus specific, care must be taken during sampling to avoid over-training on any one virus.

Structure-aware Protein Language Models

Structural features can evolve up to ten times slower than amino acid sequences[72]. As a result, some viral proteins retain structural similarity despite sharing less than 10% sequence identity[73]. This conservation of structural signal, even as sequence similarity diminishes, enables more reliable detection of distant homologs and improves protein function annotation[74–76]. We therefore hypothesize that incorporating structural information into PLMs may help mitigate the challenges posed by the sparse and highly divergent nature of viral sequence data.

To test this, we evaluated SaProt, a structure-aware PLM. SaProt encodes proteins using a hybrid alphabet of both amino acids and Foldseek 3Di tokens (learned from local 3D structural environments)[19,77]. SaProt-AF2 was trained on 40 million AlphaFold2-predicted structures from UniRef50 (AF2DB, excluding most viruses), and SaProt-PDB (hereafter SaProt) continued training on experimentally resolved PDB structures. Across our viral benchmark, SaProt consistently outperformed SaProt-AF2, with diminished advantage for viruses with few unique structural representatives in the PDB, such as Lassa and Nipah (Fig. S10). These results highlight the value of incorporating even the limited number of viral PDB structures in training (Fig. S11).

Additionally, SaProt outperformed sequence-only PLMs, ESM-1v and Tranception, on more than half of the viral DMS benchmarks despite their training on larger datasets (UniRef90 and UniRef100, respectively; Fig. 3A). All three models share transformer-based architectures: both SaProt and ESM-1v use a BERT-style MLM objective, whereas Tranception is an autoregressive model[11]. SaProt showed particularly strong performance on DMS assays measuring stability, expression, and other properties where structural context may be especially informative (Fig. 4A, Fig. S10).

In contrast to alignment-based methods, we find that PLM performance improves with the number of relevant training points (Fig. 4B,C). This suggests that PLMs can more effectivity prioritize relevant sequences than alignment-based models, and that their performance is likely to improve as more viral sequences and structures become available. Notably, structural training data have a greater impact on performance than sequence data, as evidenced by three observations. First, SaProt's performance is better predicted by the combined number of sequence and structure training points than by either alone (Fig. S12). Second, despite being trained on fewer relevant amino acid data points than ESM-1v (whose limited UniRef90 data leads to poor performance), SaProt achieves superior performance (Fig. 4B). Third, while SaProt and Tranception had comparable numbers of related training points overall, SaProt exhibited greater performance gains as the number of homologs increased—implying that each additional training point is more informative for SaProt than each sequence-only training point for Tranception (Fig. 4C).

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 (Fig. 4C). 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[78] and membership in the same CATH and SCOP superfamilies[79,80] (Fig. S12). 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.

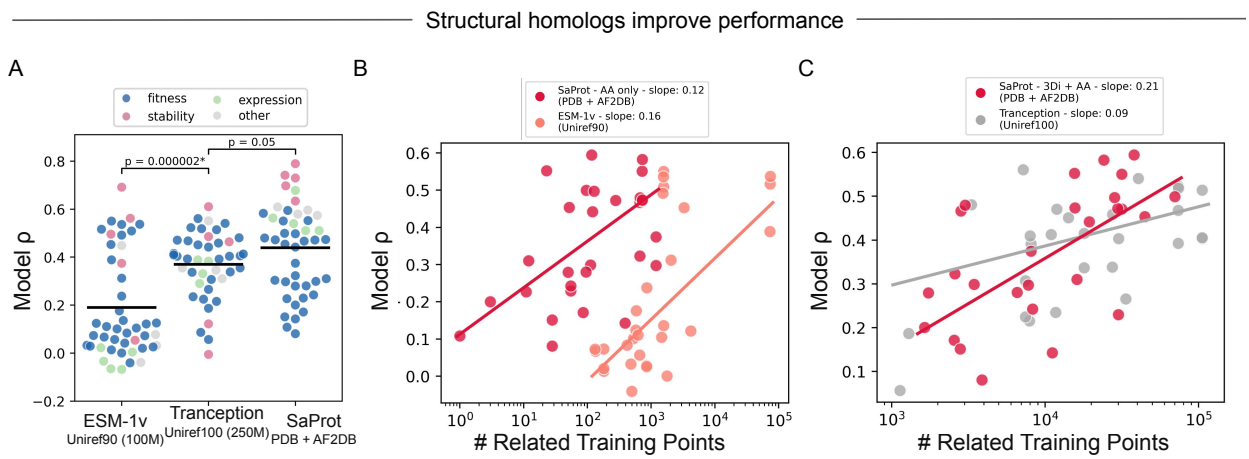


Figure 4: Remote structural homologs compensate for limited sequence diversity.
A. Spearman correlation between PLM and DMS scores across viral protein DMS benchmark. DMSs colored by assay type. Tranception (Uniref100) is a significant improvement over ESM-1v (Uniref90) ($P < 0.00001$; paired t-test).
B. PLM performance on fitness assays 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. MMseqs chosen for comparability to Foldseek.
C. PLM performance on fitness assays heavily corresponds to number of relevant sequences and structures in training datasets (via Foldseek search using amino acids and 3Di tokens). SaProt (trained on PDB + AF2DB) has a greater slope than Tranception (trained on Uniref100 amino acid sequences only, training points via MMseqs) in terms of performance boost from number of related training points. Slopes calculated with robust linear model estimation using RANSAC.

Together, these results suggest that structural information–either the local structural context or its ability to link relevant sequences with remote structural homology–drives the improved performance of SaProt on viral proteins that are currently well-represented (Fig. 4C). While viral structure data in general-purpose structure databases is insufficient for some viruses, this can likely be mitigated by training on newer predicted viral-specific structural databases[75,81–83]. However, the extent to which errors arise in predicted viral structures resulting from low sequence diversity[84] 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.

EVEREST: A Framework for Reliability Estimation

Among alignment-based and PLM approaches, EVE and SaProt were the top-performing models with comparable average Spearman correlations across our viral benchmark (ρ = 0.44 for both; Fig. S13, Table S4). However, their relative performance varied systematically. SaProt-PDB excelled on proteins with abundant structural representations in the PDB and AF2DB, such as phage capsid proteins, influenza hemagglutinin, and viral proteases, as well as for stability assays, but had more variable performance overall. On the other hand, EVE was more consistent, had a higher average correlation for antigen fitness assays (ρ = 0.43 vs. ρ = 0.39), and performed better for viruses with sparse structural data (Fig. S13, Table S4).

Reliability estimation has been useful in other biological domains when presented alongside model predictions, most notably pLDDT predictions for AlphaFold2[84], but such estimates have not yet accompanied variant effect predictions. To address this, we calibrate reliability metrics from our viral-specific DMS benchmark tailored to each modeling approach.

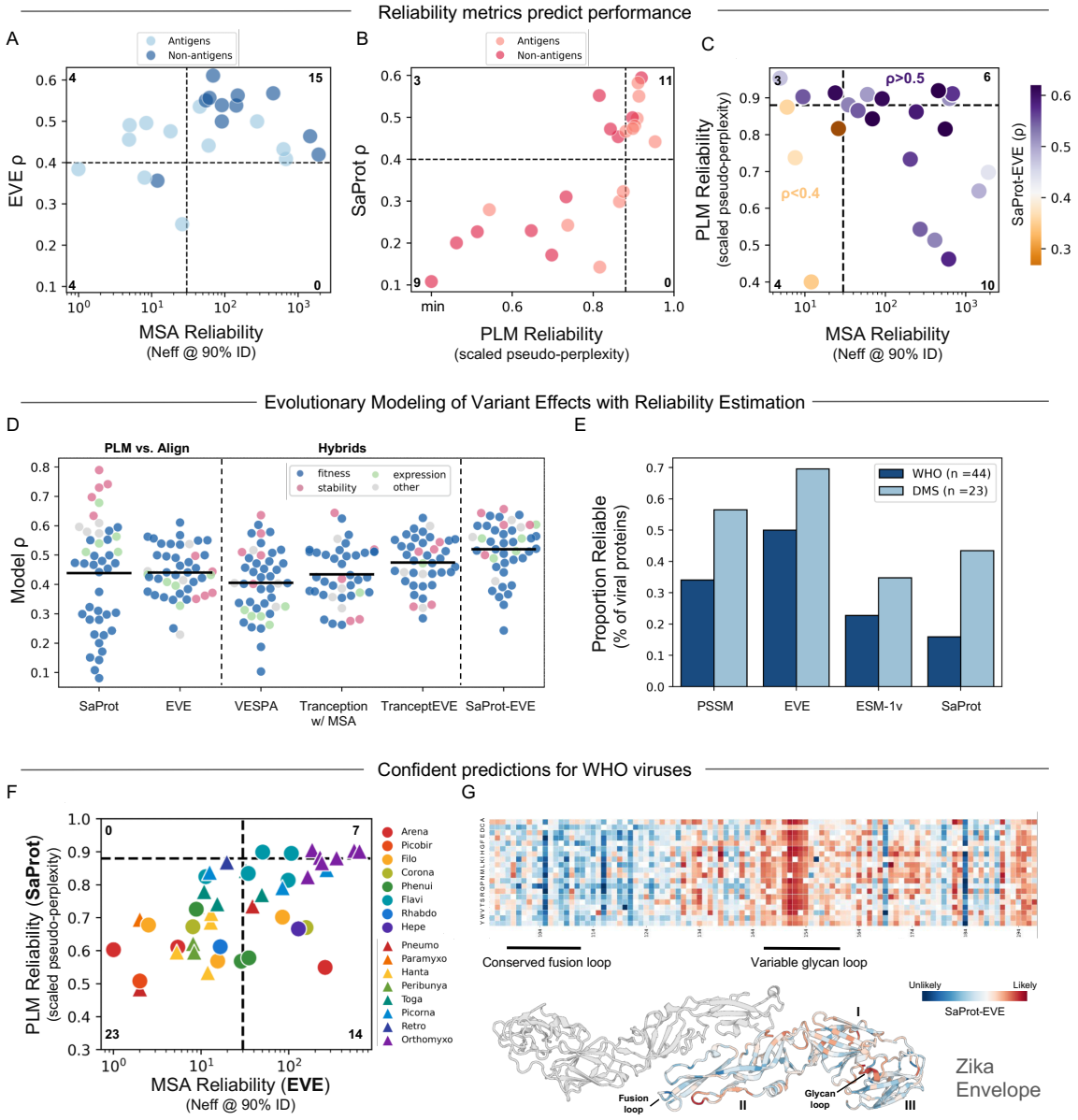


Figure 5: EVEREST produces reliability estimates across priority viruses.
A. Number of effective sequences (Neff) with greater than 90% identity to the query (Neff @ 90% ID) predicts EVE performance for a given viral protein. The optimal threshold that maximizes sensitivity and specificity for predicting high EVE performance (defined as Spearman $\rho > 0.4$) was 30 (dotted line). For comparability, we subset to the set of 23 fitness assays, with length <1024 (context size of PLMs) and date post-2015. See methods for more details.
B. SaProt reliability score, calculated as scaled pseudo-perplexity, predicts model performance. The optimal PLM reliability threshold that maximizes sensitivity and specificity for predicting high SaProt performance (defined as Spearman $\rho > 0.4$) was 0.88 (dotted line). PLM confidence less than 0.4 shown as minimum value.
C. SaProt-EVE has correlation below 0.4 only where both reliability metrics are below their reliability threshold. SaProt-EVE also has correlation greater than 0.5 when both reliability metrics are above their reliability threshold.
D. Hybrid model SaProt-EVE has improved performance relative to either model alone (EVE or SaProt), as well as other hybrid models ($P < 0.005$, paired t-test). DMSs colored by assay type.
E. 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. S14). DMS benchmarks overestimates generalizability of model performance due to focus on well-characterized viruses.
F. PLM and MSA reliability scores for all WHO viral antigens. More than half of WHO antigens have both low MSA and PLM confidence. Most viruses do not have sufficient PLM confidence using a structural-aware PLM.
G. SaProt-EVE predictions for priority virus Zika envelope, showing *in silico* mutational scan and site-average fitness scores mapped on the structure (AF3 structure and PDB 5ire). Fusion loop (98-107) and glycan 150 loop (147-161).

For alignment-based models, we found that the number of effective sequences within 90% identity (Neff @ 90% ID) was a robust predictor of EVE's reliability (Fig. 5A). While fraction of sequences within 90% identity is effective for alignment selection, it is less comparable across different viral proteins. Yet, the importance of highly relevant sequences for alignment selection, informs this choice of Neff tailored to this highly relevant subset of sequences. For PLMs, we used pseudo-perplexity (PPPL), a measure of uncertainty in a model's predictions given a specific input[14,85] (Fig. 5B). 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 (ρ = 0.75), outperforming even the number of homologous training sequences and structures (ρ = 0.68; Fig. S12).

We chose stringent optimal reliability thresholds based on maximizing sensitivity and specificity for predicting high model performance (defined as Spearman $\rho > 0.4$): 30 effective sequences within 90% identity for EVE MSA reliability and 0.88 scaled PPPL (interpretable as an average of ≈3 amino acid choices per residue) for SaProt PLM confidence. We similarly identify cutoffs using these metrics for other alignment-based models and PLMs (Fig. S14). For some models other metrics were more predictive of performance—for example, Tranception performance was better explained by number of relevant training datapoints; Fig 4C). These MSA and PLM reliability scores also predicted which model would perform best for a given viral protein (Fig. S11). 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, S11). However, many of the highest PLM reliability scores are for common viral antigens in the PDB, despite potential for the greater mutability of antigens being interpreted as greater model uncertainty (Fig. 5B, S11).

To leverage the complementary strengths of alignment-based and structure-aware protein language model approaches, we present a new hybrid model, SaProt-EVE, as a more consistent model to use across pandemic-threat viruses. This model serves as an optimal default choice when experimental validation data is unavailable for performance assessment or when a single, robust model is preferred over multi-model evaluation. We find that SaProt-EVE significantly outperformed either EVE or SaProt alone, as well as other hybrid methods, including VESPA, Tranception with MSA retrieval, and TranceptEVE ($P < 0.005$, paired t-test, Fig. 5D, S13, S15), even when restricted to fitness assays and viral antigens (Table S4).

When applied to our viral-specific DMS dataset, SaProt-EVE only underperforms (Spearman correlation <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 the most reliable estimates of variant effects representative of both the model's internal confidence in its predictions and the richness of evolutionary information supporting those predictions. Our results demonstrate that structural-aware PLMs and alignment-based evolutionary models each capture synergistic facets of viral protein space. We integrate these perspectives, providing interpretable reliability metrics and state-of-the-art accuracy, and offering an ideal framework for modeling understudied viral families.

Evaluating 40 WHO Priority Pathogens

The WHO recently prioritized 40 RNA viruses 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[86]. Leveraging the EVEREST framework, we sought to identify the most effective modeling strategies for the antigenic proteins of each WHO-prioritized virus—especially those lacking experimental data—and to assess whether additional sequence or structural information would significantly improve prediction accuracy.

We use our EVEREST reliability metrics to test which WHO priority viruses current models can reliably predict. We find that existing state-of-the-art models fail to reliably predict the functional impacts of mutations in large fractions of these antigens (Fig. 5E-F). Of the 44 WHO viral antigens, only 21 exhibited sufficient sequence diversity to achieve high MSA confidence for EVE, while only 7 antigens—mostly influenza hemagglutinins—demonstrated high PLM confidence for SaProt (all 7 of which also met the high MSA confidence threshold). Based on our viral benchmarks, antigens with high model confidence are expected to yield Spearman correlations greater than 0.4 with DMS experiments (Fig. 5F, dashed lines; Table S6). These findings highlight a key limitation of current structure-aware PLMs: they often lack sufficient viral diversity for confident predictions, underscoring the need for viral-specific structural databases[75,81–83] in training. This will be particularly important for under-sequenced viral families or structural classes, many of which cluster together with low reliability (Fig. S16), reflecting biases in available training data and structural coverage.

EVEREST uniquely enables us to extrapolate model reliability assessments to critical, understudied viruses without large-scale experimental data. When we compare the fraction of viral proteins where modeling would be reliable

between the DMS and the WHO priority viruses, we find that the fraction of reliable predictions for WHO viruses is lower than what would be expected from DMS-based benchmarks, even after accounting for possible false negatives (Fig. 5E). Currently, reliability levels are masked when considering only DMS data, as DMSs are biased to focus on well-studied, well-sequenced viruses. So, while 66% of the DMS dataset are reliably predicted by EVE, only 48% of WHO priority antigens meet the same reliability threshold. The difference between performance on the viral DMS benchmark and the WHO viruses is further exasperated for PLMS with only 16% of WHO viruses having reliable SaProt predictions. However, the disproportionate inclusion of many influenza subtypes (7 of 44 antigens) within the WHO priority virus list does inflate ESM-1v's performance on the WHO subsets. EVEREST applied to WHO viruses therefore provides a more balanced estimate of where current models fall short.

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 to prioritize mutations or identify conserved regions in the absence of experimental data, as many "unreliable" predictions still exhibit correlations greater than 0.3 in our viral benchmark, and increase even further for the hybrid SaProt-EVE. We provide SaProt-EVE predictions for all 44 antigens across the 40 prioritized viruses, alongside estimated model reliability scores as a community resource. These mutation effect predictions recapitulate known aspects of viral proteins. For the Zika envelope protein dimer, a class II fusion protein, model predictions reflect the known evolutionary constrained fusion loop and variable glycan loop[87] (Fig. 5G). These predictions offer a scalable framework for guiding experimental efforts and improving preparedness for both well-studied and emerging viral threats.

Biosecurity Assessment

By providing calibrated reliability estimates for the 40 WHO viruses, EVEREST not only assesses the promise of current models for surveillance and vaccine design, but also their implications for dual-use biosecurity risk[88]. Biosecurity involves carefully balancing risks against benefits, and these considerations are inherently linked: as models improve, they offer greater benefits to public health while simultaneously increasing potential risks. The same capabilities that enable rational vaccine and therapeutic design could potentially be misused to engineer more virulent or immune-evasive pathogens.

We focus here on one component of the larger viral design problem relevant to biosecurity assessment. Our analysis is limited to models predicting single amino acid substitutions in individual proteins, excluding nucleotide-level effects, mutation combinations, insertions/deletions, and full viral genome contexts. We also concentrate on functions relevant to viral surveillance and therapeutic design rather than harm-enhancing mutations (such as increased pathogenicity or virulence), and we exclude viral antibody escape DMSs from our benchmark. Additionally, we focus on functions learnable from natural evolutionary sequences and structures rather than those important under misuse-relevant conditions like non-natural temperature or pH ranges. While the general protein engineering models presented here are already open source, future viral-specific models may require different approaches. Large-scale protein language models trained on viral-specific data that are costly and resource-intensive to train may warrant tiered access or governance frameworks.

We show for the first time how current mutation effect models are unreliable for more than half of pandemic-threat viruses, hindering also the biosecurity risk of these models generalizing to many viruses with harm potential. However, viral-specific performance has improved over time (Fig. S14), and will likely continue to improve with better models, and more viral sequence and structure data. Efforts such as new virus-specific structural databases–such as BFVD, the Viral AlphaFold Database, and Viro3D[75,81–83]–as well as finetuning on specific viruses[25,89–91], are likely to result in better viral models in the near future.

EVEREST uniquely provides new opportunities to evaluate dual-use concerns for emerging viral-specific and general protein engineering models, and we position it as an objective and transparent approach to dynamically assessing current and future capabilities. AI biosecurity assessment–and the AI tools and capabilities themselves–is a quickly evolving domain, where best practices are in the process of being defined. As viral prediction tools become increasingly capable and widely adopted, technical innovation must be matched by cross-domain collaboration between policy, public health, and biosecurity experts. While recent progress has articulated principles for responsible AI development in protein design[92], there is a critical need to operationalize these principles. This requires developing new biosecurity risk frameworks, mitigation practices, and global norms around viral AI tools; we position EVEREST as part of this solution.

CONCLUSION

Our comprehensive evaluation of viral protein mutation effect predictors reveals key limitations and opportunities for advancing current modeling approaches. For alignment-based methods, our results suggest that re-weighting sequences based on their proximity to the target could yield better performance. For PLMs, training on clustered sequence datasets reduces the already limited representation of viral sequences and disproportionately hampers prediction accuracy for underrepresented protein families, and variable performance has meant that predictions could not necessarily be trusted for a new virus. Commonly used PLMs like ESM-2 are therefore unsuitable for many viral applications, despite their increasing deployment in this domain[23–25]. This may be ameliorated by training on unclustered datasets, domain-specific adaptations or finetuning on viral-specific datasets[25,89–91]. Similar insights could be extended to underperforming RNA language models for viruses[93] which likely suffer from similar training dataset problems.

Incorporating additional data modalities offers a complementary strategy. Structure-aware modeling, in particular, mitigates the sparsity of viral sequences by integrating more conserved structural constraints. While this benefit of remote homologs in PLM training may seemingly contradict our finding that alignment-based models should focus on more relevant sequences, it may be that alignment-based models (besides redundancy-weighting) treat all sequences identically regardless of relevance, while PLMs by learning across the protein universe must by necessity learn to reason about sequences with varying degrees of relevance. Moreover, the fact that both alignment-based models and PLMs benefit from less sequence clustering could allow for more leakage of test mutations in the training data, but the previous success of these models even when trained on structured time splits like for SARS-CoV-2 alleviates this concern.

Our EVEREST framework provides reliability estimates for the first time for mutation effect models that accurately predict model performance. By curating the first viral-specific mutation effect benchmark, and then calibrating reliability scores, we can generalize estimates of model performance across a greater set of often under-studied pandemic-threat viruses. This has uniquely allowed us to show that the best models can reliably predict the fitness effects of mutations for only half of WHO priority pathogens, most of which lack extensive experimental fitness data. Conversely, our reliability estimates and model evaluations have allowed us to provide the first large-scale mutation effect predictions across WHO viruses as a community resource, and create an objective, transparent framework for analyzing dual-use biosecurity risk of current and future models. Overall, EVEREST tackles current challenges in viral mutation effect prediction, offering practical heuristics for alignment and model selection, methods for reliability estimation and calibration, and guidelines for improving future models.

DATA AND CODE AVAILABILITY

Data, results, and code are available at https://github.com/debbiemarkslab/priority-viruses for 45 viral DMSs and 44 WHO-priority viral antigens. Code for model training and for mutation effect scoring is available through ProteinGym[22] at https://github.com/OATML-Markslab/ProteinGym/. DMS data used for model evaluation collected and standardized from papers listed in Table S2[30–56].

ACKNOWLEDGMENTS

The authors thank Aaron Kollasch, Sarrah Rose Mikhail Leung, 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 S2. 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).

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] Aarushi Mehrotra, Noor Youssef, Sarah Gurev, and Debora Susan Marks. Forecasting h1n1 influenza pandemic and seasonal evolution. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] Phil Huss, Kristopher Kieft, Anthony Meger, Kyle Nishikawa, Karthik Anantharaman, and Srivatsan Raman. Engineering bacteriophages through deep mining of metagenomic motifs. *Science Advances*, 11(16):eadt6432, 2025.

[59] Josephine ES Strange, Pimlapas Leekitcharoenphon, Frederik Duus Møller, and Frank M Aarestrup. Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. *Scientific Reports*, 11(1):1600, 2021.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

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

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

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] Dianzhuo Wang, Marian Huot, Zechen Zhang, Kaiyi Jiang, Eugene I Shakhnovich, and Kevin M Esvelt. Without safeguards, ai-biology integration risks accelerating future pandemics.

[89] 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.

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

[91] 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.

[92] ResponsibleAIxBiodesign. Community values, guiding principles, and commitments for the responsible development of ai for protein design. *https://responsiblebiodesign.ai/*, 2023.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

A METHODS

A.1 VIRAL DEEP MUTATION SCANS

We curated and standardized a set of 45 viral DMSs from 25 papers[30–56], more than doubling the number of viral datasets in previous benchmarks[22]. We searched for all viral deep mutational scans, focusing here on single substitution mutations and phenotypes that are relevant for pandemic-preparedness, measuring expression, host receptor binding, replication, or infectivity. Several DMSs (including from ProteinGym) were excluded from this benchmark depending on the assayed phenotype, for example drug inhibition assays, viral escape, or difficulties with the data, but may be included in the future. The viruses included are relevant for vaccine design (e.g., SARS-CoV-2 and other sarbecoviruses, seasonal and pandemic Flu, and pandemic-threat Lassa and Nipah viruses) as well as for viral vector design (e.g., AAV).

A.2 SEQUENCE AND STRUCTURE DATASETS

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

UniRef100, is a database of a non-redundant protein sequence database containing all UniProt Knowledgebase records plus selected UniParc records[57]. 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)[57], Metaclust[70], and the Soil and Marine Eukaryotic Reference Catalog[61]. MGnify includes proteins from metagenomic assemblies[62]. AlphaFold databse (AF2DB) is comprised of 40 million AlphaFold predicted structures for UniRef50 sequences, which notably excludes all viral proteins.

A.3 ALIGNMENT-BASED MODELS

A.3.1 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) [5]. This involved five search iterations of the profile HMM homology search tool JackHMMER[94] 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 S3), 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) [6]) 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 instead of 80% sequence identity[2,6,7] 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. See Confidence Metrics below for more details on sequence clustering and alignment selection.

A.3.2 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) [5].

A.3.3 EVmutation

To predict the effects of mutations that explicitly captures pairwise residue dependencies between positions, we used EVmutation as implemented in Hopf et al. (2017) [5].

A.3.4 EVE

To predict the effects of mutations capturing high-order dependencies between positions, we used EVE, a Bayesian VAE model architecture, as implemented in Frazer et al. (2021) [6]. 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.3.5 Imputing missing data

For accurate comparisons across the same sets of mutations in DMSs, we impute missing model values with the mean value across the target protein.

A.4 Protein language models

A.4.1 Tranception

Tranception[11] 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[22].

A.4.2 ESM-1v

ESM-1v[10] 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[22] to handle sequences that are longer than the model context window (i.e., 1023 amino acids).

A.4.3 SaProt

SaProt[19] introduces a structure-aware vocabulary, into protein language modeling by training on Foldseek[77] 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[14] but expands the embedding layer to encompasses 441 structurally-aware tokens instead of the original 20 amino acid residue tokens. We use both SaProt-650M-AF2, trained on approximately 40 million AF2 sequences/structures (from UniRef50) which notably excludes all viral proteins, and SaProt-650M-PDB, which continuously pre-trains the SaProt-650M-AF2 model on the PDB.

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

A.5 Mutation effect scoring

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

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

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

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

For Tranception[11], an autoregressive model, the likelihood of x factorizes via the chain rule and is calculated as:

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

In the masked language model setting, for ESM-1v[10] and SaProt[19], 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[6], 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:

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

A.6 EVEREST - Evolutionary Variant Effect prediction with Reliability ESTimation

A.6.1 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)[22]. Following prior work[2,5], 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[5].

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

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) [5], 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[22].

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:

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

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). 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[94],

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.

A.6.2 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 [85].

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 [19].

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 new 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.

A.6.3 RELIABILITY THRESHOLDS

We choose an optimal reliability threshold that maximizes the average sensitivity and specificity 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.

A.7 HYBRID MODELS

A.7.1 SAPROT-EVE

We create a new hybrid model, SaProt-EVE that combines the advantages of alignment-based, family-specific models with 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. As part of EVEREST, we have defined confidence metrics for both alignment-based models and PLMs, which we provide along with mutation effect mutations. Here we use EVE for the alignment-based model (with a homology sequence diversity-based confidence metric) and SaProt-PDB (with an intrinsic pseudo-perplexity-based model confidence metric) for the PLM, but this method can be adapted to other models with comparable confidence calculations.

A.7.2 VESPA

VESPA[29] combines the embeddings from ProtT5[60] with a per-residue conservation prediction and supplements with BLOSUM substitution scores[28]. ProtT5 uses a T5 architecture which uses an encoder and decoder and was first trained on BFD and then finetuned on UniRef50.

A.7.3 TRANCEPTION WITH MSA RETRIEVAL

Tranception[11] 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[22]. Sequences are re-weighted as in Hopf et al. (2017) [5]. 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.7.4 TRANCEPTEVE

TranceptEVE[16] is a hybrid method that combines Tranception[11] with an MSA-based EVE log prior[6]. 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 PROTEINGYM

ProteinGym is a collection of benchmarks for comparing the ability of models to predict the effects of protein mutations[22]. 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[22]. Multiple sequence alignments used in ProteinGym were created for each assay by performing five search iterations of the profile HMM homology search tool JackHMMER[94] on the UniRef100 database of non-redundant proteins[57]. 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[95], detailed in Notin et al. (2022) [11].

B SUPPLEMENTARY TABLES

Model Type	Model	Avg Spearman	Viral Avg Spearman
Alignment-based model	GEMME	0.455	0.469
	EVE (ensemble)	0.439	0.428
	EVE (single)	0.433	0.424
	DeepSequence (ensemble)	0.419	0.344
	DeepSequence (single)	0.407	0.323
	EVmutation	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
	RITA XL	0.372	0.402
	CARP (640M)	0.368	0.273
	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

Supplementary Table 1: Model performance previously available in ProteinGym[22]. 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.

ID	Virus	Protein	Paper
Fitness			
LASSA_GP_Carr	Lassa	GP	Carr et al. (2024) [56]
SARS2_MRPO_Flynn	SARS-CoV-2	MRPO	Flynn et al. (2022) [53]
SARS2_BA1_SPIKE_Dadonaite	SARS-CoV-2 BA.1	SPIKE	Dadonaite et al. (2023) [51]
SARS2_DELTA_SPIKE_Dadonaite	SARS-CoV-2 Delta	SPIKE	Dadonaite et al. (2023) [51]
DENV_POLG_Suphatrakul	Dengue virus	POLG	Suphatrakul et al. (2023) [33]
IAV_H1_HA_Doud	Influenza A virus H1N1	H1	Doud and Bloom (2016) [39]
IAV_H1_HA_Wu	Influenza A virus H1N1	H1	Wu et al. (2014) [40]
IAV_PB2_Soh	Influenza A virus H1N1	PB2	Soh et al. (2019) [46]
IAV_H3_HA_Lee	Influenza A virus H3N2	H3	Lee et al. (2018) [41]
IAV_H5_HA_Dadonaite	Influenza A virus H5N1	H5	Dadonaite et al. (2024) [42]
IAV_H3_NP_Doud	Influenza A virus H1N1	NP	Doud et al. (2015) [43]
IAV_H1_NP_Doud	Influenza A virus H1N1	NP	Doud et al. (2015) [43]
IAV_PA_Wu	Influenza A virus H1N1	PA	Wu et al. (2015) [45]
IAV_RDRP_Li	Influenza A virus H1N1	RDRP	Li et al. (2023) [47]
NIPAH_F_Larsen	Nipah Henipavirus	F	Larsen et al. (2025) [55]
AAV2_CAPSD_Sinai	AAV2	CAPSD	Sinai et al. (2021) [30]
CVB3_2A_Alvarez	Coxsackievirus B3	2A	Álvarez-Rodríguez et al. (2024) [32]
CVB3_2B_Alvarez	Coxsackievirus B3	2B	Álvarez-Rodríguez et al. (2024) [32]
CVB3_2C_Alvarez	Coxsackievirus B3	2C	Álvarez-Rodríguez et al. (2024) [32]
CVB3_3A_Alvarez	Coxsackievirus B3	3A	Álvarez-Rodríguez et al. (2024) [32]
CVB3_3B_Alvarez	Coxsackievirus B3	3B	Álvarez-Rodríguez et al. (2024) [32]
CVB3_3C_Alvarez	Coxsackievirus B3	3C	Álvarez-Rodríguez et al. (2024) [32]
CVB3_3D_Alvarez	Coxsackievirus B3	3D	Álvarez-Rodríguez et al. (2024) [32]
CVB3_VP1_Alvarez	Coxsackievirus B3	VP1	Álvarez-Rodríguez et al. (2024) [32]
CVB3_VP3_Alvarez	Coxsackievirus B3	VP3	Álvarez-Rodríguez et al. (2024) [32]
EV_CAPSD_Bakhache	Enterovirus A	CAPSD	Bakhache et al. (2024) [35]
EV_REP_Bakhache	Enterovirus A	REP	Bakhache et al. (2024) [35]
CVB3_POLG_Mattenberger	Coxsackievirus B3	POLG	Mattenberger et al. (2021) [31]
HIV1_BF520_ENV_Haddox	HIV-1 BF520	ENV	Haddox et al. (2018) [36]
HIV1_HV1B9_ENV_DuenasDecamp	HIV-1 HV1B9	ENV	Duenas-Decamp et al. (2016) [37]
HIV1_BG505_ENV_Haddox	HIV-1 BG505	ENV	Haddox et al. (2018) [36]
Stability			
PESV_POLG_Tsuboyama	Porcine enteric sapovirus	POLG	Tsuboyama et al. (2023) [34]
BPP22_COAT_Tsuboyama	Salmonella phage P22	COAT	Tsuboyama et al. (2023) [34]
BPT7_RPOL_Tsuboyama	Enterobacteria phage T7	RPOL	Tsuboyama et al. (2023) [34]
LAMBDA_HCP_Tsuboyama	Enterobacteria phage lambda	HCP	Tsuboyama et al. (2023) [34]
BP434_RPC1_Tsuboyama	Enterobacteria phage 434	RPC1	Tsuboyama et al. (2023) [34]
Expression			
SARS2_PRD0038_RBD_Starr	Bat coronavirus PRD0038	RBD	Starr (2024) [49]
RmYN02_RBD_Starr	Bat coronavirus RmYN02	RBD	Starr (2024) [49]
RsYN04_RBD_Starr	Bat coronavirus RsYN04	RBD	Starr (2024) [49]
SARS2_XBB15_RBD_Taylor	SARS-CoV-2 XBB.1.5	RBD	Taylor and Starr (2023) [52]
SARS2_RBD_Starr_expression	SARS-CoV-2	RBD	Starr et al. (2020) [50]

Table S2 continued on next page

Table S2 continued from previous page

ID	Virus	Protein	Paper
Binding			
SARS2_RBD_Starr_binding	SARS-CoV-2	RBD	Starr et al. (2020) [50]
IAV_NA_Jiang	Influenza A virus H3N2	NA	Jiang et al. (2016) [44]
Activity			
SARS2_PLPRO_Wu_activity	SARS-CoV-2	PLPRO	Wu et al. (2024) [54]
Abundance			
SARS2_PLPRO_Wu_abundance	SARS-CoV-2	PLPRO	Wu et al. (2024) [54]

Supplementary Table 2: Deep Mutational Scanning studies of 45 viral proteins across 25 papers, with fitness, stability, expression, and other assays.

Bitscore	UniRef90	UniRef100	UniRef+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 3: Number of successful models given the equal memory constraints at different bitscores across UniRef90, UniRef100, and UniRef+BFD datasets. Failed models with high numbers of sequences would mostly be excluded by selected Neff cutoff.

	Model	All Viral DMS	Fitness Only	Fitness+Antigens Only
Alignment	PSSM	0.38	0.41	0.37
	EVmutation	0.38	0.39	0.36
	EVE	**0.44**	**0.46**	**0.43**
PLM	ESM-1v - Uniref90	0.19	0.19	0.31
	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**

Supplementary Table 4: Model Performance (Spearman ρ) on new viral 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.

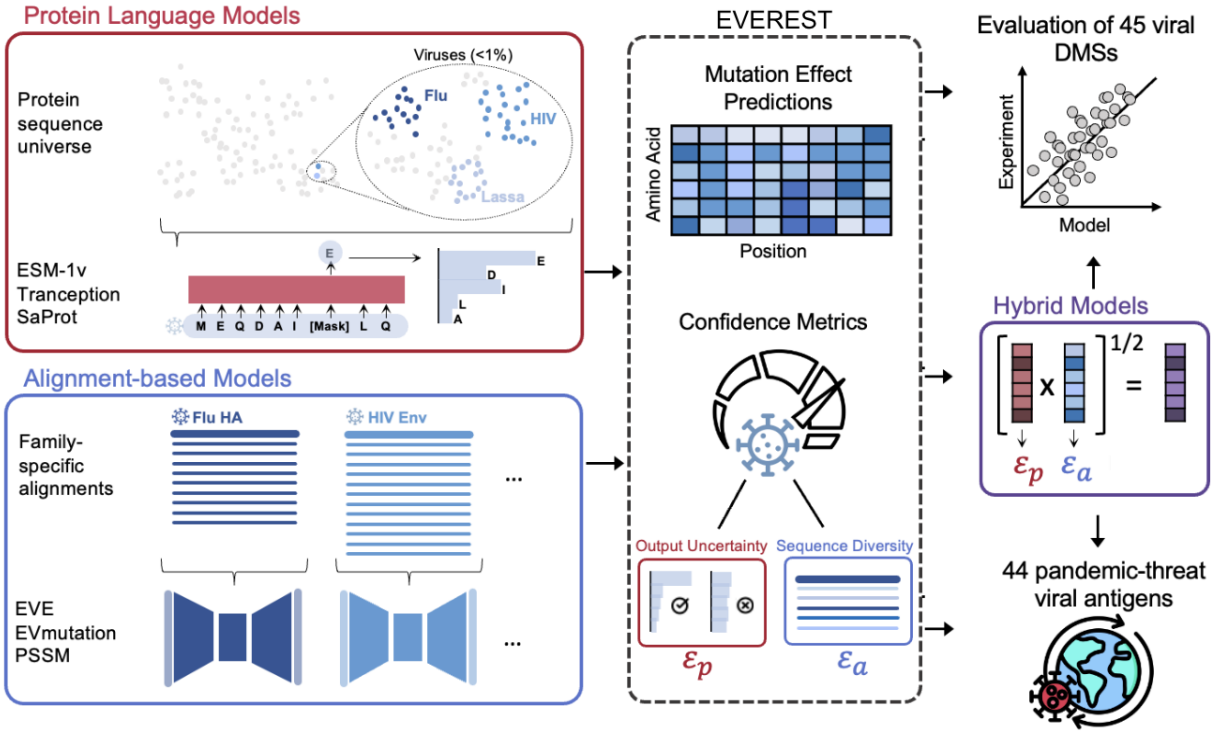
Model	Parameters	Data Size	Objective	Cutoff	Context	All Taxa (ρ)	Virus (ρ)	Reference
Uniref100								
RITA XL	1.2B	280M	AR	2022	1024	0.372	0.402	Hesslow et al. (2022) [12]
Tranception L	700M	250M	AR	2022	1024	0.374	0.395	Notin et al. (2022) [11]
RITA L	680M	280M	AR	2022	1024	0.365	0.391	Hesslow et al. (2022) [12]
RITA M	300M	280M	AR	2022	1024	0.350	0.385	Hesslow et al. (2022) [12]
RITA S	85M	280M	AR	2022	1024	0.304	0.358	Hesslow et al. (2022) [12]
Tranception M	300M	250M	AR	2022	1024	0.348	0.349	Notin et al. (2022) [11]
Tranception S	85M	250M	AR	2022	1024	0.303	0.315	Notin et al. (2022) [11]
Uniref90+BFD30								
Progen2 XL	6.4B	230M	AR	2022	1024	0.391	0.391	Nijkamp et al. (2023) [13]
Progen2 L	2.7B	230M	AR	2022	1024	0.380	0.333	Nijkamp et al. (2023) [13]
Progen2 M	764M	230M	AR	2022	1024	0.379	0.342	Nijkamp et al. (2023) [13]
Progen2 Base	764M	230M	AR	2022	2048	0.378	0.328	Nijkamp et al. (2023) [13]
Progen2 S	151M	230M	AR	2022	1024	0.336	0.285	Nijkamp et al. (2023) [13]
Uniref90								
ESM-2 (650M)	650M	65M	MLM	2022	1024	0.414	0.238	Lin et al. (2023) [14]
ESM-1v (ens.)	650M	98M	MLM	2020	1024	0.407	0.279	Meier et al. (2021) [10]
ESM-2 (3B)	3B	65M	MLM	2022	1024	0.406	0.274	Lin et al. (2023) [14]
ESM-2 (15B)	15B	65M	MLM	2022	1024	0.401	0.313	Lin et al. (2023) [14]
ESM-2 (150M)	150M	65M	MLM	2022	1024	0.387	0.137	Lin et al. (2023) [14]
ESM-1v	650M	98M	MLM	2020	1024	0.374	0.245	Meier et al. (2021) [10]
ESM-2 (35M)	35M	65M	MLM	2022	1024	0.321	0.102	Lin et al. (2023) [14]
ESM-2 (8M)	8M	65M	MLM	2022	1024	0.226	0.078	Lin et al. (2023) [14]
Uniref50								
ESM-1b	650M	27M	MLM	2018	1024	0.394	0.241	Rives et al. (2021) [96]
CARP (640M)	640M	40M	MLM	2023	4096	0.368	0.273	Yang et al. (2024) [15]
CARP (76M)	76M	40M	MLM	2023	4096	0.328	0.150	Yang et al. (2024) [15]
CARP (38M)	38M	40M	MLM	2023	4096	0.279	0.125	Yang et al. (2024) [15]
ProtGPT2	738M	40M	AR	2022	1024	0.188	0.141	Ferruz et al. (2022) [97]
Unirep	18.2M	24M	AR	2018	280	0.190	0.057	Alley et al. (2019) [98]
CARP (600K)	600K	40M	MLM	2023	4096	0.106	0.042	Yang et al. (2024) [15]

Supplementary Table 5: Single-sequence Protein Language Model performance of models in ProteinGym[22] 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[60]). 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.

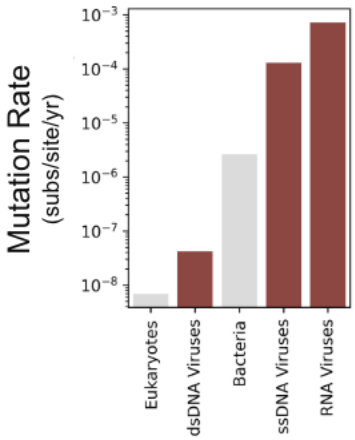
ID	Family	Structural Class	MSA confidence	PLM confidence
Junin_GPC	Arena	Class I Fusion	5	0.61
Lassa_GPC	Arena	Class I Fusion	259	0.55
Lujo_GPC	Arena	Class I Fusion	1	0.60
MERS_Spike	Corona	Class I Fusion	8	0.67
SARS2_Spike	Corona	Class I Fusion	157	0.67
Ebola_GP	Filo	Class I Fusion	85	0.70
Marburg_GP	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.

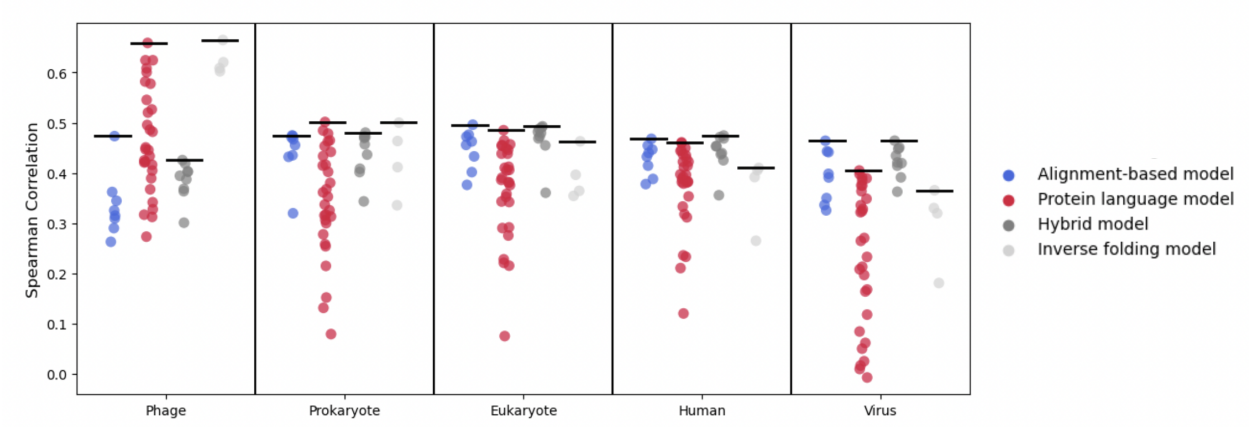
C SUPPLEMENTARY FIGURES



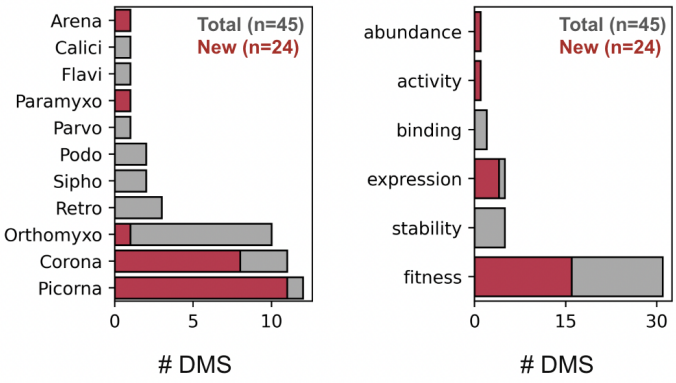
Supplementary Figure 1: Modeling overview of alignment-based and protein language models and analysis details.



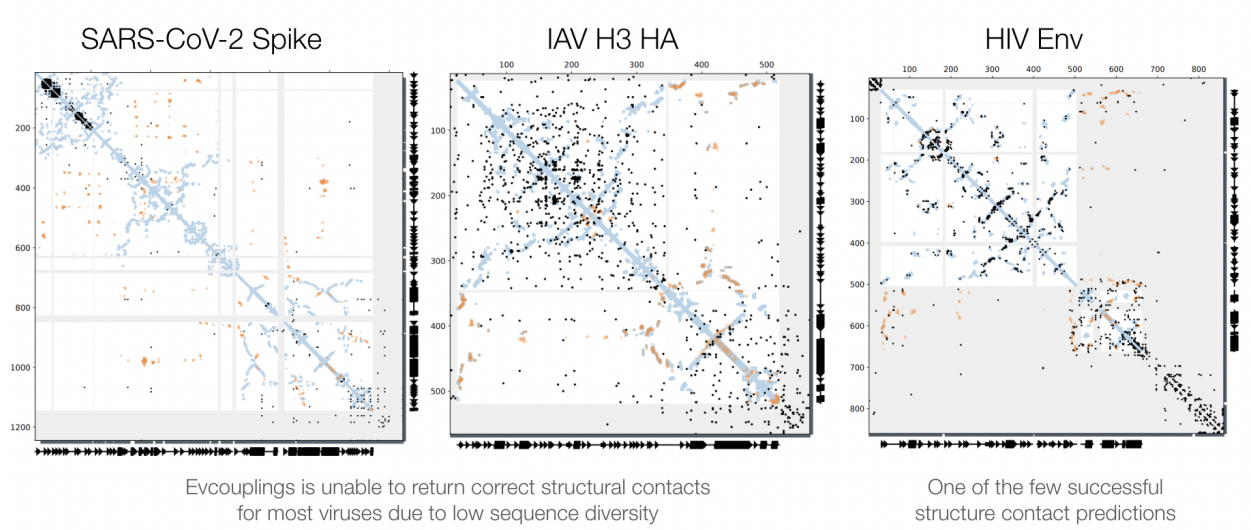
Supplementary Figure 2: Geometric average mutation rates (substitutions per site per year) from Linz et al.[26] Double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA).



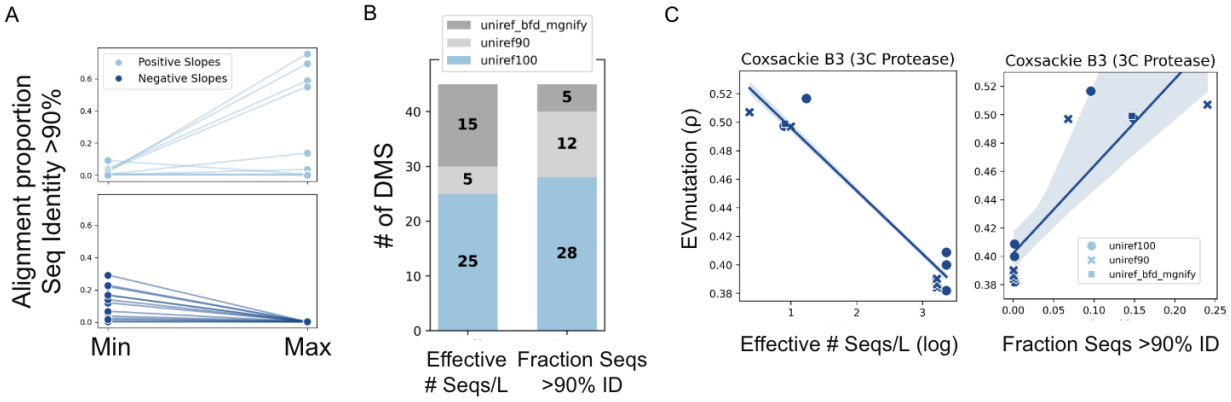
Supplementary Figure 3: Protein language models have higher performance across all taxa except for eukaryotic viruses. Each point represents the average correlation across 218 DMS assays for each of the 50 fitness models in ProteinGym, labeled with their model type. Viral performance is on the dataset provided by ProteinGym which is a more limited set of DMSs than available in this paper. Lines denote maximum Spearman correlation per model type for each taxa. For details of included models see Supp Table S1. Notably, this benchmark excludes SaProt-PDB.



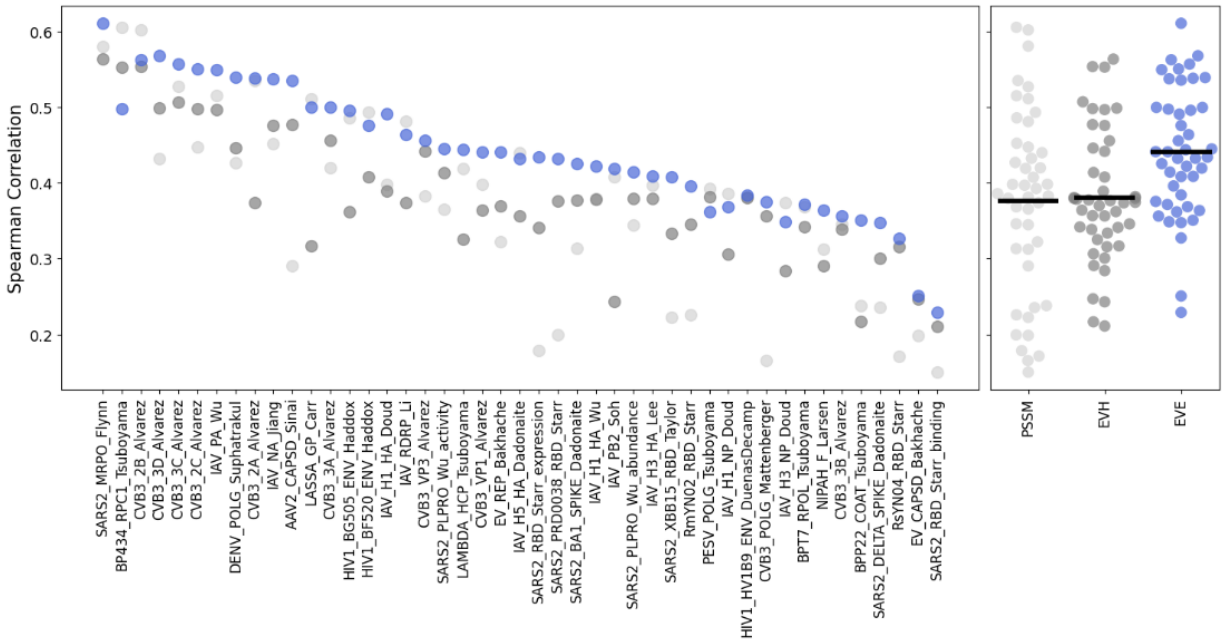
Supplementary Figure 4: Our viral-specific benchmarking dataset approximately doubles the number of viral assays included in prior benchmarks, providing a more comprehensive evaluation framework. Viral families and assay types with deep mutational scans in the new viral benchmark, counts shown for newly added scans or scans previously in ProteinGym



Supplementary Figure 5: Selected 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 with a DMS where EVmutation successfully predict structure contacts, while ability to predict structural contacts is a strong predictor of model performance for other taxa.



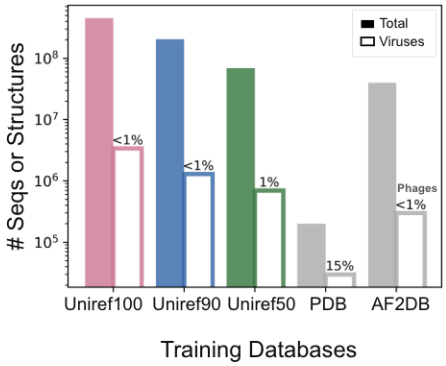
Supplementary Figure 6: A. Proteins that benefited from deeper alignments had a higher proportion of closely related sequences, while those that performed worse often contained more distantly related sequences. Deeper alignments improve (or reduce) performance when a larger (or smaller) proportion of sequences are within 90% identity of the target sequence. B. Fraction of selected alignments using each database for old and new alignment selection heuristic. C. Using fraction sequences greater than 90% identity as a confidence metric instead of Neff/L leads to a positive slope with Spearman performance for coxsackievirus protease.



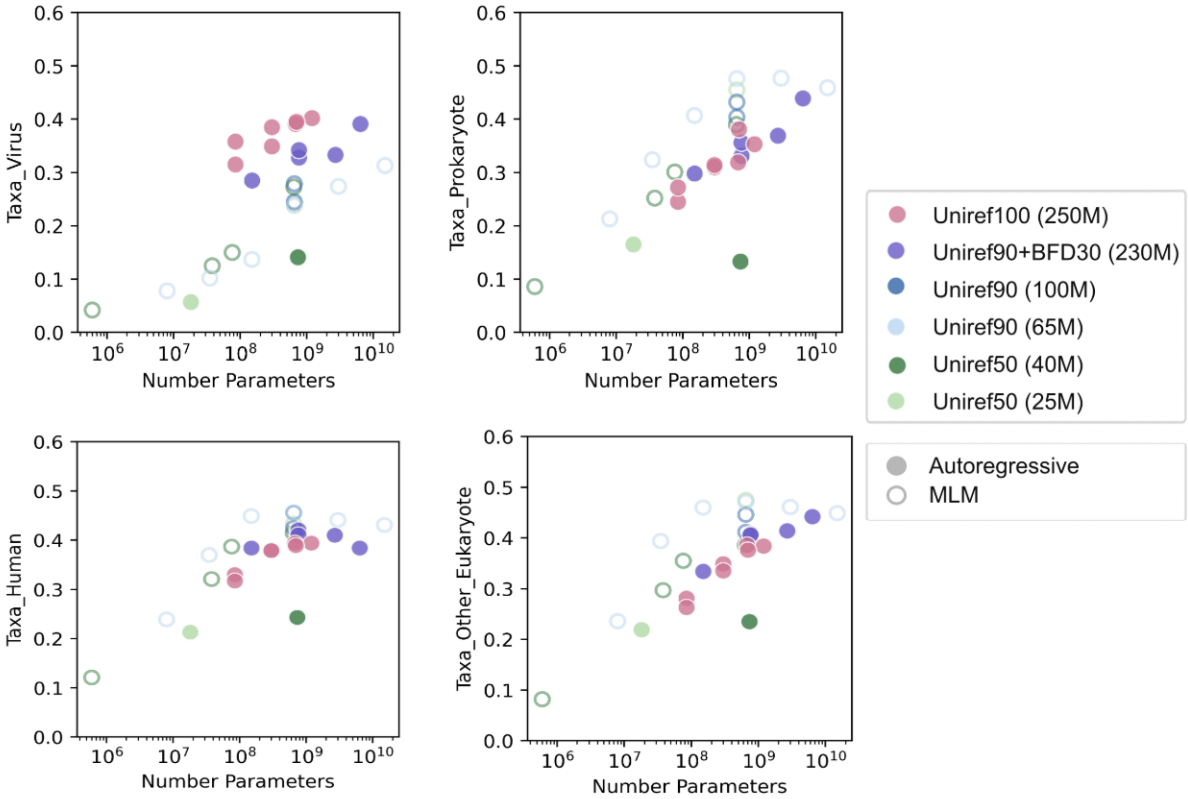
Supplementary Figure 7: Spearman rank correlation between alignment-based model score and DMS score for each curated viral protein. EVE outperforms other alignment-based models on almost all proteins.

A

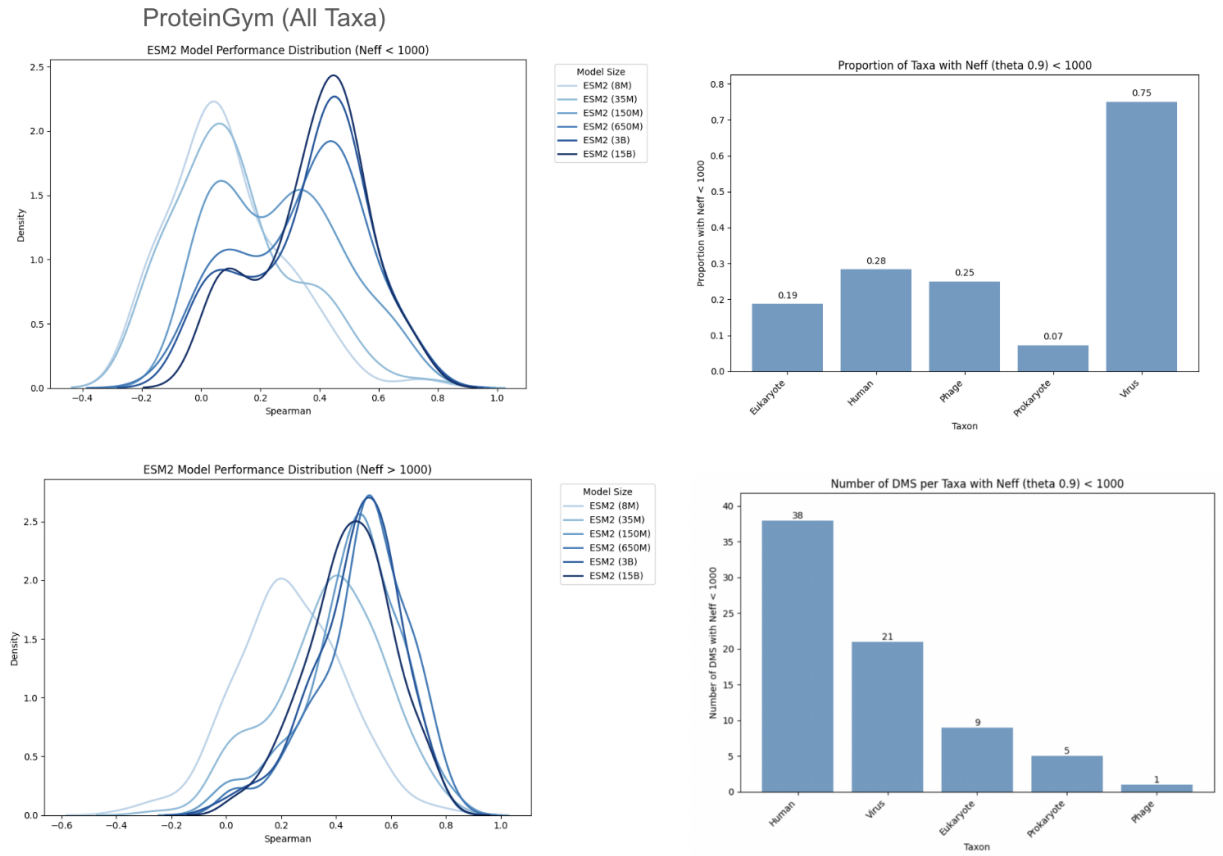
	UniRef100 (Tranception)	UniRef90 (ESM-1v)	UniRef50 (CARP)	PDB (SaProt-PDB)	AF2DB (SaProt-PDB/AF)
Total (millions)	453	205	69	0.2	40
Viral (millions)	3.4	1.3	0.7	0.03	Eukaryotic: 0 Phage: <0.3
Influenza HA	75,733	2,116	268	<800	0
Arenavirus GPC	1,381	228	46	<60	0
Paramyxo-virus F	6,280	607	147	<80	0



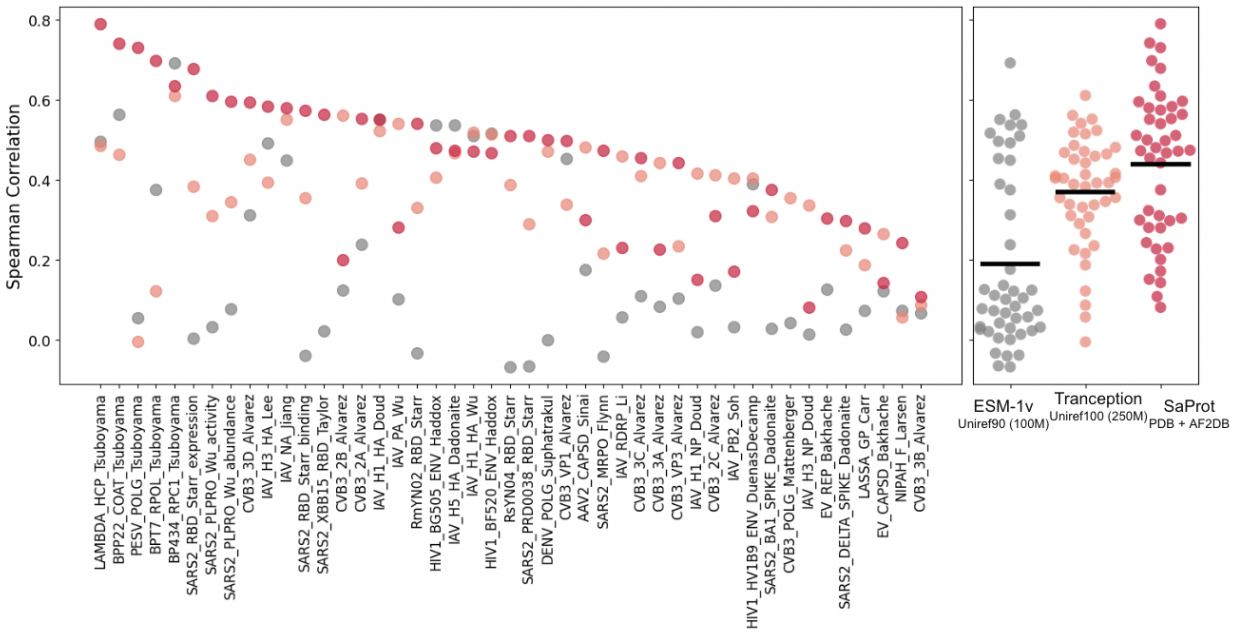
B



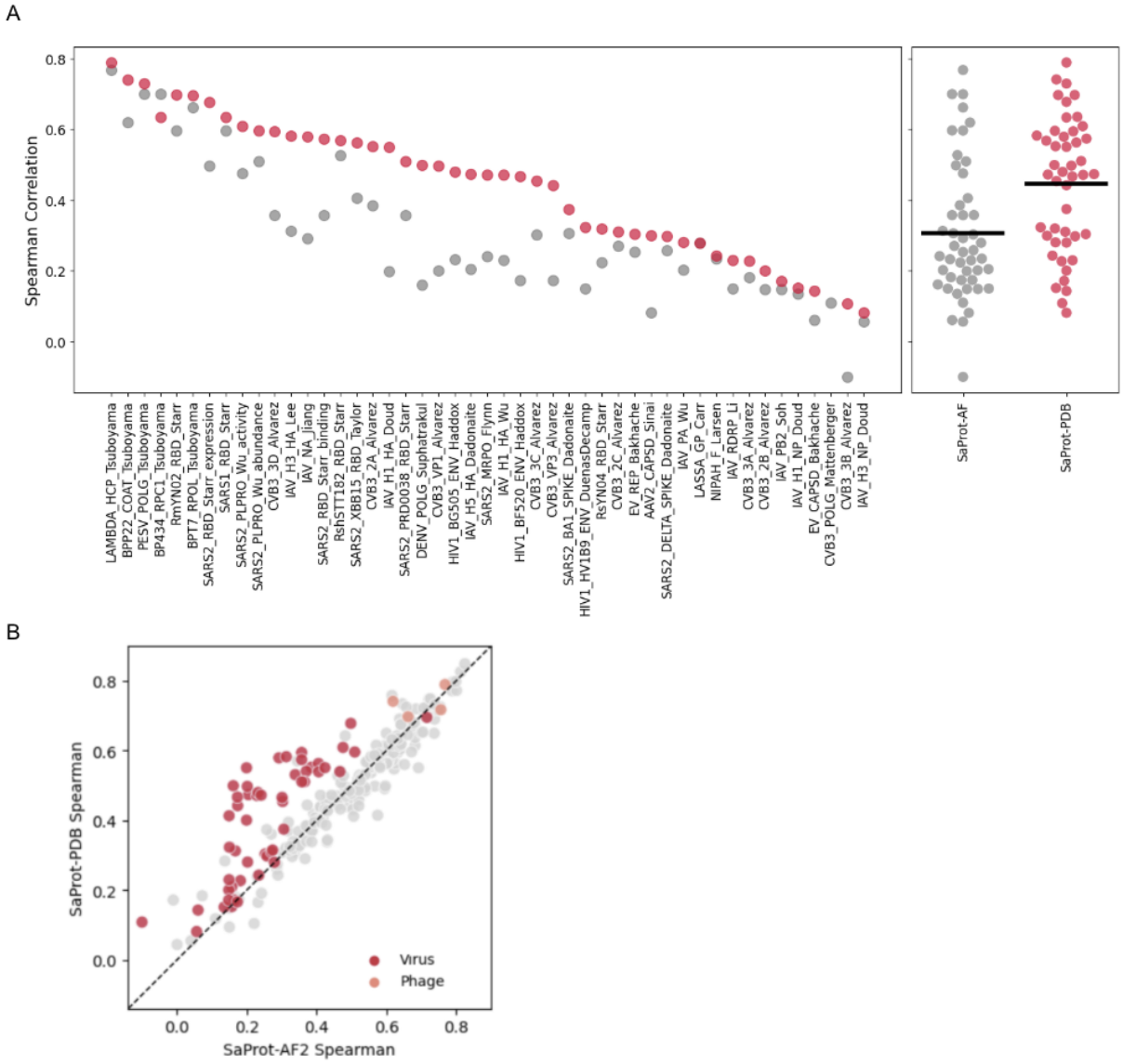
Supplementary Figure 8: A. Total number of sequences in UniRef100, UniRef90, and UniRef50, and structures in PDB and AlphaFold2 DB (AF2DB/Uniref 50), shown overall (solid) and for viral proteins (outline), and in table. B.Scaling number of PLM paramaters leads to higher average Spearman correlation with viral proteins (top left), without the plateau in top-performing model performance seen in general fitness effect prediction of other taxa, especially for MLMs. Models annotated by training dataset and by training objective. The best PLMs for viruses are autoregressive, while there is little distinction in MLM or autoregressive performance for general DMS prediction.



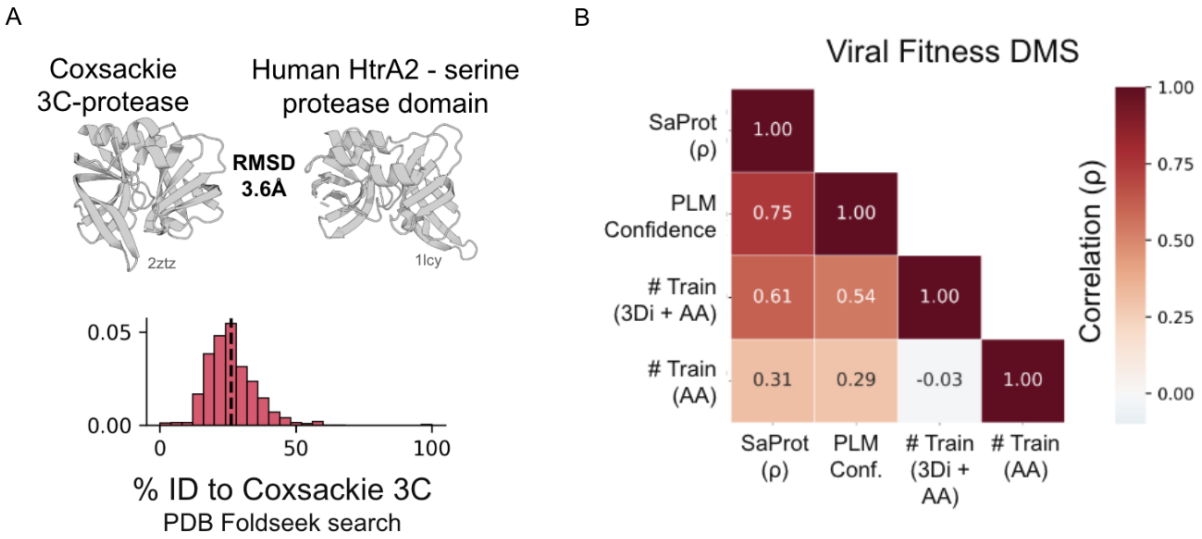
Supplementary Figure 9: The masked protein language model ESM-2 benefits from increased number of parameters especially when there is a low number of related sequences in the PLM training data (calculated as number of effective sequences in a JackHMMER alignment of Uniref100 with clustering at theta 90%, since ESM-2 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 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.



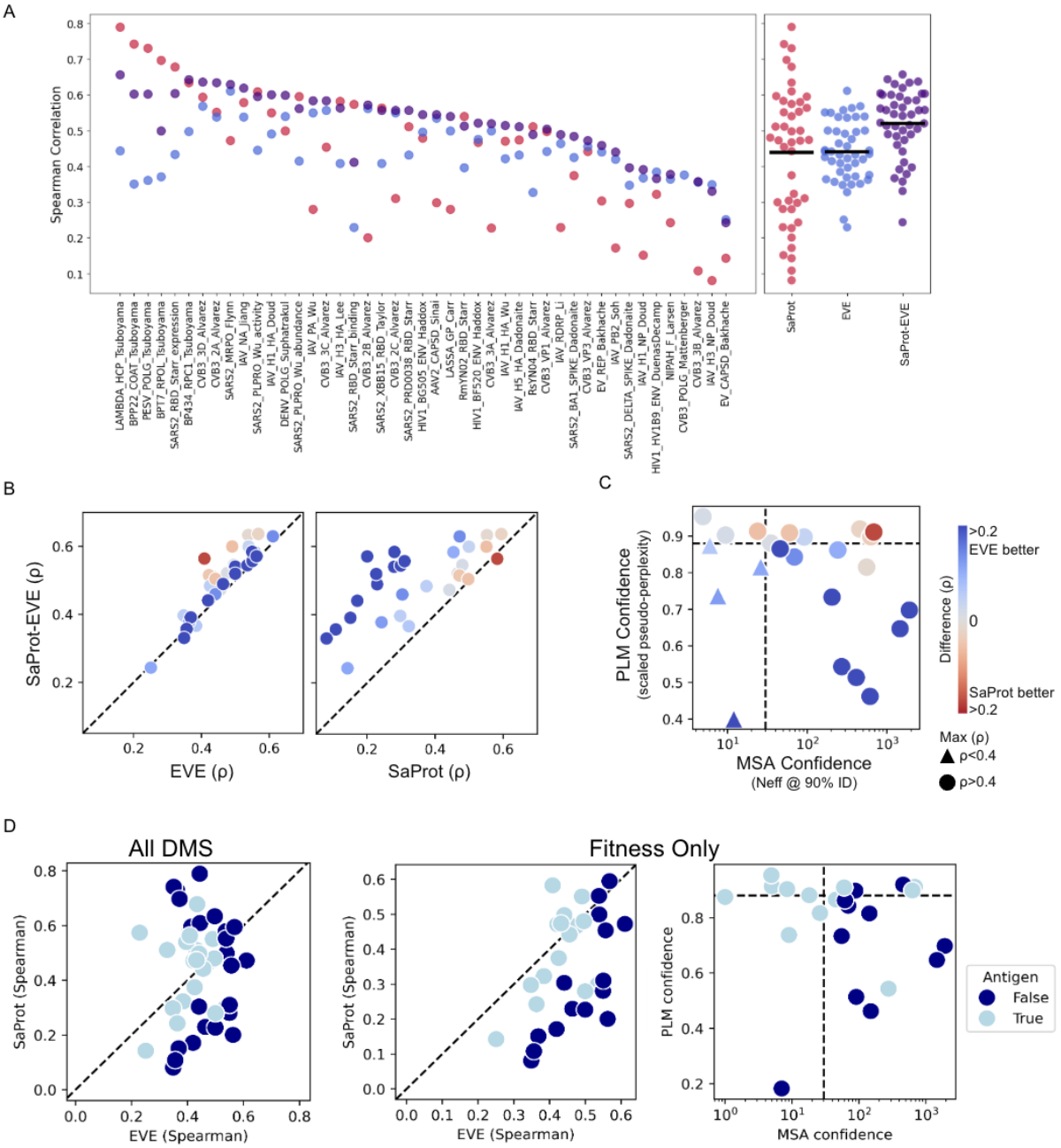
Supplementary Figure 10: Spearman rank correlation between PLM model score and DMS score for each curated viral protein.



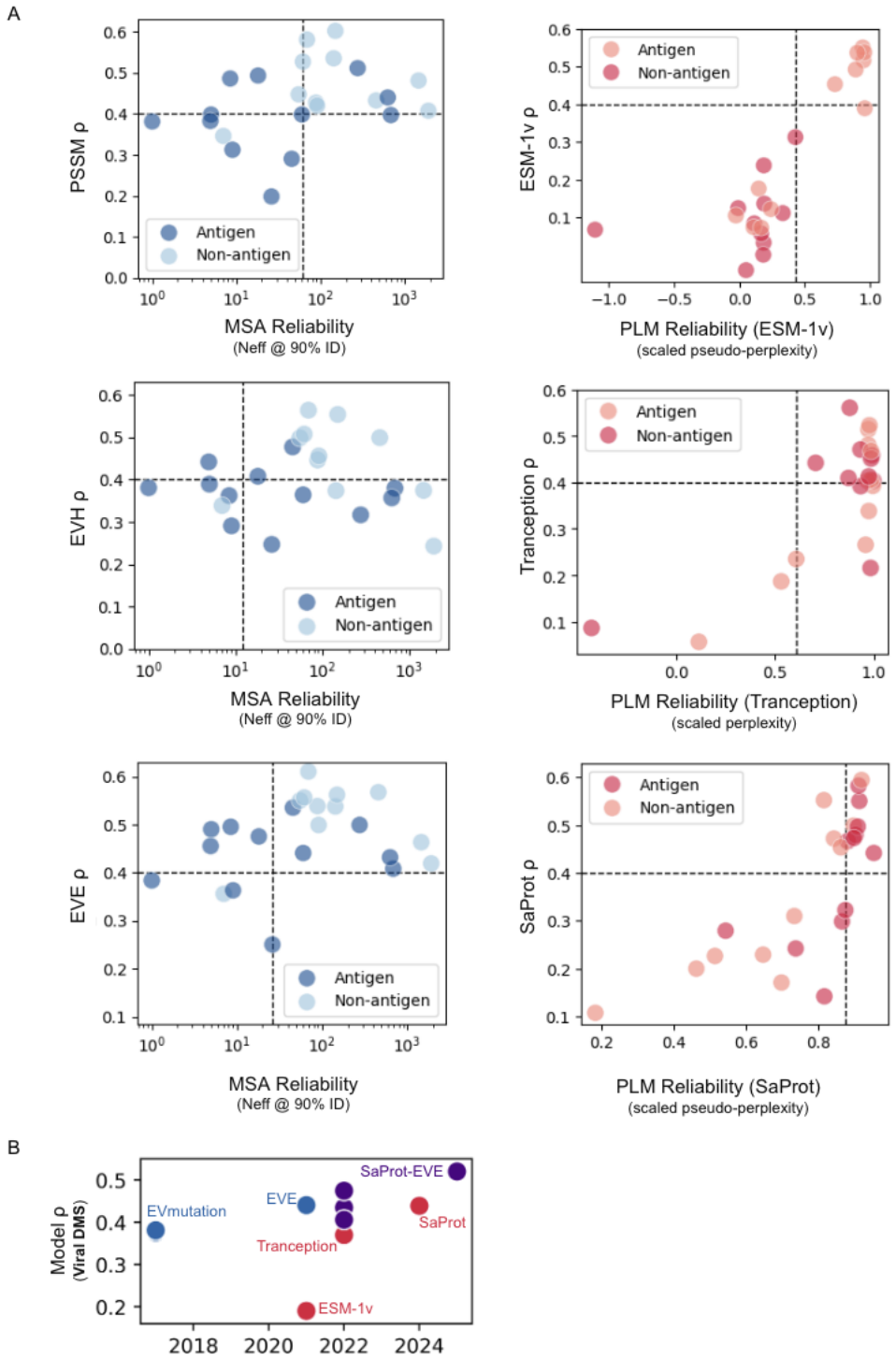
Supplementary Figure 11: A. Comparison of the Spearman rank correlation of the SaProt-PDB model to SaProt-AF2 model for all viral DMSs. SaProt-PDB performs best on almost all viruses. SaProt-PDB has an advantage over SaProt-AF2 especially for viruses like SARS-CoV-2 and Influenza that are vastly over-represented in the PDB. B. 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.



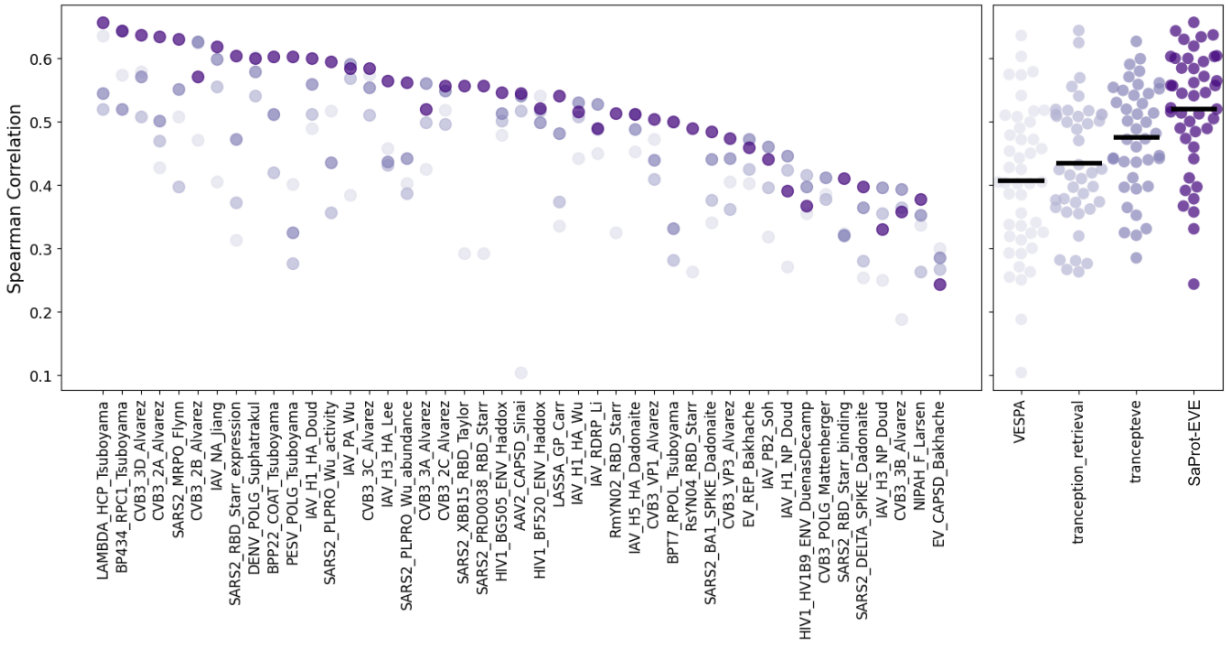
Supplementary Figure 12: A. Viral cysteine proteases like the coxsackievirus 3C protease are structurally homologous to the trypsin-like serine protease family present across the kingdoms of life, despite significant sequence divergence[78]. Whether this is an example of true homology (divergence from a common ancestor) or convergent evolution on a common biochemistry is a matter of debate[99]. Coxsackievirus 3C protease structure (2ztz chain A) and Human HtrA2 serine protease domain (1lcy, selected residues 1-210[100]) 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. 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. Interestingly, there was no significant correlation between the number of training examples retrieved by sequence similarity alone and those identified via combined structural and sequence homology ($\rho = -0.03$). This suggests that structural tokens introduce a distinct and complementary set of informative training points. However, this relationship may be partly confounded by the redundancy in PDB entries for well-studied viral proteins, as well as database-specific conventions—for instance, the PDB often separates glycoprotein subunits into distinct chains, while sequence databases treat them as a single entry.



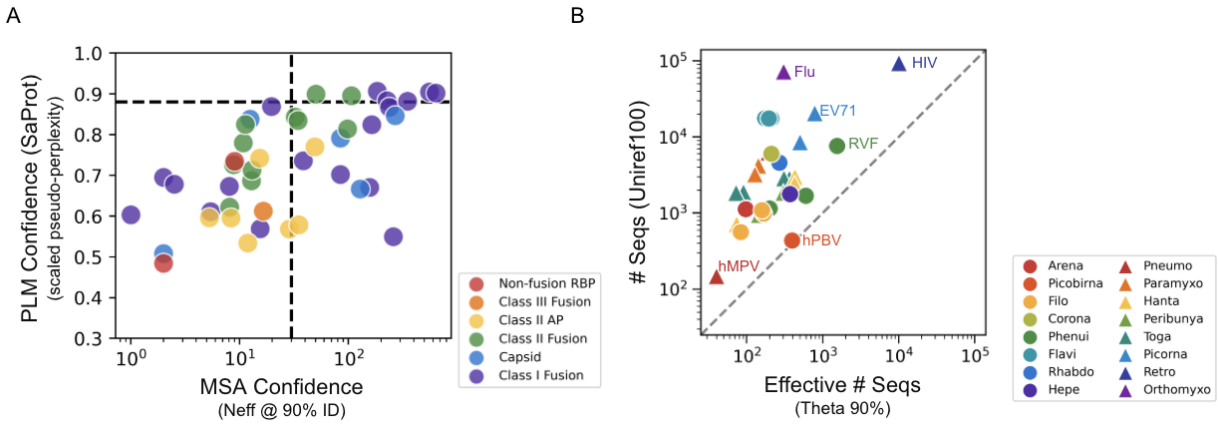
Supplementary Figure 13: A. Spearman rank correlation between best alignment based EVE or PLM model SaProt score (components of SaProt-EVE) or SaProt-EVE and DMS score for each curated viral protein. B. SaProt-EVE Spearman correlations meet or exceed EVE or SaProt correlations for all viral fitness DMSs. C. Viral proteins 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). D. 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.



Supplementary Figure 14: A.Reliability scores predict performance for a given viral protein for PSSM, EVE, ESM-1v, and SaProt models, though not for EVmutation 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.



Supplementary Figure 15: Spearman rank correlation between each hybrid model score and DMS score for each curated viral protein.



Supplementary Figure 16: 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.