SEQUENCE-BASED PROTEIN MODELS FOR THE PREDICTION OF MUTATIONS ACROSS PRIORITY VIRUSES

Sarah Gurev[*†]
EECS
Massachusetts Institute of Technology

Noor Youssef[*†]
Systems Biology
Harvard Medical School

Navami Jain
Bioinformatics and Integrative Genomics
Harvard Medical School

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

ABSTRACT

Viruses pose a significant threat to human health. Advances in machine learning for predicting mutation effects have enhanced viral surveillance and enabled the proactive design of vaccines and therapeutics, but the accuracy of these methods across priority viruses remain unclear. We perform the first large-scale modeling across 40 WHO priority pandemic-threat pathogens, many of which are under-surveilled, discovering that most have sufficient sequence or structural information for effective modeling, highlighting the potential for using these approaches in pandemic preparedness. To understand the limits of current modeling capabilities for viruses, we curate 51 standardized viral deep mutational scanning assays to systematically evaluate the performance of three alignment-based models, three Protein Language Models (PLMs), and two structure-aware PLMs with different training databases. We find marked differences in performance of these models on viruses relative to non-viral proteins. For viral proteins, we find alignment-based models perform on par with PLMs though with predictable differences in which model is better for a particular function or virus depending on data available. We define confidence metrics for both alignment-based models and PLMs that indicate when additional sequence or structural data may be needed for accurate predictions and to guide model selection in the absence of available data for evaluation. We use these metrics to inform the development a confidence-weighted hybrid model that builds on the strength of each approach, adapts to the quality of data available, and outperforms either of the best alignment or PLM models alone.

1 INTRODUCTION

Viral diseases pose a significant challenge to public health, with increasing human-animal interactions heightening the risk of novel spillover events. The rapid mutation rates of viruses further threaten the long-term efficacy of vaccines and therapeutics, necessitating robust predictive strategies for viral evolution. Recent advances in machine learning for predicting the functional effects of mutations have transformed the landscape of viral surveillance and therapeutic design. Such methods fall broadly into two categories: Alignment-based models and Protein Language Models (PLMs). Alignment-based models infer functional constraints from multiple sequence alignments (MSAs) of homologous sequences (Hopf et al., 2017; Frazer et al., 2021). PLMs use deep learning approaches, without relying on MSAs, to learn constraints from all available protein sequences (Meier et al., 2021; Marquet et al., 2022; Notin et al., 2022). Structurally-aware PLMs, such as SaProt, integrate both sequence and local structural information (Su et al., 2023). We perform the first large-scale modeling of protein fitness across pandemic-threat viruses. We evaluate model performance across all 40 2024 World Health Organization (WHO) priority and prototype RNA viral

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

pathogens across 21 viral families (WHO, 2024), identifying pathogens that are likely to benefit from additional sequence and structural data.

The best PLMs are beginning to outperform alignment-based models across proteins from diverse taxa (e.g., bacteria and eukaryotes). Yet, it is unclear if this trend holds for viruses since databases used for training PLMs often contain fewer and less diverse viral sequences compared to sequences from other species. Viral proteins have unique biophysical properties, such as lower stability, loosely packed cores, and high structural flexibility, which are hypothesized to facilitate immune evasion and host adaptation (Tokuriki et al., 2009). Alignment-based methods have been previously successful in predicting viral antibody escape and vaccine effectiveness (Thadani et al., 2023; Youssef et al., 2024). One challenge, however, is that many high-risk viral pathogens are understudied and have few homologous sequences available, potentially limiting the applicability of both alignment-based models and PLMs.

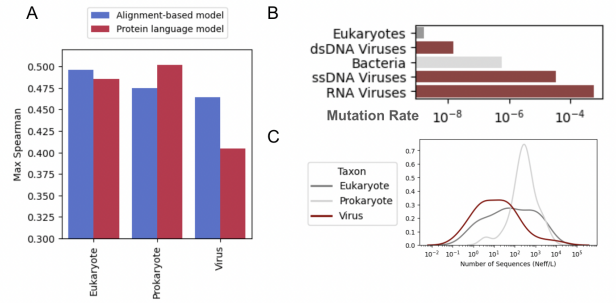
We present a comprehensive evaluation of mutation effect prediction models for viruses by curating a dataset of 51 standardized viral deep mutational scanning (DMS) assays–more than twice the number in existing benchmarks (Notin et al., 2023). We assess the performance of three alignment-based models (PSSM, EVmutation, EVE) with different epistatic assumptions, and three PLMs (ESM1v, Tranception, VESPA) and two structurally-aware PLMs (SaProt-AF and SaProt-PDB) with different training datasets. Our analysis reveals that sequence clustering hinders PLM performance on viruses, whereas incorporating local structural information can partially compensate for the limited number of viral sequences. Additionally, we show that increasing sequencing depth does not necessarily improve alignment-based model performance, rather optimizing alignments by prioritizing sequences with greater identity to the target protein yields better predictions. These findings challenge conventional assumptions about optimal modeling strategies and highlight the need for virus-specific adaptations in PLMs.

2 RESULTS

Using the ProteinGym benchmark (Notin et al., 2023), which evaluated 50+ protein fitness models across 250+ DMS assays, we found that the top-performing PLM achieved comparable or superior performance to the best alignment-based model across proteins from eukaryotic and prokaryotic organisms (Fig 1A). However, for viral proteins, the best alignment-based model (GEMME (Laine et al., 2019)) significantly outperformed the best PLM (VespaG (Marquet et al., 2022); Fig 1A, S1). The relative under-performance of PLMs for viral proteins may stem from unique evolutionary constraints of viral proteins, such as their higher mutation rates(Fig 1B), or limitations in the composition of PLM training datasets (Fig 1C,D). Most PLMs benefit generally from training on clustered subsets of UniRef, such as UniRef90 or UniRef50, to account for biased sequencing, but they contain disproportionately low numbers of viral sequences—0.6% and 1% respectively. For example, while UniRef100 contains over 6,000 paramyxovirus fusion protein sequences, less than 10% remain in UniRef90, despite UniRef90 being roughly half the size of UniRef100 (Fig 1D). This substantial loss of viral sequence diversity likely contributes to the lower performance of PLMs on viral proteins and suggests that incorporating additional viral-specific training data without as strict clustering could improve predictive accuracy. This evaluation informed the development of a viral-specific deep mutational scanning benchmark across 47 scans, with double the number of viral assays in ProteinGym, and with more in depth analysis of modeling considerations unique to viruses. Our goal in creating this benchmark is to identify the optimal models to use to forecast mutation effects on all WHO priority RNA viruses, and to make practical recommendations to improve predictions for these viruses.

2.1 ALIGNMENT-BASED MODELS

For each protein in our newly curated viral benchmark, we trained three alignment-based models (PSSM, EVmutation, and EVE, each with different epistatic assumptions) using MSAs generated from three databases (UniRef90, UniRef100, and UniRef100+BFD+Mgnify) at multiple bit scores. We found that increasing alignment depth improved performance for some proteins (positive correlation between depth and DMS agreement) but reduced performance for others (negative correlation; Fig 2A, S2)). 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



D

	UniRef100 (Tranception)	UniRef90 (ESM-1v)	UniRef50 (CARP)	PDB (SaProt-PDB)	AF2DB (SaProt-PDB/AF)
Total (millions)	453	205	69	0.5	200
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
Paramyxovirus F	6,280	607	147	<80	0

Figure 1: Understanding modeling considerations for unique viral features A. Best alignment-based model or PLM performance for eukaryotic, prokaryotic, and eukaryotic-infecting viral proteins from ProteinGym, measured as the average Spearman correlation against DMS assays. **B**. Geometric average mutation rates (substitutions per site per year) from Linz et al. (2014). **C**. Number of effective sequences (normalized by sequence length, clustered by 99% identity) in alignments from ProteinGym. **D**. Number of sequences in UniRef100, UniRef90, and UniRef50, and structures in the PDB and AF2DB, and for antigenic proteins of selected viruses or viral families.

(Fig 2B). This suggests that after a certain threshold, including additional sequences introduces noise or redundant information that can negatively impact performance, as mutations are seen on vastly different background sequences, that are likely under different functional constraints. This holds true even when considering the NDCG metric focused on top mutation ranking (Fig S3).

To generalize these findings to viral proteins without existing DMS data for evaluation, we propose a heuristic for selecting the optimal alignment: choosing the alignment with the highest number of effective sequences (Neff), the number of sequences after clustering at 99% identity, up to a threshold of $10^{3.5}$ (Fig S4)). Using this threshold, we found that EVE outperformed both PSSM and EVmutation across almost all viral proteins (Fig 2C, S5), consistent with previous findings that EVE's probabilistic approach capturing high-order epistasis offers better predictive accuracy (Frazer et al., 2021; Riesselman et al., 2018; Thadani et al., 2023). Additionally, when applying this alignment selection strategy to viral proteins, EVE performed slightly better on average for viral compared to non-viral proteins (for which alignments were selected using the protocol outlined in ProteinGym). These results highlight the benefit of using this Neff threshold, as this improvement is remarkable considering that almost none of the evaluated viral proteins have sufficient sequence diversity to accurately predict structural contacts (Fig S6), as can commonly be done for non-viral proteins. Lastly, we find that Neff normalized by the length of the protein (L) is an indicator of EVE performance across viral proteins, and can therefore be used as an estimate of model confidence (Fig 2D).

2.2 Protein Language Models

We likewise evaluated our newly compiled viral DMS datasets using three sequence-only PLMs—ESM-1v (trained on UniRef90), Tranception (trained on UniRef100 without MSA retrieval), and VESPA (trained on BFD and UniRef50)—and two structurally-aware PLMS—SaProt-AF2 (trained on the AlphaFold2 database), and SaProt-PDB (continues training on the PDB). PLMs were chosen for their relatively high performance on the ProteinGym dataset and to cover a range of training databases. PLMs trained on UniRef50 sequence alone, such as CARP (Yang et al., 2024), were excluded due to their predictably low performance on viruses (Table S1).

Our analysis revealed that sequence scale of the training dataset improves sequence-only PLM performance for viruses, with ESM-1v having the lowest performance and VESPA having the best performance (Fig 3A), unlike for other taxa (Fig S7). Moreover, including local structural information lead to the best performing PLM: SaProt-PDB outperforms all other PLMs in approximately 50% of the viral DMS datasets (Fig S8). The performance boost provided by SaProt-PDB over SaProt-AF2 is particularly pronounced for viral proteins, while SaProt-PDB's advantage over other PLMs declines for viruses with low numbers of unique strains in the PDB, such as Lassa and Nipah (Fig S8).

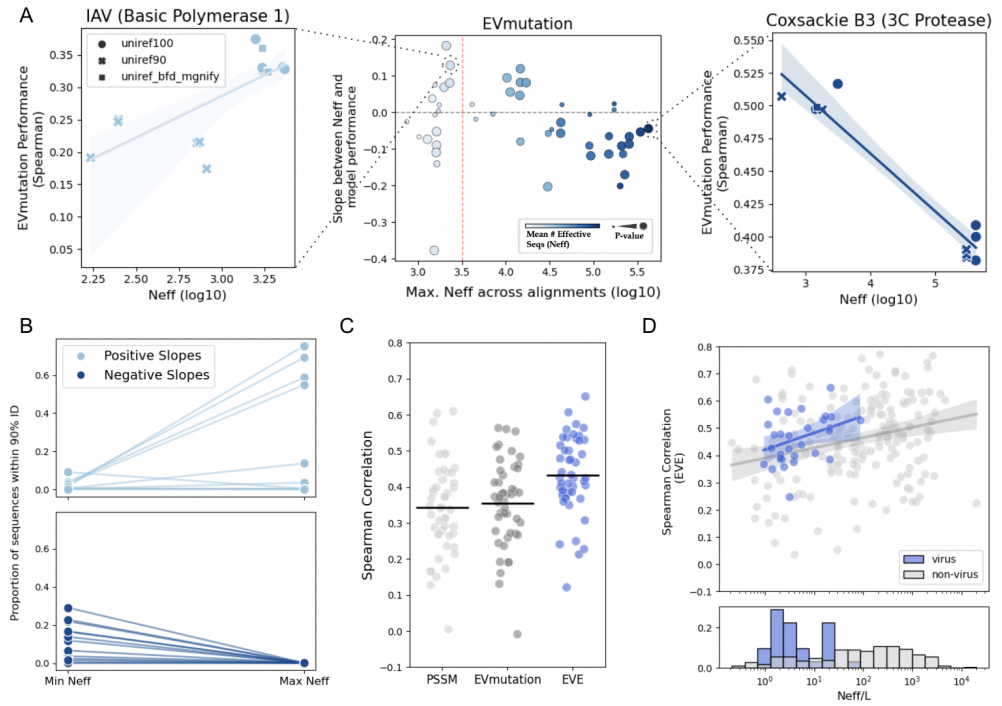


Figure 2: EVE outperforms other alignment-based models. A. Increasing alignment depth often reduces model performance. Each point represents the slope for a given viral protein of EVmutation model performance (Spearman with DMS) and Neff (across 6 bitscores and 3 databases). Positive slopes indicate performance increases with increasing Neff (as in influenza polymerase), and negative slopes indicate decreases in performance with increasing Neff (as in Coxsackie Protease). **B**. Deeper alignments improve (or reduce) performance when a larger (or smaller) proportion of sequences are within 90% identity of the target sequence. **C**. Spearman correlation between alignment-based models and DMS scores across collected viral proteins. **D**. Spearman correlations between EVE and DMS scores for viral proteins and non-viral ProteinGym datasets increase with increasing Neff/L (Neff normalized by sequence length).

We identify a model confidence score, measured as invesre pseudo-perplexity, that effectively predicts SaProt-PDB performance for viral proteins, with higher confidence scores correlating with better performance (Fig. 3B). This suggests that confidence scores may serve as a useful uncertainty quantification metric for viruses lacking experimental evaluation. As expected, SaProt-PDB had lower confidence scores for viral proteins than for non-viral proteins, likely due to the sparse representation of viral structures in the PDB. SaProt-PDB's performance was directly linked to the number of relevant structures in the training data (Fig. 3C). Therefore, incorporating even a limited number of viral structures could significantly improve predictions for viral proteins with low structural representation.

2.3 Modeling 40 WHO Priority Pathogens

The WHO recently prioritized 40 RNA viruses for research and intervention development based on transmission, virulence, and available medical countermeasures (WHO, 2024). Using insights from our prior analysis, we aimed to determine the best modeling approach for each WHO priority virus—particularly those lacking experimental data—and assess whether additional sequencing or structural information could improve predictions.

Among alignment-based and PLM approaches, EVE and SaProt-PDB were the top-performing models, with comparable average Spearman correlations (0.454 for EVE, 0.451 for SaProt-PDB; Fig. 4A). However, their relative performance varied in a predictable manner: SaProt-PDB excelled for phages with abundant structures in AF2DB and PDB, while EVE performed better for viruses with

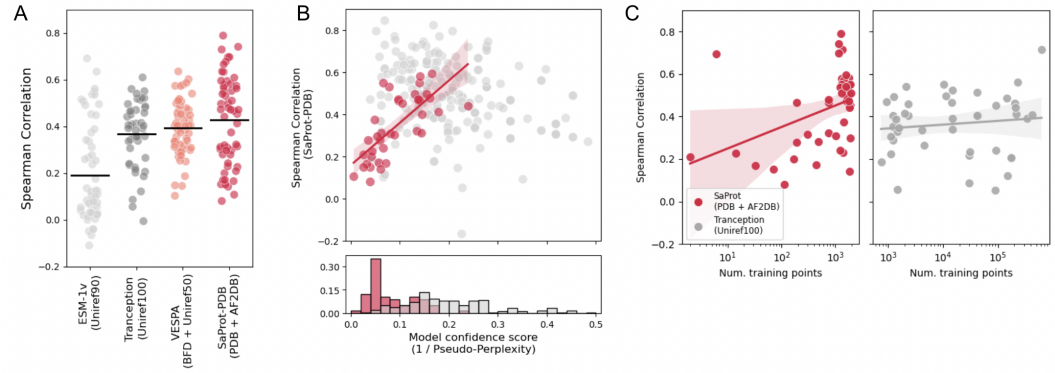


Figure 3: SaProt-PDB outperforms PLMs trained on sequences alone. **A**. Spearman correlation between PLM and DMS scores across collected viral proteins. **B**. SaProt-PDB confidence scores (1/pseudo-perplexity) are a good indicator of model performance. Viral replication or infectivity assays are highlighted in red. SaProt-PDB has lower confidence scores for viral compared to non-viral proteins. **C**. SaProt-PDB performance is related to number of related structures (via foldseek search, maximum 1000). Learning from structural information requires fewer related training points than learning from sequences (training points estimated using UniRef100 EVcouplings alignments).

limited structural data (Fig S9). Notably, our confidence measures (inverse pseudo-perplexity for SaProt-PDB and Neff/L for EVE) reliably predicted which model would perform better (Fig. 4B). Using these insights, we identified the minimum confidence score for any virus with known high model performance (Spearman greater than 0.6 against DMS). For the WHO pathogens, our analysis suggests that approximately 30% of viral antigens would benefit from additional sequencing data, while 16% would require additional structural data to improve model predictions (Fig. 4C). Four viral antigens currently lack sufficient sequences and structure information for accurate modeling and would benefit from increased sequencing and structure determination efforts: Hantavirus, Rift valley fever, Dabie bandavirus, and Sin Nombre virus Gn glycoproteins (the only Gn proteins in the WHO antigen dataset, each of which has a co-translational heterodimer Gc with sufficient data (Guardado-Calvo & Rey, 2021)).

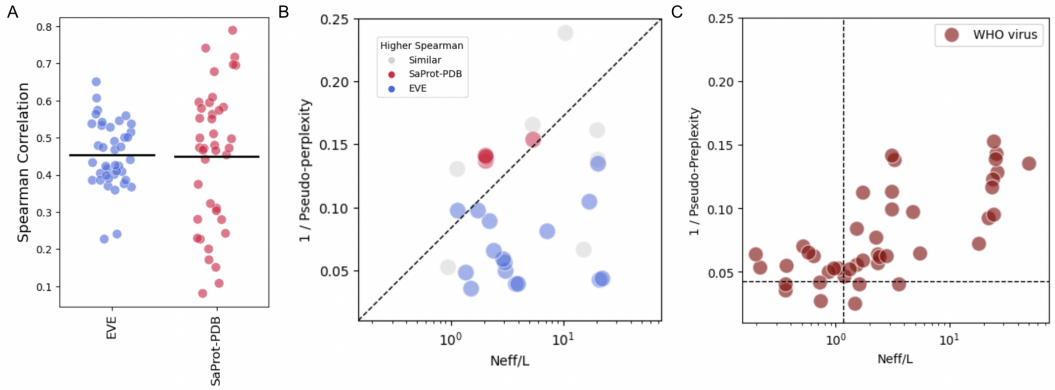


Figure 4: Evaluating EVE and SaProt utility across 40 WHO priority pathogens. **A**. EVE and SaProt-PDB show comparable performance across viruses. **B**. Viral proteins with high EVE confidence scores (Neff/L) have better performance with EVE; those with higher SaProt confidence scores (1/pseudo-perplexity) are better modeled with SaProt-PDB. Spearmans are considered similar if they are within 0.05. **C**. Few WHO priority pathogen proteins have lower EVE and SaProt-PDB confidence scores than other viral proteins with Spearman more than 0.6 (dotted lines).

3 DISCUSSION

Our comprehensive evaluation of protein mutation effect prediction models for viruses highlights key areas for improvement. While sequence clustering helps mitigate bias in PLMs (Meier et al., 2021), the low representation of viral sequences in standard databases hinders viral predictions. Fine-tuning with additional viral sequences from databases like GISAID (Shu & McCauley, 2017) may enhance performance. Structure-aware modeling also offers a promising solution by compensating for limited viral sequence diversity with structural information. Notably, SaProt-PDB requires far fewer related training structures—many fewer than 1,000—compared to sequences needed for sequence-based PLMs. This underscores the potential advantage of leveraging structural data from either experimentally-determined (e.g., PDB) or predicted structure databases (e.g., BFVD (Kim et al., 2025) and Viro3D (Litvin et al., 2024)).

We find that deeper alignments do not necessarily improve alignment-based model performance. Increasing alignment depth beyond Neff = $10^{3.5}$ introduces more distantly related sequences ($<$90% identity to the target) and reduces model performance, a trend that may extend to non-viral proteins. Future work could assess whether performance declines further with sequences in the twilight zone (20–35% identity) (Rost, 1999), whether selectively retaining closer sequences while pruning distant ones improve signal-to-noise ratio, and whether this sequence identity distribution could be used to improve the confidence metric.

Due to the high cost of training PLMs, the impact of viral sequencing depth on their performance remains unclear. Using PLMs trained on different datasets (albeit with different architectures) as a proxy, we find that PLMs trained on larger sequence databases (e.g., UniRef100 or UniRef50+BFD) perform better for viruses compared to other taxa which are well-represented in smaller databases (e.g., UniRef90). We show that inverse pseudo-perplexity is a good PLM confidence indicator. Taken together, these results suggest that for proteins with no existing experimental or sequencing data for evaluation, Neff/L and inverse pseudo-perplexity provide reliable confidence measures to determine the best modeling approach. Our results and benchmarks are available on GitHub.

Lastly, our results suggest that for the vast majority of WHO priority pathogens, most of which do not have large-scale experimental data, existing alignment-based and PLMs can accurately predict the fitness impacts of mutations. Overall, our findings provide insights for future model development tailored to viral mutation prediction, which could enhance efforts in viral surveillance, identifying immune and antibody escape mutations and aiding variant-proof vaccine design.

CODE AND DATA AVAILABILITY

Data and code available at https://github.com/debbiemarkslab/priority-viruses.

ACKNOWLEDGEMENTS

The authors thank Aaron Kollasch and members of the Marks lab. This work was supported by the Coalition for Epidemic Preparedness Innovations (CEPI).

REFERENCES

Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49(D1):D480–D489, 2021.

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*, pp. 2024–03, 2024.

Orr Ashenberg, Jai Padmakumar, Michael B Doud, and Jesse D Bloom. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by mxa. *PLoS pathogens*, 13 (3):e1006288, 2017.

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.

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.

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.

Adam S Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D Bloom. An antigenic atlas of hiv-1 escape from broadly neutralizing antibodies distinguishes functional and structural epitopes. *Immunity*, 50(2):520–532, 2019.

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.

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.

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.

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.

Jason D Fernandes, Tyler B Faust, Nicolas B Strauli, Cynthia Smith, David C Crosby, Robert L Nakamura, Ryan D Hernandez, and Alan D Frankel. Functional segregation of overlapping genes in hiv. *Cell*, 167(7):1762–1773, 2016.

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.

Filipp Frank, Meredith M Keen, Anuradha Rao, Leda Bassit, Xu Liu, Heather B Bowers, Anamika B Patel, Michael L Cato, Julie A Sullivan, Morgan Greenleaf, et al. Deep mutational scanning identifies sars-cov-2 nucleocapsid escape mutations of currently available rapid antigen tests. *Cell*, 185(19):3603–3616, 2022.

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.

Pablo Guardado-Calvo and Félix A Rey. The viral class ii membrane fusion machinery: divergent evolution from an ancestral heterodimer. *Viruses*, 13(12):2368, 2021.

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.

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.

Jeremiah D Heredia, Jihye Park, Hannah Choi, Kevin S Gill, and Erik Procko. Conformational engineering of hiv-1 env based on mutational tolerance in the cd4 and pg16 bound states. *Journal of virology*, 93(11):10–1128, 2019.

Nancy Hom, Lauren Gentles, Jesse D Bloom, and Kelly K Lee. Deep mutational scan of the highly conserved influenza a virus m1 matrix protein reveals substantial intrinsic mutational tolerance. *Journal of virology*, 93(13):10–1128, 2019.

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.

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.

Caroline Kikawa, Catiana H Cartwright-Acar, Jackson B Stuart, Maya Contreras, Lisa M Levoir, Matthew J Evans, Jesse D Bloom, and Leslie Goo. The effect of single mutations in zika virus envelope on escape from broadly neutralizing antibodies. *Journal of Virology*, 97(11):e01414–23, 2023.

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.

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.

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.

Ruipeng Lei, Andrea Hernandez Garcia, Timothy JC Tan, Qi Wen Teo, Yiquan Wang, Xiwen Zhang, Shitong Luo, Satish K Nair, Jian Peng, and Nicholas C Wu. Mutational fitness landscape of human influenza h3n2 neuraminidase. *Cell reports*, 42(1), 2023.

Ruipeng Lei, Enya Qing, Abby Odle, Meng Yuan, Chaminda D Gunawardene, Timothy JC Tan, Natalie So, Wenhao O Ouyang, Ian A Wilson, Tom Gallagher, et al. Functional and antigenic characterization of sars-cov-2 spike fusion peptide by deep mutational scanning. *Nature communications*, 15(1):4056, 2024.

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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

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.

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

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.

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.

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

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.

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*, pp. 16990–17017. PMLR, 2022.

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

Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. *PLoS pathogens*, 10(4):e1004064, 2014.

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*, 51(D1):D753–D759, 12 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1080. URL https://doi.org/10.1093/nar/gkac1080.

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.

Burkhard Rost. Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94, 1999.

Yin Xiang Setoh, Alberto A Amarilla, Nias YG Peng, Rebecca E Griffiths, Julio Carrera, Morgan E Freney, Eri Nakayama, Shinya Ogawa, Daniel Watterson, Naphak Modhiran, et al. Determinants of zika virus host tropism uncovered by deep mutational scanning. *Nature microbiology*, 4(5): 876–887, 2019.

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

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

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.

Marion Sourisseau, Daniel JP Lawrence, Megan C Schwarz, Carina H Storrs, Ethan C Veit, Jesse D Bloom, and Matthew J Evans. Deep mutational scanning comprehensively maps how zika envelope protein mutations affect viral growth and antibody escape. *Journal of virology*, 93(23): 10–1128, 2019.

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

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.

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

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.

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

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*, pp. 2023–03, 2023.

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.

Qi Wen Teo, Yiquan Wang, Huibin Lv, Kevin J Mao, Timothy JC Tan, Yang Wei Huan, Joel Rivera-Cardona, Evan K Shao, Danbi Choi, Zahra Tavakoli Dargani, et al. Deep mutational scanning of influenza a virus nep reveals pleiotropic mutations in its n-terminal domain. *bioRxiv*, pp. 2024–05, 2024.

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.

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.

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.

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*, pp. 2022–02, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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

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

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.

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.

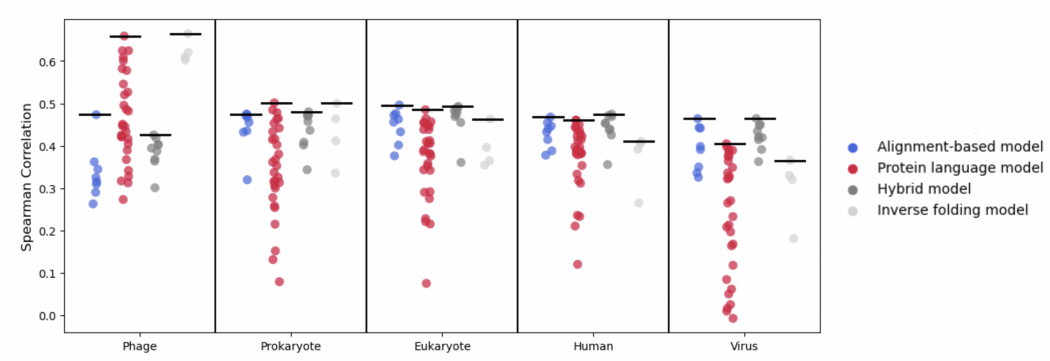
Nicholas C Wu, Yushen Du, Shuai Le, Arthur P Young, Tian-Hao Zhang, Yuanyuan Wang, Jian Zhou, Janice M Yoshizawa, Ling Dong, Xinmin Li, et al. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza a virus m segment. *BMC genomics*, 17:1–15,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

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.

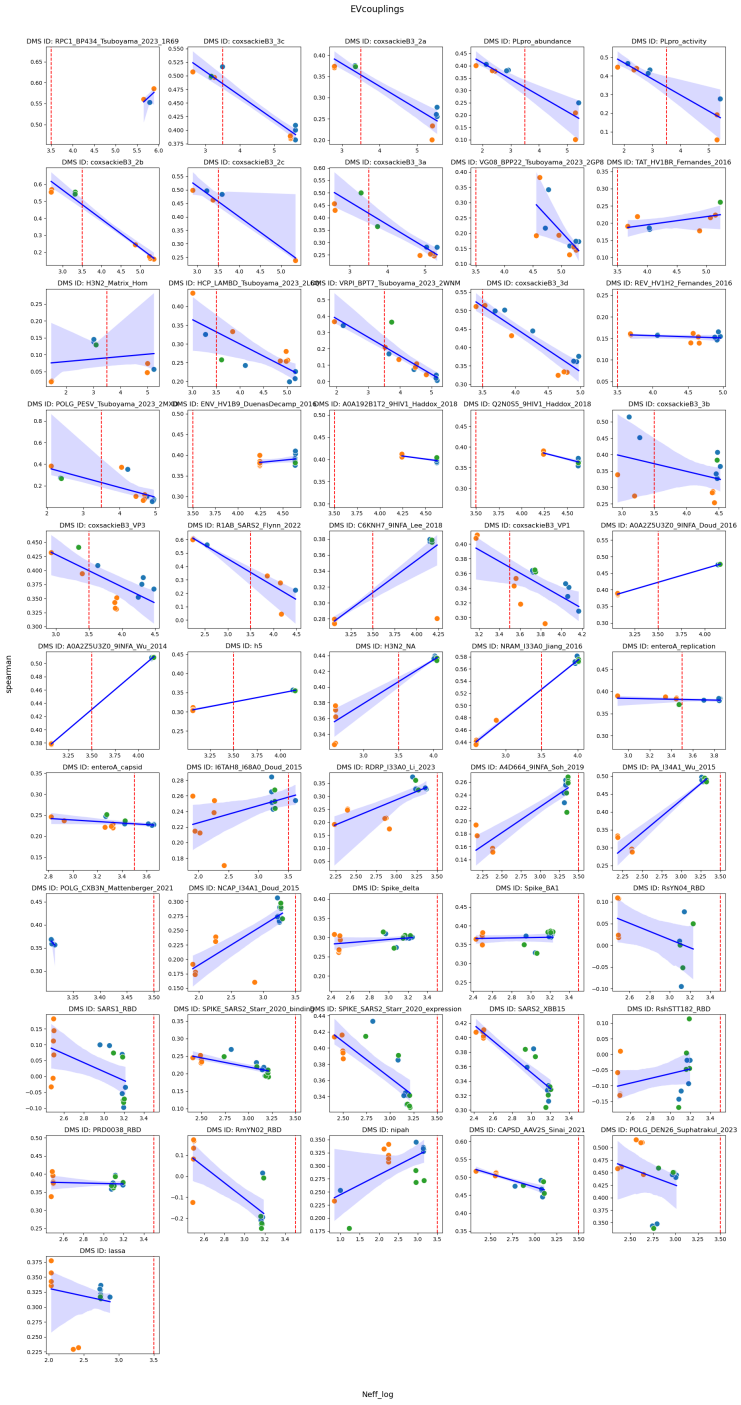
Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

Noor Youssef, Sarah Gurev, Fadi Ghantous, Kelly Brock, Javier A Jaimes, Nicole Thadani, Ann Dauphin, Amy Sherman, Leonid Yurkovetskiy, Daria Soto, Ralph Estaboulieh, Ben Kotzen, Pascal Notin, Aaron Kollasch, Alexander Cohen, Sandra Dross, Jesse Erasmus, Deborah Fuller, Pamela Bjorkman, Jacob Lemieux, Jeremy Luban, Mike Seabman, and Debora Marks. Protein design for evaluating vaccines against future viral variation. *BioRxiv*, 2024.

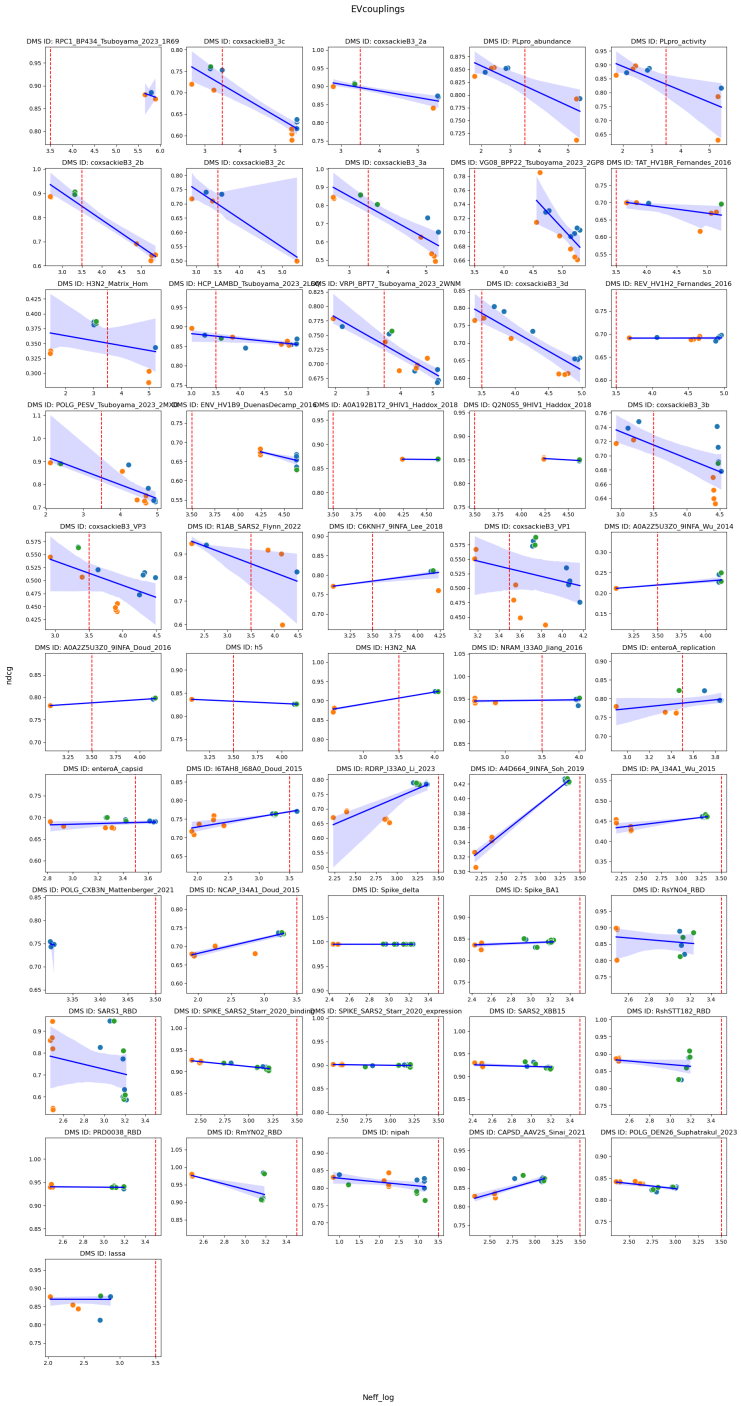
A SUPPLEMENTARY FIGURES



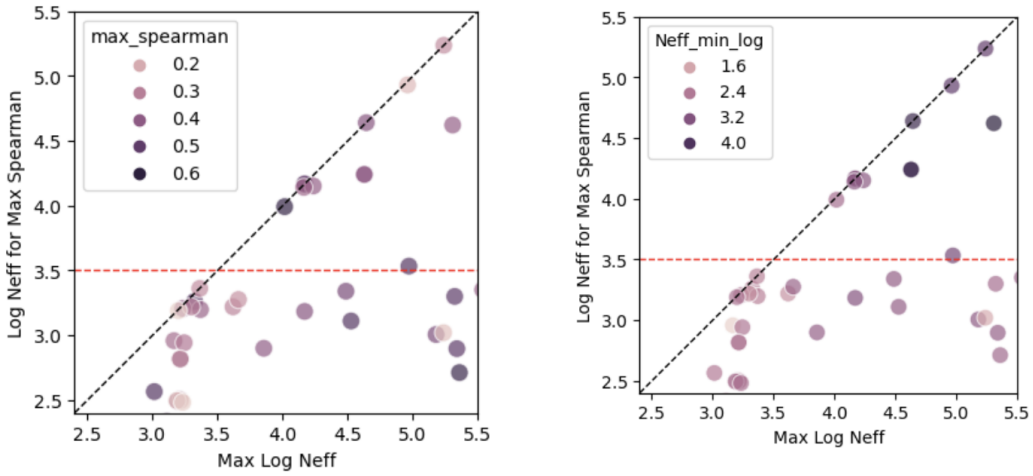
Supplementary Figure 1: Protein language models have higher performance across all taxa except for eukaryotic viruses. Each point represents the average correlation across 250 DMS assays for each of the 50 fitness models in ProteinGym, labeled with their model type. Viral performance is on the dataset 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.



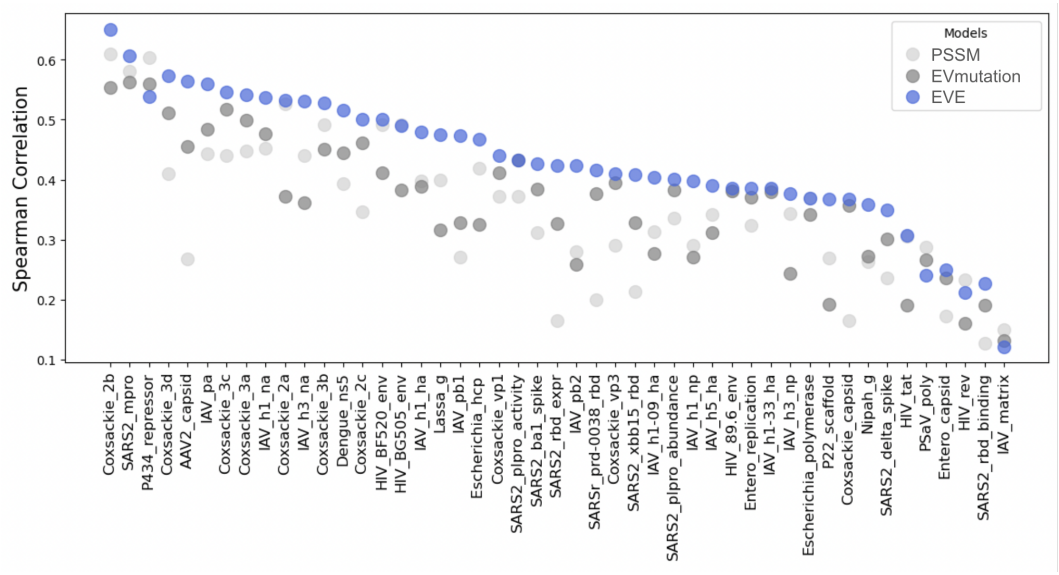
Supplementary Figure 2: Relationship between Neff and model performance (Spearman rank correlation) between model scores and DMS. Selected Neff cutoff (red line).



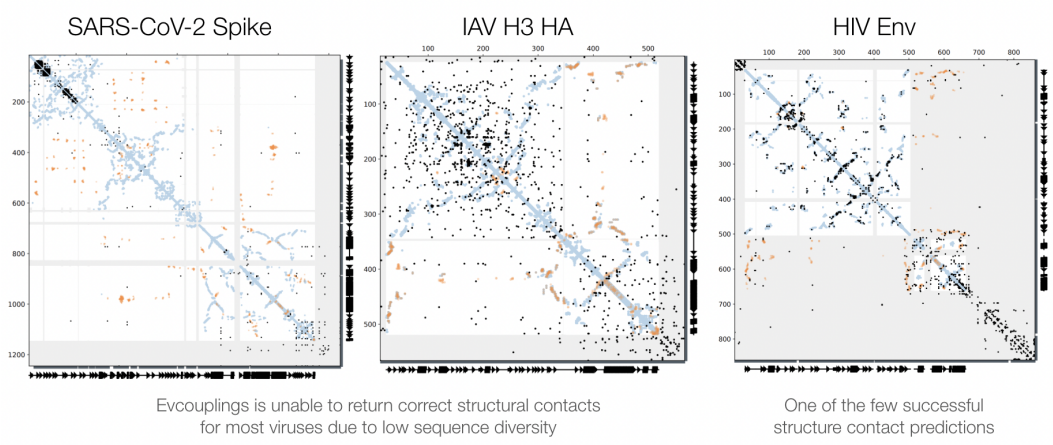
Supplementary Figure 3: Relationship between Neff and model performance (NDCG rank correlation) between model scores and DMS. Normalized Discounted Cumulative Gain (NDCG) measures between 0 (low) and 1 (high) and focuses ranking performance on the top rankings (here top 10%). Selected Neff cutoff (red line).



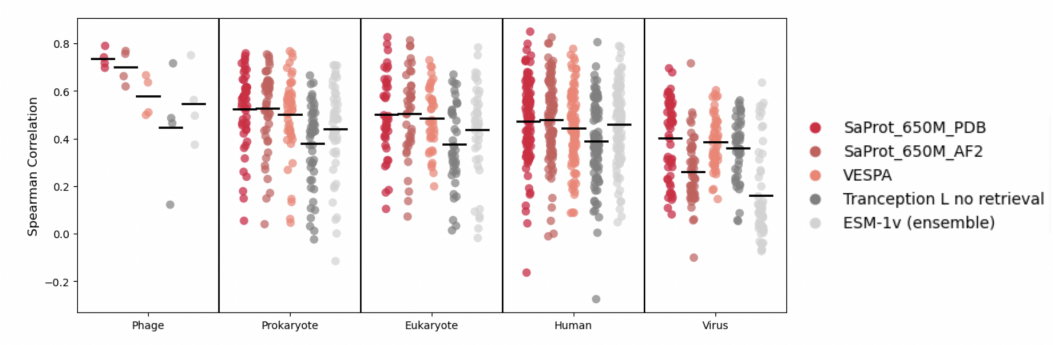
Supplementary Figure 4: Relationship between maximum log Neff of all alignments for a given viral DMS and the log Neff of the alignment with the highest EVmutation spearman. Selected Neff cutoff (red line).



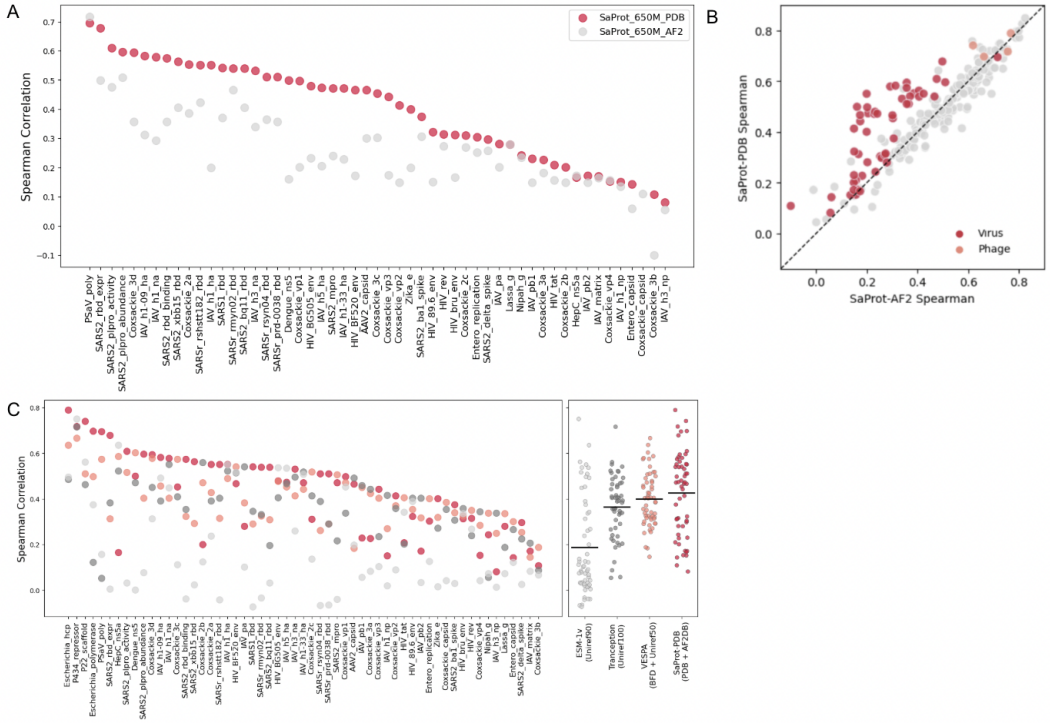
Supplementary Figure 5: Spearman rank correlation between PLM model score and DMS score for each curated viral protein. EVE outperforms other alignment-based models on almost all proteins.



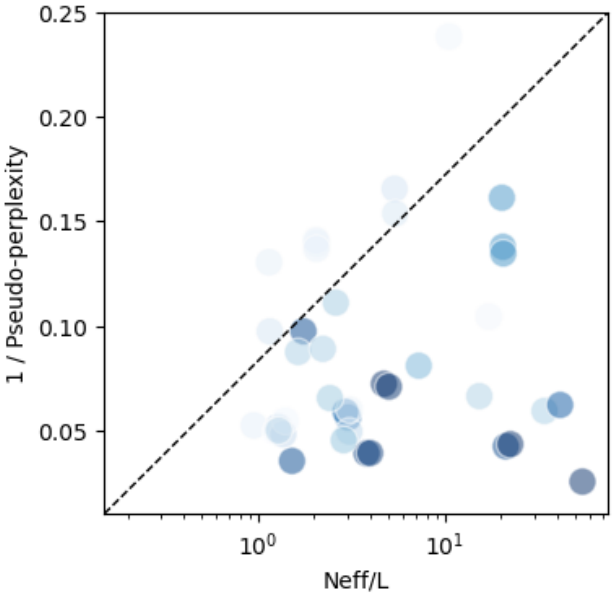
Supplementary Figure 6: Selected contact maps from UniRef100 alignments with bit score 0.1 for SARS-CoV-2 Spike, Influenza H3 and HIV Envelope proteins. HIV Env is one of the few examples of viral proteins with a DMS where EVcouplings successfully predict structure contacts, while ability to predict structural contacts is a strong predictor of model performance for other taxa.



Supplementary Figure 7: Comparison of the Spearman rank correlation of all PLM models across DMSs separated by taxa.



Supplementary Figure 8: 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 outperfoms SaProt-AF2 for most eukaryotic viral proteins, but not for phages. C. Spearman rank correlation between PLM model score and DMS score for each curated viral protein.



Supplementary Figure 9: Viral proteins with higher EVE confidence scores (Neff/L) than SaProt confidence scores (1/perplexity) tend to have fewer structures in training in AF2DB and PDB (darker blue for fewer structures).

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
	Wavenet	0.373	0.328
	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
Protein language model	VESPA	0.436	0.432
	VESPAl	0.394	0.392
	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
	ESM2 (3B)	0.406	0.274
	ESM2 (15B)	0.401	0.313
	ESM2 (150M)	0.387	0.137
	ESM2 (650M)	0.414	0.238
	ESM2 (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 (Notin et al., 2023). Note this analysis only covers half of the now curated 51 viral datasets and does not include SaProt-PDB, but can be used to contextualize our findings in terms of 50+ models.

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 2: 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.

C Extended Methods

C.1 Overview

We curated and standardized a set of 51 viral DMSs (more than doubling the number of viral datasets in ProteinGym) to evaluate eight models, which were selected to span differing modeling approaches and training on diverse sequence datasets. Assays measuring expression, host receptor binding, or replication, or infectivity. The viruses included are relevant for vaccine design (e.g., SARS-CoV-2 and other sarbecoviruses, seasonal and pandemic Flu, and pandemic-threat Lassa andNipah viruses) as well as for viral vector design (e.g., AAV).

For alignment-based methods, we tested three models: position-specific scoring matrix (PSSM), EVmutation (Hopf et al., 2017), and EVE (Frazer et al., 2021). PSSM assumes each position in the protein evolves independently and assigns a prediction score for each mutation dependent on its frequency in the alignment. In addition to capturing position-specific frequencies, EVmutation infers pairwise residue dependencies and can therefore account for epistatic interactions (Hopf et al., 2017). Meanwhile, EVE captures higher order interactions by using a variational autoencoder architecture (Frazer et al., 2021).

For these methods, we used three different sequence datasets for alignment generation: UniRef100, a non-redundant protein sequence database; UniRef90, a 90% identity-clustered version of UniRef100; and UniRef100+BFD+Mgnify which combines UniRef100 with the Big Fantastic Database (BFD) and the Mgnify database. The BFD integrates sequences from UniProt (Swiss-Prot and TrEMBL) (uni, 2021), Metaclust (Steinegger & Söding, 2018), and the Soil and Marine Eukaryotic Reference Catalog (Steinegger et al., 2019). Mgnify includes proteins from metagenomic assemblies (Richardson et al., 2022). 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. To mitigate redundancy, sequences in the MSA were clustered at 99% identity, with each sequence within a cluster assigned a weighted contribution equal to the inverse of the cluster size. 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 S2), but these would mostly be excluded by the choice of Neff cutoff. This highlights the computational intensity of alignment-based approaches, particularly when using comprehensive databases like BFD and Mgnify, with 2.1 billion and 2.4 billion sequences, respectively.

For PLMs, we evaluated Tranception (without MSA retrieval) (Notin et al., 2022), ESM-1v (Meier et al., 2021) and VESPA (Marquet et al., 2022). Tranception is a autoregressive PLM trained on UniRef100. ESM-1v has a Transformer encoder architecture and was trained on UniRef90. VESPA combines per-residue conservation prediction with the embeddings from ProtT5 (Elnaggar et al., 2021), a PLM with T5 architecture trained first on BFD and then finetuned on UniRef50.

We also evaluate a structure-aware protein language model, SaProt (Su et al., 2023), which employs the same architecture as ESM2 (trained on UniRef50) but expands the embedding layer to encompasses 441 structurally-aware tokens instead of the original 20 amino acid residue tokens. We evaluated two versions of SaProt: SaProt-AF2 which was trained on the AlphaFold2 database compromising of approx. 40 million predicted structures (without eukaryotic viruses, though including phages), and SaProt-PDB which continues pretraining of the SaProt-AF2 model on the 60,000 experimentally derived structures from the PDB. We use the ProteinGym benchmark(Notin et al., 2023) to extrapolate the results of the models tested here to the over 50 alignment-based, protein language model, hybrid, and inverse folding models evaluated previously on a more limited set of viral assays.

C.2 Viral Deep Mutation Scans

We searched for all viral fitness and escape deep mutational scans, focusing here on single substitution mutations (Sinai et al., 2021; Mattenberger et al., 2021; Álvarez-Rodríguez et al., 2024; Suphatrakul et al., 2023; Tsuboyama et al., 2023; Bakhache et al., 2024; Qi et al., 2014; Haddox et al., 2018; Duenas-Decamp et al., 2016; Haddox et al., 2016; Fernandes et al., 2016; Heredia et al., 2019; Doud & Bloom, 2016; Wu et al., 2014; Lee et al., 2018; Dadonaite et al., 2024; Doud et al., 2015; Jiang et al., 2016; Lei et al., 2023; Wu et al., 2015; Soh et al., 2019; Li et al., 2023; Ashen-

berg et al., 2017; Teo et al., 2024; Hom et al., 2019; Wu et al., 2016; Starr, 2024; Starr et al., 2020; Dadonaite et al., 2023; Taylor & Starr, 2023; Flynn et al., 2022; Wu et al., 2024; Sourisseau et al., 2019; Setoh et al., 2019; Maurer et al., 2024; Welsh et al., 2024; Dingens et al., 2019; Frank et al., 2022; Lei et al., 2024; Kikawa et al., 2023). We focus now only on fitness DMSs, but will later include escape from monoclonal antibodies and patient sera. Moreover, some DMSs were excluded from this benchmark depending on the assayed phenotype, for example drug inhibition assays, or difficulties with the data, but may be included in the future.

C.3 ALIGNMENT-BASED MODELS

C.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). Briefly, this involved five search iterations of the profile HMM homology search tool jackhmmer against the specified sequences database. We evaluated the impact of searching against three database with vastly different number of sequences: UniRef100, a database of non-redundant protein sequences; UniRef90, a database obtained by clustering the UniRef100 database based on 90% sequence identity and specifying a representative sequence per cluster; Big Fantastic Database (BFD) covering protein sequences from UniProt (Swiss-Prot&TrEMBL; uni (2021)), Metaclust (Steinegger & Söding, 2018) and Soil Reference Catalog Marine Eukaryotic Reference Catalog (Steinegger et al., 2019). We used length-normalized bit scores to threshold sequence similarity. We generated alignments across six bit scores of 0.5, 0.3, 0.1, 0.05, 0.03, 0.01 bits/residue. 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.

C.3.2 PSSM

To infer the contribution of site-specific amino acid constraints without considering explicit epistatic constraints, we used a site-wise maximum entropy model as implemented in Hopf et al. (2017).

C.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).

C.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).

C.4 PROTEIN LANGUAGE MODELS

C.4.1 SEQUENCE DATASETS

The protein language models described here do not use multiple sequence alignments, instead using variants of a Transformer (Vaswani, 2017) popularized in natural language modeling for self-supervised training on a large corpus of sequence data. In this case, the models are trained on large protein sequence datasets from across the entire protein universe, rather than only sequences specific to a given family of proteins. These datasets include BFD, UniRef100, and UniRef90, as well as the AF2 and PDB structure databases. Moroever, because these datasets are alignment-free, these models can more naturally score insertions and deletions.

C.4.2 TRANCEPTION

Tranception (Notin et al., 2022) combines an autoregressive protein language model with inference-time retrieval from a MSA. We used Tranception Large (700M parameters) trained on UniRef100 without MSA retrieval as implemented in ProteinGym (Notin et al., 2023).

C.4.3 ESM-1V

ESM-1v (Meier et al., 2021) has a Transformer encoder architecture similar to BERT [Devlin et al., 2019] and was trained with a Masked-Language Modeling (MLM) objective on UniRef90. We use the implementation presented in ProteinGym (Notin et al., 2023) to handle sequences that are longer than the model context window (ie., 1023 amino acids).

C.4.4 VESPA

VESPA (Marquet et al., 2022) combines the embeddings from ProtT5 (Elnaggar et al., 2021) with a per-residue conservation prediction. ProtT5 uses a T5 architecture which uses an encoder and decoder and was first trained on BFD and then finetuned on UniRef50.

C.4.5 SAPROT

SaProt (Su et al., 2023) introduces a structure-aware vocabulary, into protein language modeling by training on Foldseek (van Kempen et al., 2022) 3Di tokens which represent the local geometric conformation information of each residue relative to its spatial neighbors. These 3Di tokens are combined with typical amino acid residue tokens as input to the SaProt model, which utilizes an ESM-2 Transformer architecture (Lin et al., 2022). 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.