Alignment-based and protein foundation models for viral evolution, vaccines and vectors

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

Protein mutation effect prediction has advanced several fields, from designing enzymes to forecasting viral evolution. These models are typically trained on sequence data, structural data, or a combination of both. Sequence-based models learn constraints governing protein structure and function from sequence data and fall broadly into two categories: alignment-based models and protein language models (PLMs). We provide the first detailed comparison of modeling approaches specifically for viruses. We curated a dataset of over 59 standardized viral deep mutational scanning assays and assessed the relative performance of three alignment-based models (PSSM, EVmutation, and EVE), three PLMs (ESM-1v, Tranception, and VESPA), and two versions of SaProt, a structurally-aware protein language model (SaProt-AF2 and SaProt-PDB). Interestingly, deeper alignments often led to worse performance for alignment-based models. On the other hand, PLMs tended to perform better as the size of the training database increased. Overall, alignment-based models outperformed sequence-only PLMs, while the best alignment based model performed on par with SaProt-PDB. Our findings suggest that modeling strategies that are effective for other taxa may not translate directly to viruses, likely due to the limited number of viral sequences used for training. However, incorporating additional virus-specific data into PLMs could enhance their predictive power for viral mutation effects, important for understanding viral evolution and the design of vaccines and viral vectors.

1 Introduction

Predicting the functional impact of protein mutations is critical for a wide range of biological applications, including understanding human disease variants (Frazer et al., 2021), designing enzymes (Lu et al., 2022; Sumida et al., 2024), and forecasting viral evolution (Thadani et al., 2023). With advances in machine learning, computational approaches to mutation effect prediction have seen significant improvements, offering faster and more scalable alternatives to traditional experimental methods. These models can leverage protein sequence data, structural data, or a combination of both. Sequence models fall into two primary categories: alignment-based models and protein language models (PLMs). Alignment-based models use evolutionary information from multiple sequence alignments (MSAs) to predict the impact of mutations (Hopf et al., 2017; Frazer et al., 2021; Laine et al., 2019). Alternatively, PLMs leverage deep learning techniques to capture patterns

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

AI for New Drug Modalities at NeurIPS 2024.

and relationships between residues by training on large databases of sequences, without relying on MSAs (Alley et al., 2019; Elnaggar et al., 2021; Yang et al., 2024; Lin et al., 2022; Meier et al., 2021a; Rives et al., 2021; Jumper et al., 2021; Marquet et al., 2022; Notin et al., 2022). More recently, structurally-aware protein language models, such as SaProt (Su et al., 2023), incorporate protein structural information with amino acid sequences to further improve predictive performance.

Given the rapid emergence of new models for predicting the effects of mutations, benchmarks such as ProteinGym (Notin et al., 2023) were curated to aid in the comprehensive evaluation and comparison of these models. Conclusions from such comparisons revealed that across proteins from diverse taxa, PLMs often outperform alignment-based models. However, this trend does not hold true for viruses. Viral genomes often have fewer and less diverse sequences for training, which can limit the effectiveness of PLMs that rely on large datasets for optimal performance. Viral proteins also contain unique biophysical features, with lower protein stability, loosely packed cores, and an abundance of short disordered segments and coil residues (Tokuriki et al., 2009). These adaptations facilitate an increased structural flexibility, enabling diverse interactions with varying host proteins across a broad host range and avoidance of host immune response. They may also compensate for the deleterious effects of mutations that arise due to the abnormally high mutation rate of viruses or the overlapping reading frames required by the unusually compact genomes of RNA viruses, which mean single mutations impact multiple proteins. Conversely, rather than learning across the protein universe, we have previously shown that alignment-based methods trained on single viral families are successful for viral antibody escape prediction (Thadani et al., 2023) and vaccine evaluation (Youssef et al., 2024), effectively predicting SARS-CoV-2 evolution from pre-pandemic information. Thus, it remains unclear which modeling approaches are most effective for predicting viral mutation effects, where the biological and evolutionary constraints may differ from those observed in other taxa.

We present the first comprehensive evaluation of mutation effect prediction models for viruses. We assessed three alignment-based models, three PLMs, and two structurally-aware PLMs. We curated a collection of 59 standardized viral deep mutational scanning (DMS) assays, measuring a wide range of viral phenotypes including expression, host receptor binding, infectivity, antibody binding, and neutralization susceptibility. The viruses included are relevant for vaccine design (e.g., SARS-CoV-2 and broader sarbecoviruses, seasonal and pandemic Flu, and pandemic-threat Lassa and Nipah viruses) as well as for viral vector design (e.g., AAV). We found that while deeper sequence alignments do not consistently improve the performance of alignment-based models, PLMs tend to benefit from larger training databases. Importantly, structure-aware PLMs outperformed other PLMs despite more limited viral representation, suggesting that incorporating structural information is useful for viral mutation effect prediction. These findings challenge existing assumptions about optimal modeling strategies and highlight the potential for further improvements by tailoring PLMs to virus-specific data. These models The information gained from this work can inform the development of next-generation models for viral mutation effect prediction that can facilitate pandemic protection efforts, aiding in the surveillance of pandemic variants and the design and testing of variant-proof vaccines as well as deimmunized and targeted viral vectors.

2 Methods

We curated and standardized a set of 59 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. 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, while EVmutation can capturing pairwise residue dependencies (Hopf et al., 2017). Meanwhile, EVE captures higher order interactions by using a a variational autoencoder architecture (Frazer et al., 2021). For these methods, we used three different sequence datasets for alignment generation: Uniref90, Uniref100, and Uniref100+BFD+Mgnify.

For PLMs, we evaluated Tranception (without MSA retrieval) (Notin et al., 2022), ESM-1v (Meier et al., 2021b) 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 ESM- 2 (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 verions of SaProt: SaProt-AF2 was trained on the AlphaFold2 database compromising of approx. 40 million predicted structures (without eukaryotic viruses, though including phages), while SaProt-PDB 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.

3 Results

Using the ProteinGym benchmark (Notin et al., 2023), which evaluated over 50 protein fitness models across 250+ DMS assays, we observed a clear trend: PLMs consistently outperformed other models across nearly all taxa, with one notable exception—viruses (Fig 1, Supp Table S1). For viral proteins, alignment-based models frequently achieved better performance than PLMs, as measured by Spearman correlation with DMS data. This divergence highlights a critical difference in the predictive power of PLMs when applied to viral proteins, likely due to the limited sequence diversity of viral proteins in the training datasets compared to other taxa. To this end, we sought to investigate the tradeoffs of alignment-based models and PLMs for viral mutation effect prediction.

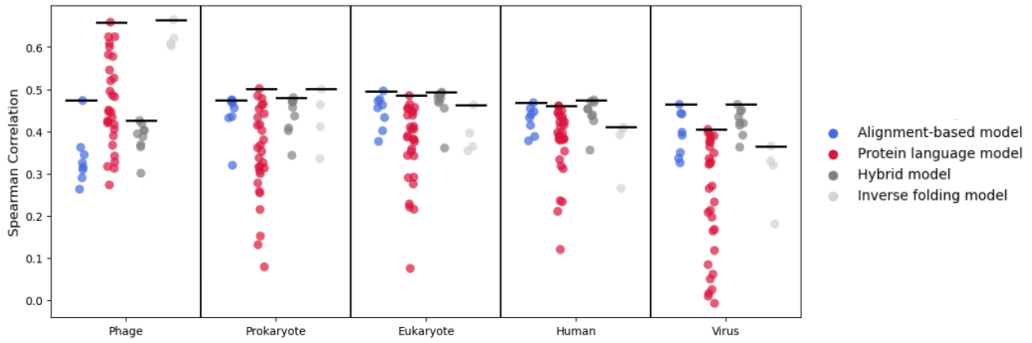


Figure 1: Protein language models have higher performance across all taxa except for eukaryotic viruses. Each point represents the average correlation across 250 DMS assays for each of the 50 fitness models in ProteinGym, labeled with their model type. Viral performance is on the dataset provided by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

3.1 Alignment-based model

For each protein in our newly curated viral benchmark, we trained three alignment-based models—PSSM, EVmutation, and EVE. We use sequence alignments derived from three databases: 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 this process was the memory required to query large sequence databases. Some models encountered memory limitations that prevented training on the largest alignments (Supp Table S2). 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.

We employed the EVcouplings pipeline (Hopf et al., 2019) to generate MSAs and model scores for both PSSM and EVmutation. Our analysis revealed that, although alignment depth varied across the different sequence databases, model performance remained largely consistent (Supp Fig S1). Note that despite adequate performance on deep mutational scans, almost no viruses have sufficient sequence diversity to accurately predict structural contacts from EVcouplings (Supp Fig S2), as can commonly be done for non-viral proteins. For training the EVE models, we selected the alignment for each protein that maximized both the effective number of sequences (Neff) and percent coverage. EVE outperformed both PSSM and EVmutation across most assays (Fig 2A, Supp Fig S3), consistent with previous findings that EVE's probabilistic approach offers better predictive accuracy than simpler alignment-based models Frazer et al. (2021); Riesselman et al. (2018); Thadani et al. (2023).

We next sought to assess whether alignment depth correlated with model performance. Overall, no consistent trend was observed between Neff and performance across all proteins (Supp Fig S4). To determine if this relationship held true for individual proteins—i.e., whether increasing alignment depth specifically improves performance—we employed EVmutation models to analyze the correlation between Neff and model performance on a per-protein basis. For some proteins, increasing alignment depth improved performance (resulting in positive slopes), while it decreases performance for other proteins (negative slopes; Fig 2B-C, Supp Fig S5). Interestingly, our results showed a general trend where increasing the size of the MSA was most often associated with decreased performance (Fig 2D). This effect was particularly pronounced in the PSSM models compared to EVmutation models (Supp Fig S6). This suggests that after a certain threshold, adding more sequences introduces noise or redundant information that can negatively impact model predictions as mutations are seen on vastly different background sequences, that are likely to be under different functional constraints.

3.2 Protein language models

Existing PLMs have underperformed for viruses when compared to other taxa or to alignment-based methods (Fig 1). This discrepancy can largely be attributed to the composition of the training data. Most PLMs are trained on subsets of UniRef, such as UniRef90 or UniRef50, which contain disproportionately low numbers of viral sequences, 0.6% and 0.9% respectively (Fig 3A). Sequence counts drop drastically between UniRef100 and UniRef90, severely limiting the viral diversity available for training. For example, while UniRef100 contains over 6,000 paramyxovirus fusion protein sequences, less than 10% of these remain in UniRef90—despite UniRef90 being roughly half the size of UniRef100 overall (Fig 3A). This stark reduction in viral sequence diversity might explain why PLMs underperform for viruses and suggests that additional viral-specific training data could improve their predictive accuracy.

To examine the impact of training data on PLM performance across viruses, we 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). Notably, some PLMs such as CARP (Yang et al., 2024) were trained on Uniref50, but were excluded due to their generally low performance on viruses (Table S1). Our analysis revealed that sequence scale significantly improves PLM performance for viruses, with ESM-1v performing markedly worse on viruses compared to other taxa (Fig 3B). However, for viruses with higher effective sequence numbers (Neff), using clustering at 99% sequence identity, such as Flu and HIV, the detrimental effect of training on UniRef90 was less pronounced. Moreover, the inclusion of BFD in the training data provided the greatest performance boost for phages (Supp Fig S7), which are overrepresented in metagenomic sequencing compared to eukaryotic viruses. Remarkably, inverse folding models also perform particularly well for phages (Fig 1), likely due to the nature of the tested datasets that focused on stability assays. These findings highlight the importance of training PLMs on diverse and comprehensive datasets, particularly for viruses, where sequence representation can vary significantly across databases.

Structure-aware modeling may offer a valuable solution to compensate for the limited viral sequence diversity in traditional sequence databases. Viral Foldseek sequences, for instance, can be used to retrieve structural information for viruses with poor sequence representation. For example, searching for the Zika virus envelope protein using Foldseek retrieves structures from other class II fusion proteins spanning multiple viral families, such as those from Semliki, Chikungunya, Sindbis, Rift Valley Fever, and Eastern Equine Encephalitis viruses. To test this hypothesis, we scored two versions of SaProt: SaProt-AF2, trained on the AlphaFold2 database (which excludes eukaryotic viral proteins,

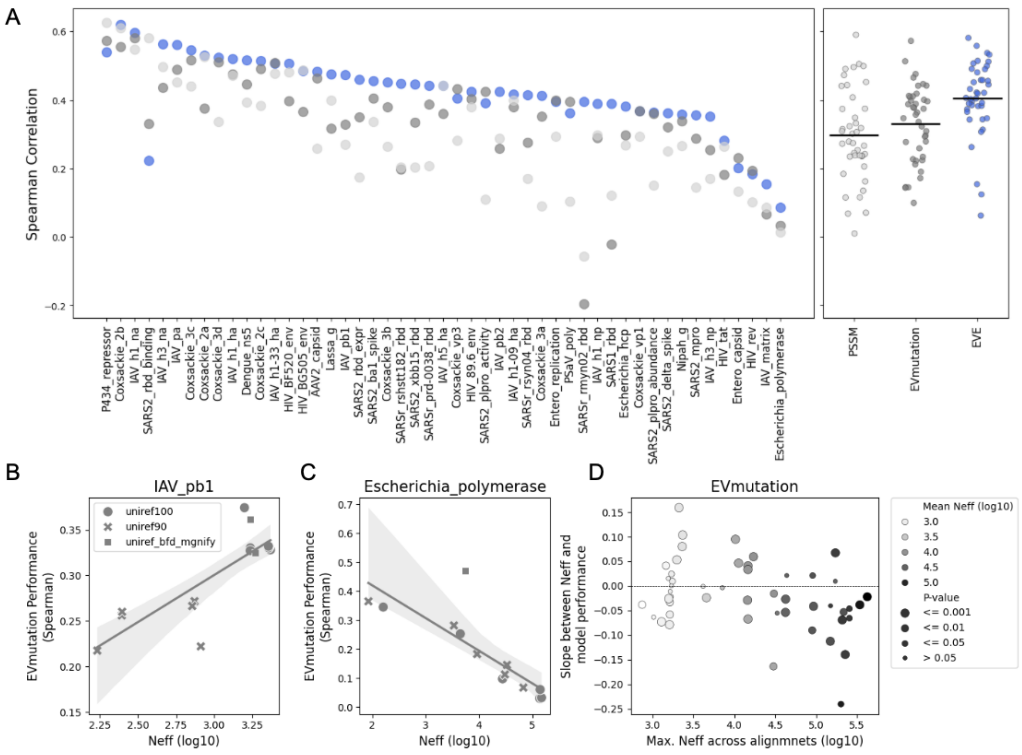


Figure 2: EVE outperforms other alignment-based models (EVmutation and PSSM) across the new benchmarks of 59 viral DMS datasets. A. Spearman rank correlation between alignment-based model score and DMS score for each of the viral proteins. B. EVmutation models for Influenza A virus polymerase basic 1 (PB1) improve in performance as Neff increases. Each point is a Spearman correlation between an EVmutation model trained on an alignments generated using one of six different bitscores and three different sequence databases. C. EVmutation models for Escherichia polymerase protein have worse performance as Neff increases. D. Increasing alignment depth results in worse performance for most viral proteins, especially those with greater maximum Neff. Each point represents the slope between model performance (Spearman) and Neff. Negative values imply a decrease in model performance with increase Neff, and positive slopes imply an increase in model performance with increasing Neff.

but does include phages), and SaProt-PDB, which continues training on the PDB (containing a limited number of viral structures). Our results show that SaProt-PDB outperforms all other PLMS in approx 50% of the viral DMS datasets (Fig 3B). Importantly, the performance boost provided by SaProt-PDB over SaProt-AF2 is particularly pronounced for viral proteins (Fig 3C, Supp Fig S7), 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 (Supp Fig S8). Additionally, SaProt and EVE–the best alignment-based model–are roughly on par, yet perform best on different DMSs (Fig 3D and Supp Fig S9). EVE is much better for viruses with no similar structure in the PDB and for fitness assays (Supp Fig S9), while SaProt is much better for phages (stability assays) which are in the AF2 structural training dataset of SaProt. SaProt pseudo-perplexity, used as a measure of uncertainty for PLMs, of the wildtypes for each of the viral fitness DMSs is a good predictor of SaProt mutation effect performance for that virus (Supp Fig S10), supporting uncertainty quantification of its use for viruses without experimental data for evaluation. Overall, this suggests that incorporating even a limited number of viral structures can significantly enhance predictive accuracy for protein language model mutation effects in viruses.

4 Conclusion

Our comprehensive evaluation of protein mutation effect prediction models for viruses reveals key differences between alignment-based models and PLMs. While alignment-based models generally

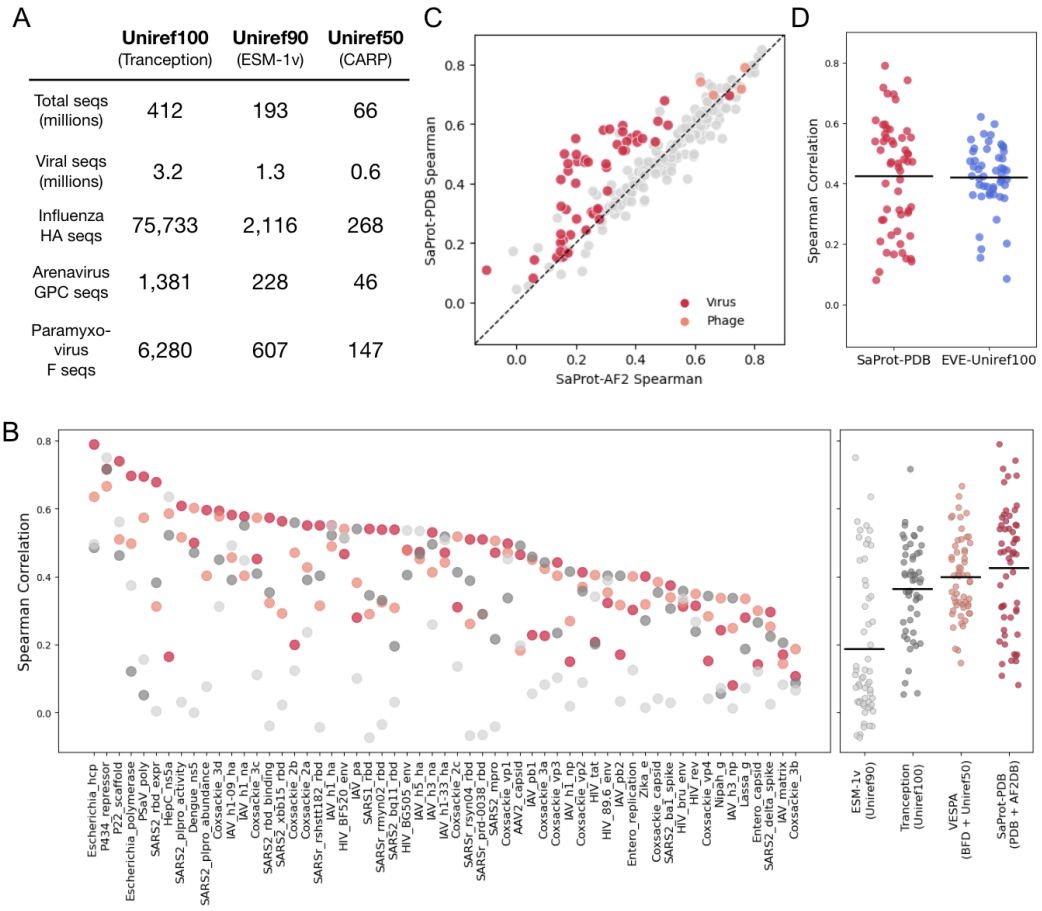


Figure 3: SaProt-PDB outperforms PLMs trained on sequences alone for viral mutation effect prediction. A. Number of sequences in Uniref100, Uniref90, and Uniref50, and for antigenic proteins of selected viruses or viral families. B. Spearman rank correlation between PLM model score and DMS score for each of 59 viral proteins. C. 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. D. Comparison of the Spearman rank correlation of SaProt-PDB to EVE trained on alignments from Uniref100 for all viral DMSs.

outperformed PLMs for viral proteins, our findings suggest that the limited viral sequence diversity in traditional training datasets, like UniRef90, may be a primary factor. Incorporating structural information through models like SaProt-PDB significantly improves prediction accuracy, especially for viruses, indicating that structure-aware training is a promising approach to overcoming the limitations of viral sequence data. These findings provide insights for future model development tailored to viral mutation prediction, which could enhance efforts in viral surveillance, aiding variant-proof vaccine and deimmunized and targeted viral vector design.

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.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Álvarez-Rodríguez, B., Velandia-Álvarez, S., Toft, C., and Geller, R. Mapping the mutational landscape of a full viral proteome reveals distinct profiles of mutation tolerability. *bioRxiv*, pp. 2024–03, 2024.

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

Bakhache, W., Orr, W., McCormick, L., and Dolan, P. T. Uncovering structural plasticity of enterovirus a through deep insertional and deletional scanning. *Research Square*, 2024.

Dadonaite, B., Crawford, K. H., Radford, C. E., Farrell, A. G., Timothy, C. Y., Hannon, W. W., Zhou, P., Andrabi, R., Burton, D. R., Liu, L., et al. A pseudovirus system enables deep mutational scanning of the full sars-cov-2 spike. *Cell*, 186(6):1263–1278, 2023.

Dadonaite, B., Ahn, J. J., Ort, J. T., Yu, J., Furey, C., Dosey, A., Hannon, W. W., Vincent Baker, A. L., Webby, R. J., King, N. P., et al. Deep mutational scanning of h5 hemagglutinin to inform influenza virus surveillance. *PLoS biology*, 22(11):e3002916, 2024.

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

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

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

Duenas-Decamp, M., Jiang, L., Bolon, D., and Clapham, P. R. Saturation mutagenesis of the hiv-1 envelope cd4 binding loop reveals residues controlling distinct trimer conformations. *PLoS pathogens*, 12(11):e1005988, 2016.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., 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.

Fernandes, J. D., Faust, T. B., Strauli, N. B., Smith, C., Crosby, D. C., Nakamura, R. L., Hernandez, R. D., and Frankel, A. D. Functional segregation of overlapping genes in hiv. *Cell*, 167(7): 1762–1773, 2016.

Flynn, J. M., Samant, N., Schneider-Nachum, G., Barkan, D. T., Yilmaz, N. K., Schiffer, C. A., Moquin, S. A., Dovala, D., and Bolon, D. N. Comprehensive fitness landscape of sars-cov-2 mpro reveals insights into viral resistance mechanisms. *Elife*, 11:e77433, 2022.

Frank, F., Keen, M. M., Rao, A., Bassit, L., Liu, X., Bowers, H. B., Patel, A. B., Cato, M. L., Sullivan, J. A., Greenleaf, M., et al. Deep mutational scanning identifies sars-cov-2 nucleocapsid escape mutations of currently available rapid antigen tests. *Cell*, 185(19):3603–3616, 2022.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021.

Haddox, H. K., Dingens, A. S., and Bloom, J. D. 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.

Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J., and Bloom, J. D. Mapping mutational effects along the evolutionary landscape of hiv envelope. *Elife*, 7:e34420, 2018.

Heredia, J. D., Park, J., Choi, H., Gill, K. S., and Procko, E. 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.

Hom, N., Gentles, L., Bloom, J. D., and Lee, K. K. 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.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P., Ingraham, J. B., Toth-Petroczy, A., Brock, K., Riesselman, A. J., Palmedo, P., et al. The evcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584, 2019.

Jiang, L., Liu, P., Bank, C., Renzette, N., Prachanronarong, K., Yilmaz, L. S., Caffrey, D. R., Zeldovich, K. B., Schiffer, C. A., Kowalik, T. F., et al. A balance between inhibitor binding and substrate processing confers influenza drug resistance. *Journal of molecular biology*, 428(3): 538–553, 2016.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

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

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

Lee, J. M., Huddleston, J., Doud, M. B., Hooper, K. A., Wu, N. C., Bedford, T., and Bloom, J. D. 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.

Lei, R., Garcia, A. H., Tan, T. J., Teo, Q. W., Wang, Y., Zhang, X., Luo, S., Nair, S. K., Peng, J., and Wu, N. C. Mutational fitness landscape of human influenza h3n2 neuraminidase. *Cell reports*, 42 (1), 2023.

Lei, R., Qing, E., Odle, A., Yuan, M., Gunawardene, C. D., Tan, T. J., So, N., Ouyang, W. O., Wilson, I. A., Gallagher, T., et al. Functional and antigenic characterization of sars-cov-2 spike fusion peptide by deep mutational scanning. *Nature communications*, 15(1):4056, 2024.

Li, Y., Arcos, S., Sabsay, K. R., Te Velthuis, A. J., and Lauring, A. S. 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.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Lu, H., Diaz, D. J., Czarnecki, N. J., Zhu, C., Kim, W., Shroff, R., Acosta, D. J., Alexander, B. R., Cole, H. O., Zhang, Y., et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022.

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

Mattenberger, F., Latorre, V., Tirosh, O., Stern, A., and Geller, R. Globally defining the effects of mutations in a picornavirus capsid. *Elife*, 10:e64256, 2021.

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

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. 2021. 2021a.

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

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

Notin, P., Kollasch, A. W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Weitzman, R., Frazer, J., et al. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023.

Qi, H., Olson, C. A., Wu, N. C., Ke, R., Loverdo, C., Chu, V., Truong, S., Remenyi, R., Chen, Z., Du, Y., 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.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L., Curtis, T., Escobar-Zepeda, A., Gurbich, T., Kale, V., Korobeynikov, A., Raj, S., Rogers, A., Sakharova, E., Sanchez, S., Wilkinson, D., and Finn, R. 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`.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, October 2018.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15), April 2021.

Setoh, Y. X., Amarilla, A. A., Peng, N. Y., Griffiths, R. E., Carrera, J., Freney, M. E., Nakayama, E., Ogawa, S., Watterson, D., Modhiran, N., et al. Determinants of zika virus host tropism uncovered by deep mutational scanning. *Nature microbiology*, 4(5):876–887, 2019.

Sinai, S., Jain, N., Church, G. M., and Kelsic, E. D. Generative aav capsid diversification by latent interpolation. *bioRxiv*, pp. 2021–04, 2021.

Soh, Y. S., Moncla, L. H., Eguia, R., Bedford, T., and Bloom, J. D. Comprehensive mapping of adaptation of the avian influenza polymerase protein pb2 to humans. *Elife*, 8:e45079, 2019.

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

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

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., 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.

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

Steinegger, M., Mirdita, M., and Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods*, 16(7):603–606, 2019.

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

Sumida, K. H., Núñez-Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I., Milles, L. F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., et al. Improving protein expression, stability, and function with proteinmpnn. *Journal of the American Chemical Society*, 146(3):2054–2061, 2024.

Suphatrakul, A., Posiri, P., Srisuk, N., Nantachokchawapan, R., Onnome, S., Mongkolsapaya, J., and Siridechadilok, B. Functional analysis of flavivirus replicase by deep mutational scanning of dengue ns5. *bioRxiv*, pp. 2023–03, 2023.

Taylor, A. L. and Starr, T. N. 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.

Teo, Q. W., Wang, Y., Lv, H., Mao, K. J., Tan, T. J., Huan, Y. W., Rivera-Cardona, J., Shao, E. K., Choi, D., Dargani, Z. T., et al. Deep mutational scanning of influenza a virus nep reveals pleiotropic mutations in its n-terminal domain. *bioRxiv*, pp. 2024–05, 2024.

Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Ritter, D., Sander, C., Gal, Y., and Marks, D. S. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, 2023.

Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., and Tawfik, D. S. Do viral proteins possess unique biophysical features? *Trends in biochemical sciences*, 34(2):53–59, 2009.

Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J. J., Mangan, N. M., Ovchinnikov, S., and Rocklin, G. J. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

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

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

Welsh, F. C., Eguia, R. T., Lee, J. M., Haddox, H. K., Galloway, J., Chau, N. V. V., Loes, A. N., Huddleston, J., Timothy, C. Y., Le, M. Q., et al. Age-dependent heterogeneity in the antigenic effects of mutations to influenza hemagglutinin. *Cell Host & Microbe*, 32(8):1397–1411, 2024.

Wu, N. C., Young, A. P., Al-Mawsawi, L. Q., Olson, C. A., Feng, J., Qi, H., Chen, S.-H., Lu, I.-H., Lin, C.-Y., Chin, R. G., et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports*, 4(1):4942, 2014.

Wu, N. C., Olson, C. A., Du, Y., Le, S., Tran, K., Remenyi, R., Gong, D., Al-Mawsawi, L. Q., Qi, H., Wu, T.-T., et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS genetics*, 11(7):e1005310, 2015.

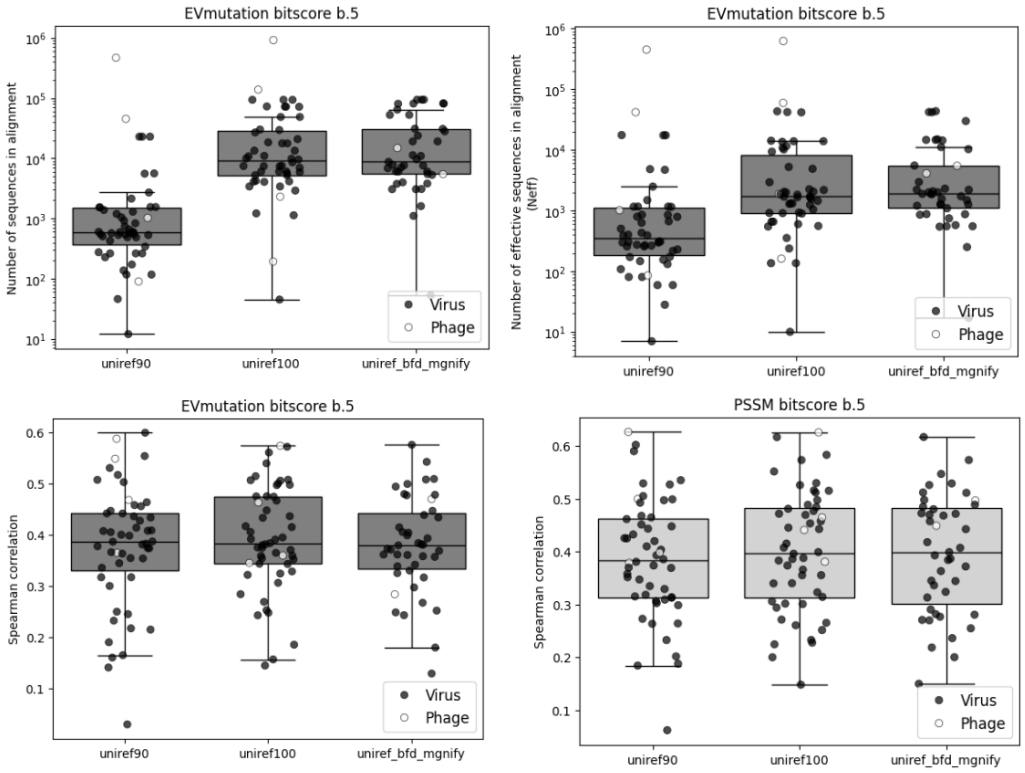
Wu, N. C., Du, Y., Le, S., Young, A. P., Zhang, T.-H., Wang, Y., Zhou, J., Yoshizawa, J. M., Dong, L., Li, X., et al. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza a virus m segment. *BMC genomics*, 17:1–15, 2016.

Wu, X., Go, M., Nguyen, J. V., Kuchel, N. W., Lu, B. G., Zeglinski, K., Lowes, K. N., Calleja, D. J., Mitchell, J. P., Lessene, G., 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.

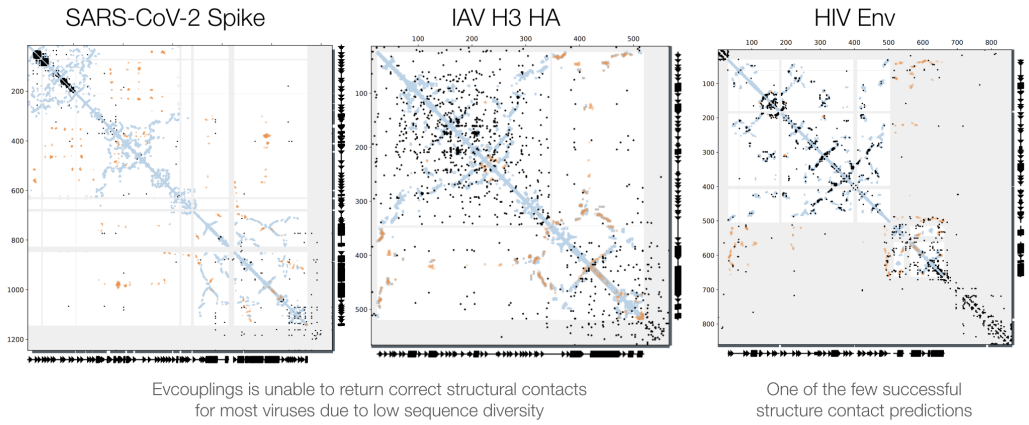
Yang, K. K., Fusi, N., and Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

Youssef, N., Gurev, S., Ghantous, F., Brock, K., Jaimes, J. A., Thadani, N., Dauphin, A., Sherman, A., Yurkovetskiy, L., Soto, D., Estaboulieh, R., Kotzen, B., Notin, P., Kollasch, A., Cohen, A., Dross, S., Erasmus, J., Fuller, D., Bjorkman, P., Lemieux, J., Luban, J., Seabman, M., and Marks, D. Protein design for evaluating vaccines against future viral variation. *BioRxiv*, 2024.

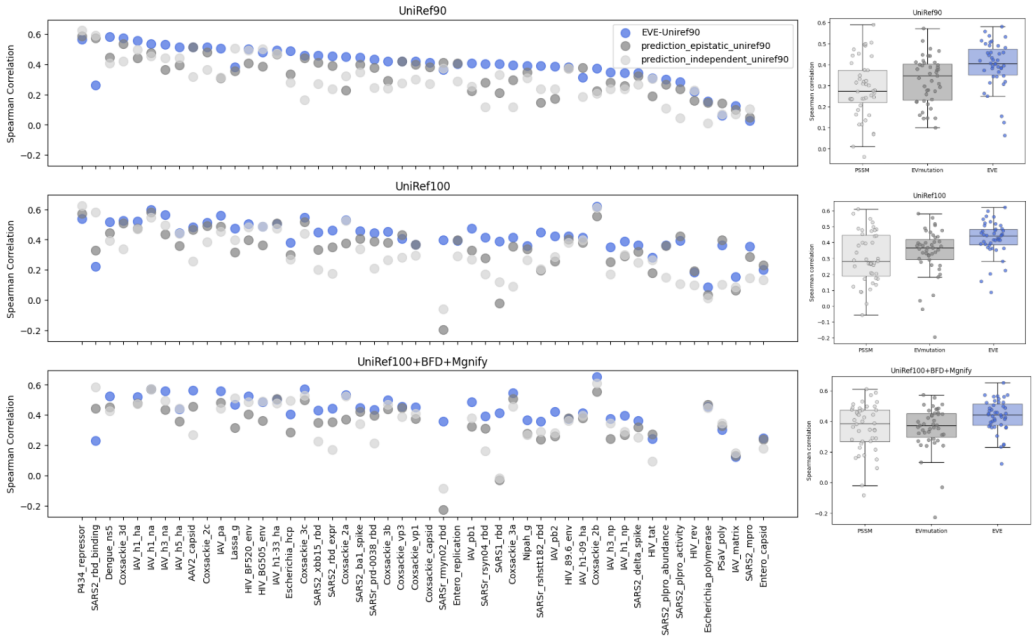
A Supplementary Figures



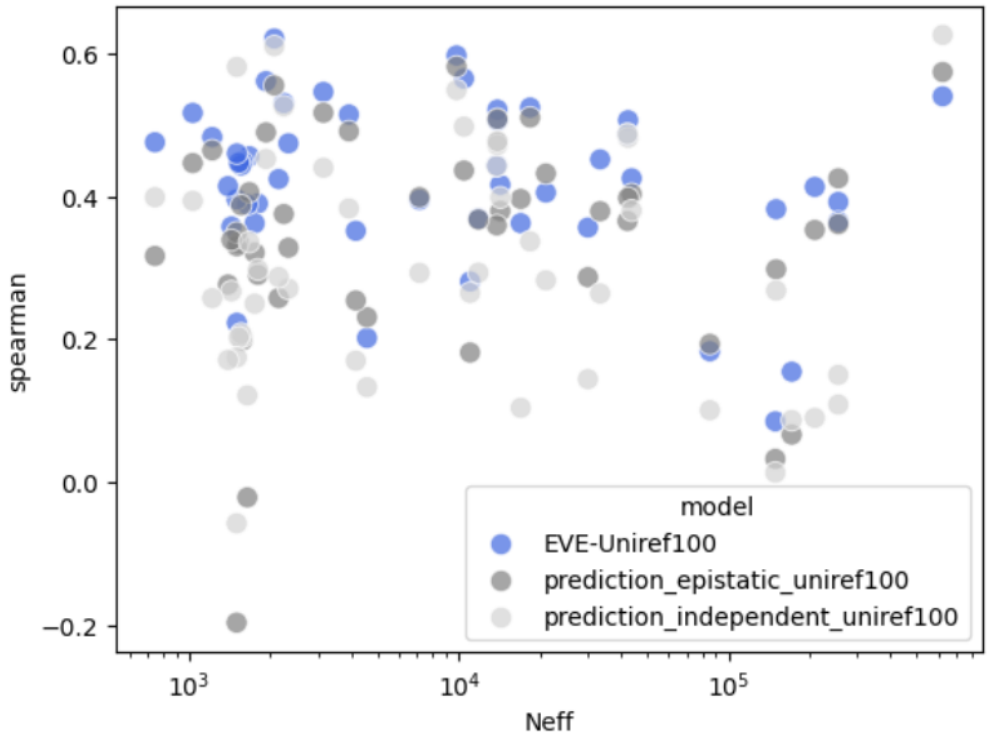
Supplementary Figure 1: Impact of sequence database choice (UniRef100, UniRef90, or UniRef100+BFD+Mgnify) on alignment depth and model performance across viral DMSs.



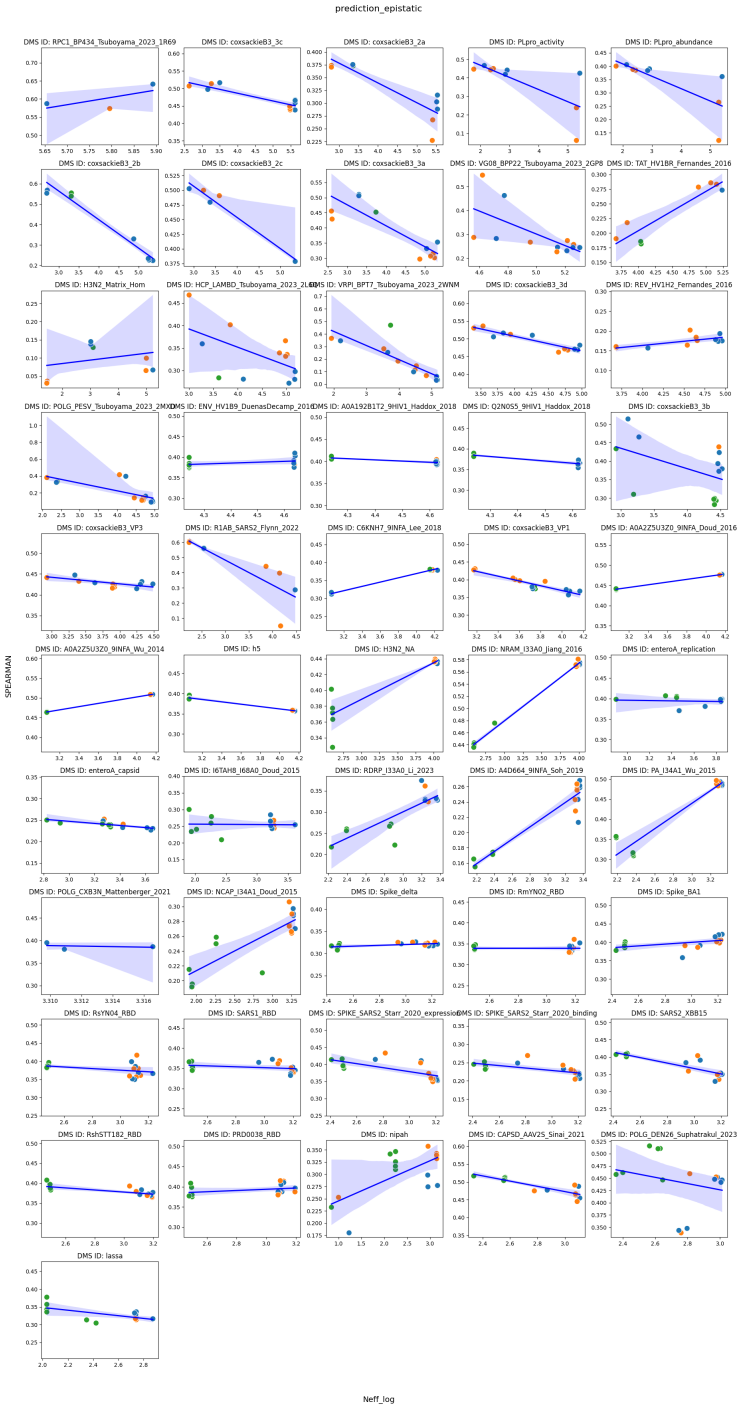
Supplementary Figure 2: Selected contact maps from Uniref100 alignments with bit score 0.1 for SARS-CoV-2 Spike, Influenza H3 and HIV Envelope proteins. HIV Env is one of the few examples of viral proteins with a DMS where EVcouplings successfully predict structure contacts, while ability to predict structural contacts is a strong predictor of model performance for other taxa.



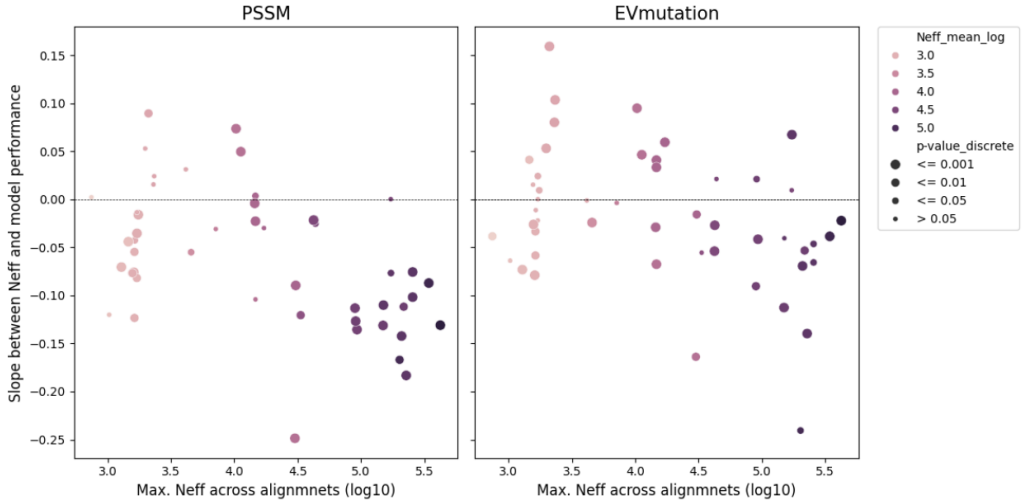
Supplementary Figure 3: EVE consistently outperforms other alignment-based models (EVmutation and PSSM) across sequence databases (rows; UniRef90, UniRef100, or UniRef100+BFD+Mgnify) for viral DMSs.



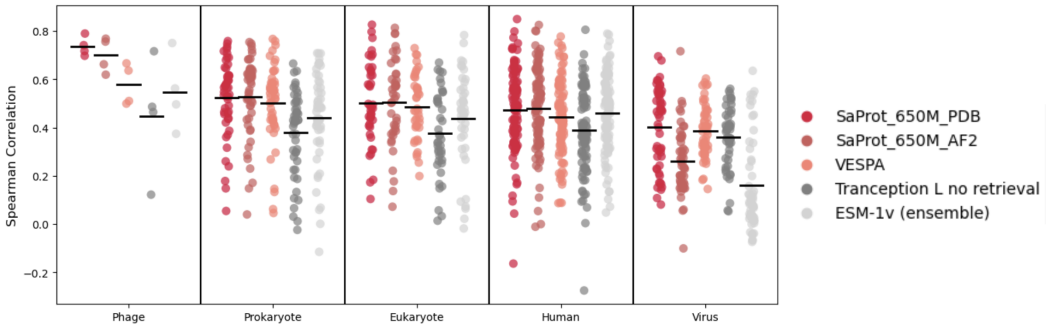
Supplementary Figure 4: Alignment-based model performance across viral DMSs does not depend on alignment depth (Neff).



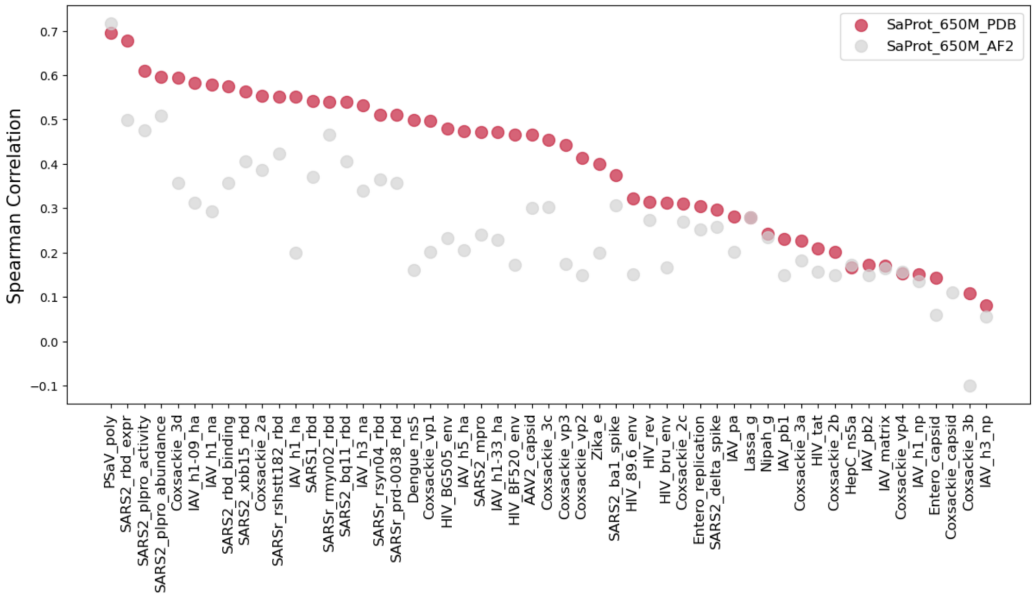
Supplementary Figure 5: Relationship between Neff and model performance (Spearman rank correlation) between model scores and DMS.



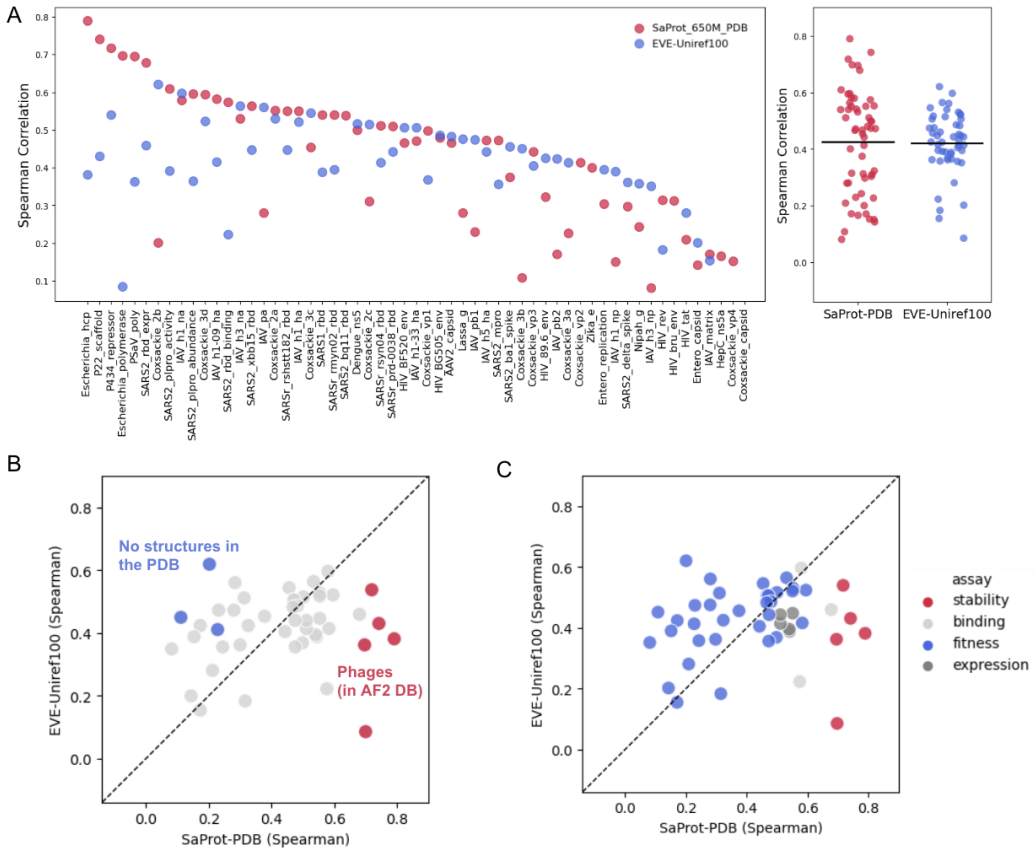
Supplementary Figure 6: Increasing alignment depth worsens model performance for some viral proteins. Y-axis plots the slope between model performance and Neff. X-axis plots the maximum Neff for a given viral protein.



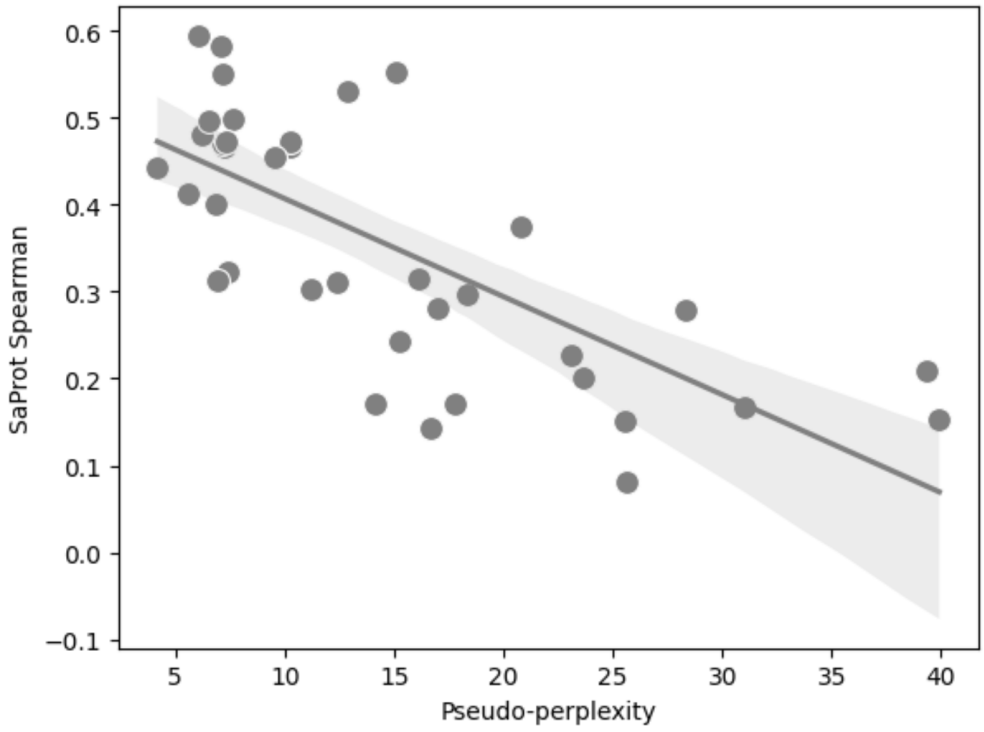
Supplementary Figure 7: Comparison of the Spearman rank correlation of all PLM models across DMSs separated by taxa.



Supplementary Figure 8: 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.



Supplementary Figure 9: A. Comparison of the Spearman rank correlation of the SaProt-PDB model to the EVE model trained on alignments from Uniref100 for all viral DMSs. Notably, some viruses do not have EVE scores given the memory constraints. The two models perform on par with the same average performance. B. The two models perform best on different DMSs. EVE is better at proteins with no structures within 90% identity in the PDB while SaProt is better on phages (with structures in the AF2 Database, while other eukaryotic viruses have been excluded). C. Models also perform best on different assay types, with EVE better at predicting overall fitness in replication or infectivity assays. SaProt's seemingly superior performance on stability assays must however be dissociated from the fact that the viral DMSs with stability assays are also the only phages, for which there is much more structural training data.



Supplementary Figure 10: SaProt pseudo-perplexity, used as a measure of uncertainty for PLMs, of the wildtypes for each of the viral fitness DMSs is a good predictor of SaProt mutation effect performance for that virus.

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 59 viral datasets.

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.

C Extended Methods

C.1 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; Ashenberg 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). 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.2 Alignment-based models

C.2.1 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.2.2 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.2.3 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.2.4 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.3 Protein language models

C.3.1 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.3.2 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.3.3 ESM-1v

ESM-1v (Meier et al., 2021b) 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.3.4 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.3.5 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.