Protein language model rescue mutations highlight variant effects and structure in clinically relevant genes

Onuralp Soylemez[†]
onuralp@gmail.com

Pablo Cordero[†]
pablo@stripe.com

Abstract

Despite being self-supervised, protein language models have shown remarkable performance in fundamental biological tasks such as predicting impact of genetic variation on protein structure and function. The effectiveness of these models on diverse set of tasks suggests that they learn meaningful representation of fitness landscape that can be useful for downstream clinical applications. Here, we interrogate the use of these language models in characterizing known pathogenic mutations in curated, medically actionable genes through an exhaustive search of putative compensatory mutations on each variant's genetic background. Systematic analysis of the predicted effects of these compensatory mutations reveal unappreciated structural features of proteins that are missed by other structure predictors like AlphaFold. While deep mutational scan experiments provide an unbiased estimate of the mutational landscape, we encourage the community to generate and curate rescue mutation experiments to inform the design of more sophisticated co-masking strategies and leverage large language models more effectively for downstream clinical prediction tasks.

1 Introduction

Understanding the effects of genetic variation in modulating disease is a central task in clinical genetics and genomic medicine, where the ultimate goal is to detect, quantify, and characterize the pathogenicity of particular mutations to elicit correct diagnosis and inform treatments. Recent advances in protein sequence and structure modeling are beginning to show promise to aid in this task. Protein language models have been shown to harbor variant effect information and high accuracy protein structure prediction [1] has vastly expanded the ways protein structure can be used in connecting genetic variation with disease effects.

More precisely, recent work ([2], [3], [4]) has shown that protein language models can effectively model deep mutational scan data from extensive genotype-phenotype mapping (e.g., fitness landscape of green fluorescent protein [5]) and are also capable of predicting pathogenicity of disease-associated mutations ([6], [7]) without further training, simply relying on the underlying patterns mined by self-supervised language modeling. Here, we expand on these trends by interrogating protein language models through *in silico* mutation.

We leverage an evolutionary insight from compensatory molecular evolution and describe a novel approach to recovering spatial features of protein structures. To ground the approach, we compare compensatory scores with compensated pathogenic deviations (CPDs): pathogenic amino acid substitutions in humans where the human pathogenic state appears to be wild-type in a functionally-equivalent protein from an orthologous species without any drastic fitness impact on the latter genetic background ([8], [9]). Additionally, we delve into the patterns of such protein language model

[†] These authors contributed equally.

36th Conference on Neural Information Processing Systems (NeurIPS 2022),
Learning Meaningful Representations of Life (LMRL) Workshop.

rescue mutations and find that they segregate guided by protein structure and can sometimes pinpoint structural features missed by structure predictors like AlphaFold.

2 Data and Methods

Clinically relevant genetic variants. We retrieved the latest list of medically actionable genes curated by the American College of Medical Genetics and Genomics (ACMG)[10]. These genes harbor high penetrance, large effect pathogenic mutations associated with clinically actionable medical conditions. We limited our analysis to 53 genes with less than 1024 amino acid residues in length in line with the ESM-1v pre-training setup [7] (see Appendix for the gene list). For each gene, we extracted ClinVar [11] variants with their corresponding clinical significance annotation, and grouped the variant impact into three categories: pathogenic/likely pathogenic (P/LP), benign/likely benign (B/LB) and variants of unknown significance (VUS). In case of conflicting interpretations of pathogenicity, annotation with the higher number of ClinVar submissions is considered.

For each ClinVar variant, we parsed the global allele frequency for the mutant pathogenic allele from the gnomAD v2.1 dataset that contains human genetic variation data from 125,748 whole exomes and 15,708 whole genomes [12]. In case of multiple nucleotide changes corresponding to the same protein change, we kept the allele frequency for the more common alternate allele.

Compensated pathogenic deviations (CPDs). We compiled a list of compensated pathogenic deviations, where the pathogenic ClinVar variant in human is conserved in an ortholog in another species. Presence of such substitutions - even in highly conserved regions - suggest that there must be other amino acid changes either within the same protein or in an interacting protein to mitigate the fitness impact, and these interdependent or epistatic interactions may represent functional or spatial constraints on the corresponding sites. Specifically, using highly conserved multiple sequence alignment from placental mammals, we identified sites where the disease associated ClinVar amino acid state (e.g., His in Arg13His substitution) appears to be wild-type in a functionally equivalent (orthologous) sequence in at least one placental mammal. To assess the relevance of sequence context in the vicinity of the pathogenic mutation, we also identified CPDs where neighboring amino acid residues are completely conserved in the placental mammal phylogeny, likely corresponding to regions of functional importance.

Fitness impact of secondary mutations. We use the ESM protein language model to score putative secondary mutations on the genetic background of a known disease mutation. Specifically, we interrogate the log odds of each amino acid in each sequence position of the protein language model under the background of the genetic variant of interest (the so-called "wild-type marginal" effect). Independently, such a score highlights the fitness impact of secondary mutations and identifies rescue mutations whose average fitness impact can compensate the fitness reduction caused by the original mutant. In aggregate, summary statistics of such scores in any given position can yield signals of gain or loss of fitness as a result of the background, pathogenic genetic variant and may lead to to insights into structure. Thus, we also consider the z-score across background mutation position of these ESM scores to compare the fitness impact of that background mutation against all others.

3 Results

Stability of predictions between different language models. For any useful clinical application, it is important that pretrained protein language models with different modeling perplexity yield robust predictions. Here we evaluate the consistency of the pathogenicity predictions from two state-of-the-art transformer protein language models. Specifically, we score the pathogenicity of ClinVar variants in *LDLR* gene using ESM-1v and ESM-2 pre-trained models, and show that there is very strong correlation among the two predictors (Pearson's correlation coefficient r=0.91) (see Figure 1). While the correlation attenuates when all ClinVar variants across the entire gene list are considered, two models generate consistent clinical significance annotations (see Supplementary Figure 1).

Notably, variants with unknown significance (VUS) and pathogenic variants (P/LP) show relatively moderate correlation (Pearson's r of 0.60 and 0.59, respectively) when compared to correlation for benign variants (B/LB) between the two language models. Moreover, we found that pre-trained

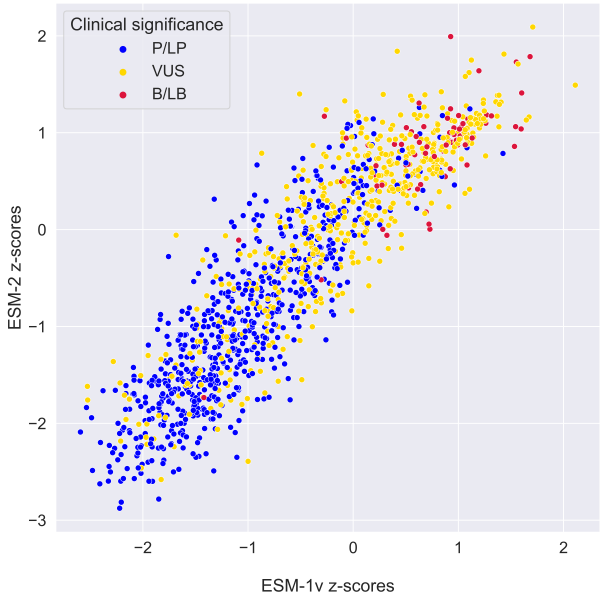


Figure 1: ESM scores are reasonably stable between model versions (Pearson's correlation coefficient (r) of 0.91). Here, we compare normalized ESM scores of ESM-1v and ESM-2 for the low-density lipoprotein receptor (*LDLR*) gene which has the highest number of mutations with ClinVar annotations available in ACMG dataset. ESM-1v scores are averaged across five models.

models with more layers scored a subset of VUS in BAG3 more pathogenic. These findings underscore the importance of model selection for *in silico* pathogenicity predictions of clinically relevant variants with unknown significance in medically actionable genes such as *BAG3* (see Supplementary Figure 3).

Variant pathogenicity predictions. Due to natural selection pressure on deleterious alleles that reduce fitness, we expect the predicted pathogenicity scores for the pathogenic missense variants (P/LP) to be inversely correlated with the allele frequency of the mutant alelle. We mapped each ClinVar variant to available large-scale human genetic variation data in gnomAD database, and found that the language model variant impact predictions are consistent with the prevalence of corresponding mutant alleles in the general population (See Figure 2). As expected, predicted scores for the benign missense variants (B/LB) do not show the same correlation. Prediction scores calculated using ESM-2 model show the same trend (see Supplementary Figure 2).

Sequence context around putative rescued sites. Structural and biochemical analysis of compensated pathogenic mutations (CPDs) found previously that CPDs are on average less deleterious than non-compensated pathogenic mutations [13]. To test whether language model predictions recapitulate this empirical observation, we identified CPDs as described in Methods and compared the prediction scores for CPDs against non-CPDs. We found that the distributions of ESM-2 prediction scores of CPDs and non-CPDs are statistically different (Mann-Whitney-Wilcoxon two-sided test p-value < 1e-03), and CPDs appear to be more tolerated than non-CPDs in line with the biochemical analysis (see Supplementary Figure 4).

Taking into consideration potential misalignment errors in multiple sequence alignments, we repeated the analysis limiting CPDs to sites where the neighboring sites are required to be fully conserved across the phylogeny. Interestingly, we did not observe a statistically significant difference between CPDs and non-CPDs when using such constraint on local homology. This finding may suggest that the existing protein language models may be limited to capture local sequence context when putative compensated mutations are present, and further hint at potential gains from more sophisticated co-masking strategies during pre-training. Modest correlation between ESM-2 model scores and Cross-Protein Transfer (CPT) model [14] scores highlight the importance of better understanding the subset of predictions where these models do not agree, and compensatory framework may be helpful

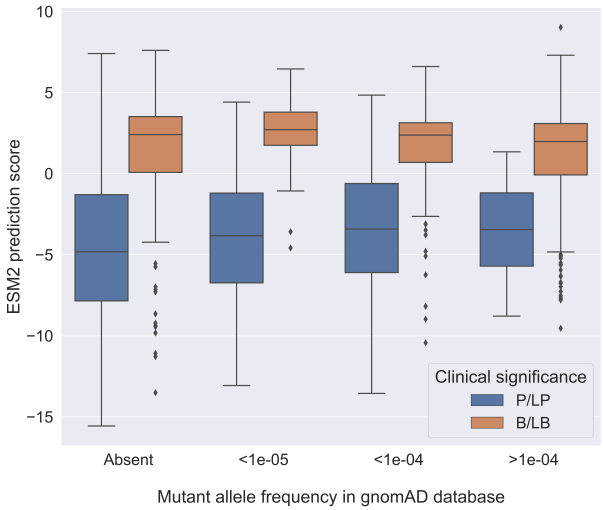


Figure 2: ESM-2 differentiates between pathogenic and benign genetic variants in clinically-relevant genes and this effect grows weaker as the variants are more common in the general population. For all pairwise comparisons, two-sided Mann-Whitney U test p-values < 1e-05.

to diagnose the inherent limitations of protein language models for resulting in such disagreements (see Supplementary Figure 5).

Rescue mutation effects reveal unappreciated structural features . We next interrogated whether simple summary statistics of mutations conditioned against a background variant along the protein sequence revealed any informative compensatory features. This follows the same intuition of residue co-evolution, where compensatory effects arise from sequence perturbations and can crucially inform downstream tasks such as structure prediction. We took the mean wild type marginal score per position, per background variation and z-scored them across all genetic variants. Plotting these scores in aggregate revealed patterns that matched predicted contact maps of the protein's AlphaFold-predicted structures, confirming the functional relevance of these compensatory effects and in line with previous results observing that protein language models can be unsupervised structure predictors. Crucially, in some cases we observed that compensatory effects of these rescue mutations predicted structural features that may have been missed in AlphaFold. For example, the myopathy-related BAG Cochaperone 3 (BAG3) gene is predicted to be mostly unstructured save for one small BAG domain by AlphaFold – the same domain that has been experimentally characterized. Rescue mutation effects reveal compensatory changes within this supposedly unordered region that harbors multiple variation of unknown significance. Further, comparing ESMfold vs AlphaFold structures confirm that the AlphaFold structure is mostly predicted to be disordered while ESMfold shows a more compact structure with more structural features.

4 Discussion and Future Directions

Our findings add additional support to the promise of protein language models as tools for interrogating possibly pathogenic genetic variation, following lines of evidence in deep mutational scanning and genome-wide scoring of genetic variants. While extensive large-scale experimental measurements in deep mutational scan data sets provide clear and robust genotype-phenotype maps, curation of clinically relevant genetic variants poses significant challenges. For example, a non-trivial fraction of pathogenic or likely pathogenic ClinVar variants may have incomplete penetrance and therefore their pathogenicity may be highly context dependent. We focused on an expert curated subset of medically actionable genes to enrich our genetic variant dataset for better studied variants with well-established genotype-phenotype associations. Likewise, medical phenotypes in ACMG gene list include diseases with highly complex diagnosis criteria, and it is challenging to establish causal associations with every single genetic variant in these genes. Population scale biomedical databases and biobanks such

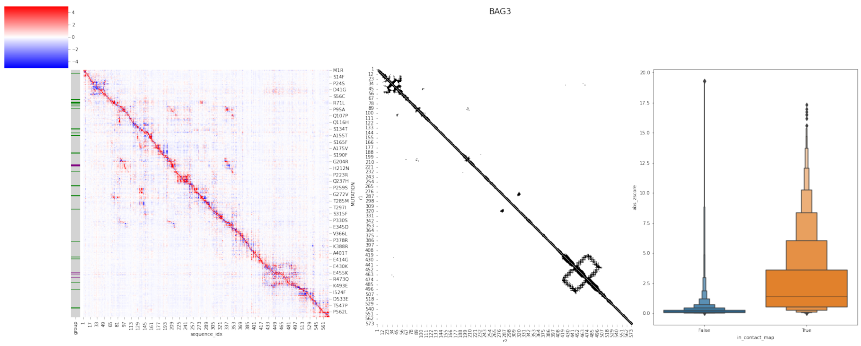


Figure 3: Summary statistics of secondary mutation effects on a variant genetic background segregate in structural features and resemble a contact map. Normalized mean of mutation effects in each position of the BAG Cochaperone 3 (*BAG3*) gene (left) bring out structural patterns consistent with the 10 angstroms contact map of the predicted AlphaFold structure (center; see square pattern corresponding to the BAG domain) and highlight additional, potentially missed structural patterns in regions deemed disordered by AlphaFold. High and low effects tend to segregate within the contact map (right)

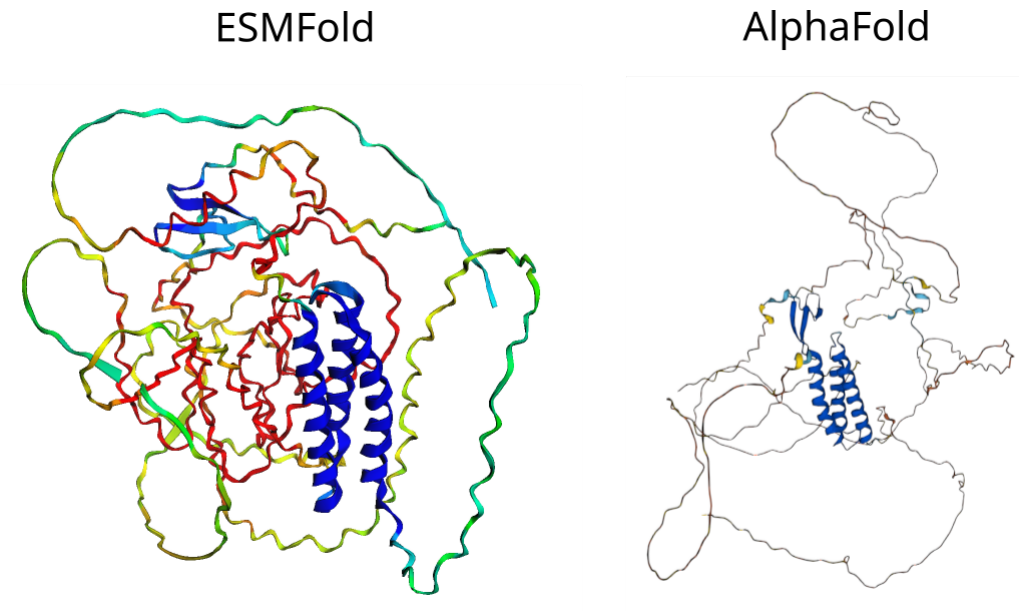


Figure 4: *BAG3* predicted structures via ESMfold (left) and AlphaFold (right). ESMfold picks up more structural features than AlphaFold. Both structures are colored by confidence, from blue (high confidence) to yellow (medium confidence) to red (very low confidence)

as UK Biobank ([15]) provide an opportunity to refine the curation the genetic and phenotypic data from hundreds of thousands of individuals.

Additionally, we explore an *in silico* search of compensatory mutations using protein language models as a means to further characterize the effect of genetic variation. This compensatory score maps reveal potentially unappreciated structural features in some cases. For this work, we only considered single rescue mutations, however, it is conceivable that compensatory interactions may involve more than one amino acid substitution within the same protein, co-evolving changes in the interacting protein partners, or more subtle synonymous changes affecting the secondary structure. It remains elusive to what extent incorporating higher order dependencies between sites can help improve the predictive accuracy of protein language models or highlight any potential limitation. Masked language models provide a convenient extended co-masking strategy to probe the relevance of higher order interactions. Rescue experiments offer a powerful framework to diagnose the limitations of large language models to capture clinically relevant aspects of complex fitness landscapes. While deep mutational scan experiments provide an unbiased estimate of the mutational landscape, we encourage the community to generate and curate rescue mutation experiments to inform the design of more sophisticated co-masking strategies and leverage large language models more effectively for downstream clinical prediction tasks.

5 Data and code availability

All the data used in this paper are publicly available. Details on datasets, models and analysis code can be found at `https://github.com/dimenwarper/llm-for-clinical-variants`.

References

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[2] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, pp. 2102–2110, 02 2022.

[3] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[4] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks, "Disease variant prediction with deep generative models of evolutionary data," *Nature*, vol. 599, no. 7883, pp. 91–95, 2021.

[5] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, *et al.*, "Local fitness landscape of the green fluorescent protein," *Nature*, vol. 533, no. 7603, pp. 397–401, 2016.

[6] N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos, "Genome-wide prediction of disease variants with a deep protein language model," *bioRxiv*, 2022.

[7] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *bioRxiv*, 2021.

[8] A. S. Kondrashov, S. Sunyaev, and F. A. Kondrashov, "Dobzhansky–muller incompatibilities in protein evolution," *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 14878–14883, 2002.

[9] O. Soylemez and F. A. Kondrashov, "Estimating the Rate of Irreversibility in Protein Evolution," *Genome Biology and Evolution*, vol. 4, pp. 1213–1222, 11 2012.

[10] D. T. Miller, K. Lee, N. S. Abul-Husn, L. M. Amendola, K. Brothers, W. K. Chung, M. H. Gollob, A. S. Gordon, S. M. Harrison, R. E. Hershberger, *et al.*, "Acmg sf v3. 1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the american college of medical genetics and genomics (acmg)," 2022.

[11] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, *et al.*, "Clinvar: improving access to variant interpretations and supporting evidence," *Nucleic acids research*, vol. 46, no. D1, pp. D1062–D1067, 2018.

[12] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, *et al.*, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[13] C. Ferrer-Costa, M. Orozco, and X. de la Cruz, "Characterization of compensated mutations in terms of structural and physico-chemical properties," *Journal of Molecular Biology*, vol. 365, no. 1, pp. 249–256, 2007.

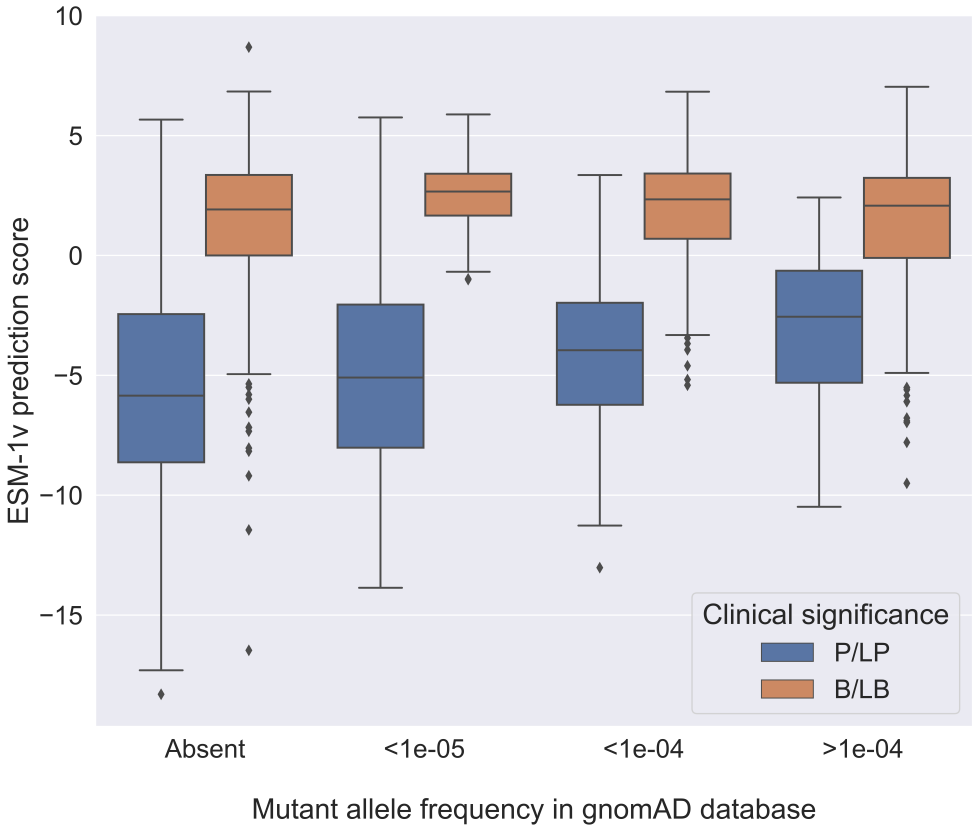
[14] M. Jagota, C. Ye, R. Rastogi, C. Albors, A. Koehl, N. Ioannidis, and Y. S. Song, "Cross-protein transfer learning substantially improves zero-shot prediction of disease variant effects," *bioRxiv*, 2022.

[15] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, *et al.*, "The uk biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

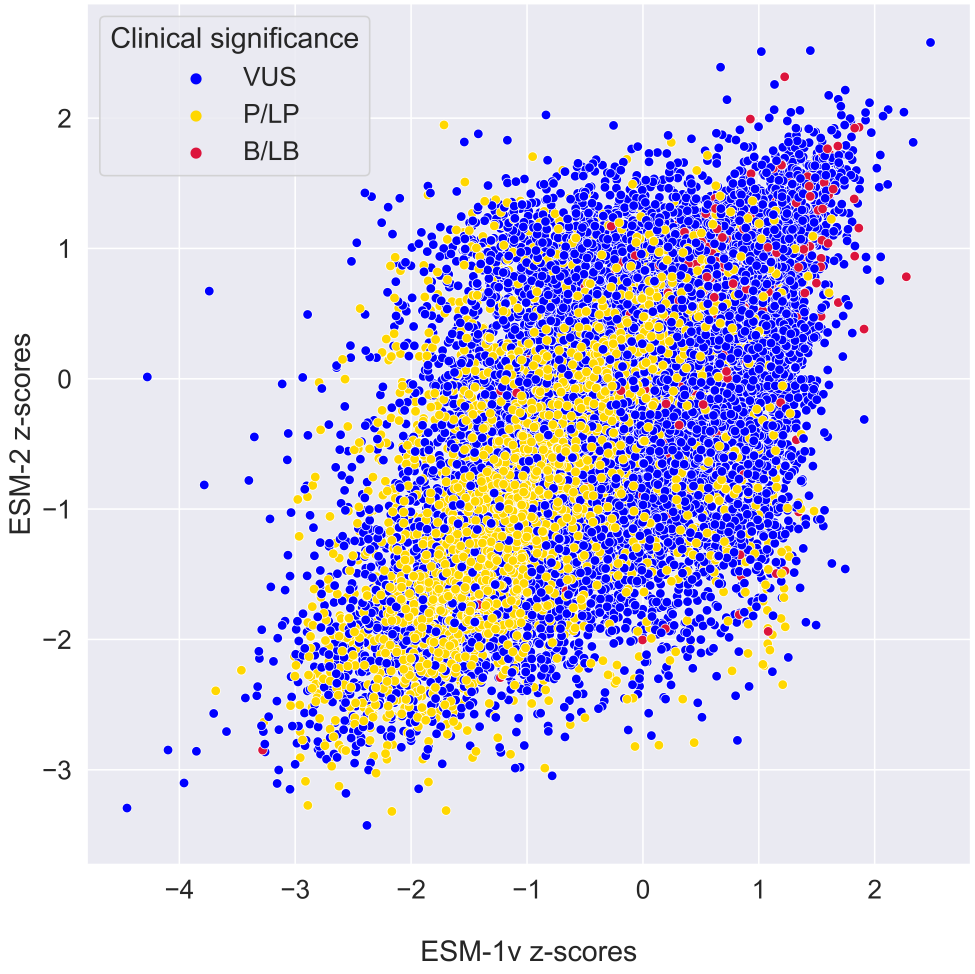
A Appendix

List of ACMG genes considered in this study:

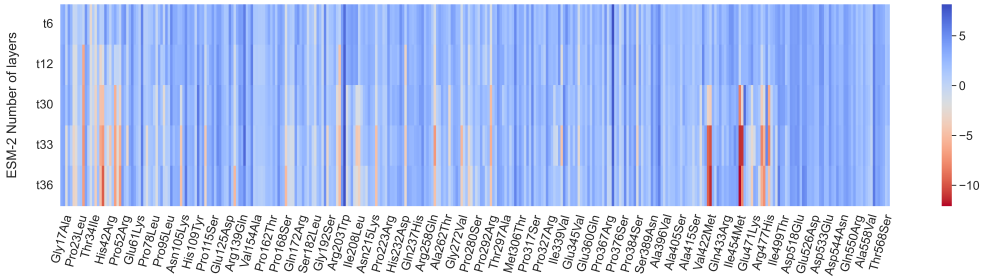
ACTA2	ACTC1	ACVRL1	BAG3	BMPR1A	BTD
CASQ2	DES	DSC2	ENG	GAA	GLA
HFE	HNF1A	KCNQ1	LDLR	LMNA	MAX
MEN1	MLH1	MSH2	MUTYH	MYL2	MYL3
NF2	OTC	PCSK9	PKP2	PRKAG2	PTEN
RB1	RPE65	SDHAF2	SDHB	SDHC	SDHD
SMAD3	SMAD4	STK11	TGFBR1	TGFBR2	TMEM127
TMEM43	TNNC1	TNNI3	TNNT2	TP53	TPM1
TRDN	TTR	VHL	WT1		



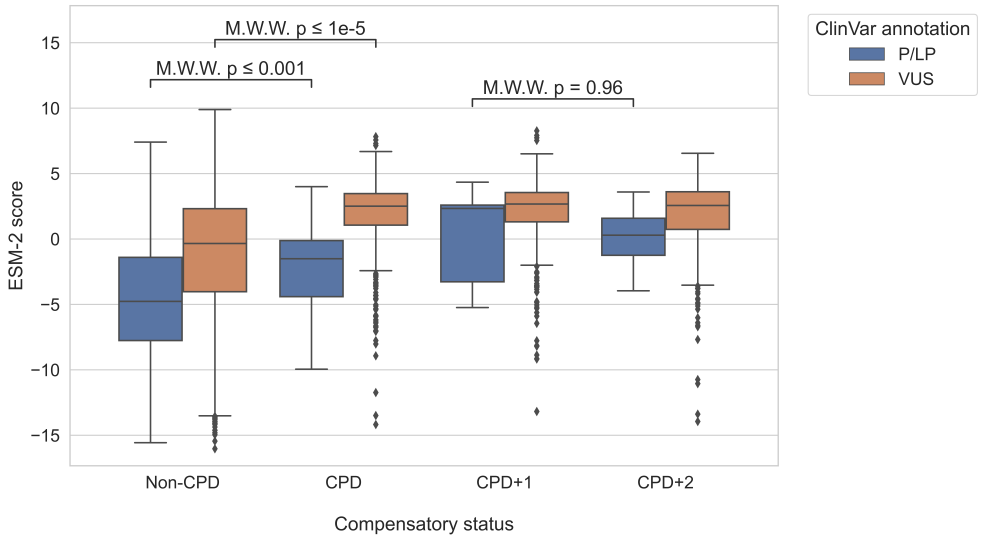
Supplementary Figure 1: Distribution of ESM-1v prediction scores for pathogenic (P/LP) and benign (B/LB) ClinVar variants group by their respective population allele frequencies in the gnomAD genetic variation database. ESM-1v prediction scores are the average ensemble score of five models. For all pairwise comparisons, two-sided Mann-Whitney U test p-values < 3.5e-05.



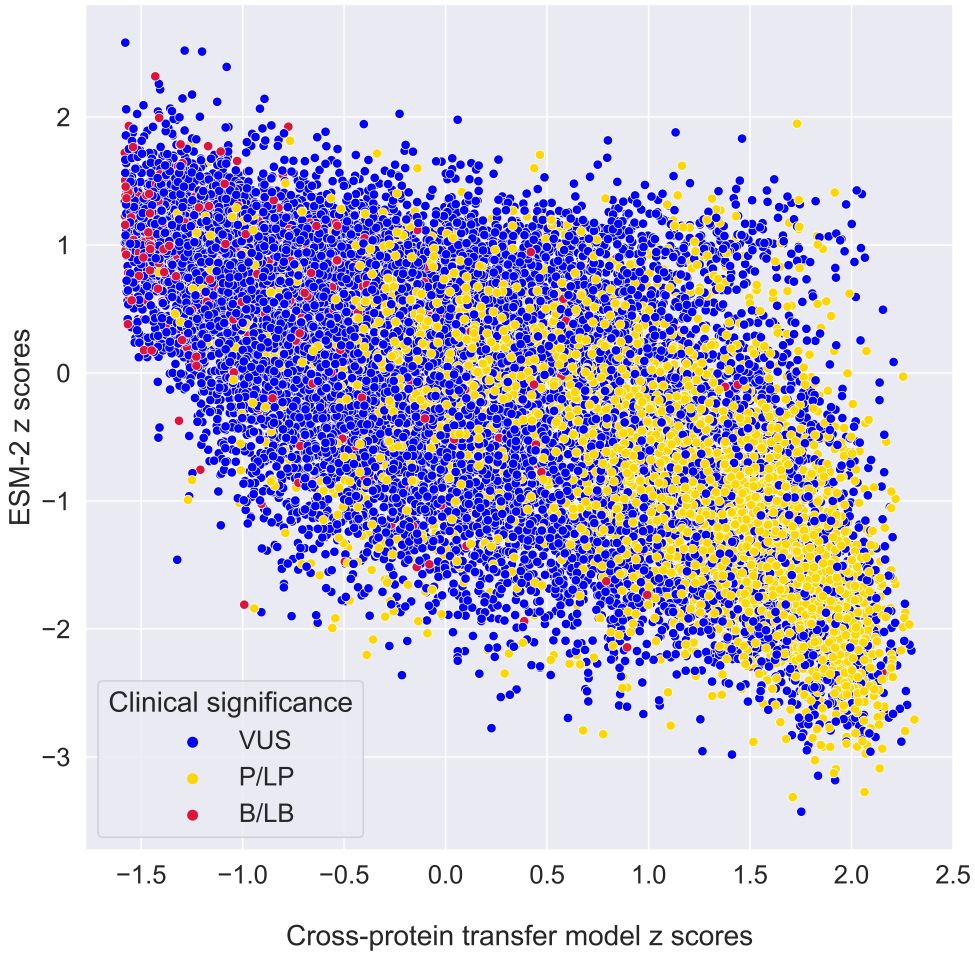
Supplementary Figure 2: Comparison of ESM-1v and ESM-2 prediction scores as normalized across the entire predictions among ClinVar variants in 53 ACMG genes. Pearson's correlation of r=0.66



Supplementary Figure 3: Comparison of ESM-2 prediction scores for *BAG3* variants with unknown significance (VUS) across ESM-2 pre-trained models with varying number of layers.



Supplementary Figure 4: ESM-2 scores for pathogenic variants (P/LP) and variants of unknown significance (VUS) at putative compensated sites. CPD refers to sites where the mutant residue is present in at least one non-human species. CPD+1 and CPD+2 refer to CPDs where the neighboring residue or two residues, respectively, are required to be fully conserved. Mann-Whitney-Wilcoxon (M.W.W) two-sided test p-values are shown.



Supplementary Figure 5: Comparison of ESM-2 and Cross-Protein Transfer (CPT)[14] model prediction scores as normalized across the entire predictions among ClinVar variants in 51 ACMG genes for which both models have predictions available. Spearman's rank correlation between the two predictors is ρ=0.69.