Cross-protein transfer learning substantially improves zero-shot prediction of disease variant effects

Milind Jagota[1,*], Chengzhong Ye[2,*], Ruchir Rastogi[1], Carlos Albors[1], Antoine Koehl[2], Nilah Ioannidis[1,3], and Yun S. Song[1,2,†]

[1]Computer Science Division, University of California, Berkeley, CA 94720
[2]Department of Statistics, University of California, Berkeley, CA 94720
[3]Chan Zuckerberg Biohub, San Francisco, CA 94158

Abstract

Genetic variation in the human genome is a major determinant of individual disease risk, but the vast majority of missense variants have unknown etiological effects. Various computational strategies have been proposed to predict the effects of missense variants across the human proteome, using many different predictive signals. Here, we present a robust learning framework for leveraging functional assay data to construct computational predictors of disease variant effects. We train cross-protein transfer (CPT) models using deep mutational scanning data from only five proteins and achieve state-of-the-art performance on unseen proteins across the human proteome. On human disease variants annotated in ClinVar, our model CPT-1 improves specificity at 95% sensitivity to 64%, from 31% for ESM-1v and 50% for EVE. Our framework combines general protein sequence models with vertebrate sequence alignments and AlphaFold2 structures, and it is adaptable to the future inclusion of other sources of information. We release predictions for all missense variants in 90% of human genes. Our results establish the utility of functional assay data for learning general properties of variants that can transfer to unseen proteins.

*These authors contributed equally to this work.
†To whom correspondence should be addressed: yss@berkeley.edu

Introduction

Variation in the human genome across individuals is a major determinant of differences in disease risk, and exponential decreases in sequencing costs have made it feasible to measure genotypes of individual patients. To be able to make accurate and targeted medical decisions based on genetic information, we need to understand the etiological consequences of human genome variants. Missense variants, which modify the amino acid at a single position of a protein, are of particular interest because of the central role of proteins in cellular function. Tens of millions of missense variants are likely to exist in the human population, and the vast majority of these have unknown consequences [1, 2]. Efforts to collect population genomic data relating variants to disease phenotypes have made progress on this problem [3, 4]. However, many of the missense variants in the human population only exist in a tiny fraction of individuals and may not clearly present their disease consequences, thus limiting the ability of population genomic data to inform variant effects. Functional assays via deep mutataional scanning (DMS) experiments have been used to measure the effects of missense variants in non-human systems at higher throughput [5–7]. However, these approaches still do not directly scale to the whole human proteome, and depend on the ability to design a relevant assay for each protein of interest.

There has also been significant interest in developing computational methods to predict the effects of missense variants [8–15]. Computational methods can provide predictions for all possible mutations across the human proteome and have proven to be effective predictors of variant pathogenicity. Recently, the methods EVE [8] and ESM-1v [9] have been demonstrated to achieve state-of-the-art performance in human disease variant prediction and functional assay prediction, respectively. EVE and ESM-1v achieve strong performance despite not training on human clinical data or functional assays. Instead, the underlying principle of these models is to collect large databases of natural protein sequences, then model the probability distribution of how protein sequences vary. However, these models also have limitations. EVE and ESM-1v model protein sequence variation across all known species, and employ redundancy filtering so that variation between highly similar protein sequences (such as from related species) is not considered. This approach allows these methods to effectively capture broad constraints of protein families, such as those imposed by a common structural fold [16–19]. However, the effects of missense mutations often involve aspects of protein function that are specific to humans or species similar to humans. EVE, ESM-1v, and related methods can struggle to model such system-specific details, as shown through the large positive impact of integrating direct functional data for a protein and system of interest [20–22].

We postulated that methods such as EVE and ESM-1v could be improved by combining them with more human-specific sources of information. Previous methods have trained ensemble models using clinical annotations or population genomic information [10–13], but such models can be inaccurate due to unknown biases in the collection of these data. It is also unclear how such models will generalize to unseen proteins that had no variants in the training data, since these models typically use data from many or most human proteins. To produce a broadly applicable model that would generalize to diverse, unseen target proteins, we instead developed a robust learning framework to train disease variant predictors using functional assay data from a small number of proteins (Figure 1). Functional assay data measure many variants per protein, allowing us to obtain sufficient data from very few proteins while saving the vast majority of the human proteome as a fully unseen evaluation. These data are also typically exhaustive (as in DMS) or generated with a well-defined process for choosing variants, avoiding biases of clinical data. In our work, we trained cross-protein transfer (CPT) models using DMS data from only five proteins, all from the same functional assay relevant to human pathogenicity, and achieved significantly improved performance

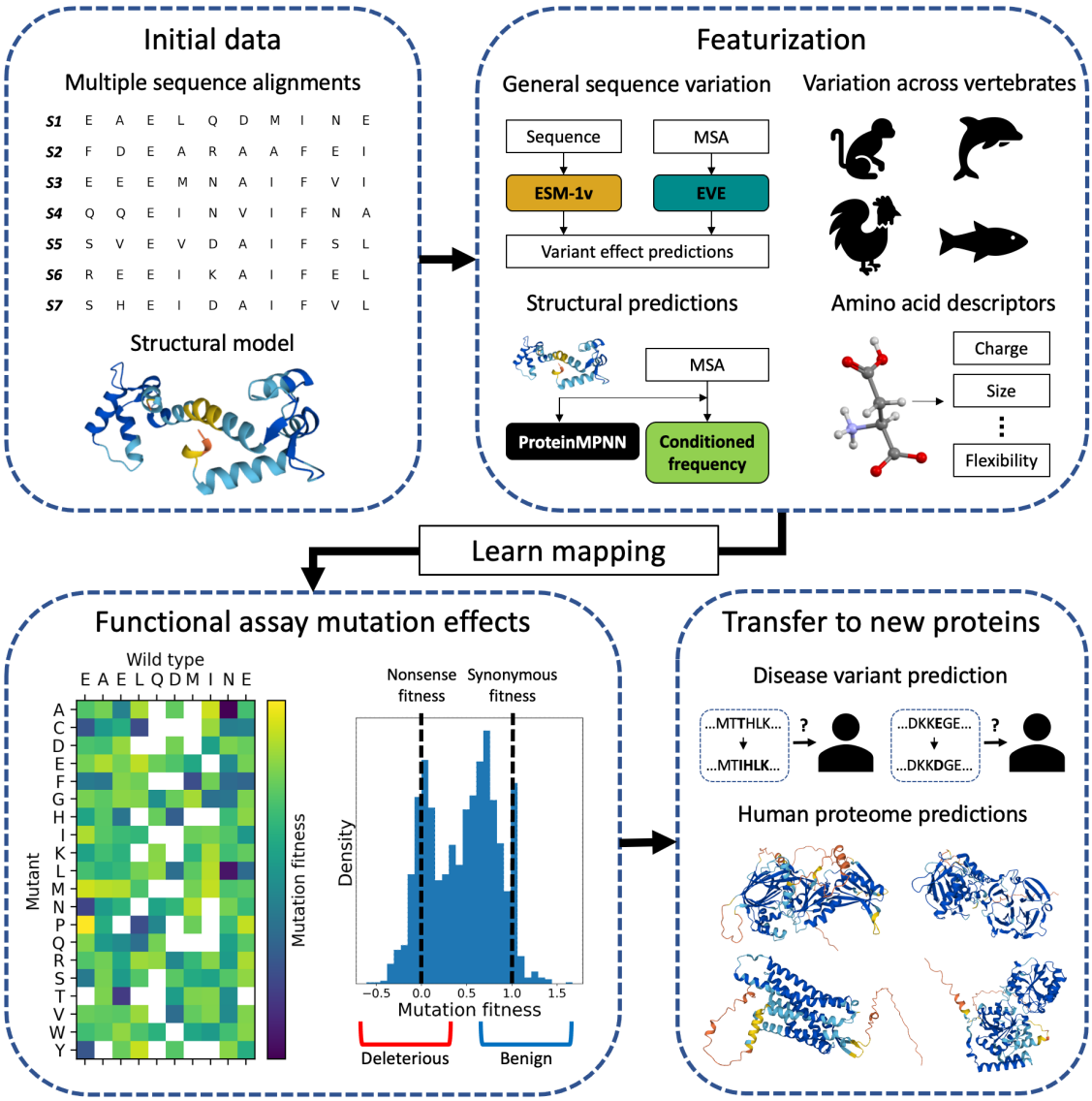

Figure 1: **Method Overview.** We develop computational missense variant effect predictors by training on functional assay data from very few proteins and achieve substantially improved performance over the state-of-the-art. We combine general protein sequence variation (EVE, ESM-1v), sequence variation at local evolutionary timescales (vertebrate alignments), protein structure (AlphaFold2, ProteinMPNN), and amino acid representations. We assess our models on unseen proteins across the human proteome and release predictions for all missense variants in 90% of human genes.

over EVE and ESM-1v. We publicly release predictions for all missense variants in 90% of human genes.

Our framework allows us to integrate multiple sequence alignments (MSAs) at local evolutionary timescales and explicit protein structure models together with state-of-the-art zero-shot variant effect predictors such as EVE and ESM-1v. To prevent data leakage, we did not use any features which were previously trained on clinical or functional assay data of other proteins. We used alignments of 100 vertebrates and 30 mammals which were constructed using whole-genome align-

ment, providing small collections of orthologous sequences that have a higher degree of functional conservation compared to the data used by EVE and ESM-1v [23, 24]. We leveraged AlphaFold2 structure models to provide the specific structure of proteins in a representation that allows use of protein geometry [25, 26]. Our framework is adaptable to the future inclusion of other predictors. While functional assay data do not readily scale to the whole proteome directly, our results demonstrate the utility of relatively small amounts of such data for enhancing computational predictors of disease variants.

Results

State-of-the-art accuracy on clinical variants and functional assays

We trained a model, CPT-1, to classify missense variants as benign or deleterious, using only DMS data from five human proteins (Figure 1, Methods). These proteins (CALM1, MTHR, SUMO1, UBC9, and TPK1) were studied using the same fitness assay by the same lab [7,27], which provided a controlled, high-quality training dataset. CPT-1 integrates the general protein sequence models EVE [8] and ESM-1v [9] with conservation features from vertebrate alignments and structural features calculated using AlphaFold2 structures. Starting from a large list of candidate features, we performed feature selection using cross validation on DMS data and selected nine features for the final model (Supplementary Table S1, Methods). We assessed CPT-1 for clinical disease variant prediction using ClinVar missense variants in human genes that are annotated as benign or pathogenic [1] (Figure 2A-C, Supplementary Table S2). We used the EVE dataset [8] of clinically relevant human proteins, with the five proteins that were used for training removed, leaving 52,239 variants in 2,918 genes (Methods). We report comparisons to EVE and ESM-1v which, like CPT-1, do not train on clinical or functional assay data from evaluation proteins. EVE has been comprehensively evaluated against other widely-used methods and shown to achieve competitive or superior performance [8].

CPT-1 achieves substantially improved performance over EVE and ESM-1v, despite not training directly on any proteins in our assessment dataset. CPT-1 has an improved sensitivity (or true positive rate) for any given specificity (or $1 -$ false positive rate) over both EVE and ESM-1v, and significantly improves the overall area under ROC curve (AUROC) (Figure 2A). Performance increases are particularly large in the clinically relevant high-sensitivity regime (Figure 2B), where a good computational predictor is expected to flag almost all pathogenic variants with as high specificity (or as few false positives) as possible [15]; for example, CPT-1 improves specificity at 95% sensitivity to 64%, from 31% for ESM-1v and 50% for EVE. We also examined per-protein performance in our dataset for those proteins with at least four benign and four pathogenic Clinvar missense variants (Figure 2C, Supplementary Table S2). CPT-1 achieved improved or equal AUROC compared to EVE and ESM-1v on 72% and 79% of genes, respectively (strictly greater AUROC on 52% and 57% of genes, respectively).

We also assessed our cross-protein transfer framework for zero-shot functional assay (DMS) prediction on the five proteins we used to train our ClinVar variant effect predictor. To this end, we trained a separate regression model for each protein, using only the other four proteins as training data (Methods). Our models achieved a higher average Spearman ρ than either ESM-1v or EVE, and also achieved the best performance on three out of five proteins (Figure 2D).

We note that EVE declines to make predictions on variants at positions where alignment quality is low, which make up 13% of variants in the ClinVar assessment dataset. We imputed EVE predictions at these variants using a nearest-neighbors approach within each gene (Methods). EVE scores are less accurate at these imputed positions but are still high-performing and improve model

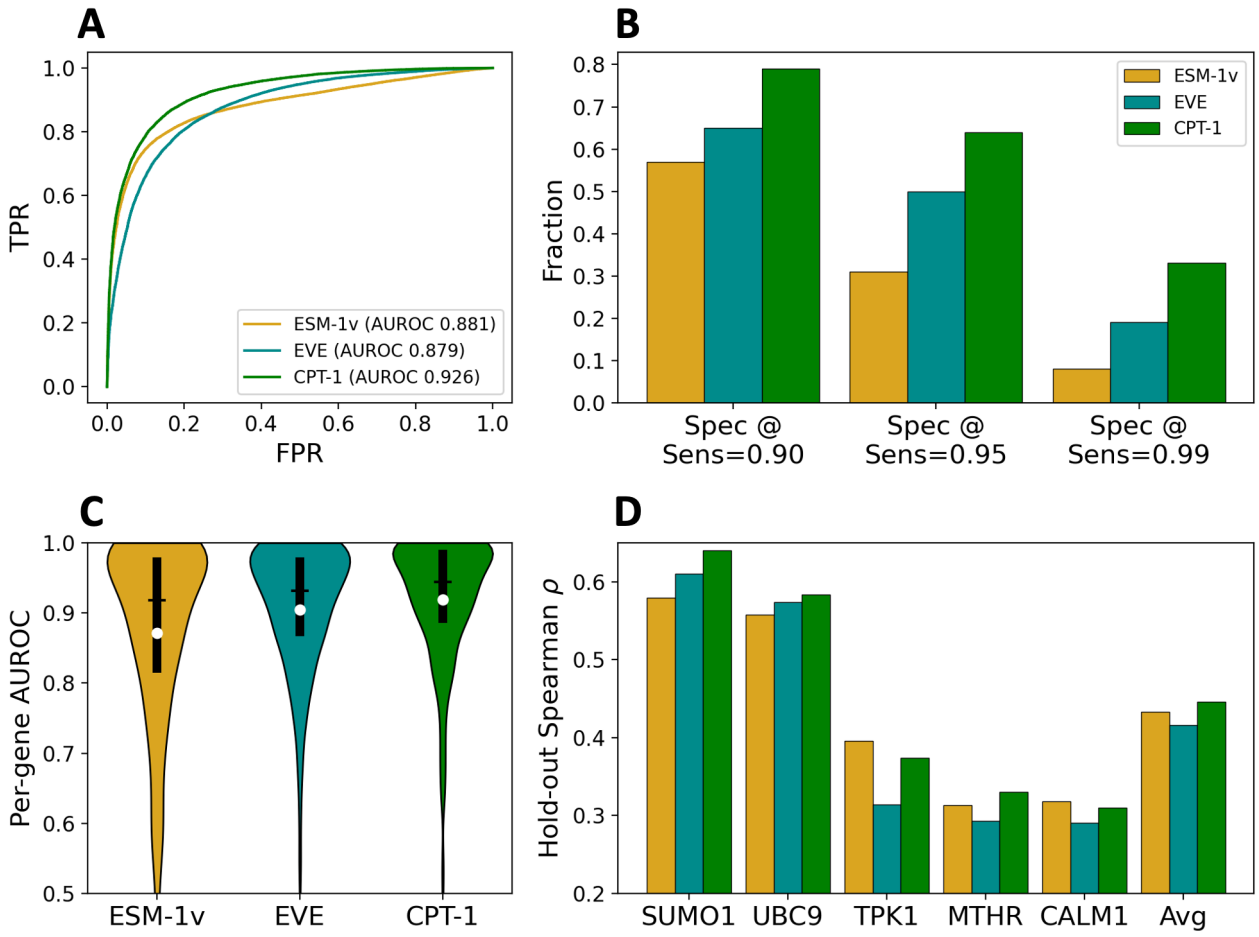

Figure 2: **CPT-1 achieves state-of-the-art performance on clinical variant and functional assay prediction**. (A) Receiver-operating characteristic (ROC) curves for ESM-1v, EVE, and our transfer model CPT-1 on annotated missense variants in ClinVar. CPT-1 improves the true positive rate at all false positive rates over both baselines and has a significantly higher AUROC. (B) Specificity in the clinically relevant high sensitivity regime on ClinVar missense variants. When all models are constrained to recall nearly all pathogenic variants, CPT-1 improves on EVE and ESM-1v by large margins. (C) Per-gene AUROC on ClinVar missense variants in 886 genes with at least four benign and four pathogenic variants. Interquartile range and median shown in black, mean shown in white. CPT-1 improves or equals the per-gene AUROC on 72% of genes for EVE and 79% of genes for ESM-1v. (D) Spearman ρ on functional assay datasets. With the test protein held out in all cases, CPT-1 achieves superior average performance to both baselines and is the best model on three out of five proteins.

performance (Supplementary Figure S1). Also, Frazer *et al.* reported performance for EVE with low confidence predictions removed [8]. In our assessments, we include all predictions for all models. ESM-1v, meanwhile, does not accept proteins longer than 1,022 amino acids by default. We developed and implemented a scheme to apply ESM-1v to longer proteins, which make up 15% the genes in our evaluation dataset (Methods). ESM-1v scores perform worse on these long genes, but EVE and CPT-1 do not suffer a loss in performance (Supplementary Figure S2).

Vertebrate alignments are key to improved performance

EVE and ESM-1v achieve impressive performance using only protein sequence variability at the scale of the whole tree of life. Concretely, these models collect protein sequences from the set of all known proteins and employ redundancy filtering for sequences with high similarity. While this approach models broadly recurring constraints well, it also has limitations. Constraints across a protein family in the entire tree of life, such as structural constraints of a fold, are generally not specific to particular organisms. On the other hand, mutation effects in a particular organism often depend on functional sites with highly variable, system-specific configurations.

We integrated two sets of local-timescale alignments into CPT-1 to address this gap, extracted from 100 vertebrates and 30 mammals via whole-genome alignment, referred to generally as vertebrate multiple sequence alignments (vtMSA) (Methods) [23, 24, 28]. These alignments are shallow, but provide sequences that are orthologous to the target human protein and from species that are close to humans in the context of the full tree of life. The protein sequences are also generally within the redundancy filtering criteria of EVE and ESM-1v. These traits mean that conservation in these alignments is likely to be non-redundant with EVE and ESM-1v while being more specific to function and organism. Conservation in local-timescale alignments has previously been studied as a powerful predictor of variant effects [29–31]; we use very simple features that are particularly well-suited to missense variants.

Features from vertebrate alignments are powerful predictors of the pathogenicity of human disease variants on their own (Figure 3A-B). The frequency of the wild type amino acid in the aligned column of the 100-vertebrate MSAs, for example, achieves a global ClinVar AUROC of 0.845. This is close to the performance of ESM-1v and EVE, and better than single conservation features calculated from the much larger EVE MSA. In particular, 100-vertebrate wild type frequency alone is competitive with EVE and ESM-1v in the high sensitivity regime (Figure 3A). However, 100-vertebrate wild type frequency has a lower overall AUROC than EVE and ESM-1v, and is much worse in the high specificity regime (Supplementary Table S2). Furthermore, if a mutant allele occurs at all in either the 100-vertebrate or 30-mammal MSA, it can be inferred to be benign with high probability (Figure 3B). This aligns with the clinical practice of expecting human mutations with high allele frequency to be benign. A benign mutant allele can be at low frequency in humans, but its presence in 30-mammal or 100-vertebrate alignments likely suggests sufficiently high frequencies in the corresponding species carrying the allele, and hence provides support for non-pathogenicity.

We also measured the importance of vertebrate alignments by training a CPT model that does not use them (Methods). This model performs substantially worse than our full ensemble on clinical data (Figure 3A, Supplementary Table S2). We observe the same trends in per-gene disease variant prediction; removing vertebrate MSAs from CPT-1 hurts performance significantly, while 100-vertebrate wild type frequency is a powerful single-feature baseline (Figure 3C).

We highlight a few human genes where vtMSA features make a large impact (Figure 3D). CAN3 and MLC1 have 59 and 22 annotated ClinVar variants, respectively, with 7 and 10 pathogenic variants. For these genes, 100-vertebrate wild type frequency is a better individual predictor than EVE or ESM-1v, and CPT-1 achieves high performance by combining signal from vertebrate alignments with EVE and ESM-1v. CPT-1 without vtMSA does not improve on EVE and ESM-1v for these genes.

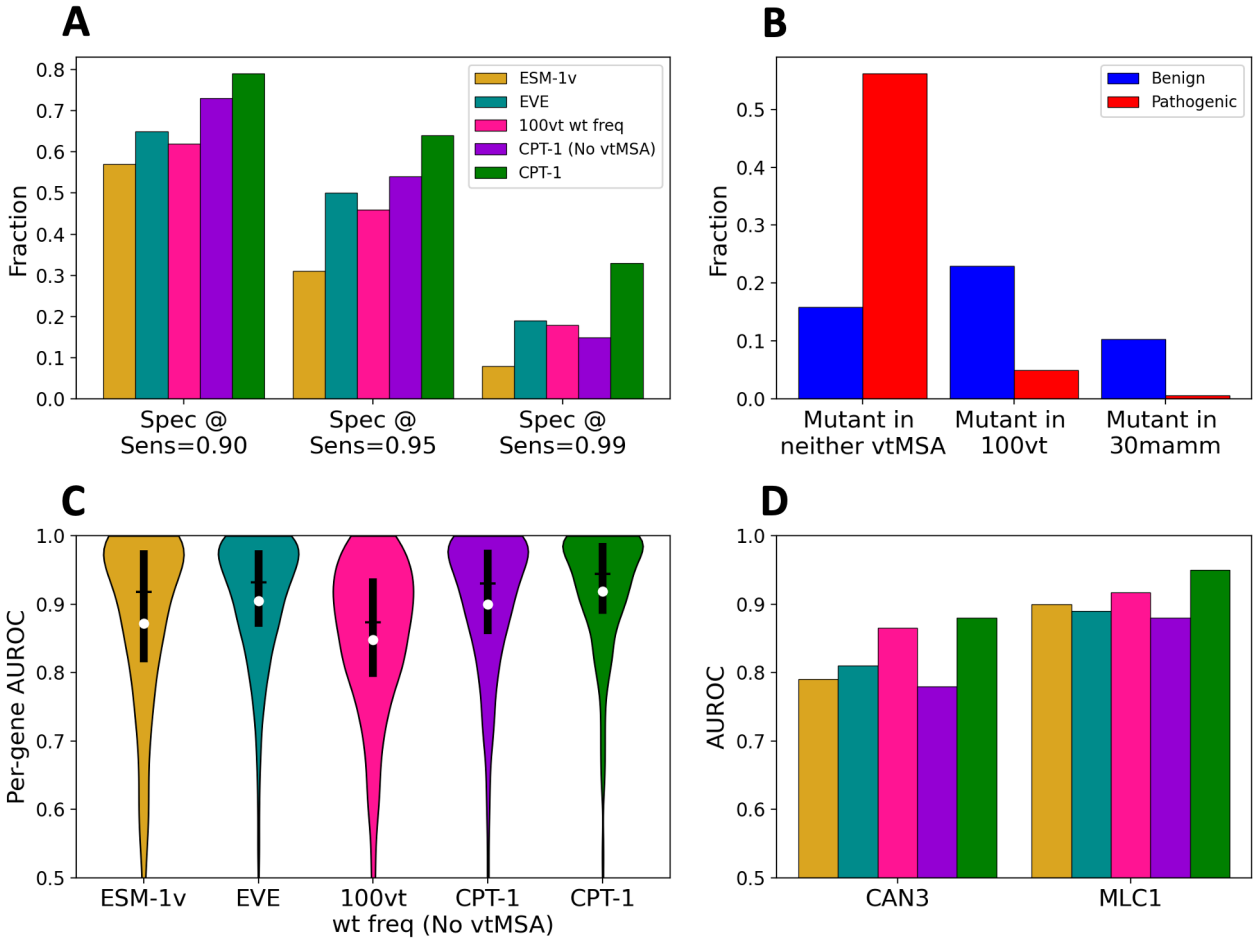

Figure 3: **Vertebrate alignments are key to improved performance and a powerful baseline.** (A) Specificity in the clinically relevant high sensitivity regime on ClinVar missense variants. Removing vertebrate alignments from CPT-1 significantly decreases the margin of improvement over baseline. Conservation among 100 vertebrates is a powerful single feature baseline, and is competitive with much more complex models in the high sensitivity regime. Vertebrate alignments are much less powerful in the high specificity regime (Supplementary Table S2). (B) If a missense variant from ClinVar appears in a vertebrate alignment, it is highly likely to be benign. Of the variants that do not occur in any of our studied vertebrates, 22% are benign. Of the variants that occur in a vertebrate, 82% are benign. Of the variants that occur in a mammal (subset of vertebrates), 94% are benign. (C) Per-gene AUROC on ClinVar missense variants in genes with at least four benign and four pathogenic variants. Conservation among 100 vertebrates is a powerful single feature baseline. Removing vertebrate alignments from CPT-1 hurts performance. (D) In genes such as CAN3 and MLC1, general protein sequence methods are outperformed by vertebrate alignments, and CPT-1 uses this signal to achieve high performance.

Insights from AlphaFold structures

Structural features provide a direct representation of protein geometry that can be informative of function. We used AlphaFold2 predicted structures from the AlphaFold2 human proteome database for all proteins in this study (Methods) [26]. There has been considerable interest in using AlphaFold2 structures for missense variant effect prediction [32–35]. The specific features we tested include predicted amino acid distributions from multiple versions of the deep neural network ProteinMPNN (which takes structure as input) [36], as well as a hand-designed feature that combines a known structure with conservation in the EVE MSA (Methods).

Structural features modestly improve performance of CPT-1 (Figure 4A, Supplementary Table S2, Methods). These performance increases hold even though ProteinMPNN, which depends the least on sequence variability out of our major features, has low accuracy on its own. AlphaFold2 structures thus appear to encode useful information that is not captured from sequence variation alone. However, improvements from adding structure are much smaller than from adding vertebrate alignments. This is consistent with previous results showing that large protein sequence variability methods like EVE and ESM-1v model protein structure implicitly [16–19]. AlphaFold2 structures can be retrieved and analyzed very rapidly, which is an advantage over the extremely slow process of MSA generation. However, AlphaFold2 uses MSAs itself to make predictions; the release of these MSAs at scale would likely enable further increases in whole-proteome performance.

The AlphaFold2 structural models for our five training proteins are all high-quality and informed by experimental structures, while many proteins in our clinical dataset do not have high quality AlphaFold2 models or have disordered regions. We analyzed the correlation of AlphaFold2 structure quality with the performance of structural features in disease variant prediction but did not observe a clear signal (Supplementary Figure S3). Even the performance of ProteinMPNN, which is trained exclusively on experimental protein structures, does not deteriorate dramatically when it is applied to structures with disordered or poorly modeled regions.

We observed that ProteinMPNN performs better on genes where ClinVar variant positions have many contacts in the AlphaFold2 structure (Supplementary Figure S4, Methods). ClinVar contact count and AlphaFold plDDt are also directly predictive of variant pathogenicity (AUROC 0.69 for both), indicating that ClinVar variants in the structural cores of proteins are more likely to be pathogenic (Figure 4B). We highlight the human gene MMP13 as a particularly interesting example (Figure 4C-D). Here, ProteinMPNN predictions outperform EVE and ESM-1v, and structural features are key to high performance by CPT-1 (Figure 4D). The AlphaFold2 structure for this gene is highly accurate and informed by many experimental structures. The gene includes 12 ClinVar variants at 11 positions, with pathogenic variants at 7 different positions and benign variants at 4 different positions. Most of the positions of pathogenic variants (5 out of 7) cluster in a single structural site in the core of the protein (Figure 4C), which seems to be well-understood by ProteinMPNN.

We noticed that ESM-1v perplexity in a gene is somewhat predictive of whether structural features improve CPT-1 performance in that gene (0.6 AUROC). Protein language models like ESM-1v do not rely on sequence alignments, but as a result can struggle to model protein families that do not have many members in large protein sequence databases [37]. Analogously, we found that the quality of ESM-1v mutation effect predictions in a gene is correlated with the depth of the EVE MSA for that gene (Supplementary Figure S5). Perplexity of protein language models is a more direct predictor of how well they have modeled a protein family [37], and we observed that ESM-1v perplexity in a gene is strongly negatively correlated with ESM-1v prediction accuracy for clinical variants of that gene (Supplementary Figure S5, Methods).

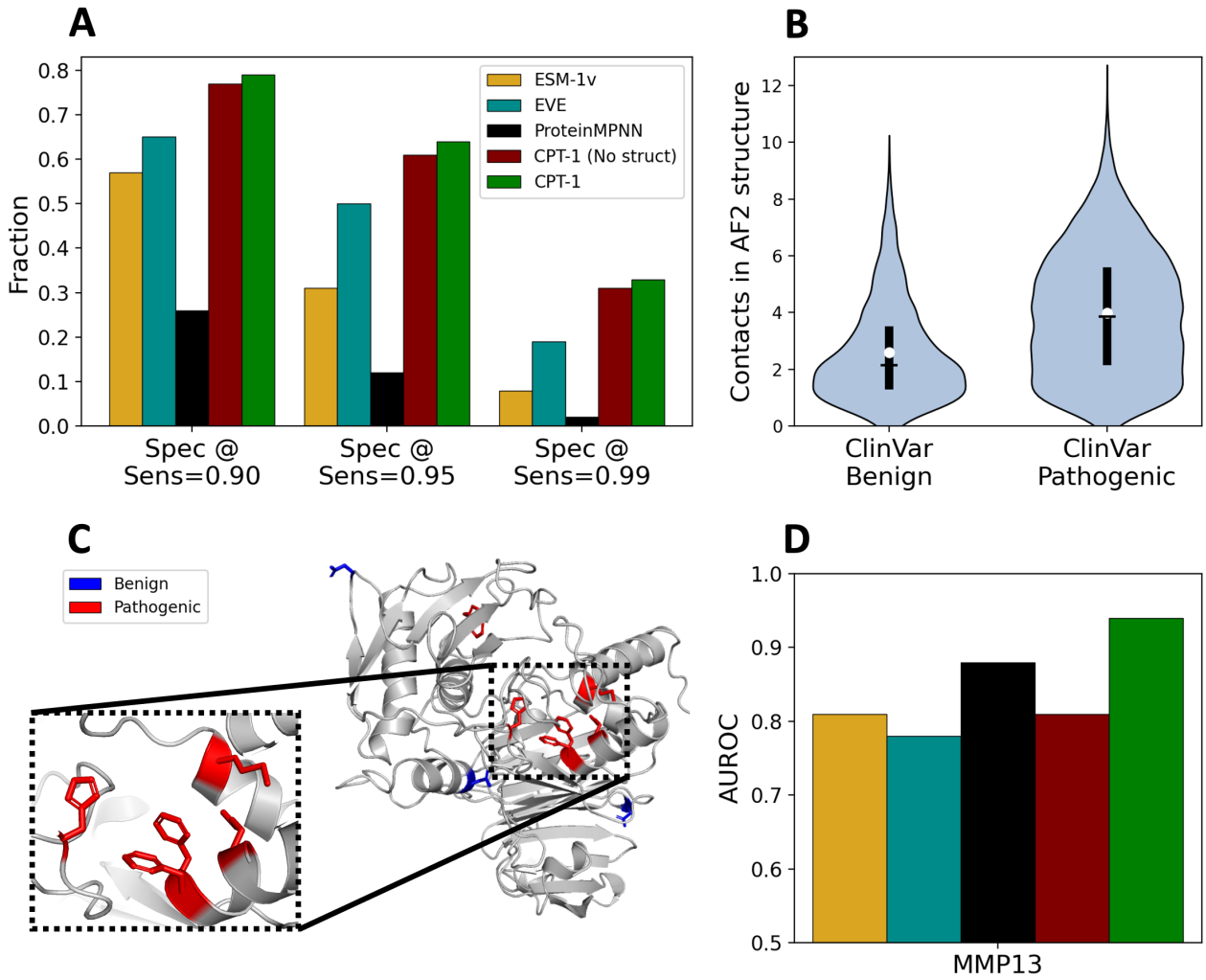

Figure 4: **Insights from AlphaFold structures.** (A) Specificity of CPT-1 in the clinically relevant high sensitivity regime on ClinVar missense variants. Structural features modestly improve CPT-1 performance even though ProteinMPNN alone has poor performance. (B) ClinVar variants with more contacts in the AlphaFold2 structure are more likely to be pathogenic. (C) ProteinMPNN performs better in proteins where ClinVar variants have many contacts (Supplementary Figure S4). MMP13 is a particularly interesting example, with 5 sites of pathogenic ClinVar variants in one structural cluster. (D) CPT-1 outperforms ESM-1v and EVE significantly on MMP13 and relies on structure to achieve this margin. ProteinMPNN is the best single feature.

Predictions across the human proteome

We looked to produce predictions from our method at whole-proteome scale. EVE MSAs and predictions are not available for the vast majority of human genes and are highly computationally intensive to compute. We therefore imputed all features that depend on the EVE MSA across genes (Supplementary Table S1, Methods). Then, using the aforementioned five functional DMS datasets, we refit coefficients of CPT-1 for use on cross-gene imputed features. We assessed this version of CPT-1 in the EVE dataset of preprocessed ClinVar variants and found that the model still outperformed EVE and ESM-1v (Supplementary Table S2). Using cross-gene imputed features instead of the true values reduces performance, but imputation increases performance compared to

removing features entirely (Supplementary Figure S6). With this version of CPT-1, we were able to produce predictions for all missense variants in 90% of human genes. We used features based on the true EVE MSA in genes where this MSA was available and the imputed features in all other genes. The excluded 10% of genes were mostly due to not being contained in our vertebrate alignment dataset; many of these genes have not been clearly shown to produce a protein product.

Discussion

The development of functional DMS assays and computational predictors have each been important to progress in missense variant effect prediction. We demonstrated that, although functional assays do not readily scale to the whole proteome directly, they can be a vital source of information for creating improved computational predictors. Using functional assay data of only five human proteins, we trained CPT-1, a computational predictor that significantly enhances the previous state-of-the-art. Our model is tested on a diverse set of proteins unseen during training and achieves improved performance by integrating vertebrate alignments and predicted structures with general protein sequence models. We used CPT-1 to release predictions for all missense variants in 90% of human genes.

A potential future direction is to test the integration of a larger set of heterogeneous functional assay datasets into the training scheme. In this study, we restricted the scope to five proteins to ensure high-quality and uniform data. Expanded training data could enable increased scope, such as modeling the effects of multiple mutants. Recent work has demonstrated progress modeling multiple mutants in the setting of functional assay prediction [38].

We found that vertebrate alignments provide strong signal for variant effect prediction that is non-redundant with EVE and ESM-1v. The utility of integrating vertebrate alignments across the human proteome points to exciting future directions. There are ongoing efforts to sequence a large number of vertebrate genomes [39]; as these data become available, more powerful models could be applied to deeper vertebrate alignments. Features calculated from AlphaFold2 structures also improve performance of our model. This result is interesting in light of the fact that AlphaFold2 primarily relies on the same evolutionary signals as EVE and ESM-1v to make structure predictions [25]. Recent work has discovered that AlphaFold2 has also learned an accurate representation of protein biophysics [40]. This additional signal may be responsible for the non-redundant information in AlphaFold2 structures.

CPT-1 mostly relies on general protein sequence variation models and sequence variation within vertebrates to predict the effects of missense variants. However, aspects of protein function that have emerged since the evolutionary divergence of vertebrates are still unlikely to be modeled well by CPT-1. Sequence variation may be insufficient to model such effects due to the sparsity of such data at very recent evolutionary timescales. Integrating experimental knowledge of the human protein interactome may help develop even more human-specific models [41], further increasing our understanding of various human diseases.

Methods

Datasets

We trained our models on data from the same functional assay on five human proteins, generated by the same research group [7, 27]. These proteins are CALM1, MTHR, SUMO1, UBC9, and TPK1. The assay measures relative yeast fitness with different variants of the human protein of

interest. We restricted ourselves to these proteins to ensure a high-quality, controlled training set while having enough diversity to transfer to entirely different proteins. The functional assay is also well-aligned with transferring to human clinical effects, since it measures the overall fitness of yeast as opposed to a specific biophysical property of the proteins. We also used these datasets to assess performance on functional assay prediction, by training on four proteins at a time and testing on the fifth, held-out protein.

We assessed our model for disease variant prediction on the EVE dataset of preprocessed ClinVar variants [8]. This dataset includes 2921 clinically relevant human proteins with at least one annotated ClinVar missense variant. The dataset also includes precomputed EVE MSAs and EVE predictions, which we used in our model. We included variants annotated as "Benign" or "Likely Benign" and variants labeled as "Pathogenic" or "Likely Pathogenic" for our benign and pathogenic labels.

We derived features from MSAs of orthologous sequences from 100 vertebrate and 30 mammalian species. These MSAs are available from the UCSC genome browser for most of the human genome, and were constructed using whole-genome alignment [23, 24, 28]. For some genes (228 in the EVE dataset), different isoforms were used in the vertebrate MSAs compared with other features, which are mainly based on UniProt. To resolve the discrepancy, we ran pairwise alignments between the vtMSA protein sequences and the UniProt sequences and only retained fragments of vtMSAs that can be aligned to Uniprot sequences. We used the Bio.Align.PairwiseAligner implemented in the Biopython package with the setting: `model = 'local', match_score = 5, mismatch_score = -4, open_gap_score = -4, extend_gap_score = -0.5`.

We obtained predicted structures for all proteins from the AlphaFold2 human proteome database [26]. For proteins with a known experimental structure, the AlphaFold2 structure is generally highly accurate because the known structure has been provided as a template to AlphaFold2. Using all AlphaFold2 structures makes the input structures have a homogeneous format. We used contact statistics from AlphaFold2 structures for certain features and analyses. We extracted contacts using the Probe software [42], which notably identifies only sidechain-sidechain contacts.

Features used in our transfer model

We initially considered a large set of potential features to include in our transfer model. Notably, we excluded predictors which were previously trained on clinical or functional assay data, to prevent data leakage. We also did not use features that do not extrapolate in an obvious manner to all 19 possible amino acid variants at a protein position (as opposed to all 9 possible single nucleotide variants in a codon). Our training functional assay data have a large amount of amino acid variants that are not expressible as single nucleotide variants. Restricting to full amino acid variant coverage allowed us to use more training data per protein and preserves applicability to functional assay prediction.

First, we included scores from the general protein homology models EVE and ESM-1v. We used the EVE scores which were normalized to a zero to one scale by Gaussian mixture modeling [8]. Precomputed EVE scores are included in the EVE dataset. We additionally computed EVE scores for SUMO1 and UBC9, which are part of our five training proteins but not included in the EVE dataset. By default, EVE does not provide scores at positions where the quality of the MSA is low; we imputed EVE scores at these positions using a within-gene K-nearest neighbors approach, which will be described in detail in the next section (see **Weighted KNN Imputation**). Imputed EVE predictions within a gene are less accurate than true EVE predictions but still increase the performance of our model (Supplementary Figure S1). We also generated predictions for human

proteins outside of the EVE dataset where no EVE predictions were available. For these genes, we imputed EVE scores (and other features relying on the EVE MSA) using a cross-gene K-nearest neighbors approach (see **Weighted KNN Imputation**). Imputed EVE predictions across genes also increased the performance of our model (Supplementary Figure S6).

We computed ESM-1v scores for all proteins in the UniProt collection of canonical transcripts for the human proteome (downloaded May 2022) [43]. We used the ESM-1v log probability difference to the wild-type amino acid, as in Meier *et al.* [9]. By default, ESM-1v does not accept proteins longer than 1022 amino acids. We developed a scheme to use ESM-1v on longer proteins using multiple sliding windows and used this scheme to compute ESM-1v scores for long proteins. Concretely, we calculated ESM-1v predictions with overlapping 1000 amino acid windows, with starting positions 250 amino acids apart on the protein sequence. Then for each mutation in the sequence, we used the score from the window whose center is the closest to the position of the mutation, as the center positions are expected to have better ESM-1v predictions given richer contextual information available. We also analyzed the relationship between ESM-1v perplexity in a gene and ESM-1v classification performance on ClinVar variants in that gene. We calculated perplexity as the average of the ESM-1v log probability for the wild-type amino acid over the entire protein, followed by exponentiation, as in Lin *et al.* [37].

To capture conservation at closer evolutionary timescales, we included features calculated using the 30-mammal and 100-vertebrate MSAs. Specifically, we obtained three types of frequencies for each mutation: the frequency of the wild-type amino acid at its position, the frequency of this mutant amino acid at this position, and the frequency of gaps in the alignment at this position. We refer to them as wild-type frequency, mutant frequency and gap frequency, respectively. These frequencies were log-transformed with offset 1. In genes where vtMSAs had different isoforms, certain regions were unmatched in pairwise alignment (see **Datasets**). Features in these regions were imputed with the weighted K-nearest neighbor (KNN) imputation strategy (see **Weighted KNN imputation**).

We included several features that explicitly use the AlphaFold2 structure of the protein. First, we calculated variant log-probabilities for all human proteins from three versions of ProteinMPNN, which was created with a focus on protein design [36]. Vanilla ProteinMPNN takes in protein structure with full protein backbone along with partial protein sequence. Cα ProteinMPNN takes in protein structure with only alpha carbons for each residue along with partial protein sequence. Cα-only ProteinMPNN uses only alpha carbons (no masked protein sequence). We normalized these scores as the log-probability difference to the wild-type allele log-probability, matching ESM-1v. Only Vanilla ProteinMPNN is used in CPT-1 after feature selection. We also included two hand-designed structural features which we found perform well on functional assay data. Using the AlphaFold2 structure and the EVE MSA, we calculated for each position the two sidechain-sidechain contacts which are most conserved to the human amino acids. Sidechain-sidechain contacts were calculated using the Probe software (see **Datasets**) [42]. We filtered the EVE MSA to the sequences where these contacts are conserved, and then extracted the frequencies of the wild-type and mutant amino acids, each as one feature. We only allowed conditioning on residues with plDDt greater than 70 in the AlphaFold2 structure, and residues with plDDt less than 70 did not have any conditioning used. For proteins that had multiple fragment AlphaFold2 structures, we took ProteinMPNN scores and contacts for each position from the fragment that maximized the plDDt of that position.

Finally, we included amino acid descriptors, which are featurizations of amino acids that encode properties such as charge, polarity, hydrophobicity, size, and local flexibility [44]. The descriptors we used include Cruciani properties [45], VHSE [46], Z-scales [47], ST-scales [48], ProtFP [49], and Georgiev's BLOSUM indices [50]. We used the differences in the descriptor values between the mutant amino acid and the wild-type amino acid as features for each mutation.

Weighted KNN imputation

We used a strategy based on K-nearest neighbor imputation to impute missing values in the feature matrix. Take the EVE scores as an example. To train a KNN model on a given gene, we first calculated the Spearman correlation between each feature and the EVE scores at the available mutations within the gene. Then the five most highly correlated features together with the EVE scores were used to build the KNN model. When calculating the distance matrix, each feature was weighted by its correlation value with the EVE scores, which was implemented as scaling each of the features by the correlation value after standardization. EVE scores were assigned weight 1 in the scaling.

When applying the fitted model to impute missing values, we used two strategies in this study, which we refer to as within-gene imputation and cross-gene imputation. For genes included in the EVE dataset, we used within-gene imputation to directly impute missing EVE scores with the model fitted on that gene. However, for genes not included in the EVE dataset where no EVE score is available for any mutation, we first fit five KNN models on the five training genes, used them to impute the EVE scores for all the mutations and then averaged the outputs across the five models to get the imputed values.

The within-gene imputation strategy was also used to impute missing vtMSA features. The cross-gene imputation strategy was used to impute missing structure-conditioned scores. We used the implementation of the KNN model in the python sklearn package (sklearn.impute.KNNImputer) with the number of nearest neighbors as 10 leaving other parameters as default.

Model architecture and training

Our models were set up as either logistic regression to classify benign from deleterious mutants (for clinical disease variant prediction) or as linear regression to predict functional assay score (for hold-out protein functional assay prediction). We trained a separate linear model for each training protein and ensembled them by averaging the model predictions at test time; we found this to be an effective method to adjust for batch effects across each protein. We found that more complex, non-linear models did not transfer to held-out proteins well. We analyzed the impact of using variable numbers of training proteins (Supplementary Figure S7). Benefits appear to be saturating at all five proteins used; additional proteins may be more useful if diversity is increased.

Although functional assay data for our five training proteins was generated by the same research group with almost the same protocol, the distribution of scores varies significantly between proteins (Supplementary Figure S8). To remedy this, and because pathogenicity annotations are binary, we decided to binarize the functional assay scores to train the classification model for disease variant prediction. Specifically, we standardized the data by taking the top 40% of variants from each protein as benign and the bottom 40% as pathogenic. We found that this percentile-based binarization provided stable results, and these results did not depend much on the exact binarization threshold (Supplementary Table S3).

We used a global feature rescaling for all proteins, calculated from our five training proteins. We scaled the features to unit standard deviation, but calculated these standard deviations by reweighting all samples from each training protein to total weight one, so that each protein has the same total weight in calculating rescaling weights. This prevents the rescaling terms from being biased towards larger proteins in our training dataset. The MSA features are frequencies and do not behave well under the scaling and were therefore left as is.

We initially considered a large list of candidate features and employed feature selection to reduce this list before fitting linear models. We used average AUROC from cross-validation on the

training functional assay data as performance metrics to select features (for regression, we instead use Spearman correlation). Specifically, in each fold we leave out one protein in the training set as the validation set. We use the remaining proteins to fit the model and then evaluate the performance using the validation protein. For the disease variant classification model, a 5-fold cross-validation was performed for all 5 training proteins. For the functional assay score regression model, a 4-fold cross-validation was performed excluding the held-out protein. We always include the two protein homology models as features, i.e. ESM-1v and EVE scores. For the remaining features, we used a two-step scheme for feature selection: the first step selects features from the 100-vertebrate MSA, 30-mammal MSA, ProteinMPNN and structure-conditioned score categories and the second step selects amino acid descriptor features. In the first step, we exhaustively searched all combinations of features within each of the four categories above, restricting that at least one feature was selected from each category. Then in the second step, given the large number of amino acid descriptor features, we used forward selection to reduce the computational burden. To go through the procedures, we start with the two features of ESM-1v and EVE. Then in the first step, we add the set of 100-vertebrate MSA features that achieves the best performance on the validation protein. Then we add the best sets of 30-mammal MSA features, ProteinMPNN features and structure-conditioned score features in the same way. Then in the second step, we greedily select the first few amino acid descriptor features that make the largest improvements in performance. Final selected features for CPT-1 are reported in Supplementary Table S1. To examine the effects of removing vertebrate alignments, 100-vertebrate and 30-mammal mutant frequencies were removed and models were retrained. To examine the effects of removing structure, ProteinMPNN and conditioned mutant/wildtype frequencies were removed and models were retrained.

Feature selection on functional assay data alone is not sufficient to produce an improved predictor, and unweighted averaging of selected features in fact performs worse than unweighted averaging of ESM-1v and EVE (Supplementary Figure S9). Learning coefficients for these features is necessary to achieve improved performance.

Data and Code Availability

We release CPT-1 predictions for all missense variants in 90% of human genes at https://github.com/songlab-cal/CPT. Model code will be released upon publication.

Competing Interest Statement

The authors declare no competing interests.

Acknowledgements

We would like to thank Sanjit Batra, Gonzalo Benegas, Chloe Hsu, Sergey Ovchinnikov, and members of the Song Lab for helpful discussion. This research is supported in part by an NIH grant R35-GM134922 and a grant from the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics.

References

[1] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 2016.

[2] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[3] Peter D Stenson, Edward V Ball, Matthew Mort, Andrew D Phillips, Jacqueline A Shiel, Nick ST Thomas, Shaun Abeysinghe, Michael Krawczak, and David N Cooper. Human gene mutation database (HGMD®): 2003 update. *Human Mutation*, 21(6):577–581, 2003.

[4] Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua D Hoffman, Daren Liu, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Bin Ye, Nilanjana Banerjee, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature*, 586(7831):749–756, 2020.

[5] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, 2014.

[6] Benjamin J Livesey and Joseph A Marsh. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular Systems Biology*, 16(7):e9380, 2020.

[7] Jochen Weile, Song Sun, Atina G Cote, Jennifer Knapp, Marta Verby, Joseph C Mellor, Yingzhou Wu, Carles Pons, Cassandra Wong, Natascha van Lieshout, et al. A framework for exhaustively mapping functional missense variants. *Molecular Systems Biology*, 13(12):957, 2017.

[8] 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*, 599(7883):91–95, 2021.

[9] 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.

[10] Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016.

[11] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

[12] Daniele Raimondi, Ibrahim Tanyalcin, Julien Ferté, Andrea Gazzo, Gabriele Orlando, Tom Lenaerts, Marianne Rooman, and Wim Vranken. DEOGEN2: prediction and interactive

visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Research*, 45(W1):W201–W206, 2017.

[13] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

[14] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, 2018.

[15] Karthik A Jagadeesh, Aaron M Wenger, Mark J Berger, Harendra Guturu, Peter D Stenson, David N Cooper, Jonathan A Bernstein, and Gill Bejerano. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics*, 48(12):1581–1586, 2016.

[16] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, 2012.

[17] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

[18] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[19] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, 2020.

[20] Bruce J Wittmann, Yisong Yue, and Frances H Arnold. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems*, 12(11):1026–1045, 2021.

[21] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, pages 1–9, 2022.

[22] Bruce J Wittmann, Kadina E Johnston, Zachary Wu, and Frances H Arnold. Advances in machine learning for directed evolution. *Current Opinion in Structural Biology*, 69:11–18, 2021.

[23] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[24] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

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

[26] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[27] Jochen Weile, Nishka Kishore, Song Sun, Ranim Maaieh, Marta Verby, Roujia Li, Iosifina Fotiadou, Julia Kitaygorodsky, Yingzhou Wu, Alexander Holenstein, et al. Shifting landscapes of human MTHFR missense-variant effects. *The American Journal of Human Genetics*, 108(7):1283–1300, 2021.

[28] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, 2004.

[29] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[30] Adam Siepel, Katherine S Pollard, and David Haussler. New methods for detecting lineage-specific selection. In *Annual International Conference on Research in Computational Molecular Biology*, pages 190–205. Springer, 2006.

[31] Ritika Ramani, Katie Krumholz, Yi-Fei Huang, and Adam Siepel. PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastcons and phylop. *Bioinformatics*, 35(13):2320–2322, 2019.

[32] David T Jones and Janet M Thornton. The impact of AlphaFold2 one year on. *Nature Methods*, 19(1):15–20, 2022.

[33] Mehmet Akdel, Douglas EV Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, pages 1–12, 2022.

[34] Axel Schmidt, Sebastian Röner, Karola Mai, Hannah Klinkhammer, Martin Kircher, and Kerstin U Ludwig. Predicting the pathogenicity of missense variants using features derived from AlphaFold2. *bioRxiv*, 2022.

[35] Bian Li, Dan M Roden, and John A Capra. The 3D mutational constraint on amino acid sites in the human proteome. *Nature Communications*, 13(1):1–15, 2022.

[36] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, page eadd2187, 2022.

[37] 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.

[38] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[39] Arang Rhie, Shane A McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, 2021.

[40] James P Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using AlphaFold. *BioRxiv*, 2022.

[41] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[42] J Michael Word, Simon C Lovell, Thomas H LaBean, Hope C Taylor, Michael E Zalis, Brent K Presley, Jane S Richardson, and David C Richardson. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology*, 285(4):1711–1733, 1999.

[43] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019.

[44] Daniel Osorio, Paola Rondón-Villarreal, and Rodrigo Torres. Peptides: a package for data mining of antimicrobial peptides. *Small*, 12:44–444, 2015.

[45] Gabriele Cruciani, Massimo Baroni, Emanuele Carosati, Monica Clementi, Roberta Valigi, and Sergio Clementi. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *Journal of Chemometrics*, 18(3-4):146–155, 2004.

[46] HU Mei, Zhi H Liao, Yuan Zhou, and Shengshi Z Li. A new set of amino acid descriptors and its application in peptide QSARs. *Peptide Science: Original Research on Biomolecules*, 80(6):775–786, 2005.

[47] Maria Sandberg, Lennart Eriksson, Jörgen Jonsson, Michael Sjöström, and Svante Wold. New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, 41(14):2481–2491, 1998.

[48] Li Yang, Mao Shu, Kaiwang Ma, Hu Mei, Yongjun Jiang, and Zhiliang Li. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids*, 38(3):805–816, 2010.

[49] Gerard JP van Westen, Remco F Swier, Jörg K Wegner, Adriaan P IJzerman, Herman WT van Vlijmen, and Andreas Bender. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *Journal of Cheminformatics*, 5(1):1–11, 2013.

[50] Alexander G Georgiev. Interpretable numerical descriptors of amino acid space. *Journal of Computational Biology*, 16(5):703–723, 2009.

Supplementary Figures

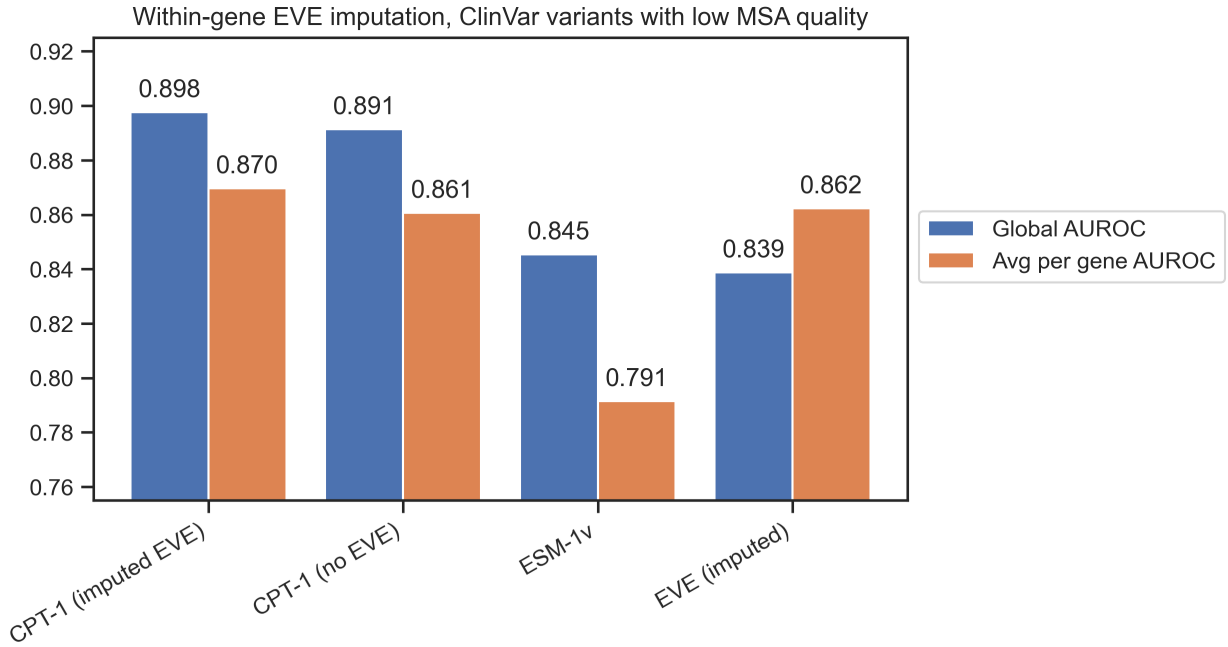

Figure S1: **Within-gene imputation of EVE predictions.** EVE declines to make variant effect predictions at positions with low alignment quality in the EVE MSA. We imputed EVE predictions at these positions using a nearest-neighbors approach within the gene. Imputed predictions increase CPT-1 performance both in terms of overall AUROC and per-gene AUROC (calculated on 47 genes with at least 4 benign and 4 pathogenic variants with low alignment quality). ESM-1v performs worse on these low-alignment quality variants as well, compared to ESM-1v performance on other variants.

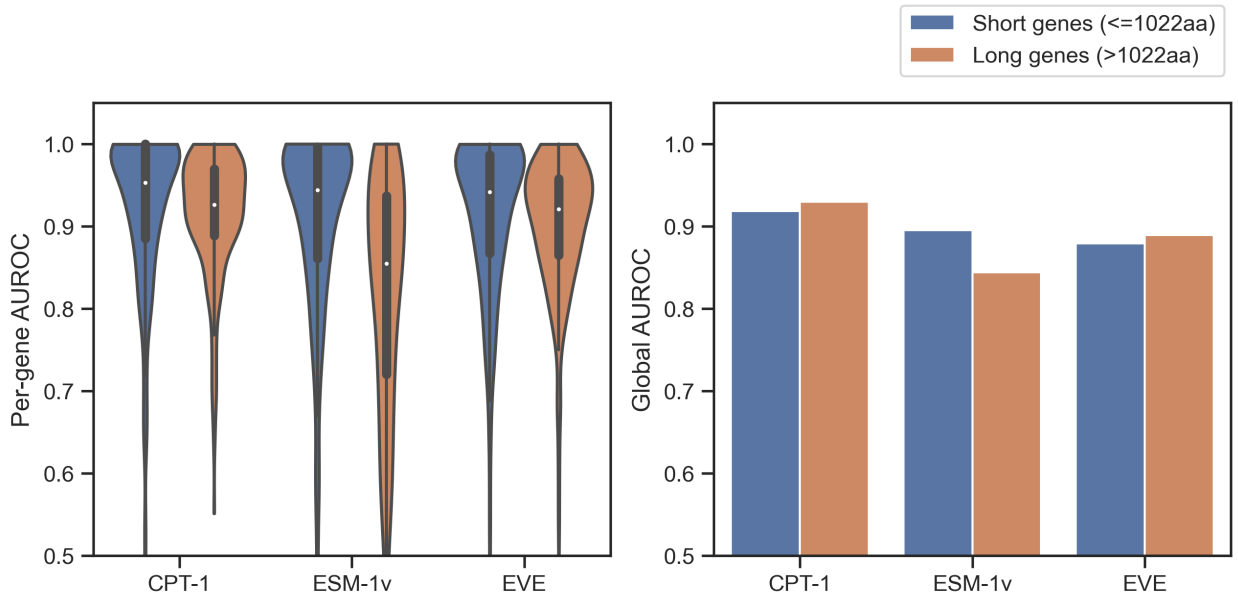

Figure S2: **Extending ESM-1v to long genes**. ESM-1v by default does not accept proteins longer than 1022 amino acids. To enable full coverage of human proteins, we developed a scheme to apply ESM-1v to longer proteins. ESM-1v with this extension performs worse on long genes, but CPT-1 does not suffer a performance drop. EVE performs as well on long proteins as on short.

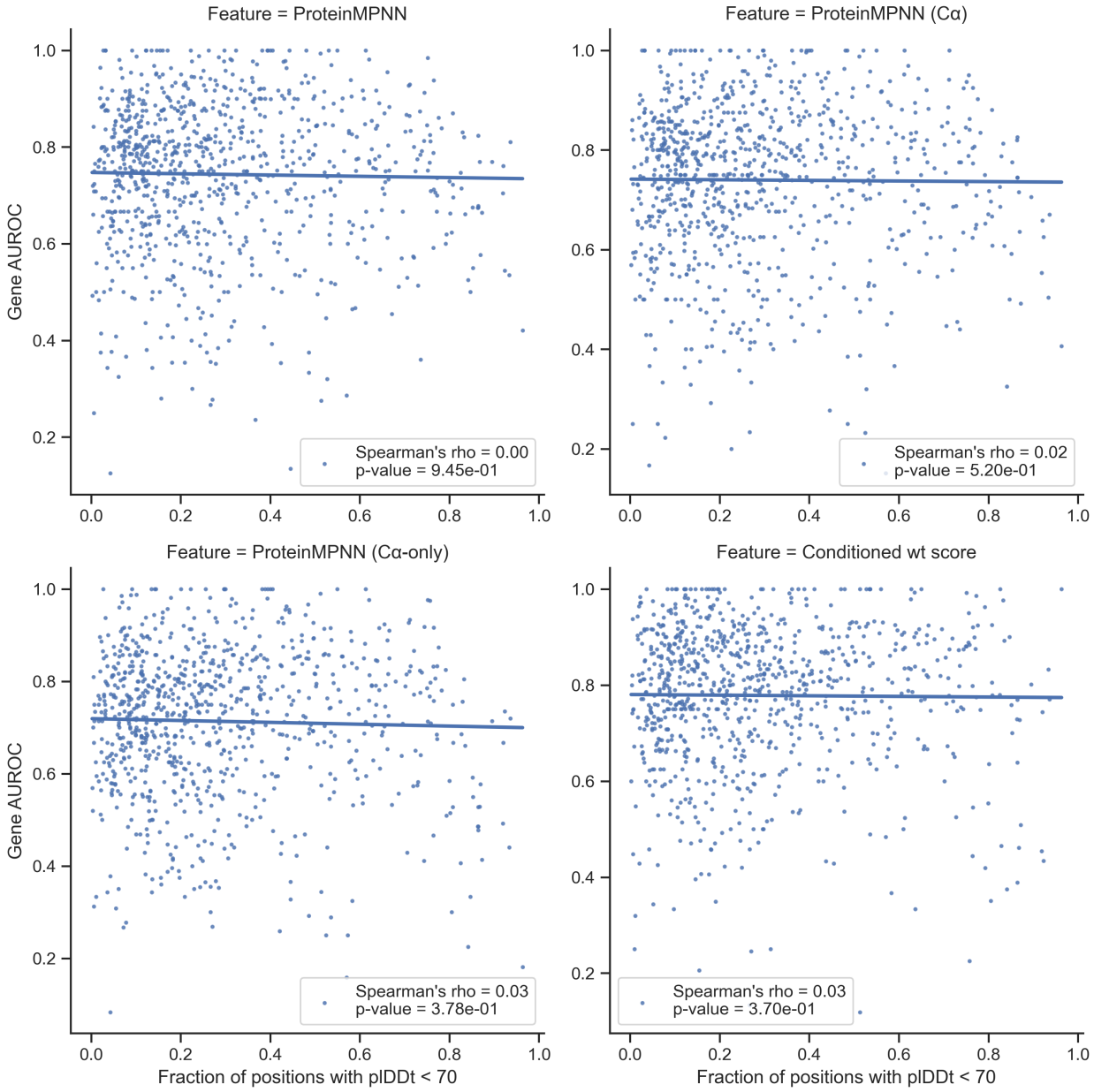

Figure S3: **Structure quality.** Our training dataset includes only high quality AlphaFold2 protein structures, whereas our assessment dataset includes many disordered or poorly modeled AlphaFold2 structures. However, we did not observe strong correlations between AlphaFold2 structure quality and the performance of individual structural features on the gene in our assessment dataset. Here we show correlation with the fraction of protein positions with AlphaFold2 plDDt less than 70; correlations are also weak if a threshold of 50 is used instead.

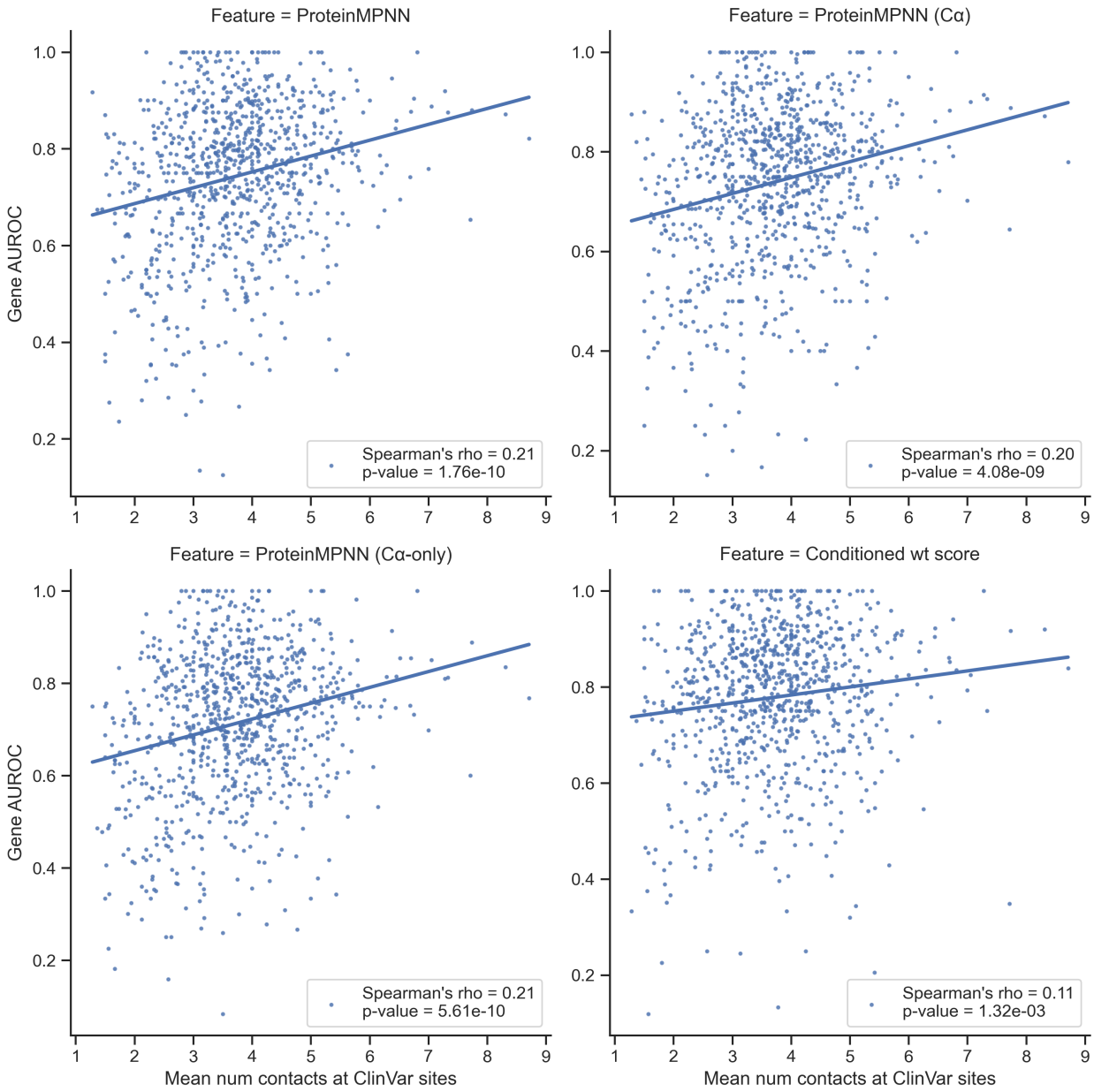

Figure S4: **Contact counts**. We calculated the number of sidechain-sidechain contacts formed by the position of each ClinVar variant and took the average of these counts in each protein. Structural features perform better in genes where ClinVar variant positions have more contacts. Contact count is itself a modest predictor of variant pathogenicity (0.69 AUROC), indicating that ClinVar variants in the core of proteins are more likely to be pathogenic.

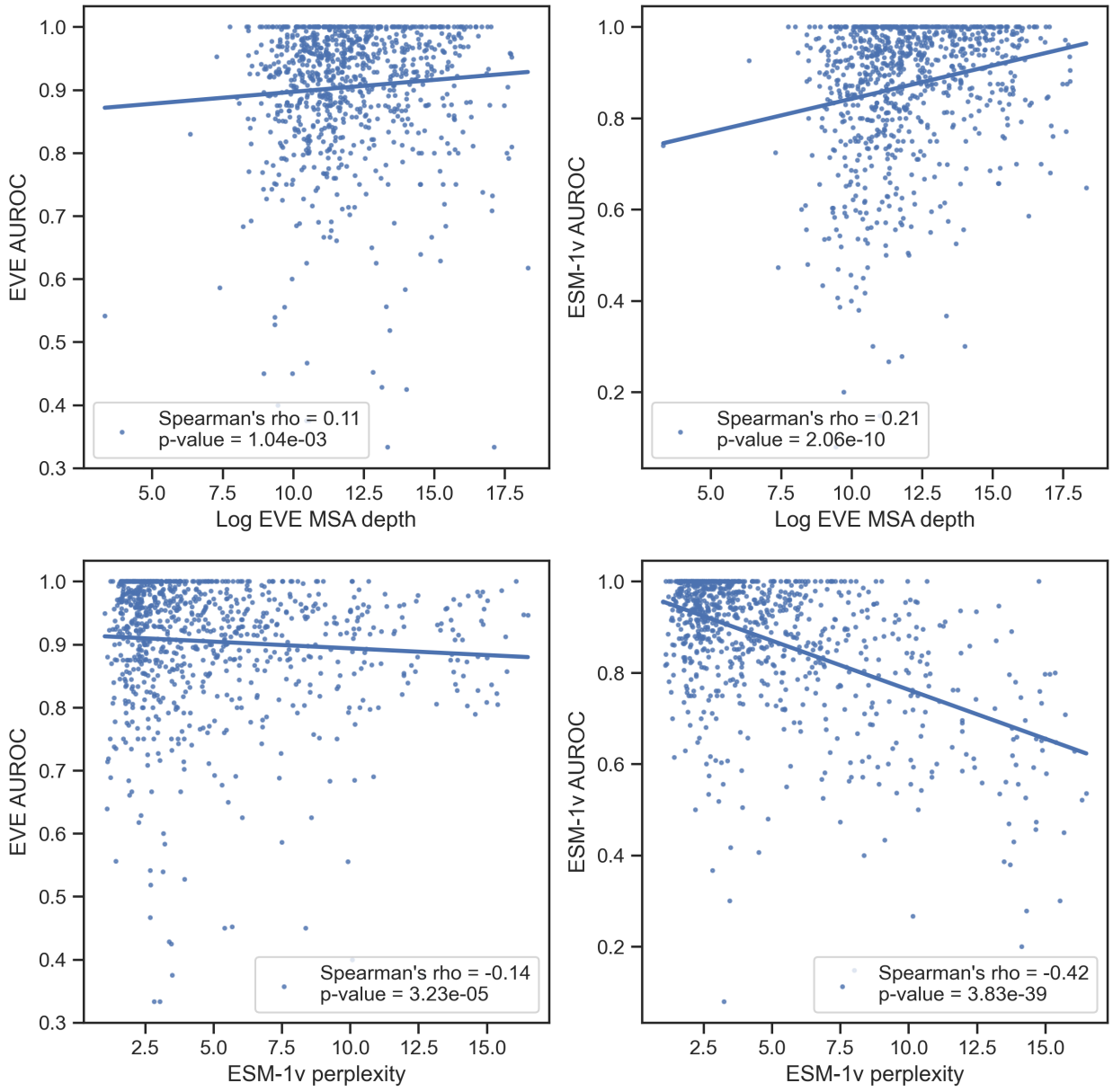

Figure S5: **MSA depth and ESM-1v perplexity**. Sequence variation methods can be sensitive to the number of available related sequences. ESM-1v and EVE performance in a gene both decrease if the EVE MSA is shallow and if ESM-1v perplexity is high. ESM-1v is more affected by both predictors, and has a particularly strong correlation with perplexity, consistent with findings applying protein language models to structure prediction.

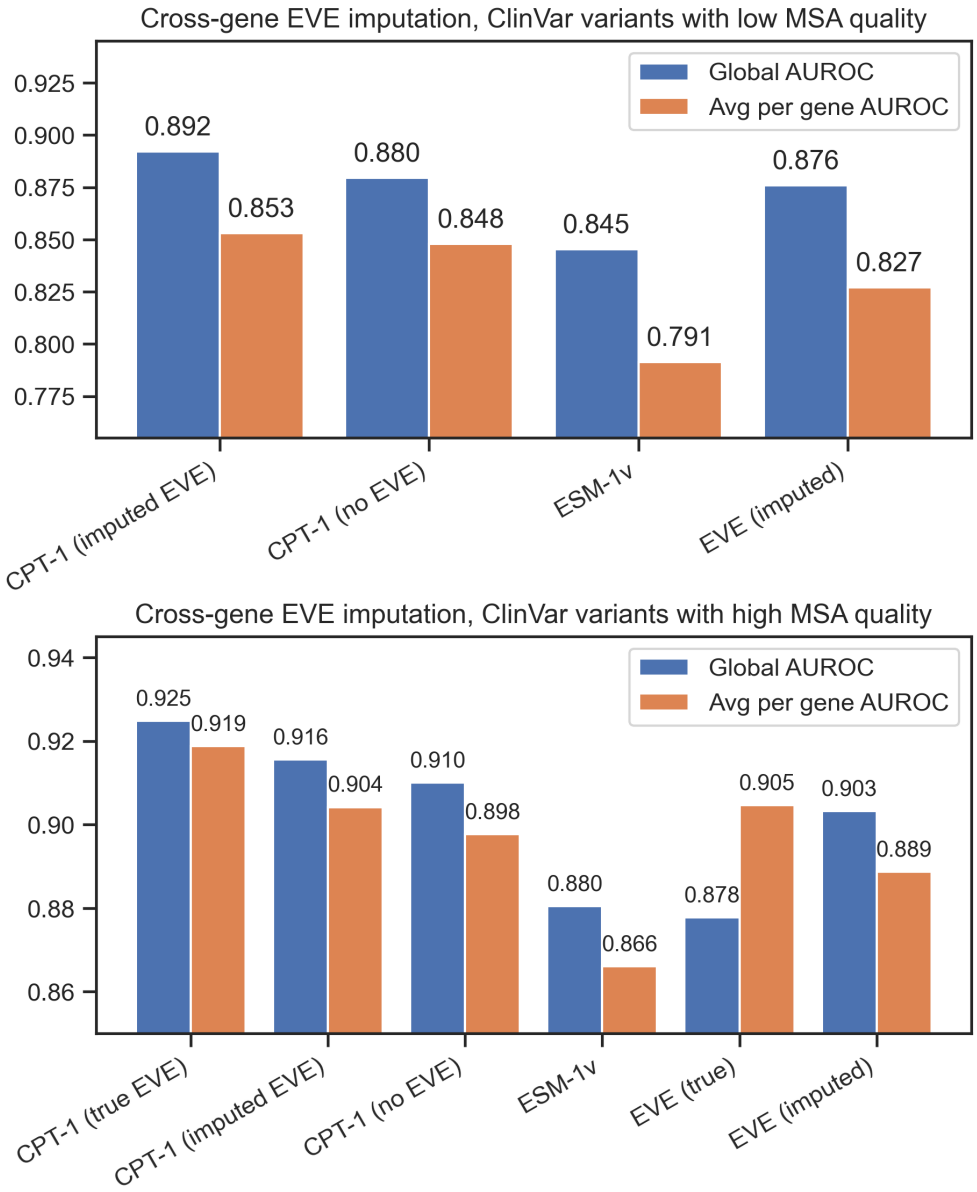

Figure S6: **Cross-gene imputation of EVE predictions.** EVE MSAs are not available for the vast majority of human proteins and are slow to compute. To scale our framework to the whole human proteome, we imputed features depending on the EVE MSA in genes where no values for these features were available. We assessed the quality of our imputation scheme within the EVE dataset. EVE MSAs are available for these genes, but we test our imputation strategy here to measure how well we can do without using them. This gives us a sense of the accuracy of our imputation for whole-proteome prediction. Within this dataset, we assessed separately on variants at positions with high EVE MSA quality and variants at positions with low EVE MSA quality. EVE predictions are not available for the latter set. Imputed features increase performance over simply removing the features in both classes of variants.

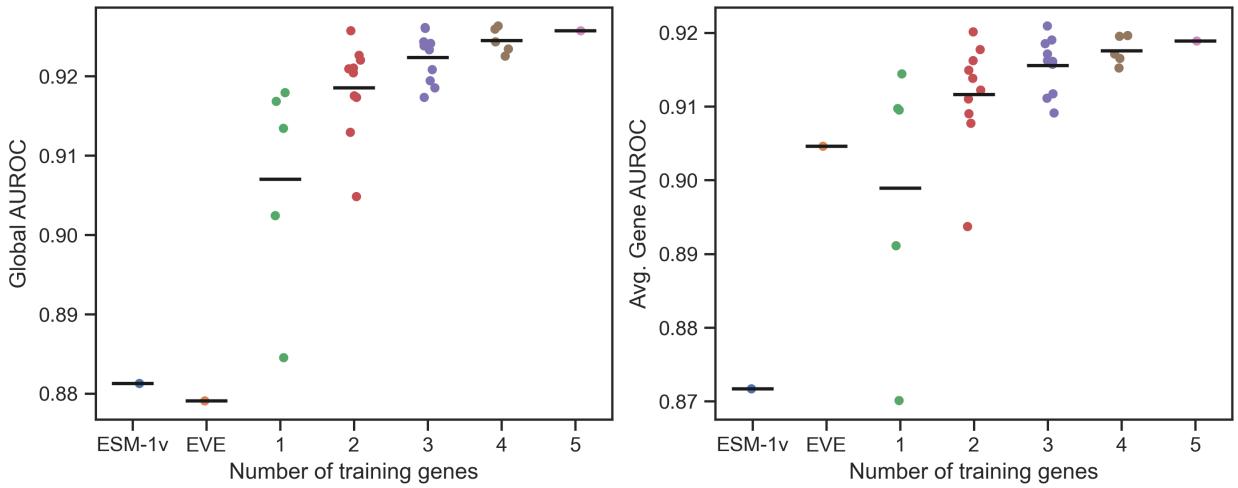

Figure S7: **Impact of number of training genes.** CPT-1 is trained on functional assay data from five proteins. We examined how performance depends on the number of training proteins used. Each dot indicates a specific choice of training genes, with the mean shown as a black bar. More training genes always increases average performance, but there is significant variance and performance increases appear to be saturating. Additional proteins may need to be more diverse to deliver significant further improvements.

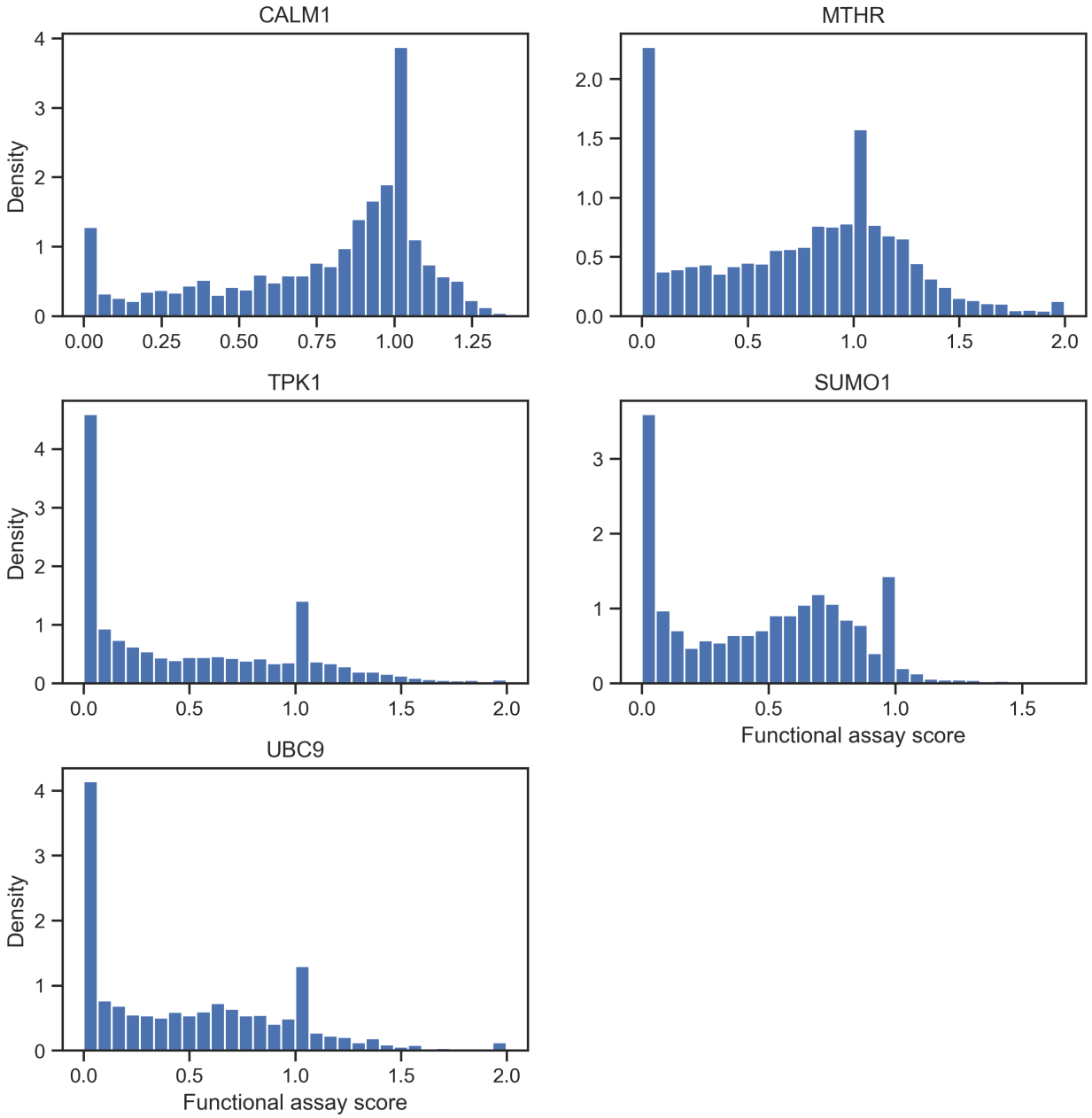

Figure S8: **Distribution of functional assay scores**. Functional assay scores for our five training proteins were generated by the same group with a standardized scale. A score of 0 is set to match the average effect of a nonsense (stopgain) variant, while a score of 1 is set to match the average effect of a synonymous variant. Some datasets were truncated at 0, so we truncated all 5 at 0. Despite the relative homogeneity between these datasets, the distribution of variant scores differs significantly between genes, which led us to use binarization for our disease variant model.

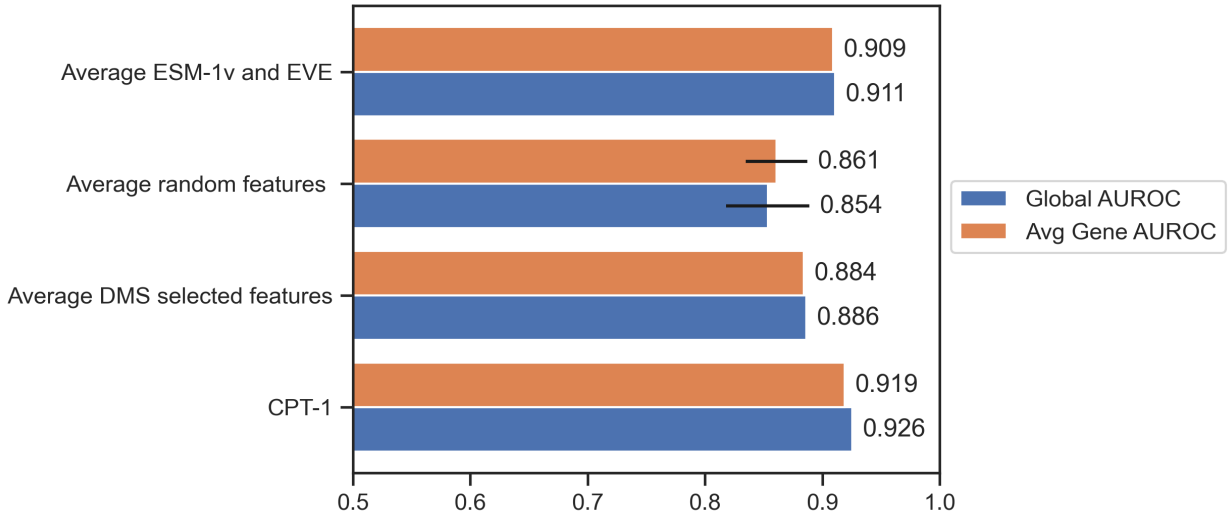

Figure S9: **Feature selection is not sufficient for increased performance.** We developed CPT-1 by first carrying out feature selection and then fitting coefficients. We tested several baselines to evaluate the importance of each of these steps. First, we tested simply averaging ESM-1v and EVE and found that this performs reasonably well. We then tested unweighted averaging of randomly selected features from our full feature set before selection; this performs much worse on average than CPT-1 with substantial variance. Finally, we tested unweighted averaging of features chosen through supervised feature selection on DMS data; this also performs much worse than CPT-1. In these experiments, unweighted feature averaging is done after setting each feature to the correct sign. These experiments indicate that fitting coefficients is a necessary step to achieve improved performance.

Supplementary Tables

Feature	Type	Description	Source
EVE	General sequence variation	Scores from neural model of protein sequences in MSA	Frazer *et al.*, 2021 [8]
ESM-1v	General sequence variation	Scores from language model of protein sequences	Meier *et al.*, 2021 [9]
100 vertebrate mut. freq.	Vertebrate alignment	Mutant frequency in orthologous proteins from 100 vertebrates	UCSC Genome Browser [24, 28]
30 mammal mut. freq.	Vertebrate alignment	Mutant frequency in orthologous proteins from 30 mammals	UCSC Genome Browser [23, 24]
Conditioned wildtype score	Structure + Sequence variation	Wildtype freq. in EVE MSA, conditioned on contact residues matching human protein	Novel
Vanilla ProteinMPNN	Structure	Scores from neural network on protein structure and partial sequence	Dauparas *et al.*, 2022 [36]
BLOSUM Index 1	Amino Acid Descriptor	The first component of BLOSUM62-derived indices.	Georgiev, 2009 [50]
ST6	Amino Acid Descriptor	The sixth principal component of ST-scales.	Yang *et al.*, 2010 [48]
Z1	Amino Acid Descriptor	The first component (lipophilicity) of Z-scales.	Sandberg *et al.*, 1998 [47]

Table S1: **Final selected features for CPT-1**. CPT-1 is trained on features based on general sequence variation, sequence variation in vertebrates, protein structure, and amino acid descriptors. Certain features are imputed across genes for the whole-proteome scale version of CPT-1 (underlined). For all amino acid descriptor features, the difference from wildtype to mutant amino acid is used.

Model	Overall AUROC	Avg gene AUROC	Median gene AUROC	Spec @ 95% Sens	Sens @ 95% Spec
CPT-1	0.926	0.919	0.944	0.641	0.665
CPT-1 (no vtMSA)	0.91	0.900	0.930	0.536	0.657
CPT-1 (no structure)	0.922	0.917	0.944	0.613	0.667
CPT-1 (cross-gene impute)	0.916	0.906	0.936	0.590	0.647
EVE	0.879	0.905	0.932	0.495	0.487
ESM-1v	0.881	0.872	0.918	0.312	0.633
100vt wt frequency	0.856	0.848	0.873	0.455	0.00
Vanilla ProteinMPNN	0.755	0.743	0.768	0.123	0.340
Conditioned mt score	0.823	0.836	0.862	0.272	0.432
Conditioned wt score	0.711	0.778	0.800	0.199	0.195

Table S2: **Detailed classification performance on ClinVar missense variants.** CPT-1 outperforms baselines on all metrics. Per-gene ClinVar AUROC is calculated in 886 genes with at least four benign and four pathogenic ClinVar variants. The margin over EVE and ESM-1v depends heavily on vertebrate alignments and, to a lesser extent, on structure. The cross-gene imputation version of CPT-1 is used to produce predictions at whole-proteome scale; it performs worse than using true feature values, but outperforms EVE and ESM-1v. 100-vertebrate MSA wildtype frequency is a powerful baseline in the high sensitivity regime. Structural features perform worse alone but have enough non-redundant signal to be worth including.

Model	Overall AUROC	Avg gene AUROC
CPT-1 (30, 70 binarize)	0.925	0.918
CPT-1 (40, 60 binarize)	0.926	0.919
CPT-1 (50, 50 binarize)	0.920	0.912
ESM-1v	0.881	0.872
EVE	0.879	0.905

Table S3: **CPT-1 performance is stable under different choices of training data binarization.** We binarized training functional assay data for CPT-1 based on percentile within each protein; variants with score below the 40th percentile were labeled benign, and variants with score above the 60th percentile were labeled pathogenic. Changing these thresholds does not change performance on ClinVar variants much, and CPT-1 always outperforms baselines.