Using deep mutational scanning data to benchmark computational phenotype predictors and identify pathogenic missense mutations

Benjamin J. Livesey and Joseph A. Marsh*

MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.

**joseph.marsh@igmm.ed.ac.uk*

Abstract

In order to deal with the huge number of novel protein-coding variants being identified by genome and exome sequencing studies, many computational phenotype predictors have been developed. Unfortunately, such predictors are often trained and evaluated on different protein variant datasets, making a direct comparison between predictors very difficult. Moreover, training and testing datasets may also overlap, introducing training bias. In this study, we use 29 previously published deep mutational scanning (DMS) experiments, which provide quantitative, unbiased phenotypic measurements for large numbers of single amino acid substitutions, in order to benchmark and compare 31 different computational phenotype predictors. We also evaluate the ability of DMS measurements and computational phenotype predictors to discriminate between pathogenic and benign missense variants. We find that DMS experiments based upon competitive growth assays tend to be superior to the top-ranking computational predictors, demonstrating the tremendous potential of DMS for identifying novel human disease mutations. Among the computational phenotype predictors, DeepSequence clearly stood out, showing both the strongest correlations with DMS data and having the best ability to predict pathogenic mutations, which is especially remarkable given that it has not been trained against human mutations. Other predictors we recommend that showed good results when tested against DMS data and human mutations include SNAP2, SNPs&GO, DEOGEN2, VEST4 and REVEL; they also benefit from being much easier for end users than DeepSequence.

Introduction

Many human genetic disorders can be attributed to sequence changes in protein-coding regions of DNA, yet damaging variants account for only a tiny fraction of the overall genetic variation seen in humans. A typical pair of unrelated individuals will differ by approximately one nonsynonymous single nucleotide variant (SNV) per protein-coding gene (1), while *de novo* mutations lead to roughly one new nonsynonymous SNV per child not observed in either parent (2–5). The vast majority of mutations identified by sequencing are 'variants of unknown significance', i.e. we are unsure if they have significant phenotypic effects or are part of the mutational background. Thus, the ability to distinguish damaging variants from benign is of tremendous importance for the diagnosis and treatment of human genetic disease.

In order to prioritise potentially pathogenic variants, a large number of computational phenotype predictors have been developed. These predictors make use of various features to produce a phenotype effect score. By far the most commonly used feature is evolutionary sequence conservation and variation (Table 2). This is the only information used by several methods such as SIFT (6) and DeepSequence (7). Other predictors integrate additional features including biophysical properties of amino acids, protein functional annotations and epigenetic data (8). Protein structural information, derived from experimentally determined models, is also used by several methods (9,10), although there is conflicting information over whether its inclusion significantly improves predictor performance (11).

While many of these approaches are able to make impressive predictions on test datasets and are widely applied in both clinical and research environments, there remain a number of unresolved sources of biases and inaccuracies. For example, when a machine-learning method is employed,

overfitting of the training set can become an issue. Instead of learning general rules, the predictor learns the niche peculiarities and noise of its training set (12). For this reason, machine-learning techniques are usually subject to out-of-sample validation, whereby data not present in the training set is used to verify that the predictor has indeed learned how to classify the data. Furthermore, when other datasets are used to benchmark these techniques, they should contain as few mutations used during training and validation as possible. Biased representation within these sets will skew the reported accuracy of methods trained and benchmarked with them (13).

Grimm et al. describe two types of data circularity that can bias the assessment of predictor accuracy (14). Type 1 circularity occurs when the data from the training set is re-used for assessing predictor performance. This can occur due to overlap between commonly used variant databases. The result is a better apparent performance than if a more appropriate validation set was used. Metapredictors (which are trained using the outputs of other predictors) amplify this issue as the methods they are built from often use differing overlapping training sets. Type 2 circularity stems from the association of certain genes with pathogenicity (e.g. many mutations in P53 will be damaging, and others being fully benign, regardless of the mutation). Tools that use this information to weight their predictions can achieve excellent results on annotated pathogenic or benign proteins, but perform poorly when faced with unannotated proteins.

An alternative to computational predictions is to experimentally characterise mutation phenotypes. While this can be extremely time consuming if a separate experiment is required for each mutation, in recent years, an assortment of approaches have been developed for the high-throughput characterization of mutation phenotypes. Deep mutational scanning (DMS) experiments combine systematic mutagenesis of a protein with a high-throughput functional test and deep sequencing (15). The result is a framework, allowing experiments to be designed that quantify the functional impact of a huge number of mutations at the same time.

DMS experiments allow direct identification of damaging human variants in a high-throughput manner (16,17). It is likely that the effects on function derived from DMS experiments will be better indicators of the clinical effect of mutations than any computational predictors. For example, a study of 179 variants in 22 human disease genes found that yeast-based functional assays could predict human pathogenicity better than current computational approaches (18). Large-scale screens of the effects of BRCA1 and P53 variants on binding using yeast two-hybrid assays showed improvements over computational algorithms (19,20). As an *in vivo* technique for protein fitness landscape analysis, DMS experiments can also be tailored to the specific definition of protein fitness required - something which computational methods are not able to account for (21). Even the best performing predictors struggle with more complex biological concepts such as allosteric regulation (22).

In addition to directly identifying damaging variants, another major benefit of DMS experiments is that they produce large, unbiased variant-effect datasets that can be used to benchmark and assess the performance of computational phenotype predictors. These are fully independent from any training and testing data used by the phenotype predictors. Previous studies have found that using unbiased DMS datasets to benchmark computational predictors results in reduced predictive power compared to other commonly used datasets, suggesting that these predictors may not be as accurate for human variants as previously reported (23). The Critical Assessment of Genome Interpretation (CAGI) experiment, which aims to drive innovations in computational phenotype predictors frequently assesses predictors against novel unseen datasets (24) including those derived from DMS experiments.

In this study, we have taken advantage of the large number of DMS experiments that have now been published for a variety of diverse proteins from different organisms. We use these datasets to perform an unbiased assessment of many different computational phenotype predictors. We also compare the ability of DMS experiments and computational phenotype predictors to directly identify pathogenic human mutations.

Results and Discussion

Overview of DMS datasets and computational phenotype predictors used in this study

To identify DMS datasets, we performed a literature search for papers presenting such experiments with available data. Using search terms such as 'deep mutational scan', 'fitness landscape', 'massively parallel mutagenesis' and 'saturation mutagenesis', we identified 29 viable DMS datasets (Table 1). As shown in Figure 1, human proteins were the most numerous targets for these DMS experiments. *Saccharomyces cerevisiae* and *Escherichia coli* were also highly represented as they endogenously produce a number of model proteins, are easy to culture and maintain, and are amenable to several effective assays for protein activity (*e.g.* growth rate and two hybrid). Proteins from viruses were also represented from studies investigating viral adaptation through massively parallel mutagenesis techniques.

There was considerable variation in functional assays applied between the DMS projects. Growth rate of yeast was the most common technique, and was applied to several human proteins by knocking out the yeast orthologue and replacing it with the human gene that is capable of rescuing the null strain (25). Viral replication assays, performed by quantitative sequencing after a certain time point, were applied to all of the viral proteins. Survival assays involved placing the organism in hostile conditions where the target protein confers an advantage such as antibiotic resistance. Two hybrid assays allow protein-protein interactions to be analysed, while fluorescence can be used to investigate enzyme activity, protein stability or transcriptional pathway activation. Phage-display assays allow a number of protein attributes to be tested *ex vivo* by externalising the protein of interest followed by selection based on its attributes. The *E. coli* toxin ccdB was assayed by reverse-survival, investigating its ability to restrict cell growth.

Each study also varied in the coverage of possible single amino acid substitutions across the entire protein (Figure 1), with only three projects investigating all possible mutations across their respective proteins (IF-1, HA-H3N2 and P53). Many of the studies included only those mutations that were possible by introducing a single nucleotide change, rather than systematically replacing each codon with every other, reducing potential coverage of all amino acid substitutions by around 70%. Some studies only focused on specific regions of the target protein. In addition, coverage was further reduced in the mutagenesis stage of most projects when specific mutants were not present in the library or if random mutagenesis was used instead of a systematic approach. For inclusion in this analysis, we required at least 5% coverage of all possible mutations in order to prevent unrepresentative low coverage data from skewing the analysis results.

The computational phenotype predictors used in this study were primarily found using the OMICtools database (26), searching for tools tagged with 'variant effect prediction'. Further tools were located by searching for 'protein variant effect prediction' and 'protein phenotype predictor' using standard internet search engines. Priority was given to tools that featured either a web interface or an API that could be queried for thousands of mutations simultaneously. We split the predictors into three broad categories, based on the features they use to make predictions:

1. *Trained predictors*. These are predictors that have been trained against a dataset of damaging and benign mutations. They can use a variety of information in their predictions including sequence conservation, Gene Ontology (GO) terms and amino acid properties. Due to our particular interest in protein structure, we also make a distinction between '*trained structural predictors*', which include information from three-dimensional protein structures (in addition to other features), and '*trained non-structural predictors*', which do not include information derived from protein structures.
2. *Untrained predictors*. These are not trained against mutation datasets, and instead make empirical calculations based on the provided input data (often a multiple sequence alignment). We also include a pair of simple amino acid substitution matrices within this group, along with a number of conservation metrics.
3. *Metapredictors*. These integrate the results of multiple other predictors and may include methods from any of the categories as features.

A number of predictors could conceivably be allocated to multiple categories, such as CADD (8), which utilizes multiple inputs from other predictors was well as sequence-derived features. In this case, the predictor is deemed to be untrained if no machine learning method is used to combine the features, trained if the majority of features are derived from protein sequence or structure, or a metapredictor if the majority of features are other predictor outputs. Note that DeepSequence is classified as an untrained predictor despite being a machine-learning method. This is because DeepSequence does not have a 'static' training set and is not trained against mutations; instead, it is re-trained using evolutionary sequence data for each new protein, essentially providing a completely new predictor (7). This approach should render DeepSequence vastly less prone to biased predictions in specific proteins than traditional supervised machine learning methods.

Among the DMS datasets, there are a few instances of the same protein being investigated in different studies by different groups. Specifically, there are four independent datasets for β-lactamase (bla) and two for UBI4. There are also two datasets for the influenza protein HA, but these were from different strains, so not directly comparable. To assess the reproducibility of DMS and its viability as a benchmark, we calculated the Spearman's correlation coefficient between the functional scores of each DMS set in the same protein. Our results (Supplemental Table S1) demonstrate a range of correlations from 0.86 (bla(a)/(c)) to 0.39 (bla(b)/(d)). The average correlations observed over all pairs of analyses was 0.58. Some level of variance is expected due to differences in experimental method, fitness assays and conditions between experiments. Overall the moderately high correlations show that DMS scores constitute a reasonably robust benchmark despite differing experimental conditions. We can also treat this correlation as a rough guide for how well we could expect a 'perfect' computational predictor to perform against DMS data from these experiments.

Assessment of computational phenotype predictors using DMS data

Where possible, we applied every computational predictor to each protein in the DMS datasets, substituting every possible amino acid at all positions. Some predictors failed to generate results for some proteins; this can occur due to an insufficiently deep multiple sequence alignment, mapping errors or other causes depending on the predictor. In order to get a measure of relative accuracy for each predictor we calculated the Spearman's rank correlation between the unbiased DMS scores for each protein and the predictions of every method (Figure 2). We also performed the same analysis using Kendall's tau (Supplemental Figure S1) which produced only minor changes in predictor ranking and lower average correlations.

Given the large number of predictors that are human-specific, we split this analysis up into human (Figure 2A) and non-human (Figure 2B). The top-performing predictor for each protein is labelled on the plot, while the full set of rankings is provided in Supplemental Tables S2-5. Figure 3 shows the relative ranking of each predictor for the human, yeast, bacterial and viral datasets using Spearman's correlation and Table S6 shows the ranking using Kendall's tau instead.

DeepSequence was the overall top performing method for predicting DMS results for the human proteins, showing the highest correlations out of all predictors for ADRB2, CALM1 and HRAS, and ranking within the top five predictors for 8/11. DeepSequence also ranked best for bacterial proteins, being the top predictor for three proteins haeIIIM, GmR and bla(c). In contrast, DeepSequence produced only a moderate rank score for yeast proteins, with a high coefficient of variance (CV). This was due to poor performance on the ubiquitin (UBI4) datasets specifically, which reduced the overall rank score considerably. If these datasets are excluded from the analysis, then DeepSequence also becomes highest ranking predictor for yeast proteins. Interestingly, DeepSequence performs poorly for viral proteins, ranking second to last out of all predictors tested. Notably, the creators report poor performance on viral proteins in the original publication, attributing this to insufficient sequence diversity within the set used (7).

Among the other predictors, certain trained non-structural approaches were particularly notable. SNPs&GO (27) ranked 2nd for human and 1st for yeast proteins, although its predictions were relatively poor for non-eukaryotic (bacterial and viral) proteins. SNAP2 (28) performed the most consistently within all groups, ranking 4th for human, 2nd for yeast, 5th for bacterial, and 1st for viral proteins. DEOGEN2 (29), a human-specific predictor, ranked 3rd for human proteins.

The predictors that incorporate protein structural information were mostly unremarkable in their performance. This is probably not due to structure unavailability, as the large majority of proteins in this study do have published structures. Three proteins only have 3D structures of orthologues available (GmR, HA-H1N1 and HA-H3N2) that were not used to make predictions, and a further three have at least 50% of the DMS measurements lying outside regions of known, high resolution (≤ 2.5 Å) structures (BRCA1, PAB1, *env*) (Supplemental Table S9). PolyPhen-2 (9) does perform notably worse on BRCA1 than any other human protein, underperforming even the substitution matrices, suggesting the lack of structural data may have impacted its predictions. However, the absence of a high quality structure for PAB1 did not appear to impact the performance of the structural predictors. Only 17% of the *env* DMS dataset was covered by a structure, which impacted both S3D-PROF and SNPs&GO3D (10) severely as they are unable to generate predictions outside areas of known structure; however, PolyPhen-2 and SuSPect (30) were not noticeably affected.

All of the metapredictors were human-specific, but were not represented among the top ranked predictors. While the REVEL metapredictor (31) produced the highest correlations with the SUMO1 and TPMT datasets, it did not consistently produce highly correlated results for other human proteins, resulting in reduced rank score (6th overall).

Of all predictors, FATHMM (32) produced the most significant outlier, generating predictions with the highest correlation of all human proteins for P53, but having low correlations for all other proteins, resulting in an overall low rank score with a high CV. The explanation for this is unclear, but it may be due to overfitting of the predictor for specific proteins, given the enrichment of P53 mutations in human disease databases compared to many of the other proteins in this study.

Different DMS datasets varied greatly in their correlations with the computational predictors. In particular BRCA1, CALM1 and TPK1 among the human proteins and Cas9, HA-H3N2 and *env* among

the non-human proteins showed low correlations, even from the best predictors. As far as we can tell, this effect appears to be unrelated to protein coverage, dataset size or DMS methodology. For example, UBE2I, SUMO1, TPK1 and CALM1 were all studied by the same group using the same methodology (growth rate in yeast) (25), yet UBE2I and SUMO1 show markedly higher correlations with all predictors than the others. Viral proteins also showed low correlations, and in fact, the BLOSUM62 substitution matrix (33) was the most highly correlated with the HA-H3N2 dataset and second highest with *env* (highest for both when using Kendall's tau). This indicates that the inclusion of typical training features are of little use when predicting the fitness of viral proteins, likely due to lack of viral representation in training sets and lack of viral sequence diversity in many databases used to generate multiple sequence alignments.

It is also interesting to note that, despite that fact that most of the predictors used in this study are human-specific, the top-ranking predictors for the human DMS datasets tend to be general predictors applicable to proteins from all species. For example, for the human proteins, only one of the top five and three of the top 12 predictors are human specific. In contrast, most of the lowest ranked predictors are human-specific. An important contributing factor to this is likely to be overfitting against human mutation datasets for some predictors, which causes them to perform poorly against unbiased experimental phenotype measurements. Three of the four worst predictors are also based upon nucleotide-level constraint (GERP++ (34), SIPHY (35) and PHYLOP (36)). These predictors ranked even lower than the simple BLOSUM62 and Grantham (37) substitution matrices, suggesting that such approaches are poorly suited to predicting the protein-level effects of mutations.

Identification of pathogenic human mutations using DMS data and computational phenotype predictors

We next investigated the ability of both the DMS experiments and the computational phenotype predictors to distinguish pathogenic human missense mutations, taken from the ClinVar database (38), from missense variants observed in the human population, taken from gnomAD (39). While some gnomAD variants may be pathogenic under certain circumstances (*e.g.* if associated with recessive, late-onset or incomplete penetrance disease), we assume that the vast majority of them should be non-pathogenic and refer to them as "putatively benign". Of the 11 human proteins with DMS datasets, 7 have known pathogenic or likely pathogenic missense variants in ClinVar (93 for BRCA1, 31 for HRAS, 189 for P53, 108 for PTEN, 9 for CALM1, 5 for TPK1 and 2 for MAPK1). For CALM1 and TPK1, we identified additional pathogenic missense mutations in the literature (40–44), leading to a total of 19 for CALM1 and 8 for TPK1. MAPK1 has too few recorded pathogenic variants to include in this analysis.

For each predictor, we plotted a receiver operating characteristic (ROC) curve for performance on every protein, classifying pathogenic ClinVar mutations as true positives and the putatively benign gnomAD mutations as true negatives (removing any ClinVar mutations from the gnomAD set). We then calculated the area under the curve (AUC) for each plot as a measure of that predictor's accuracy in classifying the data (Figure 4).

The experimental DMS data performed better than any of the 31 computational phenotype predictors for two of the six human proteins, P53 and TPK1, and ranked 6th for CALM1. For the other three human proteins, the DMS data was less useful, with several computational predictors performing much better. It is notable, however, that the three DMS experiments where the data is most predictive of disease mutations are all based upon competitive growth assays, while the other three experiments are based upon a two-hybrid approach (BRCA1 and HRAS) or fluorescence of a

GFP fusion protein (PTEN). We also repeated this analysis using only those computational predictions for which there are also DMS measurements, which had very little influence on the relative ranking of the DMS data (Supplemental Figure S2).

The above results suggest that DMS experiments using growth rate as their phenotype are the most useful for identifying disease mutations, and can perform substantially better than most or all computational predictors. It is likely that this will depend on the mechanism by which mutations cause disease. For example, if a disease were solely caused by mutations that disrupt a specific protein-protein interaction, then probing this interaction with a two-hybrid experiment would likely be a good way of assessing disease. However, if some of the BRCA1 and HRAS mutations acted via some mechanism other than by perturbing their interactions with specific binding partners (BARD1 or RasGAP, respectively), as is almost certainly the case, this can explain the relative underperformance of the DMS data: it simply is not suited to identify some pathogenic mutations. The phenotypic screen for PTEN measures protein abundance in the cell by fluorescence of EGFP bound to the protein (17). This technique called VAMP-seq identifies thermodynamically unstable variants; however, this may fail to capture disease mechanisms acting through interaction disruption and loss or gain of function unrelated to destabilisation. Thus great care must be taken when selecting an experimental phenotype. In the absence of a better phenotypic assay specifically related to a known disease mechanism, growth rate may be the most general way of probing loss of protein function, and thus the most useful for predicting disease.

Among the computational predictors, DeepSequence was again the best, ranking 1st for three proteins: CALM1, HRAS and TPK1 (although it was inferior to DMS data for TPK1). It also performed reasonably well for the remaining proteins, ranking 3rd for BRCA1, 5th for PTEN and 7th for P53. This is far better performance than any of the other computational predictors, and is especially remarkable considering the fact that it, in contrast to most of the other predictors, it has not been trained on human mutation data.

A few other predictors performed fairly well. In particular, SNAP2, SNPs&GO, DEOGEN2, REVEL and VEST4 (45) all ranked among the top five for at least half of the protein targets. These are all trained predictors and metapredictors, which means that this analysis is very likely to at least somewhat overstate their predictive performance. This is because there sure to be at least some overlap between the mutations used to evaluate the predictors here, and the mutations originally used to train the predictors. Importantly, however, all of these predictors performed reasonably well in terms of their correlations with the human DMS data, suggesting that their relative success here is not simply due to training bias. However, some of these predictors did perform poorly on certain targets. For example, DEOGEN2 ranked last for BRCA1. Interestingly, however, the relative performance of DEOGEN2 improved markedly if only predictions of mutations with DMS measurements, which covers primarily just the RING domain of BRCA1, are considered (Supplemental Figure S3). This appears to be due to DEOGEN2 assigning extremely different weights to different domains in BRCA1, thus obscuring good predictions when analysing the entire protein.

Conclusions

The number of available genome and protein sequences has increased tremendously in the last decade due to advances in next-generation sequencing technologies. In this wealth of new data, we have discovered a large number of previously unseen coding variants of unknown functional significance. To assist us in analysing this new data, computational phenotype predictors have been developed by various groups, but the training and evaluation of these predictors often suffer from biases. DMS experiments provide an ideal benchmark for testing variant effect predictors, ensuring

that none of the training data is included in the test. The availability of a large number of such experimental DMS datasets has facilitated this study.

Of the 31 different predictors evaluated in this study, we find that a single program, DeepSequence, clearly stands out from all of the others, both in terms of performance, and in terms of methodology. DeepSequence showed the strongest correlations with the DMS data in humans and bacteria and was the top predictor of human disease mutations. Most machine-learning methods make use of several features, often including some measure of sequence conservation at the site of interest, and then learn the patterns of these features that result in a mutation being classified as damaging or benign. DeepSequence makes use of deep generative models to integrate factors from the entire sequence at once rather than only one or a few sites. This type of problem is largely intractable for traditional machine learning, given the number of parameters involved; however, DeepSequence overcomes this by learning the latent factors underlying the protein sequence. This approach also produces advantages in terms of the biases inherent with training on a labelled dataset. We can expect a machine learning method confronted with an example it was trained on to correctly classify it correctly the majority of the time, producing an unrepresentative assessment of its accuracy. DeepSequence makes use of multiple sequence alignments and never sees labelled protein data, resulting in scores that are not biased by training examples. This is not to say that DeepSequence is a completely unbiased method, however. The scores which are generated depend entirely upon the database from which multiple sequence alignments are drawn. If this is under-representative of certain sequences, then predictions for those will be lower quality such as the results we observe for viral proteins drawn from the UniRef100 database. The success that DeepSequence has achieved in predicting mutations effects for human proteins show that deep generative models may well be the way forward in this field, removing the reliance on labelled datasets for making predictions.

The two most commonly used computational phenotype predictors at the moment are probably PolyPhen-2 and SIFT, which are both still very widely used in variant prioritisation. Neither showed exceptional performance in this study, ranking 14th and 16th against the human DMS data. They also both tended to rank in the middle compared to all other predictors in terms of identifying human pathogenic mutations. Therefore, we recommend other computational phenotype predictors based upon our analyses. Unfortunately, DeepSequence is very computationally intensive and quite difficult for a typical end user to run. It also does not have defined disease thresholds; these would need to be assessed on a protein-by-protein basis, likely by analysis of putatively benign (e.g. gnomAD) variants. We therefore also highlight SNAP2, SNPs&GO, DEOGEN2 and VEST4, which also tended to perform well against both the DMS and human mutation datasets, and have simple-to-use web interfaces, so these would make good choices for routine variant prioritisation. REVEL lacks a web interface, but has been pre-calculated for all human chromosomes and is available online to download. Importantly, however, these approaches all showed large variation in their performance between different proteins, suggesting that one should still not rely too much on the results of any single predictor.

It is interesting that the inclusion of protein structural models did not appear to be particularly useful for the computational phenotype predictors. In principle, since disease mechanisms can often be explained by protein structural effects (46) one would expect that protein structure should be useful. It may be that the value of evolutionary information simply dwarfs any contribution from the inclusion of structure, *i.e.* if a mutation is damaging at a structural level, this is likely to be reflected in the evolutionary conservation of that residue. Moreover, many pathogenic mutations are not highly damaging at a protein structural level, *e.g.* those associated with a dominant-negative effect

in protein complexes (47) or those that affect transcription factor binding specificity (48). It is possible that future approaches that take into consideration the diverse molecular mechanisms underlying human genetic disease and the unique structural properties of individual proteins will be able to make better use of the huge amount of protein structural data now available.

The value of DMS data for directly identifying pathogenic mutations is especially exciting, based on the results we observed here. Given the proper choice of experimental phenotype, DMS experiments are likely to be better than (or at the very least competitive with) the best computational phenotype predictors. The applicability of DMS data for direct variant prioritisation is currently limited by the small fraction of human protein residues for which DMS experiments have been performed. In the coming years, as more proteins are studied and experimental strategies are improved, we expect that the utilisation of such data for the identification of damaging variants will become routine.

Methods

DMS datasets

In total, 37 DMS datasets were identified in the literature. Four were excluded due to either lack of an associated dataset or lack of a protein reference sequence. A further four were excluded due to low mutational coverage of the target protein (<5%) on the basis that they may skew the analysis with unrepresentative results, leaving the 29 DMS datasets used in this study.

Computational phenotype predictors

Where possible, data was retrieved from web pages or APIs where these methods supported large batch submissions. Some methods were installed and run locally instead. Default or recommended settings were used for all predictors where applicable. See Table S8 for a full summary of web addresses and predictor settings used.

Additional information required by predictors were retrieved from UniProt (UniProt ID, PDB mappings), QuickGO (GO terms), EMBL (CDS) and ensemble (genome coordinates for VCF generation).

Where possible, a structure with 2.5 Å resolution or lower was used for the SNPs&GO3D and SNP-PROF predictors. The PDB ID and chain used to generate predictions for each protein are shown in Table S9.

Calculating rank scores

Rank score is defined as the mean, normalised correlation over all proteins, given by the following formula:

$$R = \frac{\sum^{m} \frac{c - c_{min}}{c_{max} - c_{min}}}{m_x}$$

where c is each correlation for a specific protein, c_{min} is the minimum correlation for each protein and c_{max} is the maximum correlation for each protein. This represents the correlation, normalised to a scale between 1 for the highest ranking method and 0 for the lowest. This is then summed across all proteins (m) for the same method, and divided by the number of proteins for which this method generated a result (m_x), in order to normalise for instances where a predictor failed to generate results for a certain protein. Where multiple DMS datasets are present for a single protein, we

averaged the normalised correlations of each predictor between these datasets, and treated the resulting values as scores from a single protein.

Coefficient of variation is calculated from the normalised correlations, before the mean is taken. It is the standard deviation of these values across all proteins, divided by the mean. This represents the variation in predictor rank between different proteins.

It should be noted that rank scores are only comparable within the set of proteins which was used to calculate it, nor does it convey any information about predictor accuracy. The rank score metric can only be used for relative ranking within a set of proteins.

Human mutation datasets

Data was retrieved from gnomAD v2.1 by searching for each of the human genes at https://gnomad.broadinstitute.org/. Data was filtered so that only missense mutations were present and a CSV file containing the mutations was downloaded.

Each gene was also searched for in the ClinVar database at https://www.ncbi.nlm.nih.gov/clinvar. Data was filtered so that only missense mutations labelled as 'pathogenic' or 'likely pathogenic' were present and downloaded as a tab-delimited file.

Plotting ROC curves and calculating AUC values

To plot the ROC curves, mutations present in the gnomAD dataset were taken as true negatives, while mutations present in the ClinVar dataset were taken as true positives. Mutations present in both sets were removed from the gnomAD set. The 'roc_curve' and 'auc' functions for the sklearn python package were used to calculate the true positive rate (TPR) and false positive rate (FPR) and the AUC. As some predictors utilise inverse metrics and thus produce an AUC under 0.5, we multiplied the predictions of all such methods by -1 to bring the value above 0.5; this is equivalent to inverting the TPR and FPR.

Acknowledgements

J.A.M. is supported by an MRC Career Development Award (MR/M02122X/1) and is a Lister Institute Research Prize Fellow. B.J.L. is supported by the MRC Precision Medicine Doctoral Training Programme.

References

1. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. The Lancet. 2012 Nov 10;380(9854):1674–82.

2. de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability [Internet]. http://dx.doi.org/10.1056/NEJMoa1206524. 2012 [cited 2019 Jul 2]. Available from: https://www.nejm.org/doi/10.1056/NEJMoa1206524?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dwww.ncbi.nlm.nih.gov

3. Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen AS, Berkovic SF, Cossette P, Delanty N, et al. De novo mutations in epileptic encephalopathies. Nature. 2013 Sep 12;501(7466):217–21.

4. Fitzgerald T, Gerety S, Jones W, van Kogelenberg M, King D, McRae J, et al. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015 Mar 12;519(7542):223–8.

5. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012 Apr 4;485(7397):242–5.

6. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W452–7.

7. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 2018 Oct;15(10):816.

8. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D886–94.

9. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

10. Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics. 2011 Jul 5;12(Suppl 4):S3.

11. Carraro M, Minervini G, Giollo M, Bromberg Y, Capriotti E, Casadio R, et al. Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. Hum Mutat. 2017;38(9):1042–50.

12. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014 Jun 14;15:30.

13. Schaafsma GCP, Vihinen M. Representativeness of variation benchmark datasets. BMC Bioinformatics. 2018 Nov 29;19(1):461.

14. Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015 May;36(5):513–23.

15. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014 Aug;11(8):801–7.

16. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet. 2016 Dec;48(12):1570–5.

17. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 2018 Jun;50(6):874.

18. Sun S, Yang F, Tan G, Costanzo M, Oughtred R, Hirschman J, et al. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 2016 May;26(5):670–80.

19. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, et al. Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci U S A. 2015 Sep 15;112(37):E5189–98.

20. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics. 2015 Jun;200(2):413–22.

21. Harris DT, Wang N, Riley TP, Anderson SD, Singh NK, Procko E, et al. Deep Mutational Scans as a Guide to Engineering High Affinity T Cell Receptor Interactions with Peptide-bound Major Histocompatibility Complex. J Biol Chem. 2016 Nov 18;291(47):24566–78.

22. Xu Q, Tang Q, Katsonis P, Lichtarge O, Jones D, Bovo S, et al. Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. Hum Mutat. 2017;38(9):1123–31.

23. Mahmood K, Jung C, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics. 2017 May 16;11(1):10.

24. Hoskins RA, Repo S, Barsky D, Andreoletti G, Moult J, Brenner SE. Reports from CAGI: The Critical Assessment of Genome Interpretation. Hum Mutat. 2017;38(9):1039–41.

25. Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, et al. A framework for exhaustively mapping functional missense variants. Mol Syst Biol [Internet]. 2017 Dec 21 [cited 2019 Jun 5];13(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5740498/

26. Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ, Desfeux A. OMICtools: an informative directory for multi-omic data analysis. Database J Biol Databases Curation [Internet]. 2014 Jul 14 [cited 2019 Jun 3];2014. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4095679/

27. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics. 2013 May 28;14(3):S6.

28. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics. 2015 Jun 18;16(8):S1.

29. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017 Jul 3;45(Web Server issue):W201–6.

30. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. J Mol Biol. 2014 Jul 15;426(14):2692–701.

31. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 6;99(4):877–85.

32. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013 Jan;34(1):57–65.

33. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915–9.

34. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Comput Biol. 2010 Dec 2;6(12):e1001025.

35. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009 Jun 15;25(12):i54–62.

36. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan;20(1):110–21.

37. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 1974 Sep 6;185(4154):862–4.

38. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D980–5.

39. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes: Supplementary Information [Internet]. Genomics; 2019 Jan [cited 2019 Aug 8]. Available from: http://biorxiv.org/lookup/doi/10.1101/531210

40. Banka S, de Goede C, Yue WW, Morris AAM, von Bremen B, Chandler KE, et al. Expanding the clinical and molecular spectrum of thiamine pyrophosphokinase deficiency: A treatable neurological disorder caused by TPK1 mutations. Mol Genet Metab. 2014 Dec 1;113(4):301–6.

41. Crotti Lia, Johnson Christopher N., Graf Elisabeth, De Ferrari Gaetano M., Cuneo Bettina F., Ovadia Marc, et al. Calmodulin Mutations Associated With Recurrent Cardiac Arrest in Infants. Circulation. 2013 Mar 5;127(9):1009–17.

42. Jensen HH, Brohus M, Nyegaard M, Overgaard MT. Human Calmodulin Mutations. Front Mol Neurosci [Internet]. 2018 [cited 2019 Jul 9];11. Available from: https://www.frontiersin.org/articles/10.3389/fnmol.2018.00396/full

43. Nomikos M, Thanassoulas A, Calver BL, Beck K, Vassilakopoulou V, Buntwal L, et al. Calmodulin Mutations Associated with Congenital Cardiac Disease Display Novel Biophysical and Biochemical Characteristics. Biophys J. 2018 Feb 2;114(3):467a.

44. Zhu L, Wu R, Ye Z, Gu R, Wang Y, Hou Y, et al. Identification of two novel TPK1 gene mutations in a Chinese patient with thiamine pyrophosphokinase deficiency undergoing whole exome sequencing. J Pediatr Endocrinol Metab. 2019;32(3):295–300.

45. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.

46. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. Trends Genet. 2003 Sep 1;19(9):505–13.

47. Bergendahl LT, Gerasimavicius L, Miles J, Macdonald L, Wells JN, Welburn JPI, et al. The role of protein complexes in human genetic disease. Protein Sci. 2019;28(8):1400–11.

48. Williamson KA, Hall HN, Owen LJ, Livesey BJ, Hanson IM, Adams GGW, et al. Recurrent heterozygous PAX6 missense variants cause severe bilateral microphthalmia via predictable effects on DNA–protein interaction. Genet Med. 2019 Nov 8;1–12.

49. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, et al. Mutational processes shape the landscape of TP53 mutations in human cancer. Nat Genet. 2018 Oct;50(10):1381.

50. Bandaru P, Shah NH, Bhattacharyya M, Barton JP, Kondo Y, Cofsky JC, et al. Deconstruction of the Ras switching cycle through saturation mutagenesis. Valencia A, editor. eLife. 2017 Jul 7;6:e27810.

51. Brenan L, Andreev A, Cohen O, Pantel S, Kamburov A, Cacchiarelli D, et al. Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. Cell Rep. 2016 Oct 18;17(4):1171–83.

52. Jones EM, Lubock NB, Venkatakrishnan AJ, Wang J, Tseng AM, Paggi JM, et al. Structural and Functional Characterization of G Protein-Coupled Receptors with Deep Mutational Scanning. bioRxiv. 2019 May 10;623108.

53. Mishra P, Flynn JM, Starr TN, Bolon DNA. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. Cell Rep. 2016 Apr 19;15(3):588–98.

54. Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J Mol Biol. 2013 Apr 26;425(8):1363–77.

55. Roscoe BP, Bolon DNA. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. J Mol Biol. 2014 Jul 29;426(15):2854–70.

56. Melamed D, Young DL, Gamble CE, Miller CR, Fields S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA. 2013 Nov;19(11):1537–51.

57. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively Parallel Single Amino Acid Mutagenesis. Nat Methods. 2015 Mar;12(3):203–6.

58. Kelsic ED, Chung H, Cohen N, Park J, Wang HH, Kishony R. RNA structural determinants of optimal codons revealed by MAGE-seq. Cell Syst. 2016 Dec 21;3(6):563-571.e6.

59. Dandage R, Pandey R, Jayaraj G, Rai M, Berger D, Chakraborty K. Differential strengths of molecular determinants guide environment specific mutational fates. PLOS Genet. 2018 May 29;14(5):e1007419.

60. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. Mol Biol Evol. 2014 Jun;31(6):1581–92.

61. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell. 2015 Feb 26;160(5):882–92.

62. Jacquier H, Birgy A, Nagard HL, Mechulam Y, Schmitt E, Glodt J, et al. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci. 2013 Aug 6;110(32):13067–72.

63. Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, et al. Deep Sequencing of Systematic Combinatorial Libraries Reveals β-lactamase Sequence Constraints at High Resolution. J Mol Biol. 2012 Dec 7;424(3–4):150–67.

64. Adkar BV, Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, et al. Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. Structure. 2012 Feb 8;20(2):371–81.

65. Rockah-Shmuel L, Tóth-Petróczy Á, Tawfik DS. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. PLOS Comput Biol. 2015 Aug 14;11(8):e1004421.

66. Spencer JM, Zhang X. Deep mutational scanning of S. pyogenes Cas9 reveals important functional domains. Sci Rep. 2017 Dec 4;7(1):16836.

67. Haddox HK, Dingens AS, Bloom JD. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. PLOS Pathog. 2016 Dec 13;12(12):e1006114.

68. Doud MB, Bloom JD. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. Viruses [Internet]. 2016 Jun 3 [cited 2019 Jun 5];8(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4926175/

69. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc Natl Acad Sci U S A. 2018 Aug 28;115(35):E8276–85.

70. Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, et al. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLOS Genet. 2015 Jul 1;11(7):e1005310.

71. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018 Feb 1;34(3):511–3.

72. Johansen MB, Izarzugaza JMG, Brunak S, Petersen TN, Gupta R. Prediction of Disease Causing Non-Synonymous SNPs by the Artificial Neural Network Predictor NetDiseaseSNP. PLoS ONE [Internet]. 2013 Jul 25 [cited 2019 Jun 5];8(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3723835/

73. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006 Nov 15;22(22):2729–34.

74. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. PLoS ONE [Internet]. 2015 Feb 3 [cited 2019 Jun 5];10(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315405/

75. Niroula A, Vihinen M. Predicting Severity of Disease-Causing Variants [Internet]. 2017 [cited 2019 Jun 5]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/humu.23173

76. Limongelli I, Marini S, Bellazzi R. PaPI: pseudo amino acid composition to score human protein-coding variants. BMC Bioinformatics. 2015 Apr 19;16(1):123.

77. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011 Sep;39(17):e118.

78. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE [Internet]. 2012 Oct 8 [cited 2019 Jun 3];7(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3466303/

79. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. Am J Hum Genet. 2011 Apr 8;88(4):440–9.

80. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016 Dec;48(12):1581–6.

Tables and figures

Table 1. Summary of all DMS datasets included in this analysis. Where multiple datasets exist for the same protein, they are indicated by adding sequential letters to the end of the protein identifier

Protein(s) (Uniprot ID)	Organism	Mutagenesis Method	Functional Assay	Reported Metric	Reference
UBE2I (P63279) SUMO1 (P63165) TPK1 (Q9H3S4) CALM1 (P0DP23)	Human	POPCode, a variant of multiple-site directed mutagenesis.	Competitive growth assay in yeast.	Normalised fitness scores.	(25)
BRCA1 (P38398)	Human	Systematic site-directed mutagenesis.	Yeast two-hybrid assay and phage display.	Inverse log of the slope of log enrichment ratio over all time points.	(20)
P53 (P04637)	Human	Systematic site-directed mutagenesis.	Competitive growth assay in the presence of P53 agonists.	Z-scores	(49)
HRas (P01112)	Human	Systematic site-directed mutagenesis.	Two hybrid assay.	Relative enrichment.	(50)
MAPK1 (P28482)	Human	Systematic site-directed mutagenesis.	Competitive growth assay.	Log enrichment.	(51)
PTEN (P60484) TPMT (P51580)	Human	Systematic site-directed mutagenesis.	Fluorescence of a GFP fusion protein.	Abundance scores.	(17)
ADRB2 (P07550)	Human	Systematic variant synthesis.	Pathway-specific reporter gene transcription assessed by RNA-seq.	Normalized pathway activity.	(52)
HSP82 (P02829)	Yeast	Systematic site-directed mutagenesis.	Competitive growth assay.	Normalised enrichment.	(53)
UBI4(a) (P0CG63)	Yeast	Systematic site-directed mutagenesis.	Competitive growth assay.	Selection coefficients.	(54)
UBI4(b) (P0CG63)	Yeast	Site directed mutagenesis by cassette ligation.	Fluorescence activated cell sorting (FACS).	Relative reactivity.	(55)
PAB1 (P04147)	Yeast	Random mutagenesis by error-prone PCR.	Competitive growth assay.	Enrichment score.	(56)
GAL4 (P04386)	Yeast	Systematic site-directed mutagenesis.	Two-hybrid assay.	Log effect size.	(57)

IF-1 (P69222)	*E. coli*	Systematic site-directed mutagenesis.	Competitive growth assay.	Relative fitness.	(58)
GmR (N/A)	*E. coli*	Systematic site-directed mutagenesis.	Antibiotic resistance.	Log enrichment.	(59)
bla(a) (P62593)	*E. coli*	Systematic site-directed mutagenesis.	Antibiotic resistance.	Normalised fitness.	(60)
bla(b) (P62593)	*E. coli*	Systematic site-directed mutagenesis.	Antibiotic resistance.	Relative fitness.	(61)
bla(c) (P62593)	*E. coli*	Random mutagenesis.	Antibiotic resistance.	Log growth rate.	(62)
bla(d) (P62593)	*E. coli*	Random and site-directed mutagenesis.	Antibiotic resistance.	ΔΔG	(63)
ccdB (P62554)	*E. coli*	Systematic site-directed mutagenesis.	Reverse survival assay (for toxin activity).	Mutational sensitivity (score from 2-9).	(64)
haeIIIM (P20589)	*H.aegyptius*	Random mutagenesis.	Competitive growth assay.	Mutation frequencies.	(65)
Cas9 (Q99ZW2)	*S. pyrogenes*	Random mutagenesis by error-prone PCR.	Survival assay.	Log enrichment.	(66)
env (P03377)	HIV virus	Systematic site-directed mutagenesis.	Competitive replication assay.	Positional amino acid preferences.	(67)
HA-H1N1 (A0A2Z5U3Z0)	Influenza virus	Helper virus approach.	Competitive replication assay.	Positional amino acid preferences.	(68)
HA-H3N2 (A0A097PF60)	Influenza virus	Helper virus approach.	Competitive replication assay.	Positional amino acid preferences.	(69)
PA (P15659)	Influenza virus	Random mutagenesis.	Competitive replication assay.	Relative fitness.	(70)

Table 2. Summary of the computational predictors used in this analysis, including the predictor category, features and number of proteins for which the predictor generated a prediction (human/yeast/bacteria/virus).

Predictor	Category	Features	Query Requirements	Total prediction sets (h/y/b/v)	Reference
CADD	Trained Nonstructural	Sequence conservation, surrounding features, epigenetic data.	Human nucleotide substitution.	11/0/0/0	(8)
DEOGEN2	Trained Nonstructural	PROVEAN, sequence conservation, protein interactions, early folding predictions, gene essentiality.	Human amino acid substitution.	11/0/0/0	(29)
Fathmm	Trained Nonstructural	Sequence conservation in Hidden Markov Models (HMMs).	Human amino acid substation.	11/0/0/0	(32)
FathmmXF	Trained Nonstructural	Sequence conservation, genomic context.	Human nucleotide substitution.	11/0/0/0	(71)
NetDisease SNP	Trained Nonstructural	Sequence conservation, predicted surface accessibility.	Amino acid substitution, any organism.	11/5/6/4	(72)
PhD SNP	Trained Nonstructural	Sequence environment, sequence profile.	Amino acid substitution, any organism.	11/5/6/4	(73)
PON-P2	Trained Nonstructural	Sequence conservation, GO terms, residue biophysical features.	Human amino acid substitution.	11/0/0/0	(74)
PON-PS	Trained Nonstructural	Sequence conservation, biophysical features, sequence properties.	Human amino acid substitution.	11/0/0/0	(75)
PseAA	Trained Nonstructural	Residue pairwise relations, sequence conservation, protein-level quantitative features.	Human nucleotide substitution.	11/0/0/0	(76)
SNAP2	Trained Nonstructural	Sequence conservation, protein structural features, residue biophysical features, predicted secondary structural features, functional annotations, SIFT predictions.	Amino acid substitution, any organism.	11/5/6/4	(28)
SNPs&GO	Trained Nonstructural	Sequence conservation, sequence environment, PANTHER output, GO terms.	Amino acid substitution, any organism.	11/5/6/4	(27)

VEST4	Trained Nonstructural	86 pre-computed features from the SNVbox database.	Human amino acid substitutions.	11/0/0/0	(45)
PolyPhen 2 (PPH2)	Trained Structural	Sequence conservation, secondary structural features, protein structural features, biophysical features.	Amino acid substitution. Any organism.	11/5/6/2	(9)
S3D-PROF	Trained Structural	Sequence conservation, structural features.	PDB structure and amino acid substitution. Any organism.	10/5/4/2	(10)
SNPs&GO3D	Trained Structural	Sequence conservation, structural features, PANTHER output, GO terms.	PDB structure and amino acid substitution. Any organism	10/5/4/2	(10)
SuSPect	Trained Structural	Protein interactions, functional annotations, sequence conservation, protein structural features.	Amino acid substitution, any organism.	11/5/6/4	(30)
BLOSUM62	Untrained (substitution matrix)	N/A	Amino acid substitution, any organism.	11/5/6/4	(33)
DeepSequence (DeepSeq)	Untrained	Sequence conservation	Amino acid substitution, any organism.	11*/5/6/3	(7)
GERP++	Untrained	Sequence conservation	Human nucleotide substitution.	11/0/0/0	(34)
Grantham	Untrained (substitution matrix)	N/A	Amino acid substitution, any organism.	11/5/6/4	(37)
Mutation Assessor Functional Impact Score (MA FI)	Untrained	Sequence conservation, conservation between subfamilies.	Human amino acids substitution.	11/0/0/0	(77)
Mutation Assessor Variant Conservation Score (MA VC)	Untrained	Sequence conservation, conservation between subfamilies.	Human amino acid substitution.	10/0/0/0	(77)

PANTHER	Untrained		Amino acid substitution. Any organism in the PANTHER database.	11/4/1/0	
PhyloP	Untrained	Sequence conservation	Human nucleotide substitution.	11/0/0/0	(36)
PROVEAN	Untrained	Sequence conservation	Amino acid substitution, any organism.	11/5/6/4	(78)
SIFT	Untrained	Sequence Conservation	Amino acid substitution, any organism.	11/5/6/4	(6)
Siphy	Untrained	Sequence conservation	Human nucleotide substitution.	11/0/0/0	(35)
CONDEL	Metapredictor	MutationAssessor, PPH2, SIFT and Pfam E-value.	Human nucleotide substitution.	11/0/0/0	(79)
M-CAP	Metapredictor	Nine other predictors, sequence and genomic conservation.	Human nucleotide substitution.	11/0/0/0	(80)
PAPI	Metapredictor	PseAA, PPH2, SIFT.	Human nucleotide substitution.	11/0/0/0	(76)
REVEL	Metapredictor	SIFT, PPH2, PROVEAN, LRT, MutationTaster, MutationAssessor, Fathmm, VEST4 and conservation scores.	Human nucleotide substitution.	11/0/0/0	(31)

* Only the first 600 amino acids of BRCA1 were run using DeepSequence due to the high computational requirements with longer proteins.

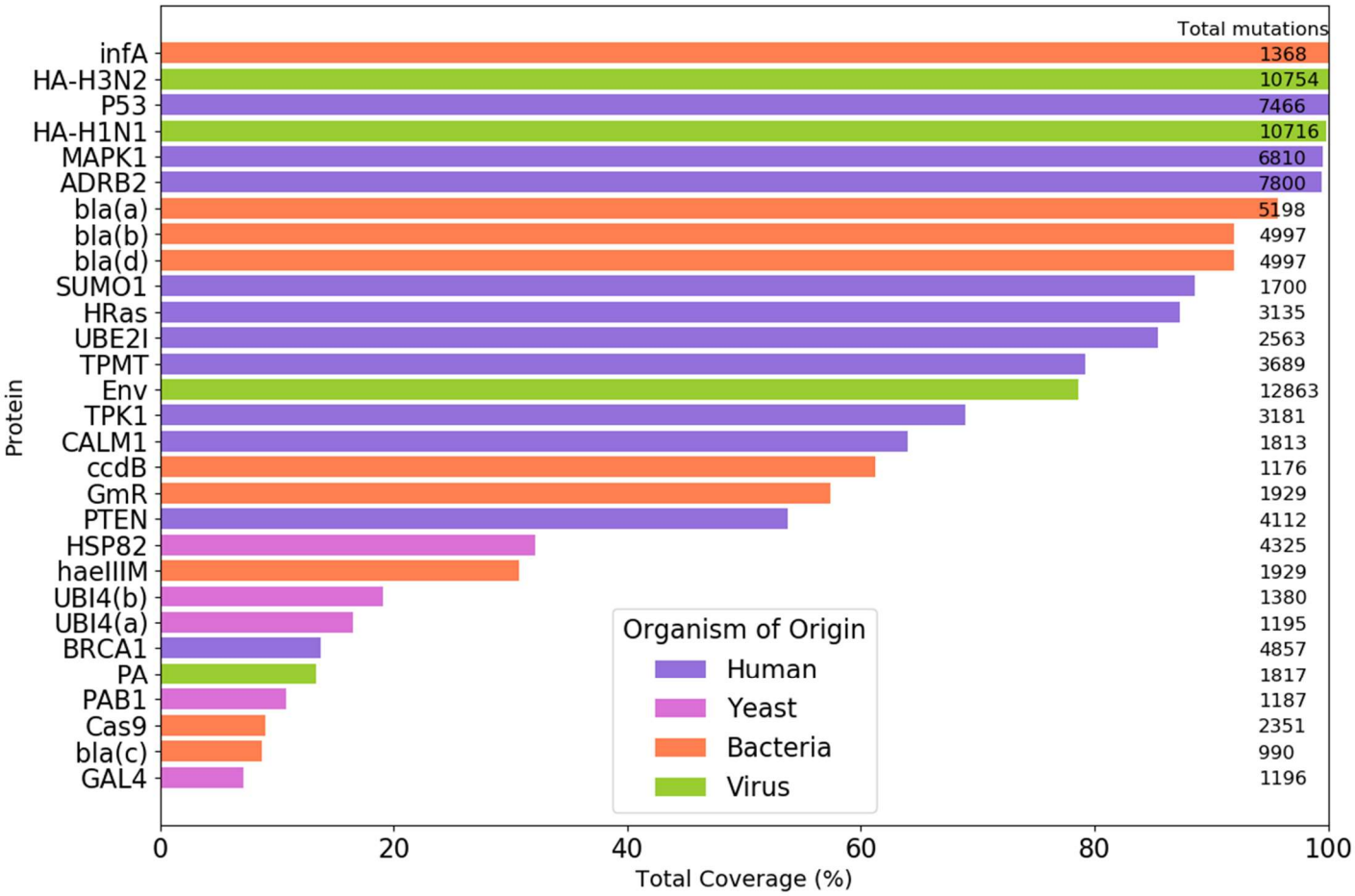

Figure 1. Coverage of the target proteins by DMS experiments shown as a percentage of the total number of single amino acid substitutions for that protein represented in the experiment (stop codons are excluded). This is under the assumption that any amino acid can occupy any position, not just those possible by SNVs. The chart is colour-coded by the organism from which each protein is derived.

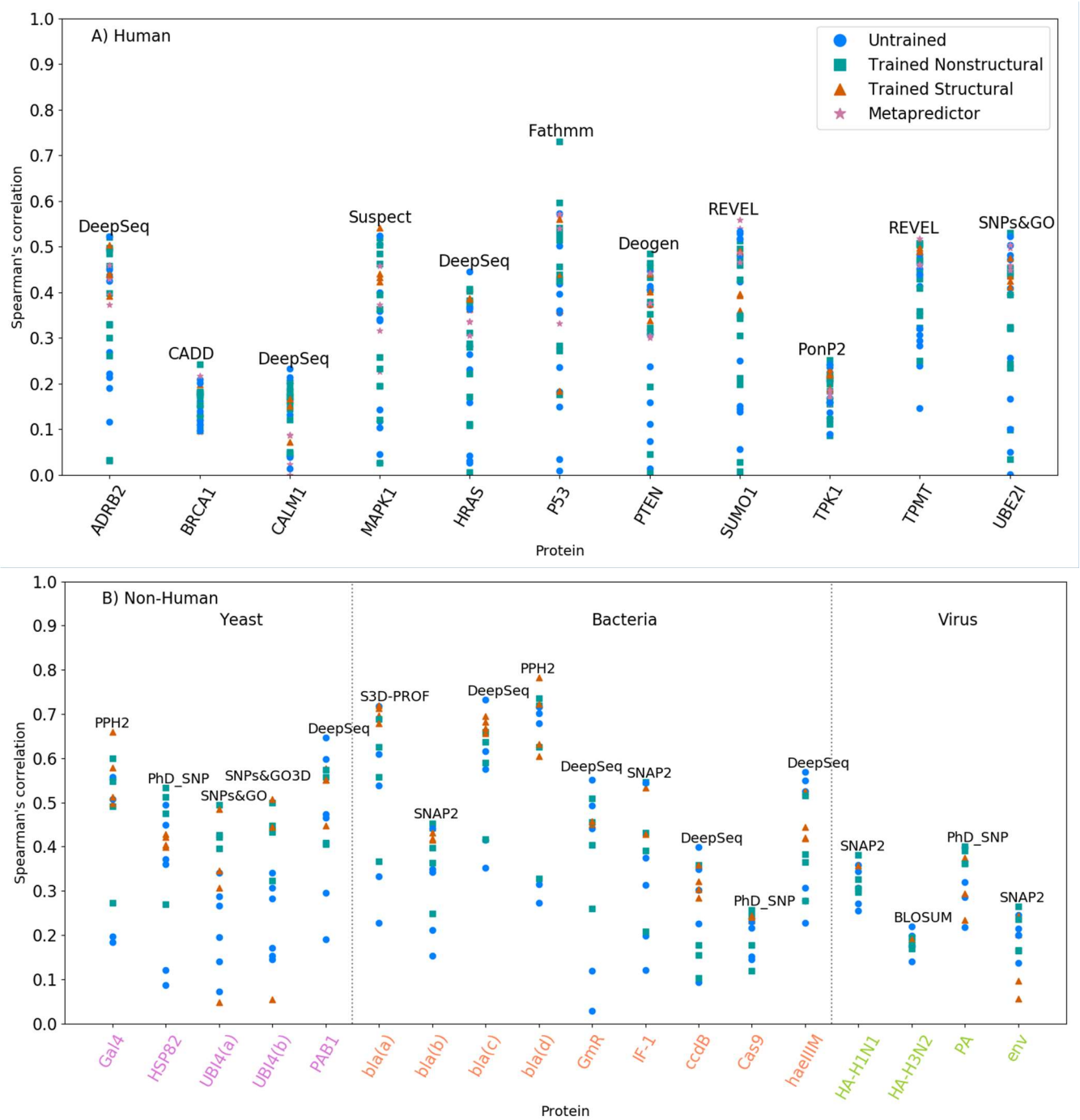

Figure 2. Comparison of the computational predictors to each DMS dataset. A) Spearman's correlations between all predictions and the DMS results for each human protein DMS dataset. B) Spearman's correlations between predictions and the DMS results for all non-human protein DMS datasets. Protein names are coloured by organism of origin as in Figure 1. The top computational predictor is labelled for each DMS experiment, while the full rankings are provided in Supplemental Tables S2-5.

A. Human Proteins

Predictor	RankScore	CV
DeepSequence	0.893	0.154
SNPs&GO	0.817	0.177
DEOGEN2	0.802	0.216
SNAP2	0.796	0.219
S3D-PROF	0.793	0.094
REVEL	0.789	0.256
SuSPect	0.788	0.257
VEST4	0.782	0.187
PhD SNP	0.774	0.133
PANTHER	0.771	0.185
SNPs&GO3D	0.740	0.250
PROVEAN	0.738	0.221
MA FI	0.728	0.305
PolyPhen2	0.704	0.347
MA VC	0.700	0.305
SIFT	0.665	0.182
M-CAP	0.647	0.337
CONDEL	0.626	0.480
CADD	0.592	0.418
NetDiseaseSNP	0.590	0.263
PAPI	0.590	0.481
PON-PS	0.564	0.343
PON-P2	0.517	0.486
BLOSUM62	0.428	0.617
FathmmXF	0.428	0.548
Fathmm	0.261	1.074
Grantham	0.251	0.811
GERP++	0.211	0.577
PseAA	0.196	0.890
SIPHY	0.186	0.854
PHYLOP	0.180	0.691

B. Yeast Proteins

Predictor	RankScore	CV
SNPs&GO	0.763	0.157
SNAP2	0.757	0.033
PolyPhen2	0.741	0.127
SNPs&GO3D	0.716	0.140
PhD SNP	0.685	0.245
PANTHER	0.672	0.132
S3D-PROF	0.637	0.137
DeepSequence	0.607	0.493
PROVEAN	0.565	0.150
SIFT	0.514	0.305
SuSPect	0.479	0.584
NetDiseaseSNP	0.390	0.405
BLOSUM62	0.183	0.965
Grantham	0.051	1.477

C. Bacterial Proteins

Predictor	RankScore	CV
DeepSequence	0.770	0.068
PolyPhen2	0.689	0.088
SuSPect	0.621	0.138
PROVEAN	0.611	0.205
SNAP2	0.606	0.452
SIFT	0.595	0.143
S3D-PROF	0.570	0.157
SNPs&GO3D	0.531	0.190
PhD SNP	0.522	0.404
SNPs&GO	0.494	0.345
NetDiseaseSNP	0.197	0.598
BLOSUM62	0.190	0.388
Grantham	0.026	2.236

D. Viral Proteins

Predictor	RankScore	CV
SNAP2	0.832	0.315
PROVEAN	0.826	0.121
SuSPect	0.799	0.131
BLOSUM62	0.748	0.322
PhD_SNP	0.701	0.263
NetDiseaseSNP	0.632	0.311
SNPs&GO	0.566	0.294
PolyPhen2	0.539	0.299
SIFT	0.533	0.200
Grantham	0.280	1.050
S3D-PROF	0.209	1.000
DeepSequence	0.172	0.930
SNPs&GO3D	0.140	0.390

Figure 3. Spearman's rank scores and coefficient of variation (CV) for each predictor, subset by organism of origin and colour-coded by predictor category (blue – untrained; orange – trained structural; green – trained non-structural; pink – metapredictor).

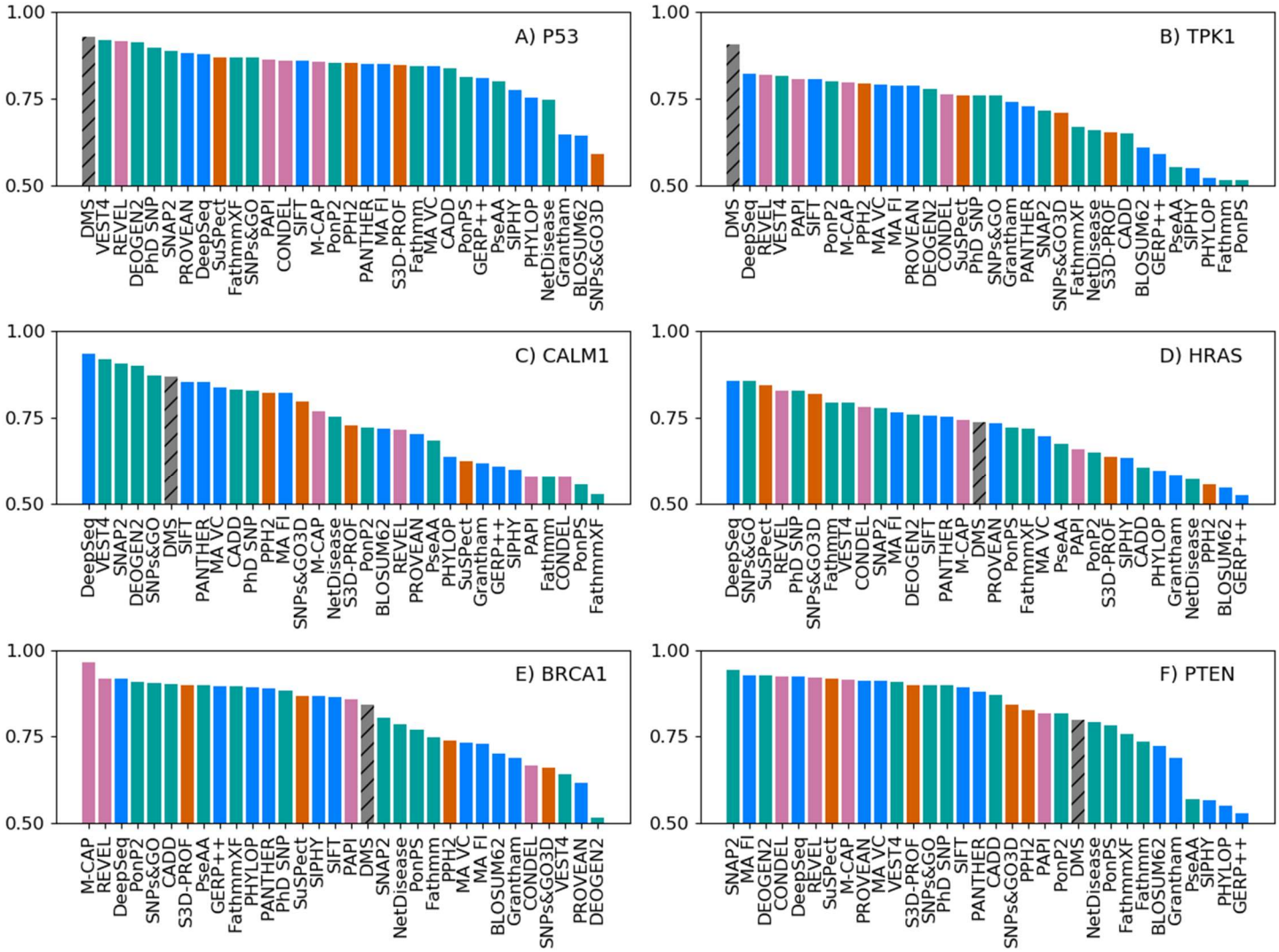

Figure 4. Area under the ROC curve (AUC) for each predictor in classifying pathogenic and putatively benign mutations. Bars are coloured by method category (blue – untrained; orange – trained structural; green – trained non-structural; pink – metapredictor), with DMS experiments being cross-hatched.