Variant effect predictor correlation with functional assays is reflective of clinical classification performance

Benjamin J. Livesey & Joseph A. Marsh*

MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom

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

Abstract

Understanding the relationship between protein sequence and function is crucial for accurate genetic variant classification. Variant effect predictors (VEPs) play a vital role in deciphering this complex relationship, yet evaluating their performance remains challenging due to data circularity, where the same or related data is used for training and assessment. High-throughput experimental strategies like deep mutational scanning (DMS) offer a promising solution. In this study, we extend upon our previous benchmarking approach, assessing the performance of 84 different VEPs and DMS experiments from 36 different human proteins. In addition, a new pairwise, VEP-centric ranking method reduces the impact of VEP score availability on the overall ranking. We observe a remarkably high correspondence between VEP performance in DMS-based benchmarks and clinical variant classification, especially for predictors that have not been directly trained on human clinical variants. Our results suggest that comparing VEP performance against diverse functional assays represents a reliable strategy for assessing their relative performance in clinical variant classification. However, major challenges in clinical interpretation of VEP scores persist, highlighting the need for further research to fully leverage computational predictors for genetic diagnosis. We also address practical considerations for end users in terms of choice of methodology.

Introduction

Deciphering the nature of the sequence-function relationship in proteins remains one of the greatest challenges in modern biology. It has profound implications for variant classification in a medical context, understanding of disease mechanisms and protein design. Computational tools for predicting variant effects, known as variant effect predictors (VEPs), can provide valuable insight into the complex relationship between protein sequence and human phenotypes. However, the profusion of new predictors has also highlighted the need for reliable, unbiased strategies for evaluating VEP performance.

Identifying a fair method to compare VEPs remains difficult due to the prevalence of data circularity in many performance evaluations (Grimm *et al*, 2015). This often results in an inflated assessment of VEP performance and can be introduced into a benchmark in two ways. Type 1 circularity can be considered as variant-level circularity, and occurs when variants (or homologous variants) used to train or tune a VEP are subsequently used to assess its performance. Type 2 circularity is essentially gene-level circularity, and occurs in cross-gene analyses when the testing set contains different variants from the same (or homologous) genes used for training. For example, if a VEP learns to strongly associated variants from a specific gene as mostly being pathogenic or benign, this can lead to excellent apparent performance if the tested variants from this gene mostly fall into the same class.

While type 2 circularity is only an issue for evaluations that combine variants from different genes, it is incredibly difficult to eliminate type 1 circularity issues. Doing so often greatly reduces the pool of

available data to compare the performance of predictors based upon supervised learning approaches, or reduces the number of VEPs that can be compared in order to expand the pool of variants. Even amongst ostensibly unsupervised predictors, some derive features from human allele frequencies, which is routinely used as strong evidence to classify variants as benign (Richards *et al*, 2015). Thus, VEPs trained with allele frequencies as a feature have likely 'seen' a large proportion of the benign data used to benchmark them. These limitations keep most independent benchmarks at a small scale, and often limited to comparing less than 10 different VEPs (Gunning *et al*, 2021; Walters-Sen *et al*, 2015; Niroula & Vihinen, 2019).

One way to address this problem came from the development of high-throughput experimental strategies, known as multiplexed assays of variant effect (MAVEs) (Weile & Roth, 2018). The technology behind MAVEs has been improving at a rapid rate, coordinated through the Atlas of Variant Effects Alliance, which aims to promote research and collaboration, and eventually produce variant effect maps across all human protein-coding genes and regulatory elements (Fowler *et al*, 2023). Datasets derived from deep mutational scanning (DMS) experiments, a class of MAVEs focusing on functional assays for measuring the effects of protein mutations (Fowler & Fields, 2014), show tremendous potential for use in comparison to the outputs of VEPs. DMS datasets provide several advantages for benchmarking over more traditional sets composed of variants observed in a clinical context. They do not rely upon any previously assigned clinical labels (*e.g.* 'pathogenic' and 'benign') that are commonly used to train VEPs, thus greatly reducing the potential for type 1 data circularity in any assessment of VEP performance. By comparing the correlations between variant effect scores from VEPs and DMS experiments on a per-protein basis, type 2 circularity is also avoided.

In previous studies, we have evaluated the performance of VEPs (Livesey & Marsh, 2020, 2023) and protein stability predictors (Gerasimavicius *et al*, 2023) using data from DMS assays. With the rapid recent growth in the field, numerous novel VEPs and DMS datasets have been subsequently released. Here we build upon this work by including 32 more VEPs and 11 more human DMS datasets, and also by improving our benchmarking methodology. Our work demonstrates an extremely high correspondence between VEP performance when benchmarked against DMS datasets and when tested for clinical variant classification when we consider those predictors that have not been directly trained on human clinical or population variants. In contrast, for VEPs trained or tuned on human variants, it is exceedingly difficult to perform a fair comparison using traditional clinical benchmarks. Therefore, we suggest that our strategy of benchmarking VEPs using a large number of diverse DMS datasets represents a reliable way of assessing their relative performance at scoring the relative clinical impacts of variants within individual proteins. Importantly, however, we recognise that we still have a long way to go in terms of clinical interpretation of VEP outputs.

Results

New DMS datasets and VEPs

The increasing popularity of MAVEs as an experimental strategy for high-throughput characterisation of variant effects has enabled us to add 11 new DMS datasets assessing the impacts of single amino acid substitutions to our benchmark compared to our previous publications (Table 1). Many of these are present in MaveDB, a valuable community resource for the sharing of MAVE datasets (Esposito *et al*, 2019). As each DMS dataset can have multiple sets of functional scores, potentially representing altered conditions, replicates, filters or even entirely different fitness assays, we chose a single DMS score set per protein to represent the overall fitness of that protein. We selected the dataset that had the highest median correlation with all VEPs in order to prevent

outliers from overtly influencing the selection process. For proteins in which multiple DMS studies were performed by different groups (*BRCA1*, *TP53*, *GDI1*, *PTEN*), we likewise only selected a single score set for each protein using the above method. A new criterion was also added that required the top 5 most highly correlated VEPs with the DMS data to have a mean Spearman's correlation greater than 0.3 in order to help exclude proteins where the DMS fitness metric is very weakly related to human fitness. The large majority of DMS datasets meet this criteria, but available datasets for *NCS1*, *TECR*, *CXCR4*, *KCNJ2* and *CD19* were therefore not included in our benchmark for this reason. The full summary of all DMS datasets, across 36 proteins and covering 198,106 different single amino acid variants, is available in Table S1

We streamlined the assignment of categories to DMS datasets by defining only two different types. Direct assays are those that directly measure the ability of the target protein to carry out one or more functions. Examples of direct functional assays include one-hybrid and two-hybrid assays, other assays that measure the interaction strength with native partners and VAMP-seq (Matreyek *et al*, 2018). Indirect assays are most commonly growth rate experiments, where the attribute being measured is not directly controlled by the target protein. Indirect DMS assays may be more representative of the biological reality of a variant's effect on cellular fitness, as the cell may be able to buffer a small or moderate loss-of-function. Direct DMS assays are more reflective of a protein's function in isolation and may be more useful for exploring protein mechanisms or for protein design.

The field of variant effect prediction has also been progressing rapidly, with many novel methods being published every year. In total, we added 32 new missense VEPs that were not used in our previous benchmarks (Table 2). These were identified by browsing new publications and , and by taking advantage of the ProteinGym resource, which benchmarks numerous VEPs and general protein language models against human and non-human DMS datasets (Notin *et al*, 2023). The large majority of VEPs from our previous analysis (Livesey & Marsh, 2023) were retained, although a small number were removed because the predictors were no longer available to run (S3D-PROF, SNPs&GO3D and PAPI), and thus we could not add predictions for the new DMS datasets. This emphasises the importance of making source code and pre-calculated variant effect scores freely available, to ensure that tools can continued to be used in the future (Livesey *et al*, 2024). In total, we included 84 different VEPs in this study, considering only those with predictions available for at least 60% of the DMS datasets in our benchmark (Table S2).

During our research we also identified a few VEPs that were difficult to accesses due to a requirement for paid subscriptions or restrictive licensing agreements. We have not included these in our benchmark, as we strongly believe that VEP methodologies and predictions need to be made freely available to enable fair, replicable assessment by the community, and ultimately, to enable the confidence required for these tools to play a greater role in making clinical diagnoses (Livesey *et al*, 2024).

Despite attempts to simplify our VEP classification system, the labels of "supervised" and "unsupervised" in our previous benchmark are imperfect for our primary concern, which is the risk of data circularity. For example, Envision (Gray *et al*, 2018) is trained by supervised machine learning, but has not seen any labelled human variants as it was trained with DMS data alone. Likewise, AlphaMissense is primarily an unsupervised method, but undergoes training with human allele frequencies as "weak labels" (Cheng *et al*, 2023), which will provide an advantage for classification of benign variants. To correct for this, we now classify methods as whether or not they include human clinical or population variants in their features ('*population-based'* and '*population-free'* predictors). To include human population data, a VEP must either draw its training data from a database of observed human variants or include a feature derived directly from human population data, for example, allele frequency.

Related to this, an advancement in VEP technology that has the potential to confound our benchmarking methodology is the increasing availability of predictors that are directly trained using DMS data. There are currently five such VEPs included in our benchmark: Envision, CPT (Jagota *et al*, 2023), DeMaSk (Munro & Singh, 2021), VARITY_R and VARITY_ER (Wu *et al*, 2021). Using DMS datasets to benchmark these VEPs carries similar caveats to benchmarking population-based VEPs using population databases. Fortunately, these methods have all been trained using only a small number of DMS datasets. Therefore, by excluding the results of these VEPs for the proteins used to train them (Table S2), we have been able to include these predictors in our benchmark.

Comparison of VEP performance using DMS data

Although there is great diversity in DMS assays, and what they measure may not always be reflective of clinical outcomes, the premise of our analysis is that VEPs that show the most consistency across a large set of DMS experiments are likely to be the most useful for predicting variant effects. We use absolute Spearman's rank correlation to assess the correspondence between VEPs and DMS, as only a monotonic relationship between the two variables is required. Thus, no transformation needs to be applied to the VEP output or DMS datasets, which can vary greatly in scale and directionality.

Figure 1 shows the Spearman's correlation between DMS results and VEP predictions for each of the 36 DMS datasets. The strongest correlations approached 0.8 for some DMS datasets (*GCH1*, *PPARG*, *GDI1*), while several others demonstrated relatively poor correlations even for the best-performing VEPs around the cut-off of 0.3 (*LDLRAP1*, *TPK1*). The average correlation of the top-performing predictors for each protein was 0.57. We also observe no clear differences between population-based and population-free methods, as the top-performing VEPs are evenly split between the two categories (18 population-based, 18 population-free). There was also little difference between the direct and indirect DMS categories: the average correlation of the top-performing VEPs was 0.58 for the direct assays compared to 0.55 for the indirect assays.

In Figure S1, we plot this data in an alternative way, by showing the distribution of Spearman's correlations for each VEP across all of the DMS datasets where it has predictions. This illustrates the wide range of Spearman's correlations each VEP can have. While this type of representation is very commonly used in published studies when assessing predictor performance, we note that it has a major limitation related to the fact that different VEPs are effectively being tested against different datasets. First, not all VEPs have predictions available across all proteins in our benchmark. In some cases, we are limited by the specific proteins for which the authors have provided pre-calculated results, or the proteins for which predictions are available for in ProteinGym. In other cases, as mentioned above, we have excluded specific proteins from the assessment of predictors that were trained using DMS data, to avoid potential data circularity.

Second, not all VEPs output scores for all possible variants in the proteins for which they are run. Some only provide predictions for those missense changes possible via single nucleotide changes, while others do not provide predictions for specific protein regions, for example, where the depth of sequence alignment is low (Frazer *et al*, 2021), or when the method can only be applied to sequences shorter than a specific size (Meier *et al*, 2021). This could potentially lead to inflated results for predictors that exclude a generally poorly predicted region of a protein. To illustrate the potential impact of this, we can observe in Figure 1 a few examples where a single outlier VEP demonstrated far higher correlations with the DMS data than other methods, such as PANTHER (Thomas & Kejariwal, 2004) for HMGCR and mutationTCN (Kim & Kim, 2020) for PTEN. Closer inspection reveals that, in both of these cases, the phenomenon arises because these VEPs provide scores for a much smaller set of variants for these specific proteins compared than other VEPs.

One solution to this problem would be to only use those DMS measurements for variants with scores available across all predictors, although this would require us to include either far fewer VEPs or far fewer DMS measurements. Moreover, it is not clear that Spearman's correlations between VEP and DMS scores should be comparable between proteins, or that they represent a good measure of the relative performance of VEPs across different proteins. For example, a protein with two functionally distinct regions, where one is highly constrained (*e.g*. a globular domain) and the other is not (*e.g.* a disordered region) might show a high correlation between VEPs and DMS, driven by this difference. In contrast, a small, highly conserved protein where mutations at most positions will have damaging effects might show a much lower Spearman's correlation, even though VEPs are not necessarily performing worse.

Therefore, to ensure that the relative ranking of VEP performance remains as fair as possible, for each DMS dataset we perform a series of pairwise comparisons in which the correlations of every possible pair of VEPs with the DMS data is calculated using only predictions for variants shared between the two VEPs and the DMS set. The percentage of the time that a VEP "won" each of its pairwise comparisons against every other VEP is then calculated across all proteins. This strategy is illustrated in Figure S2, which shows a heatmap coloured by the win rate of each VEP compared to all others. To obtain our overall ranking, we simply average the win rate of each VEP against all other VEPs. This method of ranking is more VEP-centric than DMS-centric as in our previous benchmarks, meaning it should act as a more useful basis for relative ranking, particularly accounting for cases where certain VEPs do not have predictions for all proteins.

Figure 2 shows the average win rate of the top 30 predictors. The full results, including the win percentage of each VEP against every other, are available in Table S3. The heatmap in Figure S2 is also sorted according to this same ranking, allowing for visualisation of performance across all predictors.

The top-ranking VEP using this methodology is CPT (93.4% average win rate). CPT combines both EVE (Frazer *et al*, 2021) and ESM-1v (Meier *et al*, 2021) along with structural features from AlphaFold (Jumper *et al*, 2021) and ProteinMPNN (Dauparas *et al*, 2022), and further conservation features. Importantly, although training of the model was carried out against five DMS datasets, they have all been excluded from the benchmarking of CPT, thus avoiding circularity concerns.

AlphaMissense performs only slightly worse overall than CPT in this benchmark (91.0% average win rate). AlphaMissense is a recently developed large language model with additional structural context based on the AlphaFold methodology, and 'fine-tuning' on allele frequencies from humans and other primates. While the core of the model is unsupervised, the fine-tuning process using human variants necessitates its label as a population-based predictor.

ESCOTT (Tekpinar *et al*, 2024) and GEMME (Laine *et al*, 2019) are two closely related population-free predictors that rank 3rd and 4th. GEMME is a relatively simple model, compared to the other top performers, based on epistasis between positions through evolution. GEMME also has lower computational requirements than comparably performing VEPs, and similar computational time to running a language model like ESM-1v. ESCOTT is based on a modified version of GEMME that takes into account the likely structural context of mutated residues.

VARITY_R ranked 5th overall, and was the top-ranking population-based predictor that was also included in our previous study. Interestingly, while VARITY_R previously ranked behind ESM-1v, EVE and DeepSequence, it has slightly outperformed them here. However, we also note that VARITY_R and VARITY_ER are compared to the smallest number of DMS datasets in this benchmark, necessitated by exclusion of those models on which it was directly trained, and consequently, have larger error bars.

Two variants of EVE are also amongst the top predictors. TranceptEVE (Notin *et al*, 2022b) ranks 6th overall, and is a hybrid of Tranception (Notin *et al*, 2022a) and EVE (a hybrid of Tranception (Notin *et al*, 2022a). popEVE (Orenbuch *et al*, 2023) ranks 7th and is a hybrid of ESM-1v and EVE that also performs gene-level calibration on variants from UK Biobank, with the goal of making scores from different genes directly comparable. Because of its utilisation of population data, we have classified popEVE as a population-based VEP, although it is not necessarily subject to the same type 1 circularity concerns as many other VEPs, as discussed later.

Close examination of the heatmap in Figure S2 reveals some interesting trends when considering pairwise comparisons of VEPs. Notably, AlphaMissense exceeds the performance of CPT on slightly more than half of the proteins. However, CPT ranks first overall due to equalling or exceeding AlphaMissense's win rate against most of the remaining VEPs. popEVE has a win rate greater than 50% when compared to both VARITY_R and TranceptEVE, despite ranking slightly below them overall. Wavenet (Shin *et al*, 2021) stands out with an unusual pattern due to its extreme heterogeneity in performance when tested against different proteins. For example, while it is the top performer in terms of correlation with the HRAS assay, it ranks 59th overall due to its inconsistency.

Performance of VEPs on clinical variant classification

The above DMS-based benchmarking of VEP performance might not be reflective of performance in pathogenicity prediction, which is a main practical application for which these methods are used. DMS assays are heterogeneous in their methodologies and in the meaning of their functional scores. This has been a common criticism of assessing VEP performance using functional assays (Wu *et al*, 2021; Reeb *et al*, 2020) and it is very understandable. To what extent do our DMS-based rankings reflect utility for clinical variant classification?

Traditional assessment and comparison of VEPs has typically involved testing their discrimination between known pathogenic and known benign or putatively benign variants, often using datasets such as ClinVar (Landrum *et al*, 2018) and gnomAD (Chen *et al*, 2024). However, this can be extremely difficult to do in a fair manner for population-based predictors. First, most supervised VEPs have been directly trained on pathogenic variants, so to compare performance, one needs to know the identities of all the variants used to train each predictor, and then find a set of variants not used by any of the tested VEPs. One also needs to ensure that no variants at the same positions as variants used in training, or even at homologous positions (Li *et al*, 2023), are included in the test set.

An even stricter requirement for assessment of VEP performance for variants across different genes is that one should exclude any variants from the test set from any gene for which any variants were used in training, or from any homologues of these genes. That is, supervised VEPs should only be tested on different and non-homologous genes to any used in training, not just different variants. This practise may have rarely been followed in the past, but it is necessary to avoid type 2 circularity; otherwise, predictor performance will be inflated because models will learn to associate certain genes with pathogenicity, regardless of their ability to discriminate between variant effects within those genes (Grimm *et al*, 2015; Livesey *et al*, 2024). Importantly, however, as long as performance assessment is carried out on a per-gene basis (*e.g.* the performance metric is calculated for each gene/protein separately), then it should generally be acceptable to test VEPs on genes on which they have been trained, as this avoids any risk of type 2 circularity.

There are further concerns related to the identities of variants used as the negative class. The same requirement to not use variants used in training is equally true for these. However, a critical complication arises from the fact that many VEPs now incorporate human allele frequency

information as a feature. This is severely problematic for the use of known benign variants (*e.g.* those classified as benign or likely benign in ClinVar) as the negative class. As allele frequency is routinely used in the classification of variants as benign (Richards *et al*, 2015), any VEP that includes allele frequency as a feature will suffer from circularity in these analyses. Even if allele frequency was not directly used in the clinical classification, common variants are simply more likely to be studied and receive a classification.

An alternative to using known benign mutations as the negative class is to use variants observed in the human population (*e.g.* taking all of those from gnomAD), which will mostly be very rare. We strongly advocate this approach for several reasons. First, it minimises the aforementioned circularity issue regarding the use of allele frequency to classify benign variants, although it does not completely eliminate it. Second, it is much more reflective of the actual clinical utility of variant effect predictions. The major challenge for clinicians is not in discriminating between common benign and rare pathogenic variants. Instead, it is in the much more difficult problem of distinguishing rare benign from rare pathogenic variants. Previous predictors, notably REVEL (Ioannidis *et al*, 2016) and VARITY, have acknowledged this issue and tailored their models to the problem of rare variant identification. Finally, using rare population variants allows for much larger negative classes. In many disease genes, there are no or few variants classified as benign, severely limiting the number of genes for which reliable analyses can be performed.

Here, we have assessed the performance of VEPs in distinguishing between pathogenic and likely pathogenic ClinVar missense variants, and 'putatively benign' gnomAD v4 missense variants, taking all of those not classified as pathogenic. We recognise the limitation of this, in that there is likely to be a small proportion of as-of-yet unclassified pathogenic variants in our negative class, particularly in recessive genes and those with incomplete penetrance. Nevertheless, we believe that the advantages stated above far outweigh this issue. We generated predictions for 908 proteins that had at least 10 (likely) pathogenic and 10 putatively benign missense variants using 43 population-based and 27 population-free VEPs. It was necessary to exclude a few VEPs from this analysis because predictions were not available for enough proteins, particularly those where we obtained scores from ProteinGym. For each protein, we tested the discrimination between pathogenic and benign for each VEP by calculating the area under the receiver operating characteristic curve (AUROC), which is a common measure of classifier performance that summarises the trade-off between true positive rate and false positive rate across different thresholds.

The full distribution of AUROC values for each predictor, sorted by median, is shown in Figure S3. However, for the same reasons discussed earlier in relation to the DMS ranking, this analysis has the potential to be confounded by the fact that not all VEPs provide scores for all possible variants. Therefore, we applied the same pairwise ranking strategy as above, using AUROC as our comparison metric instead of correlation. Figure 3 shows the top 30 ranking predictors in terms of their performance in clinical variant classification according to this methodology. The full ranking of all available predictors is available in Table S4.

At first glance, the rankings are strikingly different, with the top 8 all being population-based VEPs. SNPred and MetaRNN rank 1st and 2nd, respectively, in contrast to the DMS benchmark, where they ranked 14th and 15th overall. It is likely that their performance here, as well as the performance of most population-based VEPs, is highly inflated by circularity issues, as no effort has been made to exclude variants used in training. Therefore, it is interesting to note that the top-ranking population-free VEP was CPT, the same as observed with DMS. The closely related GEMME and ESCOTT models show very similar performance, with ESCOTT ranking slightly higher in the DMS benchmark and GEMME ranking slightly higher in the clinical benchmark.

Area under the precision-recall curve (AUPRC) is an alternative performance metric to AUROC. Precision-recall is considered more reflective of many real-life classification scenarios where correct identification of a minority class is more important than that of a majority class. The disadvantage of precision-recall is that relative class sizes need to remain consistent across all models in order for the AUPRC scores to be comparable. Our use of pairwise comparisons essentially cancels out this disadvantage, allowing us to use AUPRC as an alternative to AUROC. Figure S4 ranks the predictors by pairwise analysis using AUPRC as the comparison metric. The overall rankings are very similar to the ROC-based ranking in Figure 3, but with 11 population-based predictors exceeding the performance of the top population-free predictor, lending credence to the robustness of the analysis.

To compare the two benchmarks, in Figure 4, we plot the win rate from the DMS analysis *vs* the win rate from the clinical variant analysis. If we consider only the population-free models, the correlation (r = 0.98) is striking. Relative performance on the DMS benchmark appears to be highly predictive of relative performance in clinical variant classification across the entire performance range. In contrast, for the population-based models, the correlation is much lower (r = 0.86). Moreover, the population-based VEPs tend to be relatively increased in performance on the clinical benchmark compared to the population-free VEPs. This almost certainly reflects varying levels of circularity contributing to performance in the clinical benchmark. It is likely that the extent to which clinical win rates are shifted to the right relative to the population-free VEPs can be considered as measure of how overfit the models are on our pathogenic and putatively benign variants.

While most of the population-based VEPs show strong signs of circularity in their clinical variant classification performance, some appear to have much less or no bias. The two VARITY models fit perfectly with the population-free VEPs, possibly reflecting its innovative strategy to minimise training bias. SuSPect and MPC, while ranking lower overall in both categories, also appear to show little bias. AlphaMissense is also notable: it is tuned using population variants, but is not trained using labelled pathogenic variants, so any influence from circularity should be smaller. In addition, our use of mostly rare variants as the putatively benign dataset should minimise any advantage from its population tuning. While it does show relatively better performance on clinical variants, overall, it appears closer on the plot to the population-free models than to other population-based VEPs.

Finally, popEVE represents an interesting case. As noted earlier, we have classified popEVE as population-based due to the fact that it is calibrated on human variants. However, given that population variants are only used for protein-level scaling of scores, in principle, it should not be susceptible to any circularity issues when performing protein-level analyses. In other words, it should be immune to type 1 circularity issues (although it could potentially be conflated by type 2 circularity in other cross-gene analyses). Although it performs slightly better on clinical variants than DMS data, similar to AlphaMissense, overall, it is very similar to the top population-free VEPs.

Practical considerations

An often-overlooked but extremely important aspect of VEPs is how easy they are for an end-user to obtain predictions. VEPs are typically made available through a combination of three different channels.

1. A web interface that allows access either to the VEP itself (*e.g.* SIFT, SNAP2) or to a database of pre-calculated results (*e.g.* popEVE, VARITY).
2. A large compilation of pre-calculated results that usually cover either all canonical human protein positions in UniProt or all human coding region non-synonymous single nucleotide variants in genome space.
3. The method itself is made available for installation by the end user.

Of these three channels, a web interface is the most convenient for looking up single variants of interest, although most such interfaces also offer the option to view all possible variants within a given protein as well. Downloadable databases of pre-calculated results are very useful for large-scale analyses (such as this one), but may be less useful for end users than a simple web interface for searching individual variants. If such a database is formatted in genome space, then specialised software such as Tabix (Li, 2011) may be required to identify scores for variants of interest. Finally, installing and running the predictor offers the greatest degree of control over generation of the results such as the alignments and features used. However, many modern VEPs have high computational and time requirements or require significant technical knowledge to operate. We are unable to recommend such VEPs for standard day-to-day use unless the data is also obtainable through a web interface or database.

As these are all important considerations for end users, in Table 3 we provide a summary of the top 13 VEPs from the correlation-based analysis in terms of how easy it is to obtain predictions, as well as links to any online interfaces, pre-calculated results or installable packages/repositories.

Discussion

Our benchmarking strategy very much relies on comparing performance across a large number of diverse DMS datasets. We have tried to avoid making judgements about the quality of individual DMS datasets or selecting them based on what we deem to be desirable phenotypes or experimental properties, other than excluding a small number based on poor correlations with all VEPs. Although it is possible (perhaps likely) that certain types of DMS experiments will be better for VEP benchmarking than others, we feel that our approach of taking as many datasets as possible minimises the potential for bias. Although different DMS datasets differ greatly in the extent to which they reflect clinical phenotypes, they generally should show at least some relationship to fitness and thus, in general, algorithms that are better at predicting variants effects on fitness or pathogenicity should tend to show a stronger correlation with experimental measurements. The fact that we see such a strong correspondence between the relative ranking of VEPs across these diverse DMS datasets, and in the clinical classification of variants, strongly supports the utility of this approach. Importantly, however, our strategy requires comparing performance across a large number of datasets, as we observe large variability in the 'winner' from dataset to dataset. Thus, any attempts to judge performance with datasets from one or a small number of functional assays are unlikely to yield very informative results.

It is clear that much of the tendency for population-based VEPs to perform relatively better on the clinical benchmark is due to data circularity. However, one could still argue that there is some aspect of the relatively better performance that comes these VEPs learning some aspect of clinical pathogenicity that is not present in the population-free models. We think this is unlikely. For example, the strategy used by the population-based VARITY went to great lengths to minimise circularity issues in its training process, and it is highly consistent with the population-VEPs in its relative performance on DMS *vs* clinical benchmarks. AlphaMissense was not trained for pathogenicity, but even with its inclusion of allele frequencies, it is also fairly similar to the population-free methods. Finally, CPT, without any training on human pathogenic or population variants, outperforms 35/43 test population-based VEPs on the clinical benchmark, demonstrating how effective population-free methods can be on their own.

One related concern that is very difficult to address is optimisation against DMS datasets present in our benchmark. While we have excluded datasets directly used in training for the evaluation of certain predictors, it is possible that methods may have been optimised against DMS data without direct training. For example, ESM-1v was not trained on DMS data, but it was selected out of

multiple possible models based on its correlation with DMS data (Meier *et al*, 2021). Possibly this is reflected in the fact that it is slightly "left-shifted" in Figure 4, showing modestly better performance on the DMS benchmark relative to the clinical benchmark. As DMS and other functional assays are increasing used to assess performance, VEP developers will inevitably target these benchmarks and optimise for performance against them. However, currently there is little indication that DMS inclusion in VEP training or optimisation has had an impact on this benchmark, and the few methods trained directly with DMS data have proven to be relatively resistant to bias. We hope that this methodology can continue to be used to benchmark future predictors in a bias-free manner. However, there is a paradox in that, the more popular this strategy becomes, the less useful it may be.

An interesting observation from our latest DMS benchmark is that the top-three methods all use some level of protein structural information. Previously we had noted that there was no tendency for structure-aware models to perform better than those that use sequences only (Livesey & Marsh, 2020), so this represents a potentially notable advance in VEP development. One possibility is that, in the past, the large majority of performance gains have been obtained through improved elucidation of the evolutionary signal, and so the much smaller impact of structure was negligible. This is compounded by the fact that most structure-based VEPs assume that pathogenic variants will be structurally damaging and ignore non-loss-of-function effects (Gerasimavicius *et al*, 2022), and that, previously, structure models were only available for minority of human proteins. Now, however, we may be reaching the point of diminishing returns from improved evolutionary models. Combined with the availability of computational structural models for all human proteins (Jumper *et al*, 2021; Baek *et al*, 2021), we may now be reaching a point where the inclusion of structure has become more important for determining the performance of top-ranking VEPs.

Given the remarkable performance of population-free VEPs, we think that not directly including human clinical or population variants in models is the best and safest strategy for variant effect prediction. Given the desire to increase the role of computational predictions in making clinical diagnoses, it is important to minimise the potential for "double counting", *e.g.* according to ACMG/AMP guidelines (Richards *et al*, 2015). If allele frequency, or knowledge of other classified variants at the same position, have been utilised by the model, then the computational prediction cannot be considered as independent evidence. In contrast, population-free VEPs should be truly independent from the other pathogenic or benign classification criteria.

Although we believe that our relative rankings of VEP performance are reliable, the major remaining problem is still in the interpretation of their outputs. For example, how should a clinician interpret a high VEP score for making a genetic diagnosis? Recent work has attempted to establish thresholds for using variant effect scores as stronger diagnostic evidence (Pejaver *et al*, 2022). However, this work focused primarily on population-based VEPs, not really addressing the severe circularity problem discussed here. In addition, given the radically different performance of VEPs across different genes, we think that gene-specific thresholds and interpretation, as well as consideration of inheritance, will be necessary to advance the clinical utility of VEPs. To an extent, this is what the popEVE is attempting, combining population-free gene-level scoring with calibration across genes based upon the distributions of observed population variants (Orenbuch *et al*, 2023). Closely related to this, it is important to emphasise that both of our benchmarking strategies here have been based on gene-level approaches. That is, we assess the correction with individual DMS datasets, or calculate the AUROC for specific disease genes. This has advantages in terms of avoiding type 2 circularity, but does not tell us how well a VEP will perform for novel disease gene discovery. We expect this to be a major focus of research in the near future.

Overall, it is clear the variant effect prediction field is moving very fast. Along with other members of the Atlas of Variant Effects Alliance, we recently released a set of guidelines and recommendations for developers of novel VEPs, many of which related to improving the sharing and independent assessment of methods (Livesey *et al*, 2024). In addition, we strongly encourage the people to deposit new DMS datasets in MaveDB (Esposito *et al*, 2019). Making methods, predictions and DMS data freely and easily available will improve future DMS-based benchmarking. Finally, we note that, while missense variant effect prediction is reaching a level of maturity, far more work remains to be done on non-missense coding variants and on non-coding variants, both in terms of methods development and benchmarking. We hope the lessons we have learned here will prove valuable for this.

Materials and Methods

DMS datasets

In addition to the 25 human datasets retained from our previous analysis (after the exclusion of CXCR4), we identified a further 11 datasets that met our inclusion criteria of 5% minimum coverage of all amino acid substitutions and a minimum mean Spearman's correlation of 0.3 with the top three VEPs. These new datasets were primarily obtained from MaveDB (Esposito *et al*, 2019), but also by searching published works. One dataset (*GCH1*) came from an unpublished study with permission of the authors.

VEPs

The dbNSFP database, version 4.2 (Liu *et al*, 2020) served as a source for 26 VEPs. The remaining 55 VEPs were either run locally on the University of Edinburgh high performance computing cluster (EDDIE), downloaded as pre-calculated results, obtained via a web interface or obtained for a limited subset of mutations/proteins from the ProteinGym website. Table S2 provides a full list of the source used to obtain predictions from each VEP.

Spearman's correlation

Spearman's correlation was calculated between datasets using the stats.spearmanr() function of the scipy python package version 1.5.4 on Python version 3.6.8.

To calculate the correlation-based rank score displayed in Figure 2 and Table S3, for each protein the absolute Spearman's correlations between the selected DMS dataset and every pair of VEPs was calculated using only variants where the DMS and two VEPs all have available scores. The VEP that obtained the highest correlation in each pairwise comparison earned one point, while the VEP with the lower correlation earned none. The win rate of every VEP over every other VEP was then calculated across all proteins by dividing the number of wins by the number of times that particular pair of VEPs were tested. The final rank score was calculated by averaging the win rate of each VEP against all other VEPs.

AUROC and AUPRC

The area under the receiver operator characteristic curve (AUROC) was calculated using the metrics.roc_auc_score() function of the sklearn python package, while the area under the precision-recall curve (AUPRC) was calculated using the metrics.average_precision_score() function of the sklearn python package version 0.18.1. To maintain consistency between class labels, predictors that assigned low scores as pathogenic and high as benign needed to be inverted. This was done by deducting the scores from 1.

The rankings in Figure 3 were calculated by comparing the AUROC between every pair of predictors using only variants shared between them. The predictor with the highest AUC was awarded one

point. The win rate of every VEP against every other VEP was then calculated across all proteins by dividing the number of wins by the total number of times that particular pair of VEPs were tested. The final rank score was calculated by averaging the win rate of each VEP against all other VEPs. The same procedure was used to generate the AUPRC-based ranking in Figure S4, but with precision-recall instead of ROC.

Data availability

Full datasets associated with this paper are available at https://doi.org/10.6084/m9.figshare.25790610.

Acknowledgements

This project was supported by funding from the Medical Research Council (MRC) Human Genetics Unit core grant. (MC_UU_00035/9). JAM is a Lister Institute Research Prize Fellow. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

References

Alley EC, Khimulya G, Biswas S, AlQuraishi M & Church GM (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 16: 1315–1322

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, *et al* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373: 871–876

Brandes N, Goldman G, Wang CH, Ye CJ & Ntranos V (2023) Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* 55: 1512–1522

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, *et al* (2024) A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 625: 92–100

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, *et al* (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381: eadg7492

Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, Laporte J & Poch O (2020) MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. *PLOS ONE* 15: e0236962

Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, *et al* (2022) Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 378: 49–56

Dunham AS, Beltrao P & AlQuraishi M (2023) High-throughput deep learning variant effect prediction with Sequence UNET. *Genome Biol* 24: 110

Elazar A, Weinstein J, Biran I, Fridman Y, Bibi E & Fleishman SJ (2016) Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *eLife* 5: e12125

Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM & Rubin AF (2019) MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol* 20: 223

Ferruz N, Schmidt S & Höcker B (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* 13: 4348

Fowler DM, Adams DJ, Gloyn AL, Hahn WC, Marks DS, Muffley LA, Neal JT, Roth FP, Rubin AF, Starita LM, *et al* (2023) An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol* 24: 147

Fowler DM & Fields S (2014) Deep mutational scanning: a new style of protein science. *Nat Methods* 11: 801–807

Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y & Marks DS (2021) Disease variant prediction with deep generative models of evolutionary data. *Nature* 599: 91–95

Gerasimavicius L, Livesey BJ & Marsh JA (2022) Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat Commun* 13: 3895

Gerasimavicius L, Livesey BJ & Marsh JA (2023) Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Sci* 32: e4688

Gersing S, Cagiada M, Gebbia M, Gjesing AP, Coté AG, Seesankar G, Li R, Tabet D, Weile J, Stein A, *et al* (2023) A comprehensive map of human glucokinase variant activity. *Genome Biol* 24: 97

Gray VE, Hause RJ, Luebeck J, Shendure J & Fowler DM (2018) Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst* 6: 116-124.e3

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, *et al* (2015) The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Hum Mutat* 36: 513–523

Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL & Wright CF (2021) Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet* 58: 547–555

Hesslow D, Zanichelli N, Notin P, Poli I & Marks D (2022) RITA: a Study on Scaling Up Generative Protein Sequence Models. doi:10.48550/arXiv.2205.05789 [PREPRINT]

Huang Y-F (2020) Unified inference of missense variant effects and gene constraints in the human genome. *PLOS Genet* 16: e1008922

Huang Y-F & Siepel A (2019) Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Res* 29: 1310–1321

Huttinger ZM, Haynes LM, Yee A, Kretz CA, Holding ML, Siemieniak DR, Lawrence DA & Ginsburg D (2021) Deep mutational scanning of the plasminogen activator inhibitor-1 functional landscape. *Sci Rep* 11: 18827

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, *et al* (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* 99: 877–885

Jagota M, Ye C, Albors C, Rastogi R, Koehl A, Ioannidis N & Song YS (2023) Cross-protein transfer learning substantially improves disease variant prediction. *Genome Biol* 24: 182

Jiang TT, Fang L & Wang K (2023) Deciphering "the language of nature": A transformer-based language model for deleterious mutations in proteins. *The Innovation* 4: 100487

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

Kim HY & Kim D (2020) Prediction of mutation effects using a deep temporal convolutional network. *Bioinformatics* 36: 2047–2052

Kwon JJ, Hajian B, Bian Y, Young LC, Amor AJ, Fuller JR, Fraley CV, Sykes AM, So J, Pan J, *et al* (2022) Structure–function analysis of the SHOC2–MRAS–PP1C holophosphatase complex. *Nature* 609: 408–415

Laine E, Karami Y & Carbone A (2019) GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol Biol Evol* 36: 2604–2619

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, *et al* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 46: D1062–D1067

Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27: 718–719

Li N, Mazaika E, Theotokis P, Zhang X, Jang M, Ahmad M, Powell G, Heyne HO, Lal D, Barton PJ, *et al* (2023) Variant annotation across homologous proteins ("Paralogue Annotation") identifies disease-causing missense variants with high precision, and is widely applicable across protein families. 2023.08.07.552236 doi:10.1101/2023.08.07.552236 [PREPRINT]

Li S, van der Velde KJ, de Ridder D, van Dijk ADJ, Soudis D, Zwerwer LR, Deelen P, Hendriksen D, Charbon B, van Gijn ME, *et al* (2020) CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. *Genome Med* 12: 75

Liu X, Li C, Mou C, Dong Y & Tu Y (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* 12: 103

Livesey BJ, Badonyi M, Dias M, Frazer J, Kumar S, Lindorff-Larsen K, McCandlish DM, Orenbuch R, Shearer CA, Muffley L, *et al* (2024) Guidelines for releasing a variant effect predictor. doi:10.48550/arXiv.2404.10807 [PREPRINT]

Livesey BJ & Marsh JA (2020) Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol Syst Biol* 16: e9380

Livesey BJ & Marsh JA (2023) Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol Syst Biol* 19: e11474

van Loggerenberg W, Sowlati-Hashjin S, Weile J, Hamilton R, Chawla A, Sheykhkarimli D, Gebbia M, Kishore N, Frésard L, Mustajoki S, *et al* (2023) Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *Am J Hum Genet* 110: 1769–1786

Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, *et al* (2016) Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* 48: 1570–1575

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

Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, *et al* (2018) Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing. *Nat Genet* 50: 874–882

McDonnell AF, Plech M, Livesey BJ, Gerasimavicius L, Owen LJ, Hall HN, FitzPatrick DR, Marsh JA & Kudla G (2023) Deep mutational scanning quantifies DNA binding and predicts clinical outcomes of PAX6 variants. 2023.07.25.550478 doi:10.1101/2023.07.25.550478 [PREPRINT]

Meier J, Rao R, Verkuil R, Liu J, Sercu T & Rives A (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems* pp 29287–29303. Curran Associates, Inc.

Meitlis I, Allenspach EJ, Bauman BM, Phan IQ, Dabbah G, Schmitt EG, Camp ND, Torgerson TR, Nickerson DA, Bamshad MJ, *et al* (2020) Multiplexed Functional Assessment of Genetic Variants in CARD11. *Am J Hum Genet* 107: 1029–1043

Miller PG, Sathappa M, Moroco JA, Jiang W, Qian Y, Iqbal S, Guo Q, Giacomelli AO, Shaw S, Vernier C, *et al* (2022) Allosteric inhibition of PPM1D serine/threonine phosphatase via an altered conformational state. *Nat Commun* 13: 3778

Molotkov I, Koboldt DC & Artomov M (2023) SNPred outperforms other ensemble-based SNV pathogenicity predictors and elucidates the challenges of using ClinVar for evaluation of variant classification quality. 2023.09.07.23295192 doi:10.1101/2023.09.07.23295192 [PREPRINT]

Munro D & Singh M (2021) DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics* 36: 5322–5329

Nijkamp E, Ruffolo J, Weinstein EN, Naik N & Madani A (2022) ProGen2: Exploring the Boundaries of Protein Language Models. doi:10.48550/arXiv.2206.13517 [PREPRINT]

Niroula A & Vihinen M (2019) How good are pathogenicity predictors in detecting benign variants? *PLOS Comput Biol* 15: e1006481

Notin P, Dias M, Frazer J, Marchena-Hurtado J, Gomez A, Marks DS & Gal Y (2022a) Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. doi:10.48550/arXiv.2205.13760 [PREPRINT]

Notin P, Kollasch AW, Ritter D, Niekerk L van, Paul S, Spinner H, Rollins N, Shaw A, Weitzman R, Frazer J, *et al* (2023) ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. 2023.12.07.570727 doi:10.1101/2023.12.07.570727 [PREPRINT]

Notin P, Niekerk LV, Kollasch AW, Ritter D, Gal Y & Marks DS (2022b) TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction. 2022.12.07.519495 doi:10.1101/2022.12.07.519495 [PREPRINT]

Orenbuch R, Kollasch AW, Spinner HD, Shearer CA, Hopf TA, Franceschi D, Dias M, Frazer J & Marks DS (2023) Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. 2023.11.27.23299062 doi:10.1101/2023.11.27.23299062 [PREPRINT]

Pei J & Grishin NV (2021) The DBSAV Database: Predicting Deleteriousness of Single Amino Acid Variations in the Human Proteome. *J Mol Biol* 433: 166915

Pejaver V, Byrne AB, Feng B-J, Pagel KA, Mooney SD, Karchin R, O'Donnell-Luria A, Harrison SM, Tavtigian SV, Greenblatt MS, *et al* (2022) Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet* 109: 2163–2177

Petrazzini BO, Balick DJ, Forrest IS, Cho J, Rocheleau G, Jordan DM & Do R (2021) Prediction of recessive inheritance for missense variants in human disease. 2021.10.25.21265472 doi:10.1101/2021.10.25.21265472 [PREPRINT]

Ponzoni L, Peñaherrera DA, Oltvai ZN & Bahar I (2020) Rhapsody: predicting the pathogenicity of human missense variants. *Bioinformatics* 36: 3084–3092

Quinodoz M, Peter VG, Cisarova K, Royer-Bertrand B, Stenson PD, Cooper DN, Unger S, Superti-Furga A & Rivolta C (2022) Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am J Hum Genet* 109: 457–470

Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T & Rives A (2021) MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning* pp 8844–8856. PMLR

Reeb J, Wirth T & Rost B (2020) Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics* 21: 107

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, *et al* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17: 405–423

Shin J-E, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, Manglik A, Kruse AC & Marks DS (2021) Protein design and variant prediction using autoregressive generative models. *Nat Commun* 12: 2403

Takeda J, Nanatsue K, Yamagishi R, Ito M, Haga N, Hirata H, Ogi T & Ohno K (2020) InMeRF: prediction of pathogenicity of missense variants by individual modeling for each amino acid substitution. *NAR Genomics Bioinforma* 2

Tekpinar M, David L, Henry T & Carbone A (2024) PRESCOTT: a population aware, epistatic and structural model accurately predicts missense effect. 2024.02.03.24302219 doi:10.1101/2024.02.03.24302219 [PREPRINT]

Thomas PD & Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A* 101: 15398–15403

Walters-Sen LC, Hashimoto S, Thrush DL, Reshmi S, Gastier-Foster JM, Astbury C & Pyatt RE (2015) Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Mol Genet Genomic Med* 3: 99–110

Wang B, Lei X, Tian W, Perez-Rathke A, Tseng Y-Y & Liang J (2023) Structure-based pathogenicity relationship identifier for predicting effects of single missense variants and discovery of higher-order cancer susceptibility clusters of mutations. *Brief Bioinform* 24: bbad206

Weile J & Roth FP (2018) Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum Genet* 137: 665–678

Wu Y, Li R, Sun S, Weile J & Roth FP (2021) Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet* 108: 1891–1906

Yang KK, Fusi N & Lu AX (2023) Convolutions are competitive with transformers for protein sequence pretraining. 2022.05.19.492714 doi:10.1101/2022.05.19.492714 [PREPRINT]

Zhang H, Xu MS, Fan X, Chung WK & Shen Y (2022) Predicting functional effect of missense variants using graph attention neural networks. *Nat Mach Intell* 4: 1017–1028

Zhao H, Du H, Zhao S, Chen Z, Li Y, Xu K, Liu B, Cheng X, Wen W, Li G, *et al* (2024) SIGMA leverages protein structural information to predict the pathogenicity of missense variants. *Cell Rep Methods* 4: 100687

Figures

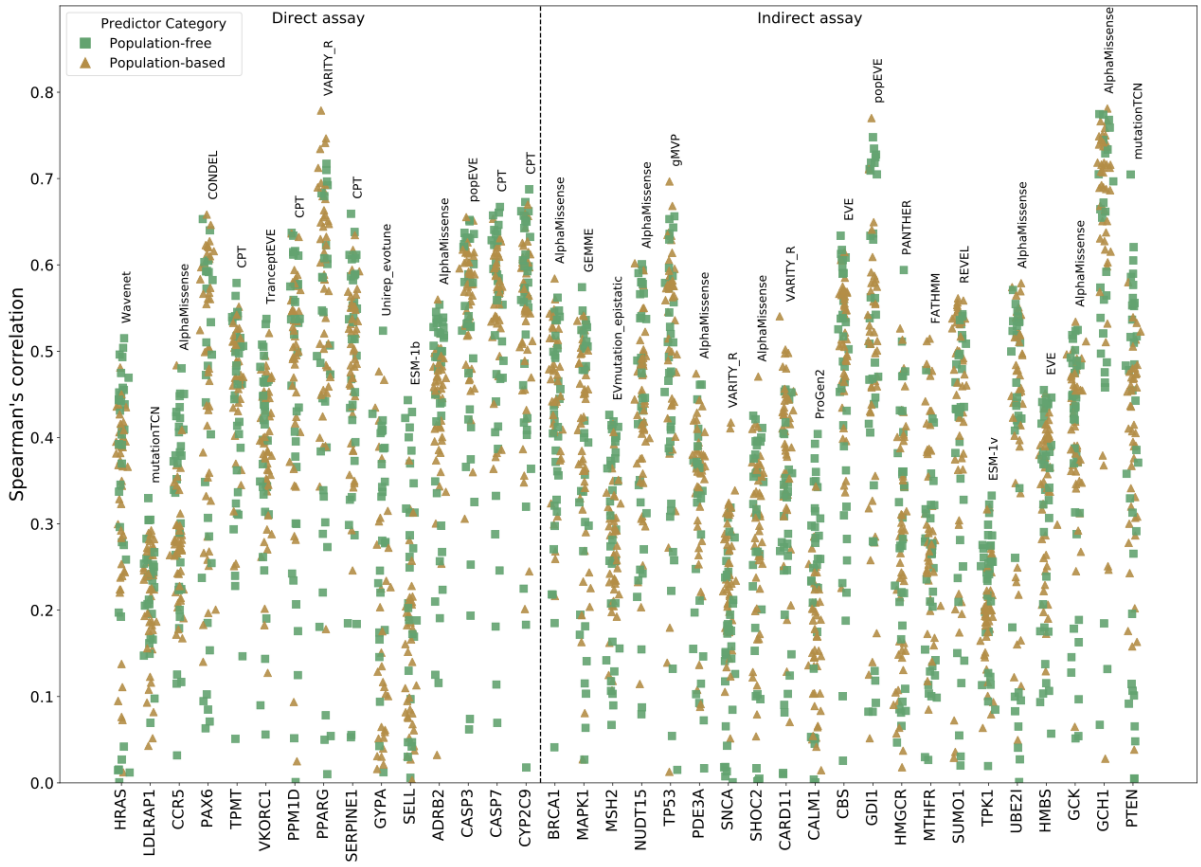

Figure 1. Correlation between variant effect scores from VEPs and DMS experiments. The Spearman's correlation between all VEPs and every selected DMS dataset. VEPs are split into 'population-based' and 'population-free' methods based on whether or not they are trained using human variants. DMS datasets are classified as 'direct' if they directly measure the ability of the target protein to carry out one or more functions, with all others being classified as 'indirect'. The VEP with the highest correlation is noted for every DMS dataset.

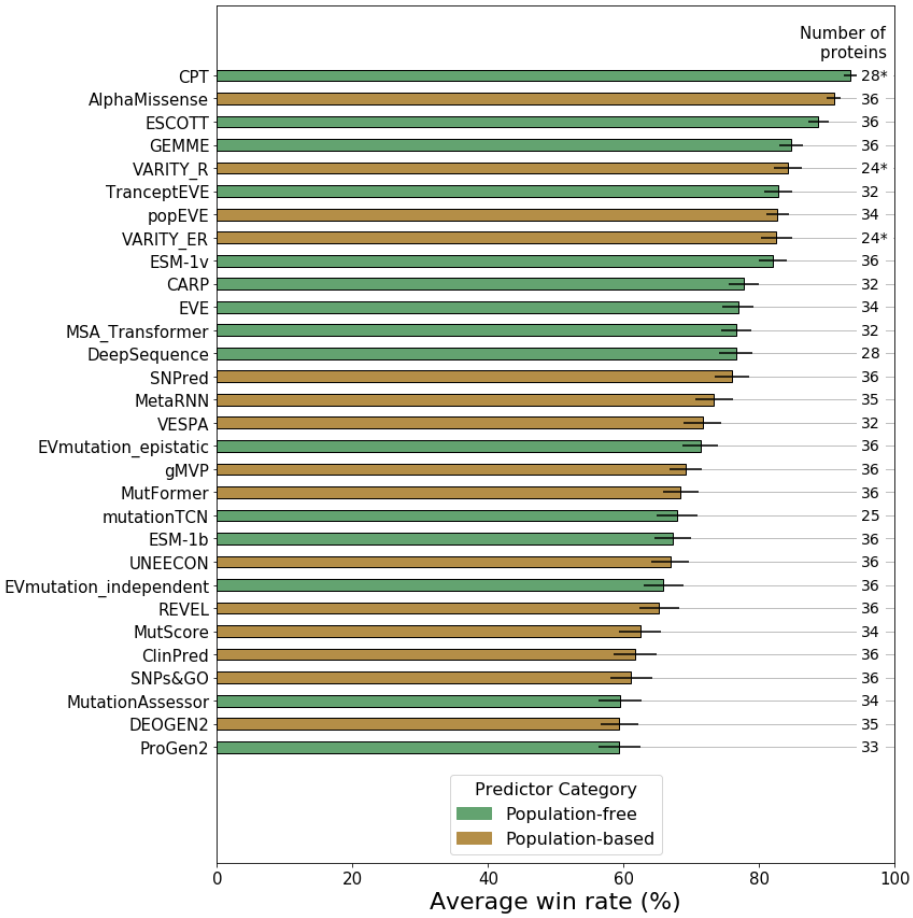

Figure 2. The top 30 out of 84 tested VEPs ranked based on performance against the DMS benchmark. VEPs are ranked according to their average win rate against all other VEPs in pairwise Spearman's correlation comparisons across all DMS datasets. The number of proteins for which each VEP had scores included is indicated in the right column of the plot. Those indicated with * had some DMS datasets excluded to avoid circularity concerns. Error bars represent the standard error across all comparisons with other VEPs. The full ranking of all VEPs and all pairwise win rates are available in Table S3.

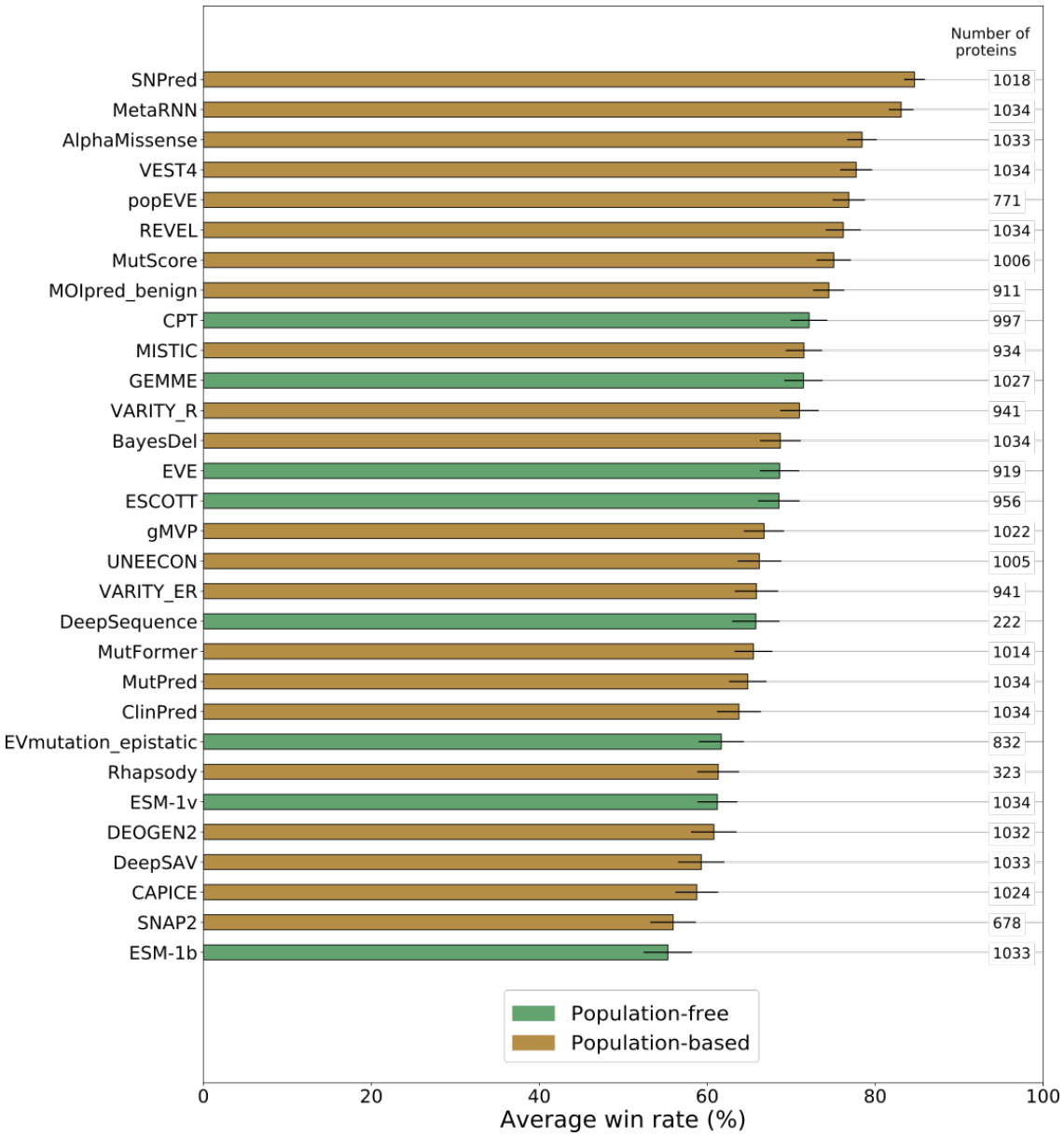

Figure 3. The top 30 out of 70 tested VEPs in terms of clinical variant classification performance. VEPs are ranked according to their average win rate against all other VEPs in pairwise AUROC comparisons across all human proteins with at least 10 pathogenic and 10 putatively benign missense variants. The number of proteins that met this condition for each predictor is indicated on the right of the plot. Some VEPs from the DMS benchmark could not be included here because predictions were not available for enough genes. Error bars represent the standard error across all comparisons with other VEPs. The full ranking of all VEPs and all pairwise win rates are available in Table S4.

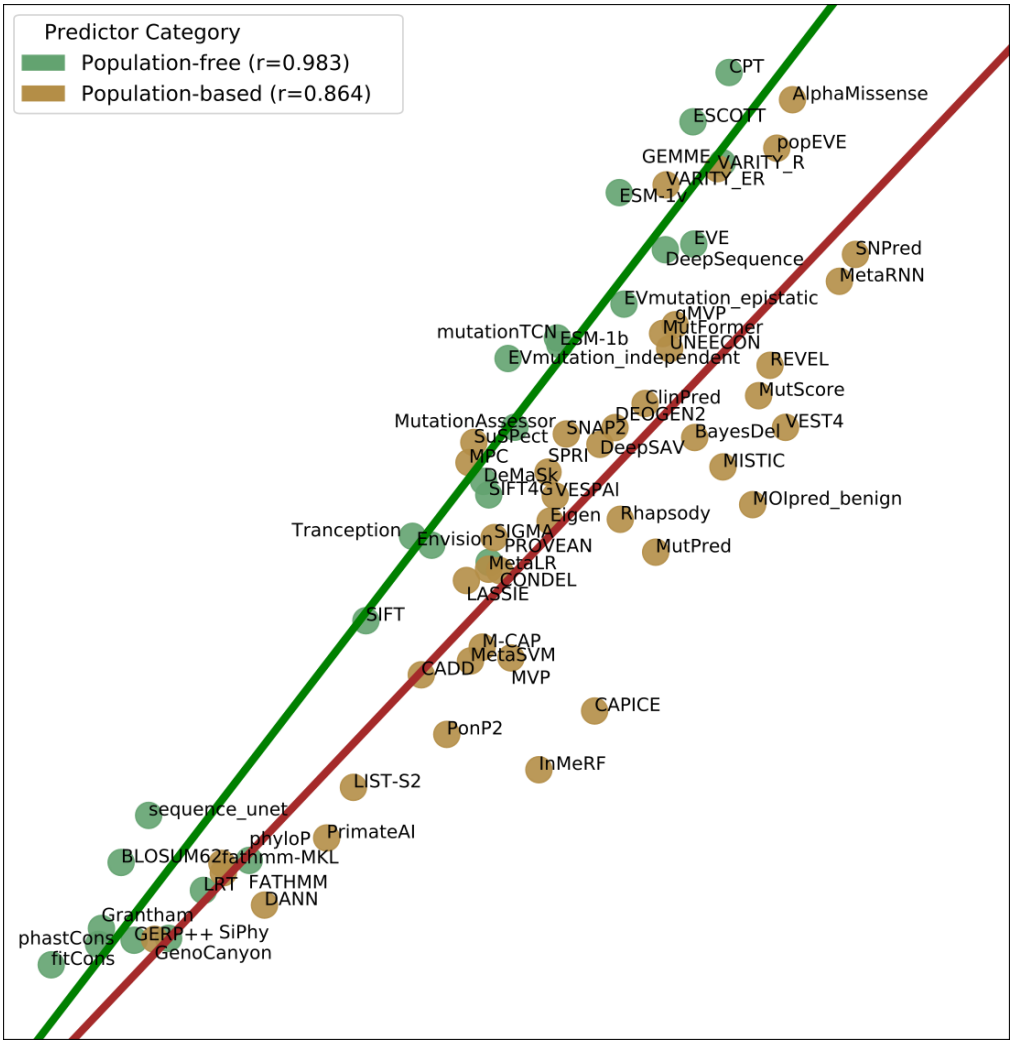

Figure 4. Strong correspondence in relative performance of VEPs on the DMS *vs* clinical benchmarks. Average pairwise win rates in the DMS *vs* clinical benchmarks are plotted. Population-free VEPs show an extremely strong correlation (Pearson's r = 0.983). In contrast, the population-based VEPs show a much weaker correlation overall (r = 0.864). The tendency for some population-based to show large rightward shifts, reflecting relatively increased performance on the clinical benchmark, is likely to be due to circularity due to training on variants and genes present in the pathogenic and putatively benign datasets.

Tables

Table 1. A summary of the 11 new DMS datasets included in this benchmark.

Gene (UniProt ID)	DMS coverage (%)	Fitness Assay	Reference
CARD11 (Q9BXL7)	10.92% 10.83%	Indirect growth assay in a human lymphoma cell line. Indirect survival assay in the presence of ibrutinib.	(Meitlis *et al*, 2020)
GYPA (P02724) ***SELL*** (P14151)	8.60% 5.19%	Direct survival assay requiring correct membrane integration (dsTβL).	(Elazar *et al*, 2016)
HMBS (P08397)	86.94%	Indirect yeast complementation assay.	(van Loggerenberg *et al*, 2023)
GCH1 (P30793)	73.31%	Indirect yeast complementation assay.	Not yet published
GCK (P35557)	97.00%	Indirect yeast complementation assay.	(Gersing *et al*, 2023)
SERPINE1 (P05121)	69.98%	Direct binding to urokinase-type plasminogen activator, assessed by phage display/antibody binding.	(Huttinger *et al*, 2021)
PAX6 (P26367)	33.95%	Direct yeast one hybrid assay (TF efficiency)	(McDonnell *et al*, 2023)
PPARG (P37231)	100.00%	Direct activation of CD36 (PPARG target) assessed by FACS.	(Majithia *et al*, 2016)
PPM1D (O15297)	68.63%	Direct suppression of a GFP-fusion gene assessed by FACS.	(Miller *et al*, 2022)
SHOC2 (Q9UQ13)	99.22%	Indirect growth assay in a SCHOC2-dependent cell line.	(Kwon *et al*, 2022)

Table 2. A summary of the 32 new VEPs included in this analysis and links to the data or source code used to produce predictions.

Predictor	Type	Source	Reference
AlphaMissense	Population-based	https://zenodo.org/records/8208688	(Cheng *et al*, 2023)
CAPICE	Population-based	https://zenodo.org/records/3928295	(Li *et al*, 2020)
CARP	Population-free	https://github.com/microsoft/protein-sequence-models Limited predictions available from ProteinGym	(Yang *et al*, 2023)
CPT	Population-free	https://zenodo.org/records/8140323	(Jagota *et al*, 2023)
DeepSAV	Population-based	http://prodata.swmed.edu/DBSAV/	(Pei & Grishin, 2021)
DeMaSk	Population-free	https://github.com/Singh-Lab/DeMaSk	(Munro & Singh, 2021)
ESCOTT	Population-free	https://zenodo.org/records/10577421	(Tekpinar *et al*, 2024)
ESM-1b	Population-free	https://huggingface.co/spaces/ntranoslab/esm_variants	(Brandes *et al*, 2023)
GEMME	Population-free	Pre-calculated results from https://datadryad.org/stash/dataset/doi:10.5061/dryad.vdncjsz1s	(Laine *et al*, 2019)
gMVP	Population-based	https://www.dropbox.com/s/nce1jhg3i7jw1hx/gMVP.2021-02-28.csv.gz?dl=0	(Zhang *et al*, 2022)
InMeRF	Population-based	https://www.med.nagoya-u.ac.jp/neurogenetics/InMeRF/download.html	(Takeda *et al*, 2020)
LASSIE	Population-based	http://compgen.cshl.edu/LASSIE/	(Huang & Siepel, 2019)
MISTIC	Population-based	https://lbgi.fr/mistic/download	(Chennen *et al*, 2020)
MOIpred	Population-based	https://zenodo.org/records/5620519	(Petrazzini *et al*, 2021)
MSA Transformer	Population-free	https://github.com/facebookresearch/esm Limited predictions available from ProteinGym	(Rao *et al*, 2021)
Mutformer	Population-based	https://github.com/WGLab/mutformer?tab=readme-ov-file	(Jiang *et al*, 2023)
MutScore	Population-based	https://mutscorebatch-wgt7hvakhq-ew.a.run.app/	(Quinodoz *et al*, 2022)

popEVE	Population-based	https://pop.evemodel.org/	(Orenbuch *et al*, 2023)
ProGen2	Population-free	https://github.com/salesforce/progen	(Nijkamp *et al*, 2022)
ProtGPT2	Population-free	https://huggingface.co/nferruz/ProtGPT2 Limited predictions available for ProteinGym	(Ferruz *et al*, 2022)
Rhapsody	Population-based	https://github.com/prody/rhapsody	(Ponzoni *et al*, 2020)
RITA	Population-free	https://github.com/lightonai/RITA Limited predictions available from ProteinGym	(Hesslow *et al*, 2022)
Sequence UNET	Population-free	https://github.com/allydunham/sequence_unet	(Dunham *et al*, 2023)
SIGMA	Population-based	https://sigma-pred.org/	(Zhao *et al*, 2024)
SNPred	Population-based	https://www.synapse.org/#!Synapse:syn52137034/files/	(Molotkov *et al*, 2023)
SPRI	Population-based	http://sts.bioe.uic.edu/spri/	(Wang *et al*, 2023)
TranceptEVE	Population-free	Limited predictions available from ProteinGym	(Notin *et al*, 2022b)
Tranception	Population-free	https://github.com/OATML-Markslab/Tranception	(Notin *et al*, 2022a)
UNEECON	Population-based	https://github.com/yifei-lab/UNEECON	(Huang, 2020)
Unirep (evotuned)	Population-free	https://github.com/churchlab/UniRep Limited predictions available from ProteinGym	(Alley *et al*, 2019)
VESPA	Population-based	https://github.com/Rostlab/VESPA Limited predictions available from ProteinGym	(Marquet *et al*, 2022)
Wavenet	Population-free	Limited predictions available from ProteinGym	(Shin *et al*, 2021)

Table 3. A summary of the top-ranking VEPs and the channels through which their results can be accessed.

VEP	Online interface	Precalculated results	Installable/Source code	Notes
CPT	https://huggingface.co/spaces/songlab/CPT	https://zenodo.org/records/7954657	https://github.com/songlab-cal/CPT?tab=readme-ov-file	CPT provides results for every amino acid substitution in 18,602 proteins as pre-calculated results.
AlphaMissense	N/A	https://zenodo.org/records/8208688	https://github.com/google-deepmind/alphamissense	Results are available for all positions in all canonical Uniprot isoforms as well as ~60,000 non-cannonical isoforms.
ESCOTT	N/A	https://zenodo.org/records/10577421	http://gitlab.lcqb.upmc.fr/tekpinar/PRESCOTT/	ESCOTT provides precalculated results for ~19,000 human proteins, indexed by Uniprot ID.
GEMME	N/A (interface currently down)	https://datadryad.org/stash/dataset/doi:10.5061/dryad.vdncjsz1s	https://hub.docker.com/r/elodielaine/gemme	GEMME is considerably easier to run locally than other top-ranking methods, and does not require a GPU.
VARITY	http://varity.varianteffect.org	Download link on web interface.	https://github.com/joewuca/varity?tab=readme-ov-file	VARITY is limited to nsSNVs and provides results for 18,239 human proteins.
TranceptEVE	N/A	N/A	https://github.com/OATML-Markslab/ProteinGym/blob/main/notebooks/TranceptEVE_example.ipynb	TranceptEVE is bundled with ProteinGym. Very high computational requirements.
popEVE	https://pop.evemodel.org/	https://pop.evemodel.org/documentation	https://github.com/debbiemarkslab/popEVE	A very comprehensive and well-constructed web interface makes it easy to retrieve results.
ESM (all models)/MSA Transformer	https://huggingface.co/spaces/ntranoslab/esm_variants (ESM-1b only)	Download for ESM-1b predictions only available through the web interface.	https://github.com/facebookresearch/esm	With the exception of ESM-1b, all ESM models must be run by the end-user. A GPU is STRONGLY recommended but calculation time is relatively short.
CARP	N/A	N/A	https://github.com/microsoft/protein-	CARP must be run by the end-user. A GPU is STRONGLY recommended.

			sequence-models	
EVE	https://evemodel.org/	https://evemodel.org/download/bulk	https://github.com/debbiemarkslab/EVE	EVE is extremely computationally intensive and takes a long time to run. Make use of pre-calculated results if available.
DeepSequence	N/A	N/A	https://github.com/debbiemarkslab/DeepSequence	DeepSequence must be run by the end-user. It can take a very long time to run even on optimal hardware. There is also very limited support for the machine learning library employed (Theano).
SNPred	N/A	https://www.synapse.org/#!Synapse:syn52137034/files/	https://github.com/ArtomovLab/SNPred	SNPred only provides results for nsSNVs. Warning – sign-up is required to access the pre-calculated results.

Supplementary Figures

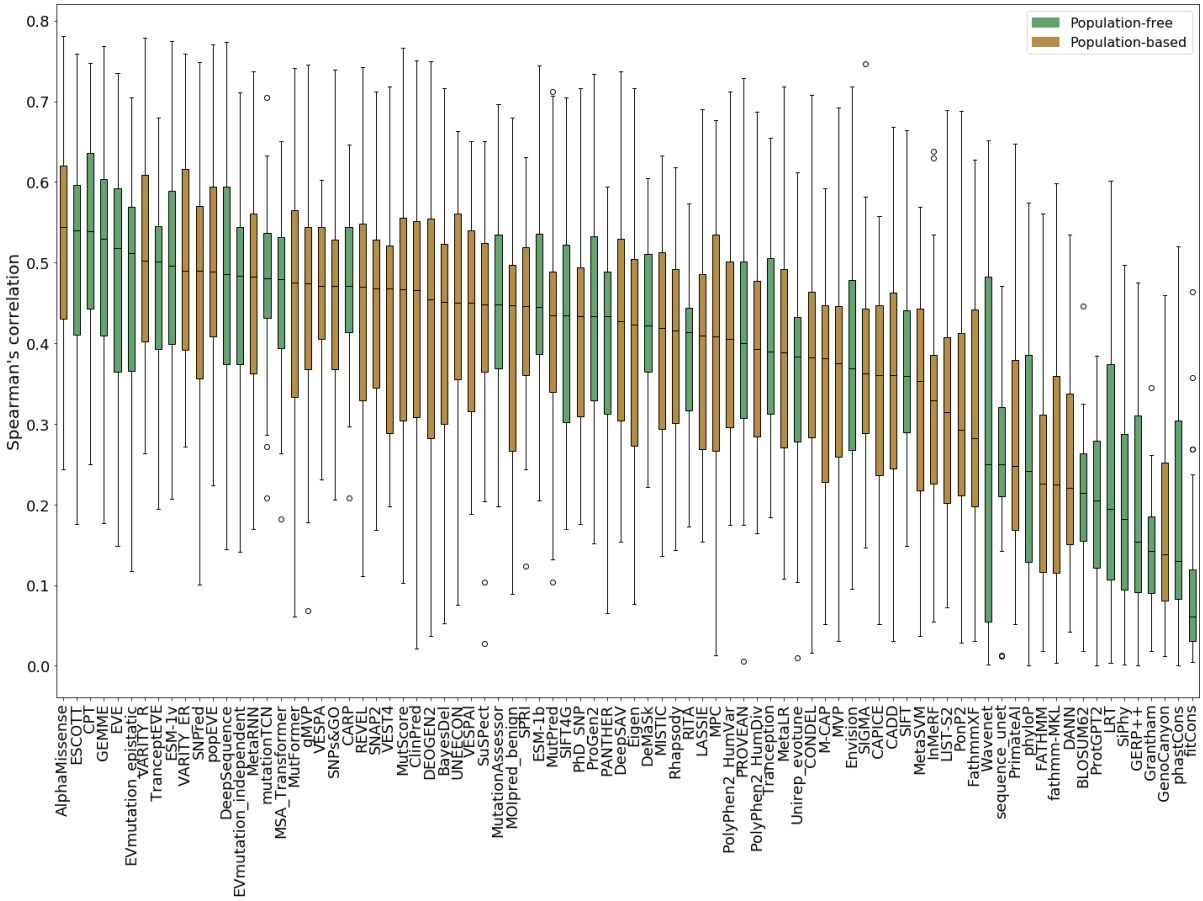

Figure S1. Distribution of Spearman's correlations of different VEPs across different DMS datasets. This is effectively the same data presented in Figure 1, but grouped by VEP instead of by protein. Correlations of different VEPs across different DMS datasets. Note that not all VEPs output scores for all DMS datasets, or all variants from individual datasets, so we can not necessarily directly compare correlations; thus motivating the pairwise ranking approach in Figure 2.

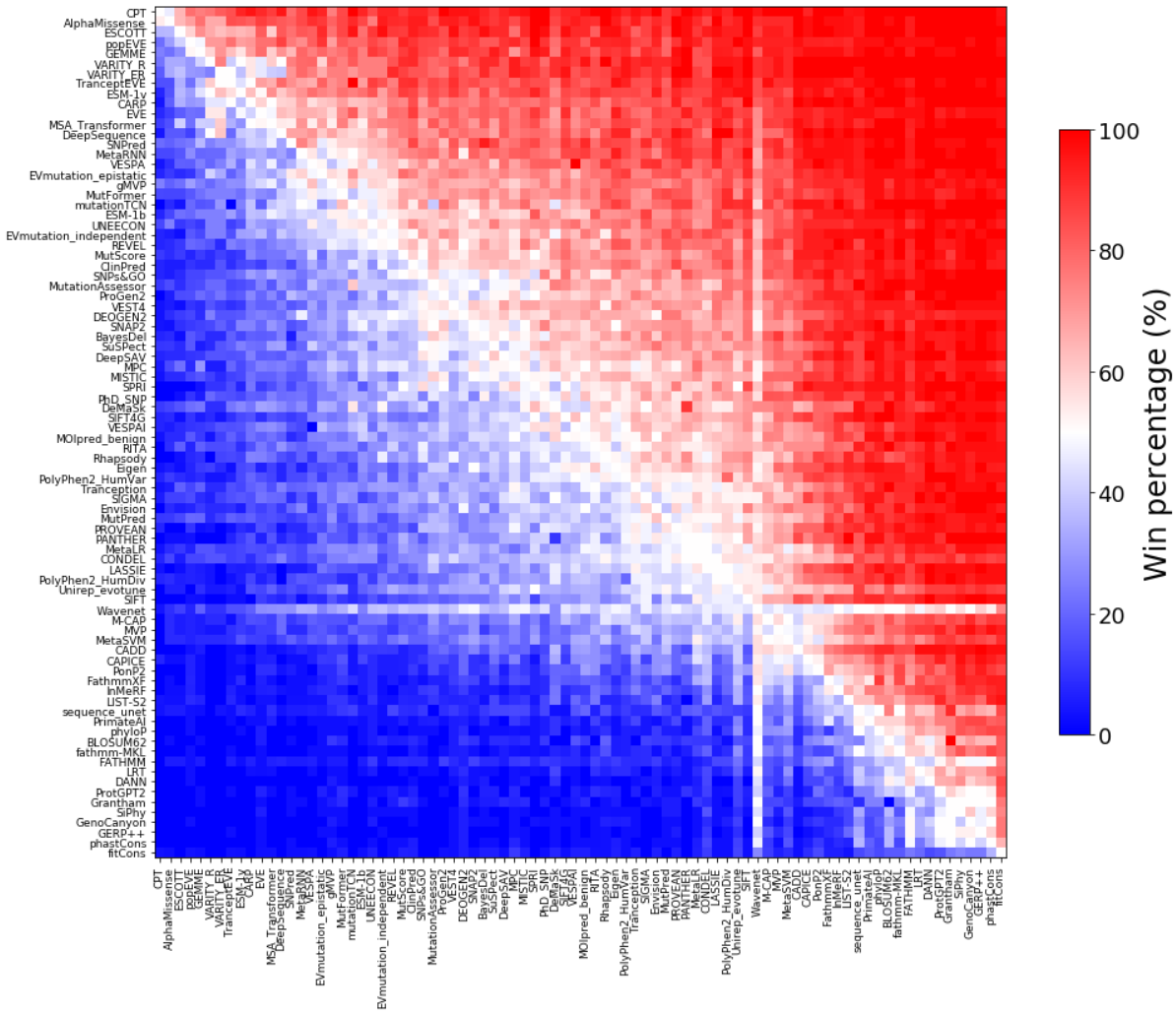

Figure S2. Heatmap of VEP vs VEP win rates across DMS datasets. For each pair of VEPs, we compare their Spearman's correlations with each DMS dataset, considering only variants for which each VEP has a prediction available. The win rate represents the percentage of proteins where the first VEP shows a higher correlation than the second. For example, red values indicate the VEP on the y-axis outperforms the VEP on the x-axis in terms of correlation with DMS measurements across most proteins. The average win rate in Figure 2 represents the mean win rate of each VEP against all other VEPs.

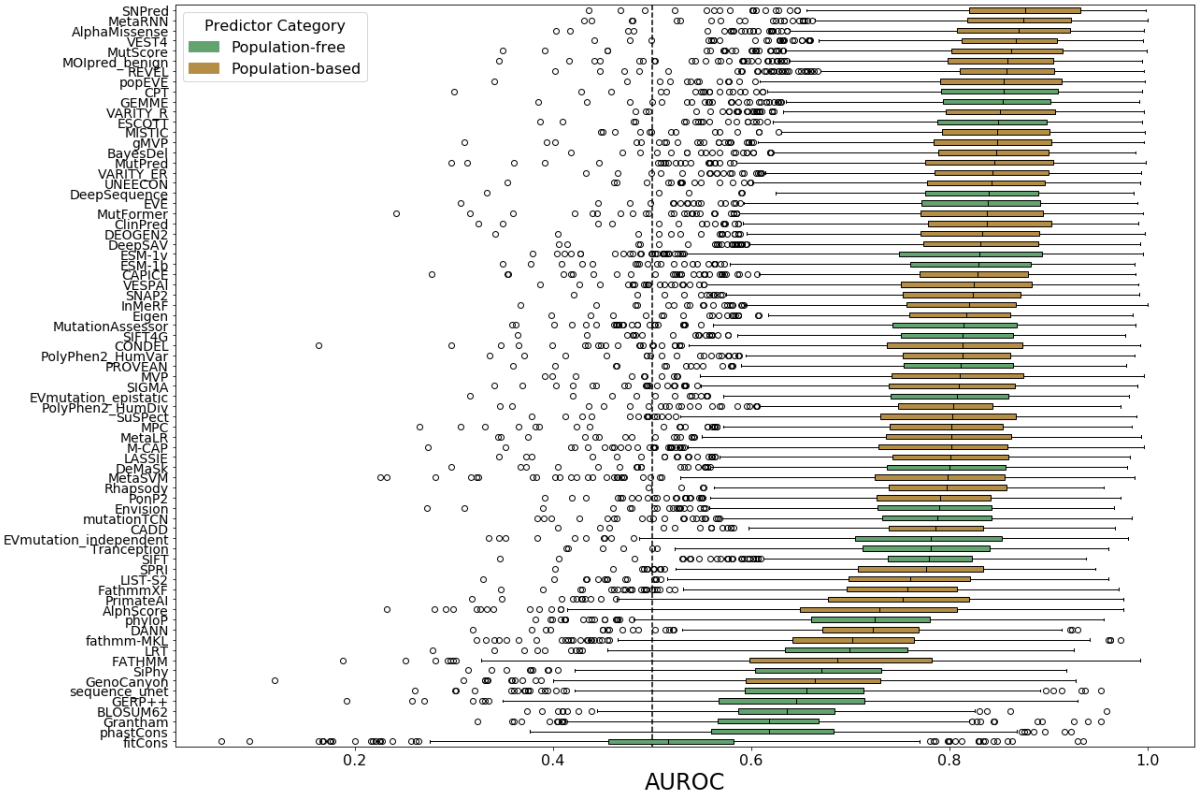

Figure S3. Distribution of AUROC representing discrimination between pathogenic and putatively benign missense variants for each VEP across different human protein-coding genes. As for the DMS comparison, not all VEPs output predictions for all proteins or all variants. Thus, the pairwise ranking represents a better reflection of relative performance.

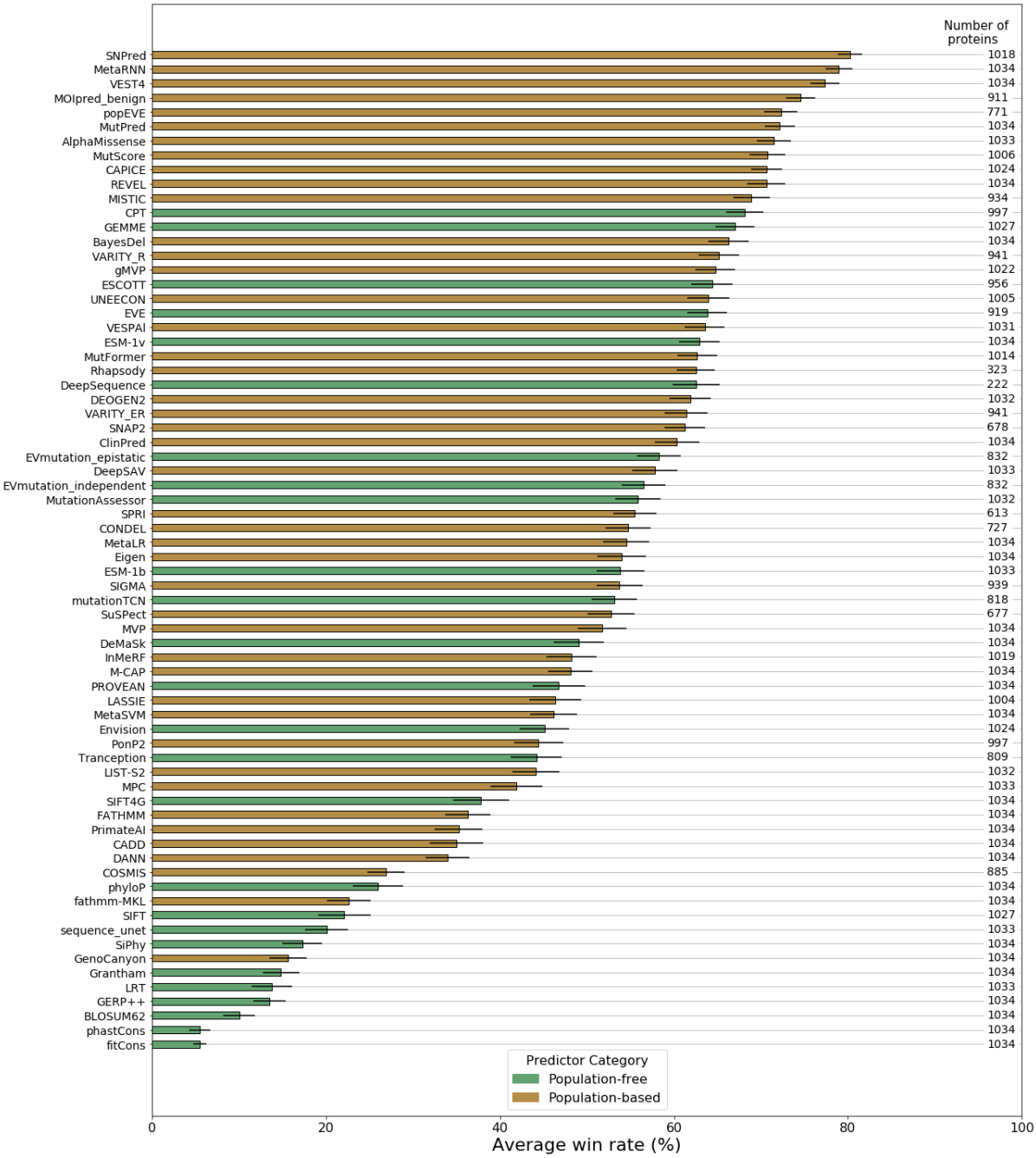

Figure S4. Ranking of VEPs on clinical variant classification using AUPRC. This utilises the same strategy as in Figure 3, but with precision-recall instead of receiver operating characteristic curves used in pairwise comparisons. Error bars represent the standard error across all comparisons with other VEPs.

Supplementary Tables

Table S1. List of all DMS datasets used in this study.

Table S2. List of all VEPs used in this study.

Table S3. Full pairwise win rate matrix for DMS benchmark.

Table S4. Full pairwise win rate matrix for clinical variant benchmark.