Interpreting protein variant effects with computational predictors and deep mutational scanning

© 2022. Published by The Company of Biologists Ltd | Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

REVIEW

GENETIC VARIANCE IN HUMAN DISEASE:
DECODING DIVERSITY TO ADVANCE MODERN MEDICINE

Interpreting protein variant effects with computational predictors and deep mutational scanning

Benjamin J. Livesey and Joseph A. Marsh*

ABSTRACT

Computational predictors of genetic variant effect have advanced rapidly in recent years. These programs provide clinical and research laboratories with a rapid and scalable method to assess the likely impacts of novel variants. However, it can be difficult to know to what extent we can trust their results. To benchmark their performance, predictors are often tested against large datasets of known pathogenic and benign variants. These benchmarking data may overlap with the data used to train some supervised predictors, which leads to data re-use or circularity, resulting in inflated performance estimates for those predictors. Furthermore, new predictors are usually found by their authors to be superior to all previous predictors, which suggests some degree of computational bias in their benchmarking. Large-scale functional assays known as deep mutational scans provide one possible solution to this problem, providing independent datasets of variant effect measurements. In this Review, we discuss some of the key advances in predictor methodology, current benchmarking strategies and how data derived from deep mutational scans can be used to overcome the issue of data circularity. We also discuss the ability of such functional assays to directly predict clinical impacts of mutations and how this might affect the future need for variant effect predictors.

KEY WORDS: Benchmarking, Circularity, Deep mutational scan, Machine learning, Multiplexed assay of variant effect, Variant effect predictor

Introduction

Rapid advances in sequencing technology over the past two decades has resulted in genomic information becoming an integral tool in both research and clinical fields. This wealth of data has helped identify thousands of human genetic variants in the population (Karczewski et al., 2020). A recent whole-exome sequencing study of the UK biobank cohort (Bycroft et al., 2018) identified a median of almost 20,000 coding variants per participant (Backman et al., 2021). However, most genetic variants are benign and unrelated to any disease. Variant effect predictors (VEPs) are computational tools that use this information to predict the phenotypic outcome of genetic variants and help highlight variants that are most likely to have clinically relevant effects. Where the phenotypic impact of a variant is uncertain, the variant is classified as a 'variant of uncertain (clinical) significance' (VUS) (Richards et al., 2015). VUSs account for a high proportion of identified variants (Balmaña et al., 2016), an issue that VEPs can potentially help to address. VEPs provide a quick, free and scalable alternative to time-consuming and expensive mutagenesis studies required to confirm the phenotypic effect of a VUS. This is particularly applicable to rare variants, where the expense of wet-lab experiments may be difficult to justify. Since the development of Sorting Intolerant From Tolerant (SIFT) (Ng and Henikoff, 2001), many other VEPs have been published with varying methodologies.

However, many of these VEPs provide contradictory results when classifying identical variants (Miller et al., 2019). This underscores the need for useful and accurate VEP benchmarking (Box 1, Glossary), which is becoming increasingly important as state-of-the-art predictors emerge. The authors of VEPs commonly assess performance against several previously published predictors using established variant databases. However, data circularity (Box 1) is an unresolved source of bias for many methods of benchmarking (Grimm et al., 2015). Comparisons between predictors often introduce bias by assessing predictor performance against the same data that were used to train them. Therefore, robust and unbiased benchmarking by independent groups is essential for assessing the performances of different VEPs.

One solution to this issue of data circularity is the use of independent variant effect datasets from deep mutational scanning (DMS; Box 1) experiments. DMS is a high-throughput technique to generate functional scores for, potentially, all variants of a protein (Fowler and Fields, 2014). For the purposes of benchmarking, DMS is fully independent from most training data. Most of the fitness scores (Box 1) from a DMS experiment are novel and not present in any sequencing dataset and, therefore, will not have been used to train or test previous predictors. Even those mutants that do exist in current datasets are scored independently of their previous categorisation.

In this Review, we will discuss the progress made in VEP methodology and assess different benchmarking strategies with an emphasis on the usage of DMS data and other large functional assays. We will also discuss the impact DMS may have as a direct independent measure of variant effects. We focus on VEPs developed to predict whether mutations are likely to be causing disease, excluding those that have been developed specifically to predict the effects of mutations on specific biophysical properties, such as protein stability (Gerasimavicius et al., 2020) or protein-protein interactions (Janin et al., 2003; Rodrigues et al., 2019). We hope that some of the issues we highlight will inform future VEP benchmarking efforts and the use of functional datasets.

Variant effect predictors

The importance of sequence conservation

Amino acid or nucleotide conservation (Box 1), calculated from alignment of related sequences, is an important feature for

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

*Author for correspondence (joseph.marsh@ed.ac.uk)

B.J.L., 0000-0001-6866-1452; J.A.M., 0000-0003-4132-0628

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

Box 1. Glossary

Bagging: A model-construction approach, most commonly used in random forest algorithms (see Box 2). Bagging creates a large ensemble of models, each of which only sees a subset of training data and a subset of input features. Although each model only has a part of the full picture, taken together they can correct each other's biases.

Benchmarking: Evaluation by comparison against a standard.

Conservation: When a particular amino acid is present in the same aligned position across a high proportion of related proteins, it is highly conserved. Conservation is an important predictive feature for many VEPs. Some predictors also use conservation at the nucleotide level.

Curated benchmarking datasets: Although large numbers of variants are available to train predictors, best results are usually obtained from high-quality datasets of validated variants. For this reason databases, such as VariBench (http://structure.bmc.lu.se/VariBench/) (Nair and Vihinen, 2013) exist, which curate the included variants to ensure high quality.

Data circularity: The re-use of data used to train a predictor, in order to assess the same predictor. Grimm et al. (2015) identified two types of data circularity that affect the assessment of VEPs.

Decision tree: A model defining a series of rules that can be used to aid in classification by splitting the samples. Data entries are split until each 'branch' of the tree contains sufficiently homogenous data, belonging primarily to a single class.

Deep mutational scanning (DMS): A high-throughput wet-lab procedure that produces fitness scores for a high number of mutations of a protein. DMS is a type of multiplexed assay of variant effect (MAVE; see below). Protein fitness is linked to cellular growth rate or other quantifiable attributes and can be assessed by measuring the abundance of each variant through sequencing counts after growth.

Dirichlet mixture model: A statistical model that can be applied to estimate the frequency of unobserved amino acids at conserved positions during multiple sequence alignment.

Fitness scores: The endpoint of DMS is the generation of quantitative fitness scores for a high proportion of possible variants in a protein. These scores are often based on the change in sequencing counts of each variant during the course of an assay, which selects against variants that reduce fitness.

Gold standard: The currently most-reliable dataset for benchmarking that best reflects real-world observations.

Holdout data: The most common approach to reduce data circularity in predictor assessment is to withhold part of a dataset from predictor training. These unseen data are then used to assess the predictor performance.

Indels: Short insertions or deletions of <1000 nucleotides within the genome. Indels in protein-coding regions can result in frameshifts.

K-fold cross-validation: A method to assess the performance of a supervised predictor on its full training data without data circularity. The dataset is split into K equal-sized subsets, with one of them being used to test the predictor performance, while all others are used to train the predictor. This process is repeated by using each data subset to test the predictor performance, producing predictions for the complete training dataset.

Meta predictors (ensemble predictors): VEPs that use the outputs of other prediction algorithms to produce their own estimates of variant pathogenicity.

Multiple sequence alignment (MSA): A data structure produced by aligning amino acid positions of related proteins. This is done with the aid of a substitution matrix that defines the likelihood of certain amino acid substitutions. MSAs form the basis of all variant effect predictors.

Multiplexed assay of variant effect (MAVE): This term describes any large-scale experimental procedure that generates fitness scores for genetic variants. MAVEs that apply to protein-coding regions of the genome are deep mutational scans.

Naïve metric: A naïve approach to a problem is one that makes a broad, probably untrue, assumption to help simplify the problem. One example is the naïve Bayes classifier (Box 2), where all inputs are assumed to be fully independent from one another, although – in reality – this is rarely the case.

NNK degenerate codons: A nucleotide NNK codon, where N is any nucleotide and K is guanine or thymine. An NNK codon can encode any amino acid but only one STOP codon. The encoded amino acids are also depleted of those comprising many possible codons as compared to entirely random, i.e. NNN, codons.

Position-specific independent counts (PSIC): An algorithm that reduces the impact of redundant sequences in position-specific scoring matrices (see below). PSIC uses a statistical approach to weight-aligned sequences when generating an alignment profile.

Position-specific scoring matrix (PSSM): A matrix of weights that can be derived from a multiple sequence alignment based on the frequency of each amino acid or nucleotide at every aligned position. PSSM alignment profiles are a way to quantify conservation within an alignment and a useful way to represent alignment features in a machine learning algorithm.

Pseudocounts: Predictors that make direct use of conservation in a multiple sequence alignment, such as SIFT, are unable to directly determine the likelihood of a residue appearing at a certain aligned position if it is never present in the alignment. To overcome this issue, SIFT makes use of amino acid pseudocounts from a Dirichlet mixture model. These are theoretical frequencies of amino acids based on substitution scores in the BLOSUM62 substitution matrix.

Receiver operating characteristic (ROC): A probability curve that represents the ability of a classifier to distinguish between binary classes. The true positive rate is plotted against the false positive rate at varying thresholds. ROC curves can be summarised by the area under the curve (AUC), which is 1.0 for a perfect classifier, 0.5 for random guessing and 0.0 for a perfect inverted classifier.

Variant effect predictors (VEPs): Computational tools that use various different sources of information to predict the phenotypic outcome of genetic variants and help highlight variants that are most likely to have clinically relevant effects.

predicting variant pathogenicity. Random mutations are continually happening across the genomes of all species; those that are detrimental to organismal fitness are removed from the gene pool, while those that have no effect are much more likely to be propagated to the next generation. As species diverge, we observe that neutral substitutions build up in homologous proteins, such that sequence similarity reduces with evolutionary time (Kimura, 1983). Since pathogenic variants result in decreased fitness, these are far less likely to be present within an alignment of homologous proteins. Thus, we can assume that substitutions frequently observed within an alignment of sufficient depth are likely to be neutral in nature, whereas those absent or rarely observed are much more likely to be pathogenic.

Sequence conservation is fundamental to every VEP (Table 1). One of the simplest tools that can be built from an alignment is an amino acid substitution matrix, such as Blocks SubstitUtion Matrix (BLOSUM) (https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt) (Henikoff and Henikoff, 1992) or Point Accepted Mutations (PAM) (Dayhoff, 1972; Jones et al., 1992). These matrices are calculated directly from alignments and contain values representing the propensities for different amino acid substitutions among related sequences. Although they were originally intended as tools to aid the alignment of protein sequences, these simple approaches have been shown to have modest ability to predict pathogenic mutations (Rentzsch et al., 2019; Shauli et al., 2021). Under some conditions, substitution matrices can even outperform specialised VEPs (Chan et al., 2007; Livesey and Marsh, 2020).

Another method for measuring sequence conservation is by comparing the rate at which each amino acid (or underlying nucleotide) appears within a column of a multiple sequence alignment (MSA; Box 1) (Ng and Henikoff, 2001). Specialised

Table 1. Summary of a selection of VEP methodologies and integrated features

			Feature groups					
Predictor	Methodology	Type	MSA	Sequence	Function	Structure	Other predictors	Reference and links
SIFT	Empirical	Unsupervised	✓					(Ng and Henikoff, 2001) https://sift.bii.a-star.edu.sg/index.html
PolyPhen	Decision trees	Unsupervised	✓	✓	✓	✓		(Ramensky et al., 2002) <Obsolete>
SNPs&GO	SVM	Supervised	✓	✓	✓			(Calabrese et al., 2009) https://snps.biofold.org/snps-and-go/snps-and-go.html
MutPred	Random forest	Supervised*	✓	✓	✓	✓	✓	(Li et al., 2009) http://mutpred.mutdb.org/
PolyPhen-2	Naïve Bayes classifier	Supervised	✓	✓	✓	✓		(Adzhubei et al., 2010) http://genetics.bwh.harvard.edu/pph2/index.shtml
Condel	Consensus	Supervised meta-predictor					✓	(González-Pérez and López-Bigas, 2011) https://bbglab.irbbarcelona.org/fannsdb/
SNPs&GO3D	SVM	Supervised	✓	✓	✓	✓		(Capriotti and Altman, 2011) https://snps.biofold.org/snps-and-go/snps-and-go-3d.html
Mutation Assessor	Combinatorial entropy optimisation	Unsupervised	✓					(Reva et al., 2011) http://mutationassessor.org/r3/
Pon-P	Random forest	Supervised meta-predictor					✓	(Olatubosun et al., 2012)
PROVEAN	Empirical	Unsupervised	✓					(Choi et al., 2012) http://provean.jcvi.org/index.php
Fathmm	HMM	Supervised	✓					(Shihab et al., 2013) http://fathmm.biocompute.org.uk/inherited.html
NetDisease SNP	Neural network	Supervised*	✓	✓		✓	✓	(Johansen et al., 2013) https://services.healthtech.dtu.dk/service.php?NetDiseaseSNP-1.0
CADD	SVM	Supervised meta-predictor	✓	✓	✓		✓	(Kircher et al., 2014) https://cadd.gs.washington.edu/
Mutation Taster2	Bayes classifier	Supervised	✓	✓	✓	✓		(Schwarz et al., 2014) https://www.mutationtaster.org/
SuSPect	SVM	Supervised	✓		✓	✓		(Yates et al., 2014) http://www.sbg.bio.ic.ac.uk/suspect/about.html
Pon-P2	Random forest	Supervised	✓	✓	✓	✓		(Niroula et al., 2015) http://structure.bmc.lu.se/PON-P2/
MetaSVM	SVM	Supervised meta-predictor	✓				✓	(Dong et al., 2015) Available vis dbNSFP https://sites.google.com/site/jpopgen/dbNSFP
SNAP2	Neural network	Supervised	✓	✓	✓	✓		(Hecht et al., 2015) https://www.rostlab.org/services/snap/
Eigen	Unsupervised spectral meta learner	Unsupervised meta-predictor	✓				✓	(Ionita-Laza et al., 2016) Available via dbNSFP https://sites.google.com/site/jpopgen/dbNSFP
REVEL	Random forest	Supervised meta-predictor	✓				✓	(Ioannidis et al., 2016) https://sites.google.com/site/revelgenomics/
M-CAP	Gradient boosting trees	Supervised meta-predictor	✓				✓	(Jagadeesh et al., 2016) Available via dbNSFP https://sites.google.com/site/jpopgen/dbNSFP
MPC	Logistic regression	Supervised meta-predictor	✓	✓			✓	(Samocha et al., 2017 preprint) Available via dbNSFP https://sites.google.com/site/jpopgen/dbNSFP
DEOGEN2	Random forest	Supervised*	✓	✓	✓		✓	(Raimondi et al., 2017) http://babylone.3bio.ulb.ac.be/MutaFrame/
Envision	Gradient boosting regression	Supervised	✓	✓		✓		(Gray et al., 2018) https://envision.gs.washington.edu/shiny/envision_new/
Fathmm-XF	Multiple kernel learning	Supervised	✓	✓	✓			(Rogers et al., 2018) https://fathmm.biocompute.org.uk/fathmm-xf/
PrimateAI	Neural network	Supervised	✓	✓		✓		(Sundaram et al., 2018) Available via dbNSFP https://sites.google.com/site/jpopgen/dbNSFP
Deep Sequence	Variational autoencoder	Unsupervised	✓					(Riesselman et al., 2018) https://github.com/debbiemarkslab/DeepSequence
ClinPred	Random forest and gradient boosted trees	Supervised meta-predictor	✓				✓	(Alirezaie et al., 2018) https://sites.google.com/site/clinpred/
MutPred2	Neural network ensemble	Supervised	✓	✓		✓		(Pejaver et al., 2020) http://mutpred.mutdb.org/
EVE	Variational autoencoder	Unsupervised	✓					(Frazer et al., 2020) https://evemodel.org/

Continued

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

Table 1. Continued

Predictor	Methodology	Type	Feature groups					Reference and links
			MSA	Sequence	Function	Structure	Other predictors	
MVP	Resnet	Supervised meta-predictor	✓	✓	✓	✓	✓	(Qi et al., 2021) Available via dbNSFP https://sites.google.com/site/jpopgen/dbNSFP
VARITY	Gradient boosted trees	Supervised*	✓	✓	✓	✓	✓	(Wu et al., 2021) http://varity.varianteffect.org/

* Despite using VEPs are predictive features these methods include far fewer than dedicated meta-predictors. Grouping was, therefore, done with supervised predictors.

Feature groups comprise:

MSA: features derived from multiple sequence alignments of related proteins, including conservation metrics.

Sequence: features derived purely from the protein sequence, including predicted secondary structure and amino acid properties.

Function: databases annotations regarding domains, active sites, protein interactions and more.

Structure: features derived from actual or predicted structures, such as accessible surface area and bond angles.

Other predictors: predictive features, such as VEP output.

HMM, hidden Markov model; MSA, multiple sequence alignment; SVM, support vector machine.

nucleotide conservation metrics, such as Genomic Evolutionary Rate Profiling (GERP++, http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html) (Davydov et al., 2010), PhyloP (http://compgen.cshl.edu/phast/) (Pollard et al., 2010) and Site-specific PHylogenetic analysis (SiPHy, https://portals.broadinstitute.org/genome_bio/siphy/index.html) (Garber et al., 2009) are often integrated into VEPs as measures of conservation. Such conservation metrics can also be used as standalone pseudo-predictors and have rivalled the performance of more-complex VEPs in some studies (Pejaver et al., 2020; Raimondi et al., 2017). Amino acid conservation is, therefore, a 'proxy' metric (Azevedo et al., 2017), but its utility for predicting pathogenicity is a testament to how effective the evolutionary process is at removing inefficient and pathogenic substitutions in nature.

Early computational predictors

Substitution matrices set the groundwork for early VEPs, such as SIFT (Ng and Henikoff, 2001), which is still frequently used today for variant effect prediction (Table 1). However, with rapid advances in computing over the past decades, capability to execute complex algorithms and process large amounts of data has increased. Although SIFT has been outperformed in multiple recent benchmarking studies (Mahmood et al., 2017; Niroula and Vihinen, 2019; Thusberg et al., 2011), it has the advantages of rapidly returning results, being easy to interpret and simple to run.

SIFT functions by generating an MSA that is based on the protein-of-interest. Each column of the alignment is scanned to determine the frequency of substitutions and the probability that a specific substitution is tolerated at each position. Substitutions at residues with high levels of conservation are the most likely to be pathogenic. This process is similar to the derivation of the BLOSUM substitution matrices but uses an MSA generated specifically for the protein-of-interest; this makes the conservation position-specific to the protein, adding more context to the value returned. 'Pseudocounts' (Box 1), calculated from a Dirichlet mixture model (Box 1) (Sjölander et al., 1996), are added to the alignment to help compensate for amino acids not observed at certain positions. Prediction quality is dependent on the depth of the alignment and can vary significantly within and between proteins. SIFT is often taken as a point of comparison for modern predictors, which is a tribute to its popularity (Pejaver et al., 2020; Raimondi et al., 2017; Sundaram et al., 2018).

Polymorphism Phenotyping (PolyPhen) is another early VEP (Ramensky et al., 2002). Unlike SIFT, PolyPhen makes use of a large amount of non-sequence protein information. PolyPhen considers protein features that are derived from the amino acid substitution site, including secondary structure and database-derived key-site annotations, such as active site and binding sites. An MSA is used to generate position-specific independent count (PSIC; Box 1) profiles (Sunyaev et al., 1999). Finally, if the sequence can be mapped to a known 3D structure, additional features – such as site-proximity, accessible surface area and secondary structure – are also incorporated into the prediction. The original PolyPhen algorithm uses a decision tree (Box 1) to calculate a score for the mutation of interest. More recently, PolyPhen-2 was released (Adzhubei et al., 2010), a version that uses more features and replaces the empirically derived classification rules with a supervised naïve Bayes classifier (Box 2, Machine learning techniques).

Comparisons between SIFT and PolyPhen have found that, although each method has relatively high sensitivity, their specificity is low (Niroula and Vihinen, 2019). They are also

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

Box 2. Machine learning techniques

Gradient-boosted trees are similar to random forest algorithms in that they use an ensemble of decision trees to make predictions. Unlike random forests, trees constructed in gradient boosting are neither random nor independent but each new tree is constructed in the attempt to correct the errors of the previous one. The output of every new tree is added to the output of all previous trees and this process continues until a pre-determined maximum tree number is reached. Independently, each tree is a weak learner, performing barely better than random guessing but, together, they can solve complex problems.

Neural Networks are composed of 'neurons' that mimic the way biological neurons in the human brain communicate. Neurons in each layer of the network take inputs from the layer above, apply a function and pass the results to the next layer; a network with many stacked layers is a 'deep' neural network. Deep networks are capable of learning more-complex relationships between the input features but are harder to train. The network learns by comparing the output to the training labels and back-propagating the errors. Deep neural networks can learn to approximate extremely complex non-linear functions in order to separate classes based on the inputs.

Naïve Bayes Classifiers are simple, supervised algorithms that classify examples on the basis of a vector of features. As a naïve method, these classifiers assume that all their input features are independent in order to simplify the problem. They are based on the Bayes theorem that is used to determine the probability of a class label given prior knowledge. Naïve Bayes methods are often fast and effective but performance can degrade when too many features violate the assumption of independence.

Random Forest (RF) algorithms construct multiple decision trees by using a process called *bagging*. Each tree is trained to use a random selection of the available features and a random subset of the training data. Owing to the bagging process, each tree within the ensemble is different and, as independent models, they have a low correlation. Having multiple decision trees acts as a safeguard against overfitting and errors made by some of the trees. Classification is performed by majority vote.

Support Vector Machines (SVMs) are algorithms that separate two classes by constructing a *hyperplane* between them. This hyperplane is a line for 2D data, a plane for 3D data and so on. The hyperplane is placed so it has the largest possible distance to any instance of training data of both classes. More classes can be separated by the addition of further hyperplanes. Data that are not linearly separable can be classified by SVMs by using a non-linear kernel. Although SVMs have excellent classification performance, they function best with low-noise non-overlapping classes.

Variational Autoencoders (VAEs) are a class of unsupervised generative models. VAEs are composed of two neural networks. One of which is an *encoder* that takes the input and compresses it to a Gaussian distribution in latent space. The distribution is then sampled and a *decoder* neural network attempts to reconstruct the original input. VAEs are 'generative' because they can generate novel outputs based on the training data they have seen.

much better at predicting loss-of-function than gain-of-function mutations (Flanagan et al., 2010). Both methods frequently appear inferior to many modern predictors in multiple benchmarking comparisons (Fig. 1) (Bendl et al., 2016; Ionita-Laza et al., 2016; Raimondi et al., 2016); however, they remain quick and straightforward to use, and their results are easy to interpret.

Machine learning

The predictors summarised in Table 1 represent only a small number of available VEPs, with machine learning being the basis of most VEPs developed since PolyPhen in 2002. Machine learning aims to find patterns in features, such as conservation, secondary structure, amino acid properties and more, then use these to make predictions about pathogenicity. Compared to empirically calculated scores, using machine learning to automatically determine feature contributions allows for the inclusion of more sources of data in the calculation. Most often VEPs use a form of supervised machine learning.

Supervised machine learning methods learn by example, by training on labelled datasets. Databases collating pathogenic variants, such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) (Landrum et al., 2014) and the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/) (Stenson et al., 2003), make ideal sources of pathogenic training examples, whereas gnomAD (https://gnomad.broadinstitute.org/) (Karczewski et al., 2020) provides a useful source of putatively benign variation observed in the human population. Several datasets exist, created specifically for the purposes of training and testing supervised VEPs, such as HumVar (http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads) (Capriotti et al., 2006) and Varibench (http://structure.bmc.lu.se/VariBench/) (Nair and Vihinen, 2013). These datasets also inevitably overlap by varying degrees (Mahmood et al., 2017). We highlight VARITY as an example of a state-of-the-art supervised predictor (Box 3, State-of-the-art predictors).

Many cutting-edge predictors use not only features calculated from protein sequences and structures but also the outputs of other VEPs. These meta-predictors, or ensemble predictors (Box 1), frequently combine features by using a supervised machine learning method, such as a random forest (Box 2) or a deep neural network (Box 2). Examples of these meta-predictors include ClinPred (Alirezaie et al., 2018), M-CAP (Jagadeesh et al., 2016), REVEL (Ioannidis et al., 2016) and MutPred2 that is highlighted in Box 3.

Unsupervised learning in VEPs has been slowly increasing in popularity, starting with the development of Eigen (Ionita-Laza et al., 2016). In unsupervised learning, the training examples are not labelled and the method makes its own decisions about how to make predictions. As the predictor does not see labelled examples during training, its predictions for specific variants are far less likely to be biased towards previous experience compared to supervised methods. Evolutionary Model of Variant Effect (EVE) is an example of an advanced unsupervised predictor (Box 3), but unsupervised learning shows even more promise as an avenue for future research.

Benchmarking the performance of VEPs

The problem of data circularity

Owing to the increasing volume of variants gathered in sequencing studies, with the majority being benign, identifying a single variant of concern in a noisy genetic background is extremely challenging, particularly if VEPs disagree as to the effect of the variant. As the amount of genetic data we generate keeps increasing, it is more important than ever to ensure we are using the correct tools for the job.

Benchmarking is most frequently performed by making predictions on sets of known pathogenic and benign variants. Relative performance is then assessed by several methods, most commonly by classification accuracy or area under the receiver operating characteristic curve (ROC AUC; Box 1). By far the most important aspect of benchmarking VEPs is the choice of variant dataset, which can substantially influence the outcome.

Grimm et al. (2015) described two types of data circularity in VEP benchmarking that can bias the assessment of predictor performance. Type 1 circularity primarily affects methods based on supervised machine learning. A method is susceptible to type 1 circularity if data used to train the model are re-used when assessing its performance. A model that is presented with data it has seen

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

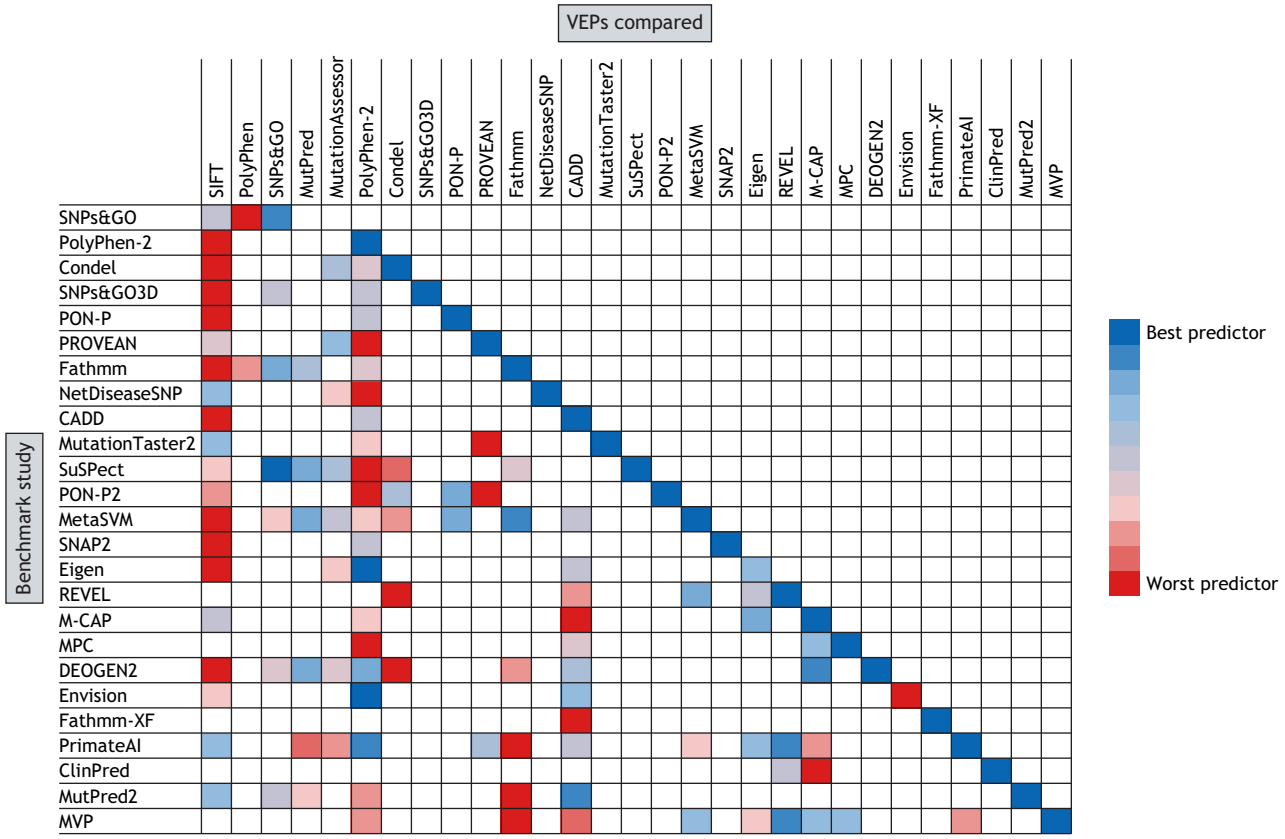

Fig. 1. Relative VEP performances in self-benchmarking analyses. The VEPs at the left are those that published a benchmark in their method paper. The VEPs at the top were compared within these benchmarks. Owing to space constraints, we could not include all VEPs compared in each study. We took the reported performance metrics, such as ROC AUC, directly from each paper. These scores were then used to rank each predictor from best to worst performance in each benchmark. Where multiple performance metrics were available, we selected a single representative measurement – i.e. ROC AUC when possible – followed by balanced accuracy and then any other presented metric. In cases where multiple benchmarks were performed, we selected one that– if available – used data independently of VEP training or, if not, the most-prominent analysis within the paper. ROC AUC, receiver operating characteristic area under the curve.

previously during training often performs better than it would if unseen data had been used. Whereas supervised machine learning methods are the most vulnerable to this form of bias, unsupervised methods are not immune. Such methods are often 'tweaked' based on their performance on a test dataset. This can lead to optimisation for that dataset and type 1 circularity if these variants are re-used while benchmarking. Type 2 circularity occurs because proteins with many variants in databases are often heavily skewed towards either pathogenic or benign outcomes. This results in deceptively good predictor performance if different variants from a single protein are used to train and test a predictor. However, in proteins that contain balanced numbers of pathogenic and benign variants, it can result in poor predictions.

Type 1 circularity is often addressed by careful curation of the variants used to benchmark predictors. However, this may limit the number of VEPs being compared if the training data of one fully overlaps the benchmarking dataset. Type 2 circularity is avoided by ensuring no variants in proteins used to train a predictor are used for benchmarking comparisons, ensuring that prior knowledge of variants in that protein is not used (Bromberg and Rost, 2007).

Self-benchmarking

When reading papers describing new computational predictors, one will tend to encounter an interesting phenomenon: the authors will almost always find their own method to be better than any others. To illustrate this, we reviewed publications describing 25 new predictors (limited to those in Table 1), where their performance in predicting coding missense variants was assessed (Fig. 1). The self-benchmarking in these papers almost exclusively finds that the novel method is superior to its predecessors for general variant effect prediction. One exception is the unsupervised predictor Eigen, which underperformed PolyPhen-2 when the authors assessed the methods on missense variants – although Eigen did perform best when assessed on missense and nonsense variants combined (Ionita-Laza et al., 2016). In addition, Envision performed the worst in its internal benchmark against missense variants; however, it is primarily intended to predict mutagenesis data rather than pathogenicity, probably explaining its relative performance (Gray et al., 2018).

Although most methods perform consistently well within their own benchmarks, benchmarks from subsequent publications often disagree markedly regarding the relative performance of earlier predictors. For example, the authors of Fathmm (Shihab et al., 2013) find it to be the top-performing method among ten benchmarked predictors, including MutPred (Li et al., 2009). However, subsequent comparisons by other groups find Fathmm to underperform MutPred (Raimondi et al., 2017; Yates et al., 2014). There are also significant differences in performance when the same author performs multiple benchmarks. For example, DEOGEN (Raimondi et al., 2016) outperformed five other methods, including SIFT, MutationAssessor (Reva et al., 2011) and PolyPhen-2, when compared using the Humsavar 11 (https://

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

Box 3. State-of-the-art predictors

We make note of the following recently developed VEPs for their innovation in either methodology or choice of training datasets.

Evolutionary Model of Variant Effect (EVE) (Frazer et al., 2021) uses a methodology similar to that of its predecessor DeepSequence (Riesselman et al., 2018). Both are unsupervised methods utilising a variational autoencoder (Box 2) to learn the latent rules that underlie an MSA. No features other than those of MSA are provided for the method, and no pathogenic or benign training data are used. In principle, the latent rules learned by EVE are those that underlie the evolutionary process responsible for generating the MSA the predictor was trained with. Variant scores are determined by comparing the probability of these rules producing the mutant sequence and the probability of them producing the wild-type sequence.

MutPred2 (Pejaver et al., 2020) uses a large number (1345) of features that can be categorised on the basis of sequence, substitution, position-specific scoring matrix (Box 1), conservation, homolog profiles or property changes. MutPred2 is a supervised, ensemble predictor composed of 30 separate neural networks that are trained using a matrix of features. A 'bagging' approach (Box 1), similar to how random forests are trained (Box 2), was used to expose each network to a random sample of training data. The predictions are the mean of the 30 neural network outputs. Training data have been derived from HGMD, SwissVar (discontinued) (Mottaz et al., 2010), dbSNP (https://www.ncbi.nlm.nih.gov/snp/) (Sherry et al., 1999) and interspecies alignments.

VARITY (Wu et al., 2021) makes use of the supervised gradient-boosted tree algorithm (Box 2) but innovates primarily in its use of unique training data. VARITY combines training examples from a large number of different sources, including functional assays. To compensate for potential low-quality data, training data are weighted based on specific metrics related to data quality, such as minor allele frequency within variant databases and internal quality metrics for functional assays.

www.uniprot.org/docs/humsavar) dataset for benchmarking (Raimondi et al., 2016). However, when using variants from an independent blind dataset, DEOGEN underperformed these three methods (Raimondi et al., 2017).

It is common for supervised machine learning-based VEPs to be benchmarked using their own training set, where predictions are generated by K-fold cross-validation (Box 1) to help prevent circularity (Adzhubei et al., 2010; Capriotti et al., 2013; Hecht et al., 2015; Pejaver et al., 2020). This may explain some of the performance discrepancies observed if it is the case that VEPs become optimised for the underlying biases of their training dataset. Furthermore, large variant datasets will inevitably have some underlying biases and structure, such as distinct proportions of variants from proteins with particular biological roles or disease mechanisms. Supervised predictors could use this information to over-perform in cross-validation or against holdout data (Box 1) (Capriotti and Altman, 2011; Carter et al., 2013; Jagadeesh et al., 2016) compared to tests using independent datasets. To overcome these issues, alternative benchmarking strategies include curated benchmarking (Box 1) datasets like Varibench (Feng, 2017; Niroula et al., 2015; Shihab et al., 2013; Yates et al., 2014) and making predictions on variants observed in relatively new studies that are unlikely to be present in any predictor training data (Dong et al., 2015; Raimondi et al., 2017; Sundaram et al., 2018).

One final issue with many benchmarking studies reported in VEP method papers is that certain well-known or innovative predictors are compared far more often than less-impactful VEPs that may still perform relatively well. SIFT and PolyPhen-2 have been compared in self-benchmarks of almost every VEP in the last 10 years (Fig. 1). SIFT, in particular, performs poorly in many of these comparisons but is still frequently used. In comparison, NetDiseaseSNP (Johansen et al., 2013) was only benchmarked in its own paper, so we have much less knowledge of how it compares to other predictors.

Independent benchmarks

From the above, it is clear that we must refine benchmarking methods of VEPs to improve their reliability. For this reason, independent benchmarks of VEPs that reflect realistic use-cases are, potentially, far more useful comparisons than self-benchmarks. One of the earliest independent comparisons of VEPs (Thusberg et al., 2011) investigated nine predictors by using variants drawn from the Phencode database (http://phencode.bx.psu.edu/) (Giardine et al., 2007), locus-specific databases and dbSNP. Although the main comparison in the paper did not exclude any training data for the supervised methods, a subsequent, smaller scale comparison that only used data from the locus-specific databases found that all methods performed worse on the limited dataset. No single predictor was superior by all outcome metrics; however, SNPs&GO (Capriotti et al., 2013) produced the highest accuracy for both the main study and on the limited dataset.

Most other early independent benchmarking studies focused on a small number of variants within a single protein or a group of related proteins. For example, VEPs were compared by using independent benchmarking in studies of 51 variants in the bilirubin uridine diphosphate glucuronsyltransferase gene (*UGT1A1*) (Galehdari et al., 2013), 74 variants in DNA mismatch repair genes (Thompson et al., 2013) and 122 RASopathy variants (Walters-Sen et al., 2015). In the *UGT1A1* study, SIFT performed best in terms of classification accuracy, despite being the oldest method. The DNA mismatch study acknowledged the issue with potential circularity, particularly in MutPred and PolyPhen-2, which were trained with variants in the target proteins. Re-training these predictors without variants in the mismatch repair genes resulted in degraded performance, particularly for MutPred. Unlike the other predictors in the study, SIFT allows the user to provide their own MSA rather than relying on the tool's native alignment. Interestingly, this group found that supplying hand-curated alignments featuring full-length homologues and sufficient variation at all positions to SIFT markedly improved performance over its native alignments. This demonstrates that the quality of the alignment used by VEPs has a significant impact on the predictions generated. The RASopathy study investigated 15 predictors, finding that the majority of programs performed below their published level. This study did not exclude any variants used to train the VEPs; therefore, even the predictors most strongly influenced by type 1 data circularity still performed poorly on this dataset.

The study by Grimm and colleagues (Grimm et al., 2015) highlighting the issue with data circularity, also contained benchmarks that adhere to the principles of minimising circularity by carefully selecting variants from Varibench, predictSNP (https://loschmidt.chemi.muni.cz/predictsnp/) (Bendl et al., 2014) and SwissVar, which were not used to train any of the assessed VEPs. Grimm et al. found that, when some of the training data for predictors are present within the benchmarking set, most supervised methods performed at their best. Performance degrades when all training data are excluded from the test. When the issue of type 1 circularity is not a factor, the empirically derived SIFT is comparable to some supervised machine learning methods, such as CONDEL (González-Pérez and López-Bigas, 2011). The approach used in this paper was highly effective at eliminating type 1 circularity by curating variant databases. However, as more predictors are trained, such datasets will need to be continuously

Disease Models & Mechanisms (2022) 15, dmm049510. doi:10.1242/dmm.049510

refined and predictor training data always made public to allow such comparisons in the future. Private datasets, such as HGMD, make these measures difficult to implement.

Modern benchmarking studies often take a similar approach to Grimm et al. (2015) in curating the assessment data to remove any VEP training variants and limit the number of VEPs being compared. A recent study comparing five predictors on clinical variants (Gunning et al., 2021) used data from ClinVar, HGMD, Online Mendelian Inheritance in Man (OMIM, https://www.ncbi.nlm.nih.gov/omim) and gnomAD, in addition to clinical and population studies to benchmark the VEPs. Although no training data were present in the variants used for the assessment, all tools – including SIFT – performed better on the dataset derived from existing database entries than on the dataset from newer clinical and population studies. Since type 1 circularity was not an issue, it is clear that the predictors are still optimised to better predict variants in the open dataset. One implication is that pathogenicity thresholds for VEPs might not be consistent between datasets, necessitating calibration studies. Such studies would involve testing VEPs against multiple datasets of variants or even individual proteins to determine the optimal prediction thresholds of each method in different contexts. Another recent study focussed on the often-overlooked ability of VEPs to predict benign variants (Niroula and Vihinen, 2019). To compare ten different predictors, this study used common variants from the former ExAC database that is now available at gnomAD (https://gnomad.broadinstitute.org/) (Karczewski et al., 2017). Training data from the supervised VEPs were filtered out of the ExAC data; however, the training data of four predictors – MetaLR, MetaSVM (Dong et al., 2015), M-CAP and REVEL – fully overlapped with the ExAC data, resulting in their exclusion from the study. This highlights that such studies are limited in the number of predictors they can assess while still accounting for data circularity.

One solution adopted by several groups to resolve the issue of data circularity is to use a gold standard (Box 1) for assessment, which is fully independent of any existing training data. This can be achieved by using datasets derived from experimental assays of variant effect. The use of independently generated functional data reduces the need to rely on databases that overlap VEP training data. Mutagenesis experiments also have the potential to assess the function of many entirely novel variants. Until recently, however, such comparisons were only possible on a small scale.

Deep Mutational Scanning

Experimental procedure

DMS is a relatively new fusion of large-scale mutagenesis and high-throughput sequencing that provides quantitative measurements of variant fitness, potentially assessing all possible variants in a protein (Fowler and Fields, 2014). This is a vast improvement over previous mutagenesis and directed evolution studies that focus on only a small subset of possible variants. DMS technology has rapidly improved, culminating in several studies that accurately recapitulate the effects of clinically validated variants (Findlay et al., 2018; Mighell et al., 2020).

All DMS studies begin with the generation of a library of mutant genes, usually accounting for all possible amino acid variants in the protein of interest (Fig. 2A). One such technique is POPCode (Weile et al., 2017), allowing replacement of each codon in the gene by using mutagenic primers with NNK degenerate codons (Box 1). Alternatives include a variety of techniques, from direct synthesis of the variant library (Jones et al., 2020) to random mutagenesis by error-prone PCR (Choudhury et al., 2020).

To assess the fitness of each variant in the library, protein function has to impact some measurable attribute of the cells expressing the mutant proteins. The exact mechanism varies greatly between experiments but, in the simplest case, the fitness of variants can be linked to cell growth rate (Fig. 2B) (Brenan et al., 2016; Giacomelli et al., 2018; Weile et al., 2017). Expressing variants that are unable to perform their function as effectively as the wild type reduces growth rates. As a final step, the growth rate effects of each variant can be determined by quantitative sequencing at different time points throughout the assay (Fig. 2C). Each variant is assigned a fitness score relative to the wild type and a null control. DMS experiments ultimately produce a heatmap of measured variant fitness over the protein or domain assessed at each position, indicating the areas that are least tolerant of substitutions (Fig. 2D).

The goal of DMS is to understand how different amino acid substitutions affect the ability of a protein to function. Where function of the protein is vital for cell growth or survival, it is simple to relate the resulting functional scores to disease risk. In proteins with a less-direct link to disease, protein function might need to be linked to cell growth rate through an artificial mechanism, such as placing an essential gene under the control of a yeast two-hybrid construct to assess protein–protein interaction fitness

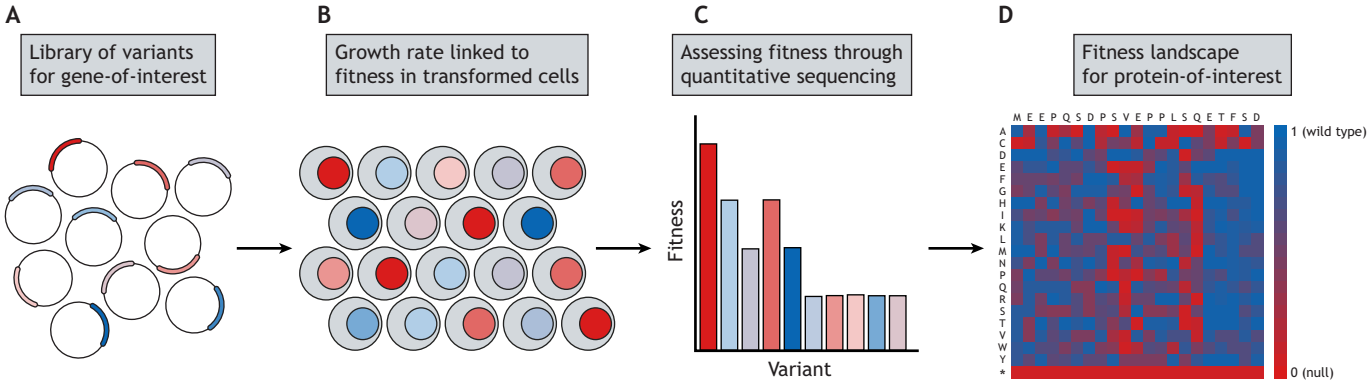

Fig. 2. Summary of a typical DMS experiment. (A) A library of variants, often representing every possible amino acid substitution in a protein, is generated and cloned into expression vectors. (B) The vectors are then introduced to mammalian or yeast cells where the function of the mutant protein is linked to the cell growth rate or some other measurable attribute. (C) Variant fitness is measured at different time points by quantitative sequencing, and compared to positive and negative controls to calculate relative fitness values. (D) A fitness map of all possible variants in the protein can be constructed from the relative fitness data.

(Bandaru et al., 2017; Starita et al., 2015). In these cases, there is no guarantee that the measured fitness metric will be correlated with human disease risk.

Using DMS to predict clinical outcomes

Potentially the most-exciting use of DMS is in directly assessing the clinical impact of novel variants, a feature it shares with VEPs. Unlike supervised VEPs, however, DMS fitness values are derived independently of previous data and, thus, are not subject to the circularity issues that VEPs are vulnerable to. DMS has shown promise in separating ClinVar pathogenic variants from putatively benign gnomAD variants, outperforming VEPs for several proteins (Livesey and Marsh, 2020).

Numerous germline variants in the tumour suppressor *BRCA1* have been found to pre-dispose women to developing breast cancer. Despite many sequencing studies, there still remains a large number of *BRCA1* VUSs and many novel variants to be observed. A recent DMS study quantified the effects of nearly 4000 *BRCA1* variants (Findlay et al., 2018). The group used a human HAP1 cell line and its growth rate as a fitness measure, owing to the essentiality of *BRCA1* for growth in this cell line (Blomen et al., 2015). The DMS results showed extremely high concordance with annotated ClinVar variants. Furthermore, Findlay et al. provided evidence that, for some variants where ClinVar and the DMS data diverge, DMS may provide a more-accurate functional assessment.

PTEN is another gene with numerous cancer-related germline variants and links to neurodevelopmental disorders. It is a tumour suppressor gene with a large number of benign variations, making it an excellent target for DMS studies. One study assessing *PTEN* variant fitness used a yeast system, in which *PTEN* activity rescues cell growth (Mighell et al., 2018). The authors found that the DMS data separated pathogenic ClinVar from putatively benign gnomAD variation with high levels of accuracy and sensitivity. This atlas of experimentally assessed variants may be useful for identifying potentially pathogenic mutations, and for distinguishing variants resulting in cancer from those causing neurodevelopmental disorders (Mighell et al., 2020).

Overall, these examples show that well-constructed DMS experiments can accurately identify known pathogenic variants, and have potential to annotate novel variants and VUS in disease-related genes. Such experiments – rather than clinical observations – may even have the potential to become a new ‘gold standard’ for variant outcome assessment.

Using DMS to assess the performance of VEPs

The Critical Assessment of Genome Interpretation (CAGI) is an ongoing experiment to assess the state of VEPs and related software (Andreoletti et al., 2019). In CAGI, software is tested in a series of variant interpretation challenges, spanning single nucleotide variants (SNVs), indels (Box 1), different molecular phenotypes, splicing effects, regulatory elements and more. CAGI assesses progress in the field by using data held back from publication, so that methods cannot be trained using these data. The 5th edition of experiments, i.e. CAGI5 challenge, included challenges derived from DMS-type experiments that included a yeast complementation assay with human calmodulin (Zhang et al., 2019), and a thermodynamic stability assay of *PTEN* and thiopurine S-methyltransferase (Matreyek et al., 2018).

Beyond CAGI, several independent groups have also applied data from DMS-style functional assays to benchmark VEPs. In an attempt to benchmark 46 VEPs, our lab previously used functional data from 31 DMS datasets from human, yeast, bacterial and viral sources (Livesey and Marsh, 2020). We calculated a relative rank score for each predictor based on the Spearman’s correlation between the continuous scores output by the VEPs and the DMS datasets. Overall, we found that the unsupervised method DeepSequence showed the best performance for human and bacterial proteins. SNAP2 (Hecht et al., 2015), DEOGEN2 (Raimondi et al., 2017), SuSPect (Yates et al., 2014) and REVEL also displayed relatively high performance, as well as ease of querying. Although variants used to train predictors were not removed from the benchmarking data, they made up only a tiny fraction of the overall DMS dataset.

Another study used three datasets composed of published *BRCA1* DMS data, *TP53* DMS data and variants in UniProt that originated from human mutagenesis experiments (UniFun). These three datasets overlap only very slightly with common training datasets (Mahmood et al., 2017). Compared to data commonly used for benchmarking, the BRCA1 and UniFun variants were poorly predicted, whereas accuracy regarding *TP53* data was relatively high for many methods. Mahmood and colleagues concluded that the difference in performance between traditional benchmarking datasets and functionally derived data is probably due to data circularity, providing an advantage on the former. An interesting consequence is that the empirical SIFT method produced the best performance on the UniFun data, outclassing multiple supervised machine learning methods.

Data from 22 DMS experiments were used to evaluate four VEPs together with conservation metrics (Reeb et al., 2020). Overall, Envision most accurately predicted variant deleteriousness determined by DMS. It was, however, unclear whether this was owing to genuine benefits of the VEP or bias due to Envision being trained directly using DMS data. The output of all VEPs tested correlated slightly with the DMS fitness values; however, all methods performed better on deleterious SNVs than on beneficial (gain-of-function) mutations. A naïve metric (Box 1) using PSI-BLAST also performed surprisingly well, outperforming some VEPs for classification.

DMS is a source of experimentally validated variant effect scores that are fully independent from existing classifications in databases most often used to train VEPs. This independence, together with the presence of large numbers of novel variants, allows for benchmarking of more predictors than traditional studies with less risk of data circularity. We can expect the popularity of using such datasets as a benchmark to increase, as DMS datasets become available for even more proteins.

Conclusions

VEPs and DMS studies can both be used to identify potentially pathogenic variants. With DMS technology constantly improving and sequencing becoming cheaper, the question arises whether we will need VEPs in the future, if DMS can provide us with direct measurements of variant effect. The precise definition of ‘fitness’ in DMS is very important as a protein’s fitness can often be defined in multiple ways. The challenge is to measure fitness in a way that correlates best with clinical outcomes. It is not always obvious how to achieve this for every protein. Although DMS results reflect the clinical outcome for many variants, for others the correlation can be poor (Livesey and Marsh, 2020). In the latter, the assessed fitness metric probably did not adequately reflect the mechanisms behind the disease caused by mutations of those proteins. However, DMS is also possible for proteins without disease association if fitness can still be assessed in some way. VEP benchmarking performance on such data is likely to dependent on whether the predictor takes the protein role and context into account for scoring.

Owing to their fast generation time and genetic tractability, many DMS studies are carried out in yeast cells. There is some concern that, because of evolutionary differences, fitness scores from yeast cells might not accurately reflect human genetic disease outcomes. Despite these differences between humans and yeast, it has been demonstrated that functional complementation assays performed in yeast systems still manage to accurately predict human disease (Sun et al., 2016).

However, DMS is resource-intensive and expensive, which can limit its use as a benchmark. A considerable level of expertise is also required to devise a suitable fitness assay and troubleshoot unforeseen issues. It is not currently feasible to perform DMS for every protein and, even if we could, there is no guarantee that the measured fitness metric would be applicable to human disease prediction. Until we have the knowledge and resources to construct assays to take into account all possible definitions of protein fitness, there will still be a requirement for VEPs in the future.

New generations of VEPs are constantly expanding their training data to broaden their experience and, hopefully, produce more accurate results. Recently, some VEPs that include DMS data as part of their training sets have been published. Benchmarking such predictors using DMS data carries the same caveats as benchmarking other supervised predictors with commonly used variant databases. For these VEPs, data circularity becomes an issue once again and, although dataset curation may help prevent circulatory, optimisation for DMS datasets may give the predictor an unfair advantage regardless. The two predictors we are aware of that make use of DMS data in their training sets are Envision and VARITY. Envision has been found to accurately predict effect magnitude in DMS datasets (Reeb et al., 2020), although our group found that Envision produced an average performance by using a different set of DMS data (Livesey and Marsh, 2020). The ability of VARITY to predict DMS data has yet to be assessed, although a study using gene-trait combinations found it to have excellent predictive performance (Kuang et al., 2021).

The question, therefore, remains whether functional assays can solve the issue of data circularity if variants used to train the VEPs are not excluded from the benchmarking dataset. The key is in the complete independence of DMS-derived datasets from variant interpretations based on clinical observation. Although there is usually a strong correlation, the measured variant effects in DMS are not necessarily identical to those in variant databases. DMS results are also often non-binary continuous values, which allows for differentiation between ‘extremely damaging’ and ‘slightly damaging’ variants, a distinction not found in traditional variant databases. Furthermore, the correlation between VEP predictions and measured intra-protein variant effects is likely to help in identifying those predictors susceptible to type-2 circularity. As previously outlined, type 2 circularity is caused by VEPs that associate particular proteins with a pathogenic or benign outcome and then applying that knowledge to new variants in the same protein. This effect provides the VEP with an advantage for binary classification of variants. In order to perform well against DMS fitness scores, a VEP has to have a strong correlation with the continuous values – which cannot be obtained solely by using knowledge of protein–disease associations. Overall DMS data stand as a potential solution to data circularity, particularly as more datasets emerge. However, this may soon become complicated by a new generation of VEPs trained with DMS data.

In this Review, we have described the recent progress made on computational predictors of variant effect and the issues with benchmarking these programs. DMS stands as not only a useful independent source of benchmarking data to limit circularity from performance estimates but also a, potentially, useful resource for direct variant classification. Outside of direct variant-effect prediction, DMS can also be applied to protein structure prediction (Adkar et al., 2012), residue contact prediction (Sahoo et al., 2015) and protein engineering (Spencer and Zhang, 2017).

We hope that, in future works, variant effect datasets from DMS-type studies will be more widely used to assess the performance of VEPs. This trend should naturally arise as further DMS experiments are carried out on human proteins and as the evolution of DMS methodology continues. With such datasets being used, we also expect to see an increase in the scope of independent benchmarking studies to include additional predictors. Such advances are likely to assist the identification of the best predictor methodologies and aid the production of more-accurate VEPs.

Acknowledgements
The authors thank Didier Devaurs, Mohamed Fawzy, Mihaly Badonyi and Lisa Backwell for their valuable feedback on this article.

Competing interests
The authors declare no competing or financial interests.

Funding
B.J.L. is supported by the Medical Research Council (MRC) Precision Medicine Doctoral Training Programme.

References
Adkar, B. V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., Swarnkar, M. K., Gokhale, R. S. and Varadarajan, R. (2012). Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* **20**, 371-381. doi:10.1016/j.str.2011.11.021

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248-249. doi:10.1038/nmeth0410-248

Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J. and Hocking, T. D. (2018). ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *Am. J. Hum. Genet.* **103**, 474-483. doi:10.1016/j.ajhg.2018.08.005

Andreoletti, G., Pal, L. R., Moult, J. and Brenner, S. E. (2019). Reports from the fifth edition of CAGI: the critical assessment of genome interpretation. *Hum. Mutat.* **40**, 1197-1201. doi:10.1002/humu.23876

Azevedo, L., Mort, M., Costa, A. C., Silva, R. M., Quelhas, D., Amorim, A. and Cooper, D. N. (2017). Improving the in silico assessment of pathogenicity for compensated variants. *Eur. J. Hum. Genet.* **25**, 2-7. doi:10.1038/ejhg.2016.129

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S. et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628-634. doi:10.1038/s41586-021-04103-z

Balmaña, J., Digiovanni, L., Gaddam, P., Walsh, M. F., Joseph, V., Stadler, Z. K., Nathanson, K. L., Garber, J. E., Couch, F. J., Offit, K. et al. (2016). Conflicting interpretation of genetic variants and cancer risk by commercial laboratories as assessed by the prospective registry of multiplex testing. *J. Clin. Oncol.* **34**, 4071-4078. doi:10.1200/JCO.2016.68.4316

Bandaru, P., Shah, N. H., Bhattacharyya, M., Barton, J. P., Kondo, Y., Cofsky, J. C., Gee, C. L., Chakraborty, A. K., Kortemme, T., Ranganathan, R. et al. (2017). Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* **6**, e27810. doi:10.7554/eLife.27810

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J. and Damborsky, J. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **10**, e1003440. doi:10.1371/journal.pcbi.1003440

Bendl, J., Musil, M., Štourač, J., Zendulka, J., Damborský, J. and Brezovský, J. (2016). PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Comput. Biol.* **12**, e1004962. doi:10.1371/journal.pcbi.1004962

Blomen, V. A., Májek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F. R., Olk, N., Stukalov, A. et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092-1096. doi:10.1126/science.aac7557

Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N. S., Zhu, C., Bagul, M., Goetz, E. M. et al. (2016). Phenotypic

characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell Rep.* **17**, 1171-1183. doi:10.1016/j.celrep.2016.09.061

Bromberg, Y. and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823-3835. doi:10.1093/nar/gkm238

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209. doi:10.1038/s41586-018-0579-z

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30**, 1237-1244. doi:10.1002/humu.21047

Capriotti, E. and Altman, R. B. (2011). Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics* **12**, S3. doi:10.1186/1471-2105-12-S4-S3

Capriotti, E., Calabrese, R. and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729-2734. doi:10.1093/bioinformatics/btl423

Capriotti, E., Calabrese, R., Fariselli, P., Martelli, P. L., Altman, R. B. and Casadio, R. (2013). WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics* **14**, S6. doi:10.1186/1471-2164-14-S3-S6

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14** Suppl. 3, S3. doi:10.1186/1471-2164-14-S3-S3

Chan, P. A., Duraisamy, S., Miller, P. J., Newell, J. A., McBride, C., Bond, J. P., Raevaara, T., Ollila, S., Nyström, M., Grimm, A. J. et al. (2007). Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). *Hum. Mutat.* **28**, 683-693. doi:10.1002/humu.20492

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688. doi:10.1371/journal.pone.0046688

Choudhury, A., Fenster, J. A., Fankhauser, R. G., Kaar, J. L., Tenaillon, O. and Gill, R. T. (2020). CRISPR/Cas9 recombineering-mediated deep mutational scanning of essential genes in Escherichia coli. *Mol. Syst. Biol.* **16**, e9265. doi:10.15252/msb.20199265

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A. and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025. doi:10.1371/journal.pcbi.1001025

Dayhoff, M. O. (1972). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125-2137. doi:10.1093/hmg/ddu733

Feng, B.-J. (2017). PERCH: a unified framework for disease gene prioritization. *Hum. Mutat.* **38**, 243-251. doi:10.1002/humu.23158

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M. and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217-222. doi:10.1038/s41586-018-0461-z

Flanagan, S. E., Patch, A.-M. and Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test Mol. Biomarkers* **14**, 533-537. doi:10.1089/gtmb.2010.0036

Fowler, D. M. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801-807. doi:10.1038/nmeth.3027

Frazer, J., Notin, P., Dias, M., Gomez, A., Brock, K., Gal, Y. and Marks, D. S. (2020). *Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning*.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y. and Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95. doi:10.1038/s41586-021-04043-8

Galehdari, H., Saki, N., Mohammadi-asl, J. and Rahim, F. (2013). Meta-analysis diagnostic accuracy of SNP-based pathogenicity detection tools: a case of UTG1A1 gene mutations. *Int. J. Mol. Epidemiol. Genet.* **4**, 77-85.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N. and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-i62. doi:10.1093/bioinformatics/btp190

Gerasimavicius, L., Liu, X. and Marsh, J. A. (2020). Identification of pathogenic missense mutations using protein stability predictors. *Sci. Rep.* **10**, 15387. doi:10.1038/s41598-020-72404-w

Giacomelli, A. O., Yang, X., Lintner, R. E., McFarland, J. M., Duby, M., Kim, J., Howard, T. P., Takeda, D. Y., Ly, S. H., Kim, E. et al. (2018). Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381. doi:10.1038/s41588-018-0204-y

Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H. et al. (2007). PhenCode: connecting ENCODE data with mutations and phenotype. *Hum. Mutat.* **28**, 554-562. doi:10.1002/humu.20484

González-Pérez, A. and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. *Am. J. Hum. Genet.* **88**, 440-449. doi:10.1016/j.ajhg.2011.03.004

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. and Fowler, D. M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst* **6**, 116-124.e3. doi:10.1016/j.cels.2017.11.003

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W. 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. doi:10.1002/humu.22768

Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L. and Wright, C. F. (2021). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J. Med. Genet.* **58**, 547-555. doi:10.1136/jmedgenet-2020-107003

Hecht, M., Bromberg, Y. and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics* **16**, S1. doi:10.1186/1471-2164-16-S8-S1

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915-10919. doi:10.1073/pnas.89.22.10915

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., 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. doi:10.1016/j.ajhg.2016.08.016

Ionita-Laza, I., McCallum, K., Xu, B. and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214-220. doi:10.1038/ng.3477

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A. and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581-1586. doi:10.1038/ng.3703

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I. and Wodak, S. J. (2003). CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2-9. doi:10.1002/prot.10381

Johansen, M. B., Izarzugaza, J. M. G., Brunak, S., Petersen, T. N. and Gupta, R. (2013). Prediction of disease causing non-synonymous SNPs by the artificial neural network predictor NetDiseaseSNP. *PLoS One* **8**, e68370. doi:10.1371/journal.pone.0068370

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275-282. doi:10.1093/bioinformatics/8.3.275

Jones, E. M., Lubock, N. B., Venkatakrishnan, A., Wang, J., Tseng, A. M., Paggi, J. M., Latorraca, N. R., Cancilla, D., Satyadi, M., Davis, J. E. et al. (2020). Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. *eLife* **9**, e54895. doi:10.7554/eLife.54895.

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B. et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840-D845. doi:10.1093/nar/gkw971

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P. et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443. doi:10.1038/s41586-020-2308-7

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310-315. doi:10.1038/ng.2892

Kuang, D., Li, R., Wu, Y., Weile, J., Hegele, R. A. and Roth, F. P. (2021). *Assessing computational variant effect predictors via a prospective human cohort*.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. and Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-D985. doi:10.1093/nar/gkt1113

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D. and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744-2750. doi:10.1093/bioinformatics/btp528

Livesey, B. J. and Marsh, J. A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380. doi:10.15252/msb.20199380

Mahmood, K., Jung, C., Philip, G., Georgeson, P., Chung, J., Pope, B. J. and Park, D. J. (2017). Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum. Genomics* **11**, 10. doi:10.1186/s40246-017-0104-8

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J. et al. (2018).

Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874. doi:10.1038/s41588-018-0122-z

Mighell, T. L., Evans-Dutson, S. and O'Roak, B. J. (2018). A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943-955. doi:10.1016/j.ajhg.2018.03.018

Mighell, T. L., Thacker, S., Fombonne, E., Eng, C. and O'Roak, B. J. (2020). An integrated deep-mutational-scanning approach provides clinical insights on PTEN genotype-phenotype relationships. *Am. J. Hum. Genet.* **106**, 818-829. doi:10.1016/j.ajhg.2020.04.014

Miller, M., Wang, Y. and Bromberg, Y. (2019). What went wrong with variant effect predictor performance for the PCM1 challenge. *Hum. Mutat.* **40**, 1486-1494. doi:10.1002/humu.23832

Mottaz, A., David, F. P. A., Veuthey, A.-L. and Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* **26**, 851-852. doi:10.1093/bioinformatics/btq028

Nair, P. S. and Vihinen, M. (2013). VariBench: a benchmark database for variations. *Hum. Mutat.* **34**, 42-49. doi:10.1002/humu.22204

Ng, P. C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863-874. doi:10.1101/gr.176601

Niroula, A. and Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLoS Comput. Biol.* **15**, e1006481. doi:10.1371/journal.pcbi.1006481

Niroula, A., Urolagin, S. and Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS One* **10**, e0117380. doi:10.1371/journal.pone.0117380

Olatubosun, A., Väliaho, J., Härkönen, J., Thusberg, J. and Vihinen, M. (2012). PON-P: Integrated predictor for pathogenicity of missense variants. *Hum. Mutat.* **33**, 1166-1174. doi:10.1002/humu.22102

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M. et al. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.*, **11**. 5918. doi:10.1038/s41467-020-19669-x

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110-121. doi:10.1101/gr.097857.109

Qi, H., Zhang, H., Zhao, Y., Chen, C., Long, J. J., Chung, W. K., Guan, Y. and Shen, Y. (2021). MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510. doi:10.1038/s41467-020-20847-0

Raimondi, D., Gazzo, A. M., Rooman, M., Lenaerts, T. and Vranken, W. F. (2016). Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. *Bioinformatics* **32**, 1797-1804. doi:10.1093/bioinformatics/btw094

Raimondi, D., Tanyalcin, I., Ferté, J., Gazzo, A., Orlando, G., Lenaerts, T., Rooman, M. and Vranken, W. (2017). DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201-W206. doi:10.1093/nar/gkx390

Ramensky, V., Bork, P. and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894-3900. doi:10.1093/nar/gkf493

Reeb, J., Wirth, T. and Rost, B. (2020). Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics* **21**, 107. doi:10.1186/s12859-020-3439-4

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-D894. doi:10.1093/nar/gky1016

Reva, B., Antipin, Y. and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118. doi:10.1093/nar/gkr407

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., 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**, 405-424. doi:10.1038/gim.2015.30

Riesselman, A. J., Ingraham, J. B. and Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816. doi:10.1038/s41592-018-0138-4

Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. and Ascher, D. B. (2019). mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338-W344. doi:10.1093/nar/gkz383

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R. and Campbell, C. (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511-513. doi:10.1093/bioinformatics/btx536

Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. C. and Varadarajan, R. (2015). Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis. *eLife* **4**, e09532. doi:10.7554/eLife.09532

Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., Neale, B. M. and Daly, M. J. (2017). Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353. doi:10.1101/148353

Schwarz, J. M., Cooper, D. N., Schuelke, M. and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361-362. doi:10.1038/nmeth.2890

Shauli, T., Brandes, N. and Linial, M. (2021). Evolutionary and functional lessons from human-specific amino acid substitution matrices. *NAR Genomics and Bioinformatics* **3**, lqab079. doi:10.1093/nargab/lqab079

Sherry, S. T., Ward, M. and Sirotkin, K. (1999). dbSNP—Database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677-679. doi:10.1101/gr.9.8.677

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M. and Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57-65. doi:10.1002/humu.22225

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. and Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* **12**, 327-345.

Spencer, J. M. and Zhang, X. (2017). Deep mutational scanning of S. pyogenes Cas9 reveals important functional domains. *Sci. Rep.* **7**, 16836. doi:10.1038/s41598-017-17081-y

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., Fowler, D. M., Parvin, J. D., Shendure, J. and Fields, S. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413-422. doi:10.1534/genetics.115.175802

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M. and Cooper, D. N. (2003). Human gene mutation database (HGMD®): 2003 update. *Hum. Mutat.* **21**, 577-581. doi:10.1002/humu.10212

Sun, S., Yang, F., Tan, G., Costanzo, M., Oughtred, R., Hirschman, J., Theesfeld, C. L., Bansal, P., Sahni, N., Yi, S. et al. (2016). An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res.* **26**, 670-680. doi:10.1101/gr.192526.115

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J. et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161-1170. doi:10.1038/s41588-018-0167-z

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G. and Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* **12**, 387-394. doi:10.1093/protein/12.5.387

Thompson, B. A., Greenblatt, M. S., Vallee, M. P., Herkert, J. C., Tessereau, C., Young, E. L., Adzhubey, I. A., Li, B., Bell, R., Feng, B. et al. (2013). Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. *Hum. Mutat.* **34**, 255-265. doi:10.1002/humu.22214

Thusberg, J., Olatubosun, A. and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32**, 358-368. doi:10.1002/humu.21445

Walters-Sen, L. C., Hashimoto, S., Thrush, D. L., Reshmi, S., Gastier-Foster, J. M., Astbury, C. and Pyatt, R. E. (2015). Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Mol. Genet. Genomic Med.* **3**, 99-110. doi:10.1002/mgg3.116

Weile, J., Sun, S., Cote, A. G., Knapp, J., Verby, M., Mellor, J. C., Wu, Y., Pons, C., Wong, C., van Lieshout, N. et al. (2017). A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957. doi:10.15252/msb.20177908

Wu, Y., Liu, H., Li, R., Sun, S., Weile, J. and Roth, F. P. (2021). Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* **108**, 1891-1906. doi:10.1016/j.ajhg.2021.08.012

Yates, C. M., Filippis, I., Kelley, L. A. and Sternberg, M. J. E. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* **426**, 2692-2701. doi:10.1016/j.jmb.2014.04.026

Zhang, J., Kinch, L. N., Cong, Q., Katsonis, P., Lichtarge, O., Savojardo, C., Babbi, G., Martelli, P. L., Capriotti, E., Casadio, R. et al. (2019). Assessing predictions on fitness effects of missense variants in calmodulin. *Hum. Mutat.* **40**, 1463-1473. doi:10.1002/humu.23857