nature portfolio

https://doi.org/10.1038/s41586-021-04043-8

Supplementary information

Disease variant prediction with deep generative models of evolutionary data

In the format provided by the authors and unedited

Supplementary Notes and Methods for Disease Variant Prediction with Deep Generative Models of Evolutionary Data

Jonathan Frazer *[1], Pascal Notin *[2], Mafalda Dias *[1], Aiden Gomez[2], Joseph K. Min[1], Kelly Brock[1], Yarin Gal †[2], and Debora S. Marks †[1,3]

* These authors contributed equally.
† Joint corresponding authors.
[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.
[2]OATML Group, Department of Computer Science, University of Oxford, Oxford, OX1 3QD, UK.
[3]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.

October 13, 2021

Contents

1 Supplementary Note: Limitations of supervised modeling methods

State-of-the-art computational methods for variant effect prediction which are trained on known clinical labels are plagued with several inference and validation issues, resulting in unreliable measures of performance and generalizability. While many of these limitations are ubiquitous in supervised models of biological sequence data in general [48], they are particularly problematic when training is done on currently known clinical labels for the following reasons:

1. Label bias – The majority of labels are concentrated in a small number of genes that are overrepresented at training and test time leading to inflated accuracy and generalizability estimates. Supervised methods train on a large proportion of the clinical labels aggregated across all genes (e.g., 90% of all labels) and estimate the generalizability of the model on the remaining hold out set (e.g., AUC /accuracy computed on the 10% of labels not used at train time). However, the distribution of the clinical labels across genes is extremely biased (e.g., 50% of all labels are in only 7% of the genes in ClinVar). Consequently, the labels used to assess generalizability performance at test time are biased towards a small number of genes that were also predominant at train time, leading to overestimated AUC/accuracy for genes with few labels [49].

2. Reduced validation set – Measuring test performance on an independent hold-out set of variants limits the scope of robust gene-by-gene evaluation. The number of test variants set aside to assess generalizability of supervised models is typically an order of magnitude smaller than the total number of labels available (e.g., 10% of all labels) – the remainder being leveraged for model training. Additionally, robust performance assessment of AUC/accuracy on a gene-by-gene basis is predicated on a sufficiently high number of labels in that test set for each gene. This thus leads overall to a much lower number of genes for which robust performance assessment can be carried out for supervised models compared with unsupervised approaches. For example, in ClinVar there are ~ 1000 genes with more than 10 labels, but a random 90-10 data split leads to only ~ 50 genes with at least 10 labels to perform validation on (Extended Data Fig. 7).

3. Label sparsity and noise – Training on sparse noisy labels leads to overfitting. Protein variant labelling in ClinVar is an error prone process. When the number of available labels is too small, supervised models risk overfitting the training data, leading to poor generalization performance [50,51].

4. Data leakage – Lack of independence between labels leads to overestimating model accuracy in the real world. Estimating model performance (e.g., AUC, accuracy) on the hold-out test set is a trustworthy measure of generalizability if the variants in that set are fully independent from the set of variants the model was trained on. This is rarely the case in practice given strong correlation between variant effects at a given residue position or, more subtly, within homologous domains. Crafting a cross-validation scheme that minimizes these dependencies between training and test sets is challenging and oftentime fully ignored, leading to inflated test set performance [50–52].

2 Supplementary Note: Comment on meta-predictors

In this work we refer to three classes of computational models: unsupervised, supervised and meta-predictors. We define meta-predictors as any model which uses as features the output from another model. The training set of a meta-predictor is the union of the training sets of the models used as features, together with any additional data used in training the meta-predictor itself. For this reason, the challenges in validating supervised models listed in Supplementary Note 1 are further exacerbated.

An additional challenge when validating the performance of meta-predictors against clinical labels, is the fact that some of these models have been used extensively in defining these same labels in ClinVar, introducing an additional form of circularity. So while in Extended Data Fig.4b and Supplementary Table 3 and 4 we compare the performance of EVE to meta-predictors (in addition to the supervised and unsupervised methods discussed in the main text, Methods), we stress that this comparison should be interpreted with care. This benchmark contains variants used in training the models as well as variants defined by the models themselves. Regardless, we find EVE to perform on par with the top-performing meta-predictors when assessing concordance with ClinVar labels. When assessing the performance of EVE to meta-predictors, in terms of concordance with high-throughput functional assays, we find EVE outperforms all methods.

On a final note, EVE and meta-predictors play distinct roles in variant classification. EVE is a single source of evidence and so should be considered in conjunction with other sources of evidence when attempting to classify the clinical significance of a variant. In contrast, meta-predictors often combine multiple distinct forms of evidence, in some cases all available sources of evidence. Hence, how these methods should be used in practice by experts differs according to the precise combination of evidence used.

3 Methods

3.1 Data acquisition

3.1.1 Choosing subset of clinically relevant proteins and building multiple sequence alignments from UniRef100 for model training

We focus on genes known to be involved in one or several diseases, that we define as the set of genes in ClinVar with at least one missense variant labeled as Pathogenic. There are 3,851 such genes in ClinVar (as of the April 2021 release). To each gene we associate a single protein, by picking the canonical transcript according to Uniprot/Swissprot [53]. To train EVE, we build multiple sequence alignments for each protein family by performing five search iterations of the profile HMM homology search tool jackhmmer [54] against the UniRef100 database of non-redundant protein sequences [55], downloaded on April 20th 2020. Following the protocol of Hopf et al. [56] and Riesselman et al. [57], we retrieve sequences that align to at least 50% of the target protein sequence, and columns with at least 70% residue occupancy. The software used is available via GitHub.

We explore a range of bit score thresholds, using 0.3 bits per residue as a reference, and select the best possible multiple sequence alignment based on the criteria of maximal coverage of the target protein sequence and sufficient, but not excessive, number of sequences in the alignment (the latter implying an alignment that is too lenient). Specifically, we prioritize alignments with coverage $L_{\text{cov}} \geq 0.8L$, where L is the length of the target protein sequence, and with a total number of sequences N such that $100,000 \geq N \geq 10L$. If these requirements cannot be met, we sequentially relax them down to $L_{\text{cov}} \geq 0.7L$ and $N \leq 200,000$. These criteria are met for 97% of alignments. For the remaining 3%, we drop the coverage constraint entirely. Following this procedure, we have so far obtained a set of 3,219 clinically relevant proteins with corresponding evolutionary training data (Supplementary Table 1). While we expect the performance of our model to depend on the quality of the multiple sequence alignments, we do not find strong correlation between performance and alignment depth N/L_{cov} (Extended Data Fig. 10).

3.1.2 Obtaining clinical significance labels from ClinVar

To obtain variants with clinical labels we make use of the ClinVar public archive [58], April 2021 release, which contains reports of the relationships between human genetic variation and phenotypes, with supporting evidence. Of particular relevance for this work is the classification of single nucleotide variants (SNVs) into five categories: Benign, Likely Benign, Uncertain Significance, Likely Pathogenic and Pathogenic. In addition, the quality of evidence provided is ranked according to a four star system, which can be summarized as follows:

- No Stars – Interpretation provided but criteria not met.
- One Star – Criteria provided, single submitter.
- Two stars – Criteria provided, multiple submitters, no conflicts.
- Three Stars – Reviewed by expert panel.

- Four Stars – Practice guideline.

Of $\sim$ 78k missense variants labeled (Likely) Benign or (Likely) Pathogenic in ClinVar, $\sim$ 63k have one star or more. In most of this work, with the exception of Extended Data Fig. 6, we only consider labels with quality rating of one star or higher.

While ClinVar contains clinical labels of SNVs, our model provides evidence at the amino acid variant level. We therefore require a procedure that selects a single label whenever more than one SNV is present in ClinVar for the same amino acid substitution. For these cases we pick the label with the most stars, and if two SNVs have labels with equal star-rating, we pick the most recent. Variants which do not match the transcript used as reference in our model are dropped from the analysis.

Finally, in order to obtain a high quality set of labels for validation, we assign all no-star labels as Uncertain and collapse the remaining (Likely) Benign and (Likely) Pathogenic into just two classes, Benign and Pathogenic, respectively. Unless stated otherwise, this is the set of benign and pathogenic labels used for benchmarking throughout the text. In total, we have $\sim$ 43k such labels across 3,219 proteins. Beside all variants labeled with Uncertain Significance and with no-star rating in ClinVar, we also define as Uncertain all variants observed in gnomAD and UK Biobank (detailed below) which do not feature in ClinVar. In total we find $\sim$ 1.3M "variants of unknown significance" across the 3,219 explored in this work.

For the purpose of comparison to predictions from high-throughput experiments, we find the above label definitions too restrictive, with only 5 genes having both a substantial number of these high quality labels, as well as high-throughput experimental data (Supplementary Table 6). To expand our analysis to include more experiments, we therefore define a second more lenient label policy. We define "Lenient" labels as the set of all labels in Clinvar – including no-star rating ones – as well as defining as Benign all variants which are more frequent in the population than the most frequent Pathogenic label in the same gene (frequencies estimated from gnomAD, see below). This policy is similar to the one used in Refs [59, 60]. It is important to stress that we expect these labels to be less trustworthy. Consistent with this, our model performance improves as we consider sets of labels with more stringent quality controls – our average AUC over all 3,219 proteins improves from 0.83 to 0.91 to 0.94 as we compute it against ClinVar lenient, 1-star or higher and 2-star or higher labels, respectively (Extended Data Fig. 6a, Supplementary Table 5).

3.1.3 Population data from gnomAD

The Genome Aggregation Database (gnomAD) [61], seeks to aggregate exome and genome sequencing data from a variety of large-scale sequencing projects, and provide summary data. We make use of both the v2 and v3 data sets, downloaded on the 23rd of April 2021. The v3 data set spans 71,702 genomes from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The v2 data set spans 125,748 exomes and 15,708 genomes from unrelated individuals, again, sequenced as part of various disease-specific and population genetic studies, totalling 141,456 individuals. The gnomAD coalition removes individuals known to be affected by severe pediatric disease, as well as their first-degree relatives, however some individuals with severe disease are potentially included in the data sets, albeit likely at a frequency equivalent to or lower than that seen in the general population [61].

Our downstream analysis makes use of this data for estimating frequency of amino acid variation over the population. For this purpose, we take amino acid variant frequencies to be the sum of frequencies of all missense SNVs coding for the same amino acid substitution.

3.1.4 Population data from UK Biobank

The UK Biobank [62] is an unprecedented large-scale biomedical database containing in-depth genetic and health information from half a million UK individuals. While it contains health-related records, bio-markers and detailed information about the lifestyle of the participants, in this work we only make use of the genomics aspect of this resource. In particular, we use the 50k whole exome sequencing cohort release from February 2020 assembled using a corrected SPB pipeline that converts raw sequencing data to a quality-controlled set of population variation. Unlike gnomAD which only provides population wide summary data, UK Biobank provides all SNVs for each individual genome, so we use this data to estimate the overall prevalence of more than one variant per gene across the population (Extended Data Fig. 9). We also use this data to extract all seen pairs of amino acid variants in the same gene, and their frequencies, across the "actionable genes" as defined by ACMG [63] (Supplementary Table 9). While we identify genes that have two or more non-synonymous mutations relative to the reference, only a fraction of these will be on the same chromosome due to the diploid nature of genomes. To highlight how distance from reference has population biases, we show results with respect to UK Biobank self-reported ethnicity background (Extended Data Fig. 9).

3.2 EVE Modeling approach

3.2.1 Learning distributions over protein sequences with VAEs

Variational Autoencoders (VAEs) [64, 65] have been shown to be effective at learning complex high-dimensional distributions across a wide range of tasks – from computer vision [66, 67], to natural language processing [68], to molecules modeling [69], and many others. In this work, we train VAEs to infer a distribution over amino-acid sequences for each protein. More formally, for a given protein family $\mathbf{p}$, we learn a distribution $p(\mathbf{s}|\theta_{\mathbf{p}})$ where $\mathbf{s}$ is a fixed-length amino-acid sequence and $\theta_{\mathbf{p}}$ are the model parameters for that protein family. Variational Autoencoders make the assumption that the data $\mathbf{s}$ are generated from a latent variable $\mathbf{z}$, and model the conditional distribution $p(\mathbf{s}|\mathbf{z}, \theta_{\mathbf{p}})$ with a neural network architecture (the "decoder", with parameters $\theta_{\mathbf{p}}$). We use amortized inference and model the approximate posterior distribution $q(\mathbf{z}|\mathbf{s}, \phi_{\mathbf{p}})$ with a neural network as well (the "encoder", with parameters $\phi_{\mathbf{p}}$). Similarly to Riesselman et al. [57], we obtain stronger performance with a Bayesian VAE in which we learn a fully-factorized Gaussian distribution over the decoder weights $\theta_{\mathbf{p}}$. We interpret this observation by the fact that in a VAE architecture the encoder is trained only in-distribution and will extrapolate arbitrarily when dealing with out-of-distribution points. Thus, for mutants far from the training data, latent positions obtained from the encoder will be less reliable. While a standard VAE will then decode that mutant to an arbitrary position, the output from a Bayesian VAE will average over decodings (the decoder being a Bayesian Neural Network), which will dampen the corresponding probability estimates over amino acid positions [70] (Extended Data Fig. 1).

Training VAEs over Multiple Sequence Alignments

The distribution of protein sequences available in genomic databases is biased by human sampling (e.g., certain species of interest are more sequenced than others) and evolutionary sampling (phylogeny). Not correcting for these biases in the Multiple Sequence Alignments (MSAs) that we extract from genomic databases to train our models will lead to learning an improper probability distribution. Following the approach described in Ekeberg et al. [71], we correct for these two biases by re-weighting each protein sequence $\mathbf{s}_i$ from a given MSA according to the reciprocal of the number of sequences in the corresponding MSA within a given Hamming distance cutoff T.

$$\pi_{s_i} = \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbb{1}[Dist(\mathbf{s}_i, \mathbf{s}_j) < T] \right)^{-1} \tag{1}$$

where N is the number of sequences in the MSA, i indexes over proteins, and $\mathbf{s}_j$ are other protein sequences in the MSA. Similarly to Hopf et al. [56], we set $T = 0.2$ for all human proteins. During model training, we sample each mini-batch element by sampling sequences according to their weight π_s.

We train the VAE models by maximizing the Evidence Lower Bound (ELBO) which forms a tractable lower bound to the log-marginal likelihood. Following the "Full Variational Bayesian" approach described in [64], the ELBO is expressed as follows:

$$ELBO(\mathbf{s}) = N.\mathbb{E}_{p(\mathbf{s})}\left[\mathbb{E}_{q(\theta_\mathbf{p}),q(\mathbf{z}|\mathbf{s})}\left(\log p(\mathbf{s}|\mathbf{z},\theta_\mathbf{p})\right) - D_{KL}(q(\mathbf{z}|\mathbf{s},\phi_\mathbf{p})||p(\mathbf{z}))\right] - D_{KL}(q(\theta_\mathbf{p})||p(\theta_\mathbf{p})) \tag{2}$$

where N is the size of the training data and $p(z)$ is the prior distribution over latent variable z (standard Gaussian), and D_{KL} is the Kullback–Leibler divergence (a measure of dissimilarity between two probability distributions). In our Bayesian VAE formulation, we learn a fully-factorized Gaussian distribution $q(\theta_\mathbf{p})$ over the decoder parameters, with standard Gaussian prior $p(\theta_\mathbf{p})$. Similar to Riesselman et al. [57] and in agreement with the sequence re-weighting scheme, we set $N = N_{eff} = \sum \pi_{s_i}$, the effective number of sequences in the MSA defined as the sum of the different sequence weights.

3.2.2 Model architecture

We performed several ablation analyses to optimize the underlying model architecture and the choice of training hyperparameters (see Extended Data Fig. 1). The main changes we suggest over the DeepSequence architecture [57] are as follows:

- Symmetrization of the encoder and decoder network architectures;
- Increased number of layers and increased layer width for the encoder and decoder networks ($2,000$ - $1,000$ - 300 and 300 - $1,000$ - $2,000$ respectively);
- Increased size of the latent space (50);

- Larger number of training steps to train the more complex architecture $(400,000$ training steps$)$;
- Lower learning rate to stabilize learning process (10^{-4});
- Removal of the group sparsity priors, responsible of significant performance drops for certain proteins;

Extended Data Fig. 2 summarizes the performance gains achieved with the changes above compared with DeepSequence, by comparing Spearman correlation of the two models with the output of 38 different benchmark MAVEs, following the same protocol as described in Riesselman et al. [57].

3.2.3 Evolutionary index

We define the *evolutionary index* of a protein variant $\mathbf{s}$ as the relative fitness of $\mathbf{s}$ compared with that of a wild-type sequence $\mathbf{w}$. Building from the probabilistic viewpoint that gave rise to the Bayesian VAE, fitness of a sequence may be measured by the difference in log-likelihood of that sequence with the wild type. Since estimating the exact log-likelihood is intractable in practice, we use the negative ELBO, a bound on the log marginal likelihood, as an approximation which leads to strong empirical results (Fig. 2). The *evolutionary index* E_s of protein $\mathbf{s}$ may therefore be expressed in terms of a tractable difference between two ELBOs:

$$E_s = -\log \frac{p(\mathbf{s}|\theta_\mathbf{p})}{p(\mathbf{w}|\theta_\mathbf{p})} \sim ELBO(\mathbf{w}) - ELBO(\mathbf{s}) \tag{3}$$

For each variant $\mathbf{s}$ of interest, since the integral over $\mathbf{z}$ is intractable in practice, we estimate ELBO values via Monte Carlo sampling of the latent space, taking a large number of samples – 20k – from the approximate posterior distribution $q(\mathbf{z}|\mathbf{s}, \phi_\mathbf{p})$.

3.2.4 Separating pathogenic and benign variants with probabilistic clustering

Evolutionary indices show very strong correlations with existing clinical labels across proteins (Extended Data Fig. 3a). We fit a Gaussian Mixture model with two components directly on the evolutionary index distribution, in order to automatically separate variants into Pathogenic and Benign clusters. Besides the advantage of being fully unsupervised, performing probabilistic clustering of variants allows us to quantify our uncertainty about the class assignment (see next section).

The Gaussian Mixture Model (GMM) [72] is a probabilistic model that assumes the data are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. We experiment with different architectures and training algorithms: single model Vs hierarchical, training with Variational Inference (VI) Vs Expectation-Maximization (EM) algorithm. We obtain the best performance on the downstream task with the following approach. We first train an overarching two-component GMM on the distribution of evolutionary indices for all single amino acid variants for all 3,219 proteins combined. We use the resulting parameters to initialize the Gaussians of the protein-specific GMMs, fit on all single amino acid variants

for each protein separately. Intuitively, the cluster with higher mean will contain sequences with higher evolutionary indices, and therefore less likely sequences under the learnt sequence distribution. For each model after convergence, we thus define the component with the highest mean as the Pathogenic cluster, and the other one as the Benign cluster. We then form a global-local mixture of GMMs to combine the cluster predictions from the overarching GMM with the predictions from the protein-specific GMM for each protein. More formally, for a given protein **s** with evolutionary index $\mathbf{E_s}$, from a protein family p, we have:

$$p(\mathbf{X_s} = 1|\mathbf{E_s}) = \alpha * p(\mathbf{X_s} = 1|\mathbf{E_s}, \theta_p) + (1 - \alpha) * p(\mathbf{X_s} = 1|\mathbf{E_s}, \theta_o) \quad (4)$$

where X_s is a binary random variable equal to 1 if **s** is pathogenic (0 if benign), α represents the relative weight of protein-level GMM in the ensemble (set to 0.3 via grid search with respect to average accuracy and AUC), θ_o and θ_p are the parameters of the overarching GMM and protein-specific GMM respectively. $p(\mathbf{X_s} = 1|\mathbf{E_s})$ is what we refer to as the *EVE score* which quantifies the propensity of a given variant to be pathogenic (Extended Data Fig. 3b).

3.2.5 Quantifying the uncertainty in the class assignment

A crucial benefit of a probabilistic clustering approach is the ability to identify the set of variants for which the classification is the most uncertain. We measure the total uncertainty on the cluster assignment for a protein **s** via the corresponding predictive entropy PE:

$$PE = -\log p(\mathbf{X_s} = 1|\mathbf{E_s})p(\mathbf{X_s} = 1|\mathbf{E_s}) - \log p(\mathbf{X_s} = 0|\mathbf{E_s})p(\mathbf{X_s} = 0|\mathbf{E_s}) \quad (5)$$

As we increase the proportion of variants excluded from our classification based on their predictive entropy (excluding higher values first), the accuracy monotonically increases (Fig. 2b). This confirms the ability of the predictive entropy metric to properly identify uncertain variants. By setting as Uncertain the 25% of variants with the highest uncertainty, the accuracyis $\sim 90\%$. In practice, we envision the user of our scores to use this uncertainty metric on a gene-by-gene basis according to the desired precision/recall.

3.3 Validation

3.3.1 Overall validation of EVE with clinical labels

We measure the aggregate and protein-by-protein performance of the EVE model in terms of its ability to properly discriminate between pathogenic and benign variants for which clinical labels already exist. As discussed above, we focus on the subset of highest quality labels for the purpose of model evaluation. Aggregate and protein-level performances are measured via the area under the Receiver Operating Characteristic curve (AUC) and the total prediction accuracy (setting a threshold that equally balances out false positives and false negatives). Our main results are summarized in Fig. 2a, b (see also Supplementary Table 2 for detailed results across all 3,219 proteins). We looked at performance across different proteins groups: the full set of 3,219 proteins and the subset thereof with at least 3, 5 and 10 Benign and at least 3, 5 and 10 Pathogenic labels.

3.3.2 Benchmarking EVE's performance against supervised models of variant effects

We compare the performance of EVE models with that of 23 of the most popular computational methods (8 supervised, 8 meta-predictors and 7 unsupervised) [73–92]. Scores for these models were obtained from dbNSFP [93]; we excluded models explicitly trained on population frequency data in order to avoid circularity when validating on known labels, as frequency is currently heavily used for classification of variants.

Performance is assessed by two metrics: first, we compute AUCs and classification accuracies over high quality labels as defined above (Fig. 2c x-axis, Extended Data Fig. 4b, x-axis; Supplementary Table 3). We compute average, weighted average and censored average of protein-level AUC and accuracy, at the variant and position levels (Supplementary Table 2). Since we do not account for the train-validation split used by supervised models and meta-predictors at train time (as some of these methods did not make this data publicly available), the reported performance for variant effect prediction is to be interpreted as an upper bound on their true performance. Furthermore, as state-of-the-art meta-predictors are used to define labels in ClinVar, their reported performance suffers especially from circularity (Supplementary Note 2).

Secondly, we compute the correlation with 40k experimentally measured variants across 10 proteins [59, 94–102] – a benchmark that we argue is less sensitive to the biases and circularity issues with ClinVar (Fig. 2c, y-axis; Extended Data Fig. 4b, y-axis; Supplementary Table 4). We chose these experiments based on the clinical relevance of the assay. Since these experiments are independent of the ClinVar labeling process, performance metrics on this benchmark provide a less biased estimate of true generalization performance, at the cost of only being available for a limited number of genes. We note nonetheless that there may be residual training bias when assessing the performance of supervised methods on assays carried out on proteins with many labels in ClinVar (e.g., BRCA1).

3.3.3 Benchmarking EVE's performance against performance of high-throughput functional assays

We compare the predictions of EVE to the predictions of a number of high-throughput experiments [59, 95–97, 100, 102–108], which were also designed with the intention of predicting the pathogenicity of variants. Not all assays are equally relevant for human disease phenotype. We compute AUCs on the intersecting set of variants common to both the assay and EVE, using multiple label definitions as described above, as well as older versions of ClinVar, to avoid comparing EVE and experiment on labels which were established by making use of data from that same experiment. On average we find EVE outperforms the experiments regardless of this choice (Supplementary Table 6). In Extended Data Fig. 6b we use "lenient" labels as well as labels with at least 1-star quality rating (defined in Data Acquisition section), selecting 2021 or 2017 ClinVar data as appropriate and noted in Supplementary Table 6.

In Fig. 3 and Extended Data Fig. 5 we present results using high quality labels only – with at least 1-star quality rating – as they are sufficient for performance estimates for P53, BRCA1 PTEN, SCN5A and MSH2. In Fig. 3 we use ClinVar 2021 data only, for simplicity. This means there is some circularity potentially inflating the estimate performance for P53 and BRCA1. For clarity, in Extended Data Fig. 5 we distinguish ClinVar 2021 data (lighter red, lighter blue) from

ClinVar 2017 (darker red, darker blue), when the experiment has been used in establishing labels in the 2021 data. Reported AUCs in Fig. 3 are computed using all available labels (as opposed to intersecting set).

3.4 Combining EVE predictions with other sources of evidence

The 2015 American College of Medical Genetics and Genomics–Association for Molecular Pathology (ACMG-AMP) guidelines [109] present steps towards a systematic approach to variant classification which can be used consistently across independent groups. They propose a classification scheme consisting of 28 criteria to classify variants into one of five categories – Benign, Likely Benign, Uncertain, Likely Pathogenic and Pathogenic. Each criteria corresponds to a different form of evidence, such as population data, or functional data, and the strength of evidence provided by a given criteria falls into one of four categories – Supporting, Moderate, Strong and Very Strong. Finally, a set of rules determines how criteria are to be combined to determine the category of a given variant. For instance, one Strong pathogenic criterion, combined with two or more Supporting criteria, would result in a variant being classified as Likely Pathogenic.

One of the defining characteristics of our model is the fact that it only uses one source of evidence – evolutionary data – to score variants according to their clinical significance. As such, it is straightforward to combine the EVE scores with other evidence in a manner similar to the strategy outlined by the ACMG-AMP. We illustrate this by using EVE scores as Strong evidence. We now detail how we collect different lines of evidence and implement them in our analysis to classify as many variants as possible.

In practice, only a small number of the other ACMG-AMP criteria are amenable to use in large N variant classification at present. Of the 28 criteria defined by ACMG-AMP, we only make use of 4 in our analysis:

- **Strong Benign criterion 1** ($BS1$) **Definition**: Allele frequency is greater than expected for disorder.
- $BS1$ **Implementation**: We make use of population data from gnomAD [61]. For a variant to satisfy this criterion, we require that it be observed more frequently in gnomAD than any known (Likely) Pathogenic variant in that gene. In other words, we require a lower bound on the frequency $\nu_{\text{var}} > \max(\nu_{\text{path}})$.
- **Supporting Benign criterion 6** ($BP6$) **Definition**: Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation.
- $BP6$ **Implementation**: We took variants in ClinVar labelled as Benign but with a zero-star rating to satisfy this criterion.
- **Moderate Pathogenic criterion 2** ($PM2$) **Definition**: Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium.
- $PM2$ **Implementation**: This criterion is the pathogenic equivalent of $BS1$. Again we use frequency data from gnomAD and this time we require that for a variant to satisfy this

criteria its frequency must be lower than any known (Likely) Benign variant in that gene, $\nu_{var} < \min(\nu_{\text{benign}})$.

- **Moderate Pathogenic criterion 5** ($PM5$) **Definition**: Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before.
- $PM5$ **Implementation**: Any variant found at the same position as a (Likely) Pathogenic variant with at least a one-star rating in ClinVar met this criterion.

Combining criteria: If we take our models as capable of providing strong evidence of a variant being either benign BS_{EVE}, or pathogenic PS_{EVE} then there are 4 ways in which our model evidence can be combined with the above criteria to reclassify a VUS as Benign, Likely Benign, or Likely Pathogenic:

- **Benign**: BS_{EVE} and $BS1$
- **Likely Benign**: BS_{EVE} and $BP6$
- **Likely Pathogenic**: (PS_{EVE} and $PM2$), or (PS_{EVE} and $PM5$)

Whenever the conclusions of this analysis are conflicting with an existing ClinVar assignment, we set the variant as Uncertain. We provide both summary statistics (Fig. 4a. Supplementary Table 2) as well as a complete list of evidence used for the classification of every variant in our analysis (available in evemodel.org).

3.4.1 Comments on criteria not used in our analysis

There are additional criteria for variant classification that we did not make use of due to concerns of "double counting" particular forms of evidence. Broadly, this problem is well known [110] but also manifests in somewhat unique ways in our analysis. As such, here we comment on how these issues would arise in our analysis should we try to make use of certain criteria.

- **Moderate Pathogenic criterion 1** ($PM1$) **Definition**: Located in a mutational hot spot and/or critical and well-established functional domain (*e.g.*, active site of an enzyme) without benign variation.
- $PM1$ **Discussion**: EVE, being a model for all possible amino acid variants in a protein sequence, is a natural hot/cold spot detector. The objective of the classification exercise, however, is to combine distinct sources of evidence, so we leave this criterion out of our analysis.
- **Supporting Benign criterion 4** ($BP4$) **Definition**: Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
- **Supporting Pathogenic criterion 3** ($PP3$) **Definition**: Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)

- $BP4$ **and** $PP3$ **Discussion**: The model we propose in this paper should in principle fall into this category but we have argued that by using an unsupervised approach it does not suffer from label biases and it can be subjected to more stringent validation. This results in a more reliable form of evidence than is provided by current state-of-the-art computational approaches, and we argue that it can stand alone as a source of strong evidence. We could in principle use other computational methods as an additional source of evidence in our classification pipeline, however this does not seem reasonable since almost all computational methods make at least some use of evolutionary data and population frequency data, both of which have already been made use of. We therefore opt for leaving this criterion out of our analysis.

- **Supporting Pathogenic criterion 5** ($PP5$) **Definition**: Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation.

- $PP5$ **Discussion**: In direct analogy to BP6, we could take variants with zero-star rating Pathogenic labels in ClinVar to satisfy this criterion. In practice this would not alter our results, since according to ACMG–AMP guidelines a single Supporting Pathogenic criterion is not sufficient to impact variant classification.

3.5 Identifying potentially incorrect labels

A benefit of our approach being unsupervised is that it provides a natural means of identifying potentially incorrect labels. While there was very good agreement between our models and the vast majority of clinical labels, there were a small number of variants for which there was strong disagreement. There are three possible reasons for these disagreements

1. The model failed in some way (*e.g.* the model was trained on a poor quality multiple sequence alignment or the training process itself had a problem.)

2. Evolutionary data does not reflect the clinical significance in these special cases.

3. The clinical label is incorrect.

In some cases, we can flag variants for which the third is the potential cause of disagreement by looking for consensus with other sources of evidence. As other sources of evidence, we used the ones described in the previous section. We identified 539 such variants (Supplementary Table 8).

3.6 Estimating the number of genes supervised models can be validated on

To estimate the number of genes on which a supervised method may be validated on a gene-by-gene basis, we repeatedly sample 10% of all ClinVar labels and count the number of genes for which there are sufficient labels for validation. Taking this to be 10 labels (of which at least one Benign and one Pathogenic), we find that the average number of genes is 52.

References

[48] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[49] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[50] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[51] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017.

[52] Dominik G. Grimm, Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G. MacArthur, Kaitlin E. Samocha, David N. Cooper, Peter D. Stenson, Mark J. Daly, Jordan W. Smoller, Laramie E. Duncan, and Karsten M. Borgwardt. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation*, 36(5):513–523, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22768.

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

[54] Sean R. Eddy. Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10):e1002195, October 2011.

[55] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, March 2015.

[56] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017.

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

[58] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, January 2018.

[59] Taylor L. Mighell, Sara Evans-Dutson, and Brian J. O'Roak. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *The American Journal of Human Genetics*, 102(5):943–955, May 2018.

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

[61] Genome Aggregation Database Consortium, Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, May 2020.

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

[63] Robert C. Green, Jonathan S. Berg, Wayne W. Grody, Sarah S. Kalia, Bruce R. Korf, Christa L. Martin, Amy L. McGuire, Robert L. Nussbaum, Julianne M. O'Daniel, Kelly E. Ormond, Heidi L. Rehm, Michael S. Watson, Marc S. Williams, and Leslie G. Biesecker. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, 15(7):565–574, July 2013.

[64] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[65] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014.

[66] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

[67] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[68] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space, 2016.

[69] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, Jan 2018.

[70] David J.C. MacKay. A practical bayesian framework for backprop networks, 1992.

[71] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1), January 2013.

[72] G. J. McLachlan and K. Basford. Mixture models : inference and applications to clustering. 1988.

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

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

[75] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(W1):W452–W457, July 2012.

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

[77] Bing-Jian Feng. Perch: a unified framework for disease gene prioritization. *Human mutation*, 38(3):243–251, 2017.

[78] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011.

[79] Daniele Raimondi, Ibrahim Tanyalcin, Julien Ferté, Andrea Gazzo, Gabriele Orlando, Tom Lenaerts, Marianne Rooman, and Wim Vranken. Deogen2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic acids research*, 45(W1):W201–W206, 2017.

[80] Hongjian Qi, Haicang Zhang, Yige Zhao, Chen Chen, John J Long, Wendy K Chung, Yongtao Guan, and Yufeng Shen. Mvp predicts the pathogenicity of missense variants by deep learning. *Nature Communications*, 12(1):1–9, 2021.

[81] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. 2012.

[82] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*, 48(2):214–220, 2016.

[83] Chengliang Dong, Peng Wei, Xueqiu Jian, Richard Gibbs, Eric Boerwinkle, Kai Wang, and Xiaoming Liu. Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Human molecular genetics*, 24(8):2125–2137, 2015.

[84] Biao Li, Vidhya G Krishnan, Matthew E Mort, Fuxiao Xin, Kishore K Kamati, David N Cooper, Sean D Mooney, and Predrag Radivojac. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21):2744–2750, 2009.

[85] Kaitlin E Samocha, Jack A Kosmicki, Konrad J Karczewski, Anne H O'Donnell-Luria, Emma Pierce-Hoffman, Daniel G MacArthur, Benjamin M Neale, and Mark J Daly. Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*, page 148353, 2017.

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

[87] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8):1161–1170, 2018.

[88] Nawar Malhis, Matthew Jacobson, Steven JM Jones, and Jörg Gsponer. List-s2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic acids research*, 48(W1):W154–W161, 2020.

[89] Hashem A Shihab, Julian Gough, David N Cooper, Peter D Stenson, Gary LA Barker, Keith J Edwards, Ian NM Day, and Tom R Gaunt. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Human mutation*, 34(1):57–65, 2013.

[90] Sung Chun and Justin C Fay. Identification of deleterious mutations within three human genomes. *Genome research*, 19(9):1553–1561, 2009.

[91] Daniel Quang, Yifei Chen, and Xiaohui Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 2015.

[92] Jana Marie Schwarz, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow. Mutationtaster evaluates disease-causing potential of sequence alterations. *Nature methods*, 7(8):575–576, 2010.

[93] Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu. dbnsfp v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. *Genome medicine*, 12(1):1–8, 2020.

[94] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109(42):16858–16863, 2012.

[95] Pradeep Bandaru, Neel H Shah, Moitrayee Bhattacharyya, John P Barton, Yasushi Kondo, Joshua C Cofsky, Christine L Gee, Arup K Chakraborty, Tanja Kortemme, Rama Ranganathan, et al. Deconstruction of the ras switching cycle through saturation mutagenesis. *Elife*, 6:e27810, 2017.

[96] Gregory M. Findlay, Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562(7726):217–222, October 2018.

[97] Andrew O. Giacomelli, Xiaoping Yang, Robert E. Lintner, James M. McFarland, Marc Duby, Jaegil Kim, Thomas P. Howard, David Y. Takeda, Seav Huong Ly, Eejung Kim, Hugh S. Gannon, Brian Hurhula, Ted Sharpe, Amy Goodale, Briana Fritchman, Scott Steelman, Francisca Vazquez, Aviad Tsherniak, Andrew J. Aguirre, John G. Doench, Federica Piccioni, Charles W. M. Roberts, Matthew Meyerson, Gad Getz, Cory M. Johannessen, David E. Root, and William C. Hahn. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nature Genetics*, 50(10):1381–1387, October 2018.

[98] Eric M Jones, Nathan B Lubock, AJ Venkatakrishnan, Jeffrey Wang, Alex M Tseng, Joseph M Paggi, Naomi R Latorraca, Daniel Cancilla, Megan Satyadi, Jessica E Davis, et al. Structural and functional characterization of g protein–coupled receptors with deep mutational scanning. *Elife*, 9:e54895, 2020.

[99] Mireia Seuma, Andre Faure, Marta Badia, Ben Lehner, and Benedetta Bolognesi. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. preprint, Genomics, September 2020.

[100] Wesley D Penn, Andrew G McKee, Charles P Kuntz, Hope Woods, Veronica Nash, Timothy C Gruenhagen, Francis J Roushar, Mahesh Chandak, Chris Hemmerich, Douglas B Rusch, et al. Probing biophysical sequence constraints within the transmembrane domains of rhodopsin by deep mutational scanning. *Science advances*, 6(10):eaay7505, 2020.

[101] Melissa A Chiasson, Nathan J Rollins, Jason J Stephany, Katherine A Sitko, Kenneth A Matreyek, Marta Verby, Song Sun, Frederick P Roth, Daniel DeSloover, Debora S Marks,

Allan E Rettie, and Douglas M Fowler. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife*, 9:e58026, September 2020.

[102] Xiaoyan Jia, Bala Bharathi Burugula, Victor Chen, Rosemary M. Lemons, Sajini Jayakody, Mariam Maksutova, and Jacob O. Kitzman. Massively parallel functional testing of msh2 missense variants conferring lynch syndrome risk. *The American Journal of Human Genetics*, 108(1):163 – 175, 2021.

[103] Eran Kotler, Odem Shani, Guy Goldfeld, Maya Lotan-Pompan, Ohad Tarcic, Anat Gershoni, Thomas A. Hopf, Debora S. Marks, Moshe Oren, and Eran Segal. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Molecular Cell*, 71(1):178–190.e8, July 2018.

[104] Kenneth A. Matreyek, Lea M. Starita, Jason J. Stephany, Beth Martin, Melissa A. Chiasson, Vanessa E. Gray, Martin Kircher, Arineh Khechaduri, Jennifer N. Dines, Ronald J. Hause, Smita Bhatia, William E. Evans, Mary V. Relling, Wenjian Yang, Jay Shendure, and Douglas M. Fowler. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, 50(6):874–882, June 2018.

[105] Lea M Starita, David L Young, Muhtadi Islam, Jacob O Kitzman, Justin Gullingsrud, Ronald J Hause, Douglas M Fowler, Jeffrey D Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of brca1 ring domain variants. *Genetics*, 200(2):413–422, 2015.

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

[107] Robert W. Newberry, Jaime T. Leong, Eric D. Chow, Martin Kampmann, and William F. DeGrado. Deep mutational scanning reveals the structural basis for -synuclein activity. *Nature Chemical Biology*, March 2020.

[108] Melissa A Chiasson, Nathan J Rollins, Jason J Stephany, Katherine A Sitko, Kenneth A Matreyek, Marta Verby, Song Sun, Fritz Roth, Daniel DeSloover, Debora S Marks, et al. Multiplexed measurement of variant abundance and activity reveals vkor topology, active site and human variant impact. *BioRxiv*, 2020.

[109] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. 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. *Genetics in medicine*, 17(5):405–423, 2015.

[110] The Invitae Clinical Genomics Group, Keith Nykamp, Michael Anderson, Martin Powers, John Garcia, Blanca Herrera, Yuan-Yuan Ho, Yuya Kobayashi, Nila Patil, Janita Thusberg, Marjorie Westbrook, and Scott Topper. Sherloc: a comprehensive refinement of the

ACMG–AMP variant classification criteria. *Genetics in Medicine*, 19(10):1105–1117, October 2017.