

# Improved pathogenicity prediction for rare human missense variants

Yingzhou Wu,<sup>1,2,3,4</sup> Roujia Li,<sup>1,2,3,4</sup> Song Sun,<sup>1,2,3,4</sup> Jochen Weile,<sup>1,2,3,4</sup> and Frederick P. Roth<sup>1,2,3,4,5,6,\*</sup>

The success of personalized genomic medicine depends on our ability to assess the pathogenicity of rare human variants, including the important class of missense variation. There are many challenges in training accurate computational systems, e.g., in finding the balance between quantity, quality, and bias in the variant sets used as training examples and avoiding predictive features that can accentuate the effects of bias. Here, we describe VARITY, which judiciously exploits a larger reservoir of training examples with uncertain accuracy and representativity. To limit circularity and bias, VARITY excludes features informed by variant annotation and protein identity. To provide a rationale for each prediction, we quantified the contribution of features and feature combinations to the pathogenicity inference of each variant. VARITY outperformed all previous computational methods evaluated, identifying at least 10% more pathogenic variants at thresholds achieving high (90% precision) stringency.

## Introduction

Fully realizing the clinical potential of human genome sequencing will require the ability to accurately determine the pathogenicity of individual sequence variants. Although genome-wide association (GWA) studies have limited power to associate rare missense variants with disease,<sup>1,2</sup> rare missense changes cannot be ignored: over 99% of the set of observed missense variants are rare (with a global minor allele frequency [MAF] below 0.5%), and 90% are extremely rare (with  $MAF < 10^{-6}$ ). Moreover, if genome sequences were known for all humans alive today, each single-nucleotide variant that is compatible with life would appear an average of  $\sim 50$  times.<sup>3</sup> Because identifying the subset of rare missense variants that are damaging represents a major unmet challenge for personal genome interpretation, we focus here on rare and extremely rare missense changes.

It has become possible to experimentally assess the functional impact of nearly all possible missense variants for a target protein, yielding “deep mutational scans” or “variant effect maps.”<sup>3–13</sup> However, the MaveDB<sup>14</sup> resource currently contains variant effect maps for fewer than 1% of the  $\sim 4,000$  human disease-associated proteins, and a high-quality fully-comprehensive experimental atlas of functional missense variation could be decades away.

By contrast, computational methods to infer variant pathogenicity can already be applied at genome scale. However, performance of current methods<sup>15–42</sup> has not fully addressed the challenge of variant interpretation. For example, using thresholds such that 90% of pathogenicity predictions are correct, such methods identified only 10%–20% of pathogenic variants.<sup>12</sup> Current guidelines for clinical variant interpretation recommend that

all computational methods be (at best) treated as “weak evidence.”<sup>43</sup>

To improve computational predictors, we would like to increase the size of the training dataset while also ensuring that the quality of training data is high, i.e., the training examples are accurately labeled and representative of the variants for which pathogenicity inference is most needed (e.g., rare variants). However, there is a tension between these competing objectives. For example, although common variants can be more confidently annotated as truly benign and are therefore likely to offer higher labeling accuracy, there is evidence that common benign variants are not representative of rare benign variants.<sup>16</sup> However, excluding common variants would substantially reduce the number of clinically annotated benign variants available for training computational models (Figure S1). While ClinVar<sup>44</sup> requires clinical interpretations to be based on a rigorous and attributable process, we can be less sure of accuracy for other resources offering pathogenicity annotations. For example, HumsaVar<sup>45</sup> offers annotations solely from literature reports and HGMD<sup>46</sup> seeks to maximize recall while accepting the risk of including false positives. The variants reported in gnomAD<sup>47</sup> have been used as “putatively-benign” training examples by many computational methods,<sup>16,17</sup> despite the potential for contamination with pathogenic variants, especially in genes associated with late-onset diseases. MaveDB<sup>14</sup> provides experimental “variant effect maps” in which the functional impact of many variants has been measured in parallel, but the set of scored variants may not be representative of the spectrum of patient variation.

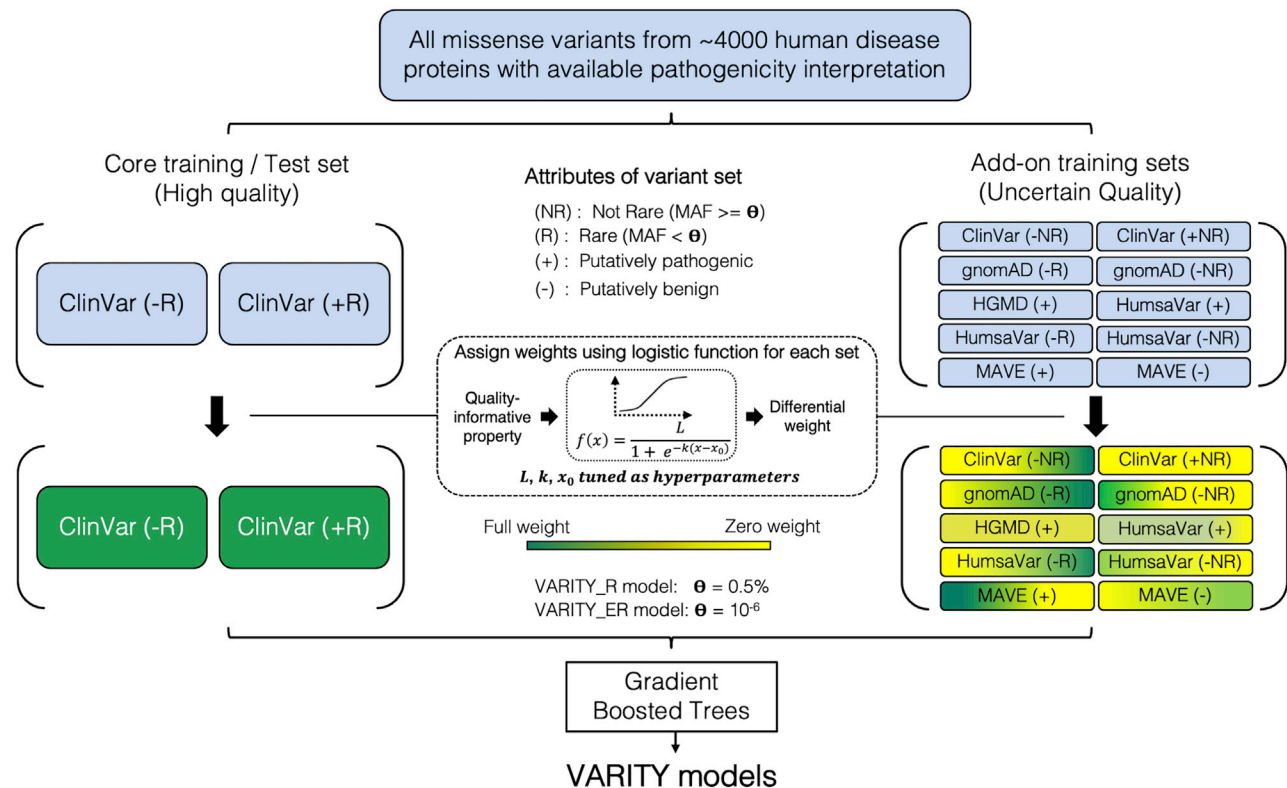
Here, we describe the VARITY approach for pathogenicity prediction, which has been specifically optimized for rare missense variation. While VARITY uses a meta-prediction

<sup>1</sup>The Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada; <sup>2</sup>Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada; <sup>3</sup>Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada; <sup>4</sup>Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, ON M5G 1X5, Canada; <sup>5</sup>Center for Cancer Systems Biology, Dana Farber Cancer Institute, Boston, MA 02215, USA; <sup>6</sup>Canadian Institute for Advanced Research, Toronto, ON M5G 1Z8, Canada

\*Correspondence: [fritz.roth@utoronto.ca](mailto:fritz.roth@utoronto.ca)  
<https://doi.org/10.1016/j.ajhg.2021.08.012>

© 2021 The Authors. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).





**Figure 1. The VARIETY framework**

VARIETY models were trained with an extension of the gradient boosted trees algorithm that allows different weights to be placed on different training examples. The VARIETY training data consist of a high-quality core set of variants (also used as test set) and a handful of diverse add-on sets with potentially lower predictive utility. For each training set (core or add-on set), one or more quality-informative properties were each used as an input to a logistic function that calculates a weight for each variant in the set. The parameters of each logistic function were treated as hyperparameters and optimized for performance on the core set of variants via 10-fold cross-validation. We generated two VARIETY models: VARIETY\_R, which included only rare (MAF < 0.5%) ClinVar<sup>44</sup> variants in its core set, and VARIETY\_ER, which only included extremely rare (MAF < 10<sup>-6</sup>) ClinVar<sup>44</sup> variants in its core set.

strategy, it limits the circularity that can arise in such approaches by excluding any feature that was informed by variant pathogenicity annotation. We also exclude features that may serve as proxies for protein identity because these may lead to predictions that are biased by the fraction of each protein's variants in the training set that are annotated pathogenic (which may be inaccurate in future application settings). VARIETY judiciously harnesses a larger set of training examples with uncertain accuracy and representativity and uses differential weighting strategies to ensure that training set expansion improves performance on a high-quality test set (Figure 1). For the task of identifying rare pathogenic variants, we find that VARIETY outperforms all other computational approaches examined.

## Material and methods

### Assembling a resource of labeled variants and their features for human proteins

To assemble a resource of labeled variants and associated features for human proteins, we first retrieved from dbNSFP (V4.0b2)<sup>48</sup> all missense variants for ~18,000 human proteins that had a

“reviewed” status in UniProt<sup>45</sup> [download date: 2019-04-08]). We next assembled variant features (summarized in Table S1) within four main categories. First, from dbNSFP<sup>48</sup> we collected scores for existing predictors based on only conservation information learned via unsupervised learning methods, e.g., the SIFT<sup>25,26</sup> method. To avoid circularity in subsequent performance testing, we excluded as features those predictors that had made direct or indirect use of annotated pathogenic and benign variants in training,<sup>16–23</sup> e.g., the PolyPhen2<sup>19,20</sup> method. Where multiple scores for a missense variant were available from a given predictor due to different coding isoforms, the score from the canonical isoform (defined in UniProt<sup>45</sup>) was chosen as the final score. Where multiple scores were available for different nucleotide-level changes yielding the same missense variant, the most deleterious score was chosen as the final score. Second, we defined a set of “delta values” by the difference in various quantitative physicochemical properties (e.g., molecular weight, polarity, or charge) between the reference and substituted amino acid. Third, we assembled features related to protein-protein interaction. For example, “maximum buried area”—defined by the largest solvent-accessible surface area around a given amino acid that is buried by any protein interaction partner—or the “maximum solvation energy change”—defined for a given amino acid by the greatest change in solvation energy provided by any interaction partner (both were estimated by PDBEPIA<sup>49</sup>). Fourth, we assembled

structure-related features. These included a set of binary variables indicating secondary structure (obtained from PSIPRED<sup>50</sup>) and accessible surface area (estimated by PDBePISA<sup>49</sup>). In addition, we used substitution matrix (BLOSUM100<sup>15</sup>) scores and whether the amino acid position is within or outside of a Pfam<sup>51</sup> (download date: 2019-04-08) domain. All 43 features and their sources can be found in Table S1. The resource of features for all missense variants from ~18,000 human proteins can be downloaded via the VARIETY web portal (see web resources).

For the purpose of training, we first identified a set of ~4,000 human disease-related proteins from all ~18,000 human proteins on the basis of whether any variant (not restricted to missense variants) had been annotated as “pathogenic” or “likely pathogenic” in the ClinVar database<sup>44</sup> or “deleterious” in the HumsaVar database<sup>45</sup> (download date: 2019-04-08). We further included a list of human proteins from multiplexed assays of variant effect (MAVE) studies<sup>6–10</sup> (see Table S2). Only variants from the set of human disease-related proteins were labeled and considered in model training. We assigned every missense variant from this set a label of either positive (putatively pathogenic), negative (putatively benign), or unknown. At this stage, our confidence in these labels might be quite low, but we will later depend on the machine-learning strategy to weight these variants. The labeling criteria depended on the source of information about the variant. For the ClinVar<sup>44</sup> database, regardless of the number of “review stars,” variants with “pathogenic,” “likely pathogenic,” or “pathogenic/likely pathogenic” annotation were labeled as positive. Similarly, variants with “benign,” “likely benign,” or “benign/likely benign” were labeled as negative for training purposes. All other variants were labeled as unknown. For the HumsaVar<sup>45</sup> database, variants with the “polymorphism” annotation were labeled as negative, variants with “deleterious” annotation were labeled as positive, and all other variants were labeled as unknown. For gnomAD,<sup>47</sup> variants for which at least one homozygous individual had been observed were labeled as negative, and other variants were labeled as unknown. For variants from MAVE studies<sup>6–10</sup> (encompassing 12 proteins, see Table S2), we used functional impact scores rescaled by a previously developed pipeline,<sup>52</sup> such that a score of 0 indicates fitness approximating that of a nonsense variant and a score of 1 indicates fitness approximating that of a synonymous variant. Here, variants with a score above 0.5 were labeled as negative, while those below 0.5 were labeled as Positive. For the HGMD<sup>46</sup> database, all variants were labeled as positive. To handle the minority of cases (~10% of all variants) where a variant appears in multiple databases, we assigned a primary source for that variant by using a somewhat arbitrary ranking of the sources from which to derive the variant label: ClinVar,<sup>44</sup> HGMD,<sup>46</sup> and HumsaVar<sup>45</sup> pathogenicity annotations, fitness scores from variant effect maps, and presence of homozygotes in gnomAD.<sup>47</sup>

### Core and add-on training sets

All missense variants labeled as positive or negative were considered as potential training examples. To train and test VARIETY models, we assembled a core training set of variants that were of high quality (i.e., accuracy and representativity). We also used various “add-on” training sets that potentially have less predictive utility (e.g., because they are less accurately labeled or less representative of the core set), but these sets were subjected to further weighting. To obtain high-accuracy core sets, we used ClinVar<sup>44</sup> missense variants. To ensure high representativity, we first separated ClinVar<sup>44</sup> missense variants into three subsets: ClinVar\_C,

ClinVar\_MR, and ClinVar\_ER with labeled missense variants that were common (MAF > 0.5%), moderately rare ( $10^{-6} < \text{MAF} < 0.5\%$ ), and extremely rare (MAF <  $10^{-6}$ ), respectively. Given evidence that common variants are not representative of rare variants,<sup>16</sup> we excluded ClinVar\_C from the core training set of the VARIETY model optimized for rare missense variants (VARIETY\_R; see Figure 1) and treated it as an add-on set. The properties of moderately rare and common variation may differ from those of extremely rare variants. Therefore, to optimize a VARIETY model (VARIETY\_ER; see Figure 1) for extremely rare missense variants, we used ClinVar\_ER as the core set and used ClinVar\_C and ClinVar\_MR as add-on sets.

All labeled missense variants outside of the core set were considered as add-on training examples and were separated into different add-on sets so that each set can be employed separately with distinct effects on training VARIETY models. Add-on variants were first grouped into add-on training sets on the basis of source of annotation (the source where the variant label was determined) and label (positive or negative). Each negative-labeled add-on set (e.g., negative-labeled variants from gnomAD<sup>47</sup>) was further split on the basis of rarity (MAF cutoff = 0.5%) except the negative set from MAVE studies in which most scored variants are rare. A full list of core set and add-on sets for both VARIETY\_R and VARIETY\_ER models can be found in Table S3.

### Performance measures

To evaluate performance of the VARIETY models, we used the standard area under the receiver operating characteristic curve (AUROC). However, we also wished to use measures based on precision and recall, which often correspond more closely to the user’s needs and intuition. Unfortunately, any measure using precision, such as the area under the precision recall curve (AUPRC), changes depending on the prior (the frequency of positively labeled examples). One strategy to compare AUPRC measures applied to different test sets with varying priors is to force each test set to be balanced, e.g., by down-sampling the most abundant label class. However, this approach destroys information by eliminating test data. Therefore, we developed two measures: area under the balanced precision recall curve (AUBPRC) and recall at 90% balanced precision (RB90P). For these measures, we simply calculate the precision for each test set and then derive the “balanced precision,” i.e., the precision that would have been expected had the prior been balanced (equal to 50%):

$$\text{balanced precision} = \frac{\text{precision} * (1 - \text{prior})}{\text{precision} * (1 - \text{prior}) + (1 - \text{precision}) * \text{prior}} \quad (\text{Equation 1})$$

Proof:

Let  $Y$  be a random variable representing the true label of an event (e.g.,  $Y = 1$  if a variant is pathogenic) and  $Y'$  represent the label predicted by a model (e.g.,  $Y' = 1$  if the model predicts a variant to be pathogenic). Precision and prior of the model can then be written as

$$\text{precision} = P(Y = 1|Y' = 1), \text{ prior} = P(Y = 1) \quad (\text{Equation 2})$$

The following equation holds according to Bayes Rule:

$$\frac{P(Y = 1|Y' = 1)}{P(Y = 0|Y' = 1)} = \frac{P(Y' = 1|Y = 1)}{P(Y' = 1|Y = 0)} * \frac{P(Y = 1)}{P(Y = 0)} \quad (\text{Equation 3})$$

Using Equation 2 and Equation 3, we can rewrite as:

$$\frac{\textit{precision}}{1 - \textit{precision}} = L * \frac{\textit{prior}}{1 - \textit{prior}}, \text{ where } L = \frac{P(Y' = 1|Y = 1)}{P(Y' = 1|Y = 0)}$$

(Equation 4)

and then from Equation 4, we obtain the likelihood ratio

$$L = \frac{(1 - \textit{prior}) * \textit{precision}}{\textit{prior} * (1 - \textit{precision})}$$

(Equation 5)

Because we know the likelihood ratio is not affected by the prior (ratio of conditional probability conditioned on prior), therefore we can write down the following, given a balanced prior of 0.5:

$$\frac{\textit{balanced\_precision}}{1 - \textit{balanced\_precision}} = L * \frac{0.5}{1 - 0.5}$$

(Equation 6)

By combining Equation 5 and Equation 6, we can solve for *balanced\_precision* and obtain Equation 1.

A balanced precision versus recall curve (BPRC) can be drawn by calculation of the balanced precision value corresponding to every precision value, and AUBPRC can be calculated just as AUPRC would be. We note that there are different ways to calculate an AUPRC (or AUBPRC). Here, we calculated AUPRC (and AUBPRC) as the weighted mean of precision (or balanced precision) at different recall thresholds and used the increase in recall from the previous threshold as the weight, which is also called average precision. It can also be shown that AUBPRC can be calculated more directly from AUPRC via

$$\textit{AUBPRC} = \frac{\textit{AUPRC} * (1 - \textit{Prior})}{\textit{AUPRC} * (1 - \textit{Prior}) + (1 - \textit{AUPRC}) * \textit{Prior}}$$

(Equation 7)

To derive RB90P, we need to first plot the BPRC and then locate (or interpolate) the recall value at which balanced precision is 90%.

### The learning algorithm and weighting of training variants

For VARITY we adopted a variant of the gradient boosted tree (GBT) machine-learning algorithm XGBoost<sup>53</sup> (Python Package V0.90). Briefly, the first stage of the GBT method is to generate an initial prediction of the target variable for each training example. In successive stages, GBT iteratively generates a series of decision trees. After each tree, the current prediction for any given training example is the initial prediction added to the sum of the outputs for all trees generated thus far. Each tree in the series is optimized to yield incremental outputs that, when added to the sum of previous outputs, tend to minimize the loss function (prediction error). The incremental value is determined by approximating both the first and second order gradient of the loss function at the current prediction stage. Optimization of each tree is accomplished by successively selecting features that split training examples into smaller groups (or nodes). The splitting feature is selected such that a single incremental output value assigned to each subgroup can most closely approximate the set of best incremental output values. For each training set (core or add-on set), each of several quality-informative properties was used as input to a logistic function that assigned a weight to each training variant within the set. The parameters of each logistic function were treated as hyperparameters that were optimized for performance (see details in the [Bayesian optimization of hyperparameters](#) section below). Subsequently, the weights for all positive training examples were scaled by a factor such that the total of

re-scaled weights for positive training examples was equal to the total weight of negative examples.

### Moving window analysis to identify quality-informative variant properties

To identify the quality-informative properties used as described above for variant weighting, candidate properties were evaluated by moving window analysis. Candidate informative properties included: allele frequency (for all training sets except the ones from MAVE studies<sup>6–10</sup>), “review stars” (for ClinVar<sup>44</sup> training sets), “number of homozygotes” (for gnomAD<sup>47</sup> training sets), “label confidence,” and “mutational accessibility” (the latter two are candidates for MAVE training sets<sup>6–10</sup>). The “label confidence” score for each variant was defined to be either the MAVE score itself (for scores < 0.5) or 1 – score (for scores > 0.5). Use of the mutational accessibility property was motivated by the fact that, although variants with more than one nucleotide change in a given codon can be generated by mutagenesis strategies used in variant effect mapping, these rarely occur in humans. The mutational accessibility of each possible amino acid change was the total probability of all possible single-nucleotide change events yielding that amino acid change, where the probability that a single nucleotide change event occurs in each codon was defined via the human codon usage frequency for that codon divided by nine (because nine single-nucleotide edits are possible for any given codon). Amino acid changes requiring more than one single-nucleotide change within the same codon received a mutational accessibility score of 0.

For each proposed candidate property, we carried out moving window analysis to evaluate whether model performance depends on the value of the property. To this end, add-on set examples were first ordered by the property. We then defined  $N_w$  moving windows, each containing a fraction  $F_E$  of the total  $X$  add-on training examples and incrementing the position of successive windows every  $I = \lceil X * (1 - F_E) / (N_w - 1) \rceil$  training examples. Where the number of examples is not evenly divisible by  $I$ , the final window may contain slightly fewer than  $X * F_E$  examples. Here, we used  $N_w = 100$  and  $F_E = 0.5$ .

To evaluate the data quality of each window, we estimated model performance (AUBPRC) on the core set by using 10-fold cross-validation via XGBoost<sup>53</sup> (with default algorithm level hyperparameters) where the training examples in each fold were supplemented by add-on examples in that moving window. To assess whether a property was informative, we calculated correlation between moving window performance and window number for each candidate informative property and also for each of ten random orderings of add-on examples. From this, we calculated a  $Z$  score for the observed correlation relative to the distribution of randomly ordered correlations and considered a property to be informative if it exhibited a  $Z$  score above 1 or less than  $-1$ . The negative and positive sign of a  $Z$  score indicates whether there was a negative or positive correlation between performance of moving windows and the property. Although moving window analysis combined one or two add-on sets at a time with the core set, we note that add-on sets may provide a contribution that is greater or lower than that observed here when they are used in conjunction with many other training sets during hyperparameter tuning. Therefore, all training sets and their associated informative properties (Table S3) were included for the weighting process that was optimized during hyperparameter tuning, even for  $Z$  scores exhibiting modest departures from zero.

## Bayesian optimization of hyperparameters

Two types of hyperparameters were employed for VARIETY models. First, there were algorithm-level hyperparameters such as number of trees, tree depth, and learning rate (see Table S4 for a full list). Second, there were hyperparameters in the logistic functions that were used for assigning weights to different training sets (see Table S3 for full list). There are three parameters in each logistic function, namely the maximum weight  $L$ , the midpoint  $x_0$ , and the growth rate  $k$ . Each logistic function takes one quality informative property as input. For the training set that has more than one associated quality-informative property, the weights were assigned as the product of multiple logistic functions.

All hyperparameters were optimized simultaneously by Bayesian optimization with HyperOpt<sup>54-56</sup> (Python package V0.2.2). This process involves random initialization of hyperparameter values (with uniform prior) followed by iteratively repeated trials (here we used 300 trials) with alternative hyperparameter values suggested on the basis of an expected improvement measure calculated via a probability distribution estimated from the performance of previously tried hyperparameter settings. For each trial, the hyperparameter set was evaluated via model performance on the core training set estimated in 10-fold cross-validation where the core training examples in each fold were supplemented with add-on training sets examples and all training examples were weighted with hyperparameter values suggested on the current trial. The metric used here for model performance (the objective function for hyperparameter optimization) was the mean AUBPRC on validation sets (average over ten validation sets).

To limit model overfitting that can arise from repeated interrogation of the evaluation set with different hyperparameter settings, we used the following procedure to automatically determine the final hyperparameter values: (1) re-order all trials from HyperOpt<sup>54-56</sup> by mean AUBPRC on training sets (averaged over ten training sets) from low to high, (2) calculate a moving window (we used window size 30 for total 300 trials) average of mean AUBPRC on validation sets, (3) define an “early stopping” point at the first moving window (the “fittest” region) for which mean AUBPRC on validation sets begins to descend, and (4) select as final the hyperparameters from the trial within this “fittest” region that achieved the highest mean AUBPRC on validation sets.

## Performance evaluation with independent test sets and nested cross-validation

To obtain an independent test dataset based on *de novo* variation in neurodevelopmental disorders, we downloaded all *de novo* variants from denovo-db 1.61<sup>57</sup> (download date 2020-10-20) that corresponded to neurodevelopmental disorder case/control studies. We extracted the subset of rare missense variants (MAF < 0.5%) from 253 candidate neurodevelopmental disease genes for which variation was enriched in neurodevelopmental disease patients.<sup>58</sup> For the purpose of performance evaluation, the variants that appeared in “control” and “case” populations were labeled as negative (putatively benign) and positive (putatively pathogenic), respectively. Because a subset of variants annotated in HGMD<sup>46</sup> had been used to train computational methods against which we wished to compare our results, variants annotated in HGMD<sup>46</sup> (version 2020) were removed, as were variants used in VARIETY\_R and VARIETY\_ER model training. There are in total 367 *de novo* missense variants (321 putatively pathogenic and 46 putatively benign) left for comparison. Considering only variants

scored by each of the 23 computational methods to be compared, the final validation dataset contained 215 *de novo* missense variants (188 putatively pathogenic and 27 putatively benign). For the validation analysis with published variant effect maps, we assembled map scores for a subset of six proteins associated with human genetic disease: CALM1 (MIM: 114180), TPK1 (MIM: 606370), CBS (MIM: 613381), PTEN (MIM: 601728), BRCA1 (MIM: 113705), and VKORC1 (MIM: 608547).

To evaluate the generalization performance of VARIETY models more broadly, we adopted 10-fold nested cross-validation strategy. For each outer loop, the hyperparameters were determined via Bayesian optimization<sup>54-56</sup> based on inner-loop 10-fold cross-validation. The optimized hyperparameters were then used for training based on all data in the outer-loop training set. Performance on VARIETY models (and other computational approaches) were assessed on the ten held-out outer-loop validation sets (Figure S2).

To compare the performance of VARIETY models to previously developed pathogenicity predictors, we collected the scores from 25 predictors. For 22 of these, scores were collected from dbNSFP (V4.0b2)<sup>48</sup>, while scores for two predictors (MPC<sup>41</sup> and EVMutation<sup>27</sup>) were assembled from links provided by the corresponding papers and scores for DeepSequence<sup>42</sup> were collected from an independent benchmarking study of 11 proteins:<sup>54</sup> UBE2I (MIM: 601661), TPMT (MIM: 187680), TPK1 (MIM: 606370), TP53 (MIM: 191170), SUMO1 (MIM: 601912), PTEN (MIM: 601728), MAPK1 (MIM: 176948), HRAS (MIM: 190020), CALM1 (MIM: 114180), BRCA1 (MIM: 113705), and ADRB2 (MIM: 109690).

## Feature contribution to output score and model performance

For each variant, the VARIETY model provides a log(odds) (“lod”) score (before subsequent transformation to the final score via a sigmoid transformation), which is more positive for variants that are inferred to be pathogenic and more negative for variants that are inferred to be benign. To provide intuition about which features of a given variant drove its score, we defined an “output contribution” score by using Shapley additive explanation values<sup>59,60</sup> retrieved from the XGBoost<sup>53</sup> output (with option “predict\_pred\_interactions” turned on).

It can be useful to estimate the contribution of individual feature groups or pairs of feature groups to a model’s performance (as opposed to a model’s output values). We therefore defined a “performance contribution” score for all features for each variant used in training. For a positive (putatively pathogenic) variant used in training, performance contribution was the output contribution multiplied by the weight given to the variant during hyperparameter tuning. For a negative (putatively benign) variant used in training, each feature performance contribution is obtained by negating the output contribution and again multiplying by the weight used for that variant. Thus, output contributions moving in the right direction are counted as beneficial to model performance.

## Results

### Developing a machine-learning and data-weighting strategy

We first assembled a large resource of input features that are potentially informative about variant function for all

possible missense variants in ~18,000 human proteins. To limit the circularity that may inflate the model performance and to limit the effects of training set bias<sup>61</sup>, we excluded features informed by variant annotation (e.g., scores predicted by supervised models) or protein identity (Table S1). For model training and performance evaluation, we next identified from ~4,000 human disease proteins a “core” set of variants with high quality annotation from ClinVar<sup>44</sup>. This set was limited to rare variants (MAF < 0.5%) for the VARIETY\_R model. To assess sensitivity of our results to this threshold, we also defined a core set of extremely rare variants (MAF < 10<sup>-6</sup>) to train and test a VARIETY\_ER model. For the core set, we labeled variants as “positive” (putatively pathogenic) or “negative” (putatively benign) on the basis of pathogenicity interpretations. We then collected various “add-on” variants, employing potentially less-reliable information about whether non-core variants should be labeled as positive or negative (see [material and methods](#)). Add-on variants were drawn from gnomAD,<sup>47</sup> HGMD,<sup>46</sup> HumsaVar,<sup>45</sup> ClinVar<sup>44</sup> (excluding the core set variants), and a handful of multiplexed assays of variant effect (MAVE) studies<sup>6–10</sup> (Table S2). Add-on variants were grouped into add-on training sets based on source (e.g., gnomAD<sup>47</sup>), label (e.g., “negative”), and rarity (e.g., “MAF > 0.5%”) so that each add-on set can be employed separately with distinct effects on training VARIETY models (see [material and methods](#); Table S3).

Although VARIETY employs a “workhorse” machine-learning method (the gradient boosted tree algorithm XGBoost<sup>53</sup>) (see [material and methods](#)), it is distinct from previous machine-learning strategies applied to variant impact prediction. Distinct *a priori* weights were assigned to different training examples, and these weights were considered in the loss function used for tree optimization. For each training set (core or add-on set), one or more quality-informative properties were used as inputs to a logistic function that assigned weights to each variant. The parameters of each logistic function were optimized as hyperparameters (see [material and methods](#); Figure 1).

### Quality-informative properties of each training set

To identify properties related to the utility of different add-on sets for inferring pathogenicity, we ordered the examples in an add-on set (or logically related pairs of add-on sets) by candidate property and carried out a moving window analysis (see [material and methods](#)). To evaluate the predictive utility of each window, we estimated the model performance by using 10-fold cross-validation on the core training set where the training examples in each fold were supplemented by all of the add-on examples in that moving window. When a candidate property correlated with moving window predictive utility in an add-on set, this property was identified as an informative property for the add-on set. For example, a moving window analysis of putatively benign variants from gnomAD<sup>47</sup> showed variants with both lower and higher allele frequency to be less

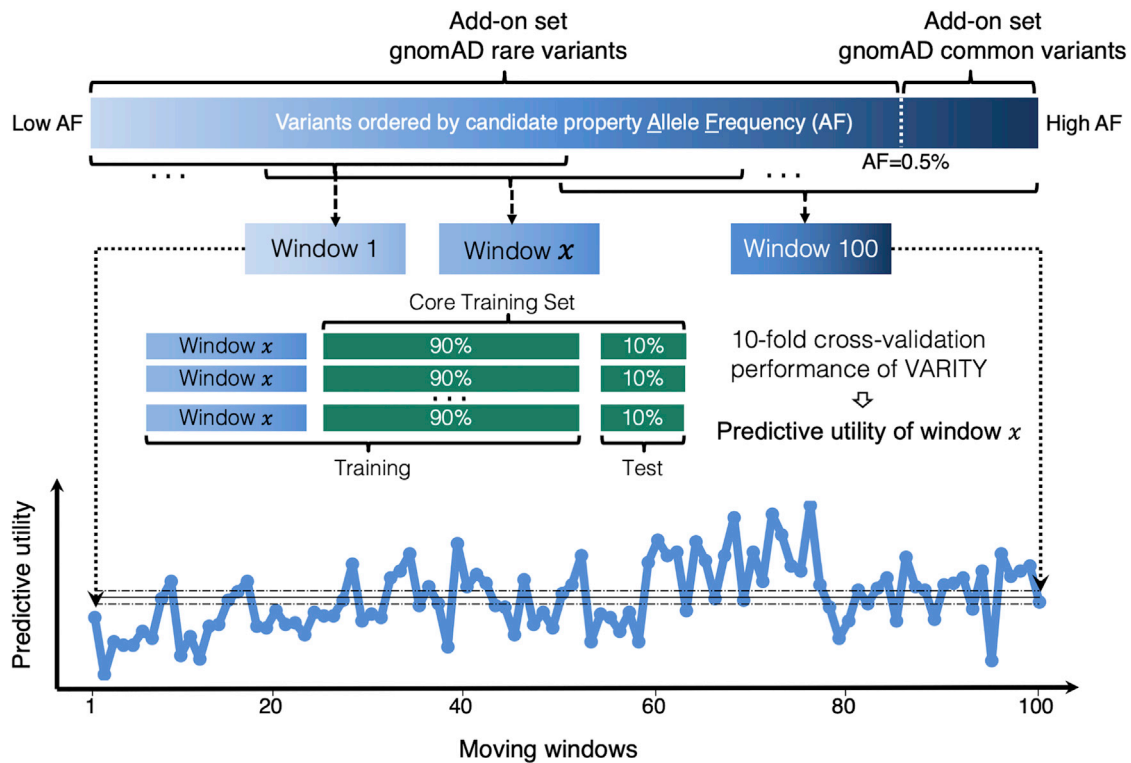
useful in training VARIETY\_R (Figure 2). This agreed with our expectation that variants with lower allele frequency are more likely to be “contaminated” with damaging variants and that common variants might not be representative of the rare variants used for testing. Allele frequency was similarly identified as a quality-informative property for negative-labeled ClinVar<sup>44</sup> and HumsaVar<sup>45</sup> variants. For several positive-labeled (putatively pathogenic) add-on sets, variants with higher allele frequency were found to be less useful. We also identified both label confidence and mutational accessibility properties (see [material and methods](#) for definitions) as quality-informative properties for variants with MAVE scores (Figure S3; Table S3). Although the core training set was considered high quality, we considered ClinVar<sup>44</sup> review stars as a quality-informative property to further weight these training variants.

### Bayesian optimization on hyperparameters

The parameters of the variant-weighting logistic function used for each training set were treated as hyperparameters subject to optimization, as were algorithm-level hyperparameters such as learning rate, number of trees, and the maximum depth of a single tree (Table S3 and Table S4). All hyperparameters were tuned simultaneously with Bayesian optimization,<sup>54–56</sup> which consists of many consecutive trials, each making use of a different set of hyperparameter values suggested by the previous trials. For each trial, the hyperparameter set was evaluated via VARIETY model performance on the core set examples in 10-fold cross-validation where the core training examples in each fold were supplemented with add-on training sets examples and all training examples were weighted with hyperparameter values suggested on the current trial. To avoid overfitting to the validation set during hyperparameter tuning, we selected a hyperparameter set by using a scheme akin to “early stopping” (see [material and methods](#); Figure S4). With thus-optimized hyperparameters, all 34,087 core set variants were given high weight ( $\geq 70\%$ ). While nearly all (123,591 of 123,621) add-on set variants were given non-zero weights for VARIETY\_R, only 1.8% of add-on variants received  $\geq 70\%$  weight (see Figure 3 for the optimized weights of gnomAD<sup>47</sup> add-on sets and Figure S5 for all training sets). The effective number of training examples (the sum of weights of all core and add-on set examples) for VARIETY\_R was 41,898 (24,842 positive examples and 17,056 negative examples) as compared with the 157,708 total of core and add-on examples. A subsequent reweighting step equalized the effective number of positive and negative examples.

### An alternative VARIETY\_ER model optimized for extremely rare variants

Although the threshold that we used to define rare variants (MAF < 0.5%) in the VARIETY\_R model is commonly used,<sup>1</sup> we considered the possibility that moderately rare variants (10<sup>-6</sup> < MAF < 0.5%) may not be representative of extremely rare variants (MAF < 10<sup>-6</sup>, for which no allele



**Figure 2. Identifying properties informative of predictive utility with moving window analysis**

Here, we illustrate evaluation of whether the allele frequency property can identify less useful subsets of two add-on variant sets (common and rare gnomAD<sup>47</sup> variants). Variants from both common and rare add-on gnomAD<sup>47</sup> sets were ordered from low to high allele frequency, and a series of moving windows (each capturing the same number of examples) was generated. To evaluate the predictive utility of each window, we estimated the model performance on the core set by using 10-fold cross validation where the training examples in each fold were supplemented by examples in that moving window. One solid and two dashed horizontal black lines indicate the mean  $\pm$  standard error of the predictive utility of all moving windows. Here, moving window analysis showed that allele frequency correlated with moving window predictive utility, and lower and higher allele frequency gnomAD<sup>47</sup> variants exhibited less utility as a source of negative variants for VARITY\_R performance.

count was detected in gnomAD<sup>47</sup>). We therefore developed an alternative VARITY\_ER model by using only extremely rare variants from ClinVar<sup>44</sup> as the core set (but including moderately rare ClinVar<sup>44</sup> variants as add-on variants). Like VARITY\_R, VARITY\_ER modeling proceeded via moving window analysis and hyperparameter tuning, and all 18,792 core set variants and 133,848 out of 138,916 add-on set variants received non-zero weights. The effective number of training examples was 24,074 for VARITY\_ER, well below the total of 157,708 training examples considered (Figure S6, Figure S7, and Figure S8).

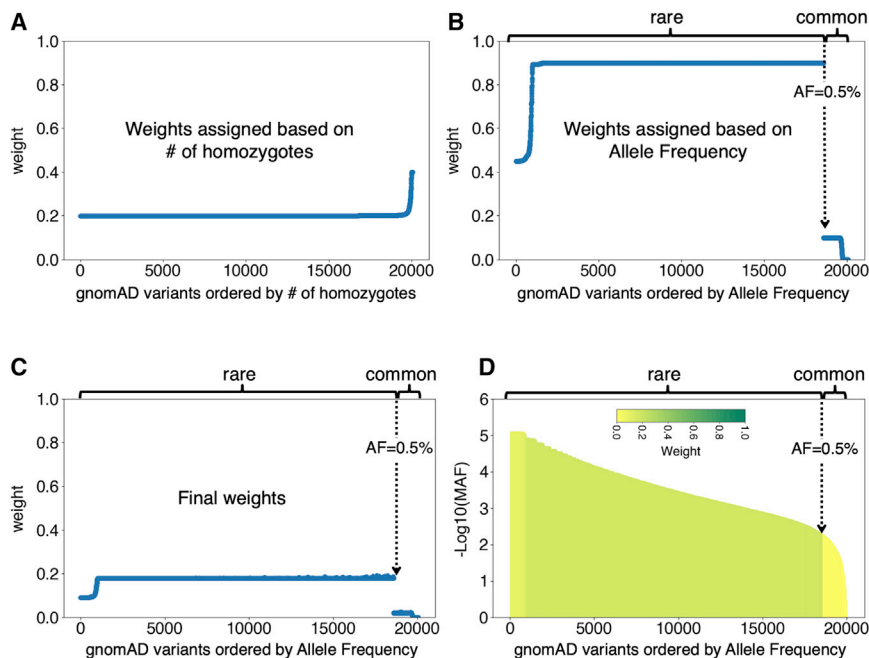
#### Leave-one-variant-out (LOO) prediction for ~18,000 human proteins

The final VARITY\_R and VARITY\_ER models were each trained with both weighted core and add-on sets with optimized hyperparameters and were each used for inference of functional impact for all possible missense variants for each of ~18,000 human proteins. We observed that cross-validation performance for VARITY\_R and VARITY\_ER models was only slightly higher for the training set than for the validation sets using the optimized hyperparameter settings (See Figure S4 and Figure S7), suggesting

that overfitting, if any, was modest. However, to avoid inflating performance estimates as a result of overfitting, we applied a leave-one-variant-out (LOO) strategy such that each variant used (i.e., having non-zero weight) in training was excluded in turn from the VARITY\_R (or VARITY\_ER) model used to score that variant, yielding a set of VARITY\_R\_LOO (or VARITY\_ER\_LOO) scores.

#### Extracting intuition from VARITY models

To better understand the rationale for each VARITY score, we used Shapley additive explanation values<sup>59,60</sup> to assess the contributions made by features and their pairwise combinations. For each variant, contributions were estimated both to VARITY score and, if the variant was used in training, to model performance (see material and methods). For each feature group (Table S1), we estimated the total contribution (including independent and pairwise contribution) to model performance as weighted average of the contribution from all training examples by using weight of each training example that was optimized during hyperparameter tuning (Figure 4). Contributions to model performance at the level of individual features and feature pairs were also evaluated (Figure S9).



**Figure 3. Assigning weights to variants in training add-on sets**

Here, we illustrate weight assignment to variants in common and rare gnomAD<sup>47</sup> add-sets of negative (putatively benign) training examples based on the allele frequency and number of homozygotes observed in gnomAD.

(A) Weights of all gnomAD<sup>47</sup> variants (the union of the rare and common gnomAD<sup>47</sup> add-on sets) assigned via logistic function that takes number of homozygotes as an input.

(B) Weights of variants from either rare or common gnomAD<sup>47</sup> add-on set assigned via logistic function that takes allele frequency as an input.

(C) The final weights of gnomAD<sup>47</sup> variants by taking the product of the weights assigned by (A) and (B) for each variant.

(D) Another view of the final weights of gnomAD<sup>47</sup> variants with  $-\log_{10}(\text{allele frequency})$  as the y axis and showing the final weights by color.

For the VARITY\_R model, we found that conservation-based features were the most important for model performance, which is perhaps unsurprising given that they are foundational to most computational predictors of variant function. Among these, Proven<sup>24</sup>, SIFT,<sup>25,26</sup> EVMutation,<sup>27</sup> and LRT<sup>29</sup> provided the greatest contribution. The next most important feature group was “delta AA properties,” a set of features capturing differences in physico-chemical properties between the missense and wild-type amino acid. Among these, the “delta cyclic property” (capturing whether a proline residue has been substituted for a non-proline residue or vice versa) was the most important. The “delta sulfur property” (capturing whether a cysteine or methionine residue has been substituted for a non-sulfur-containing residue or vice versa) and the “delta hydrophathy index” were also very helpful. Next were features related to “surface-accessible area,” representing the extent to which the amino acid is exposed to solvent, and “PPI,” a group of features modeling the extent to which the environment of an amino acid changes upon protein-protein interaction (see [material and methods](#)).

Features and feature groups will of course combine to contribute to VARITY models in a non-additive way. For example, the “IN/OUT Pfam domain” feature (capturing whether an amino acid position falls within a conserved protein domain), which is derived from multiple sequence alignments, contributed to the performance of the “conservation scores” feature group. It is therefore understandable that this feature group masks the predictive value of “IN/OUT Pfam domain” (as indicated by negative pairwise performance contribution values for VARITY\_R in [Figure 4](#)). A negative pairwise performance contribution for individual features was also observed among nine conservation scores. Because these conservation scores are similar, par-

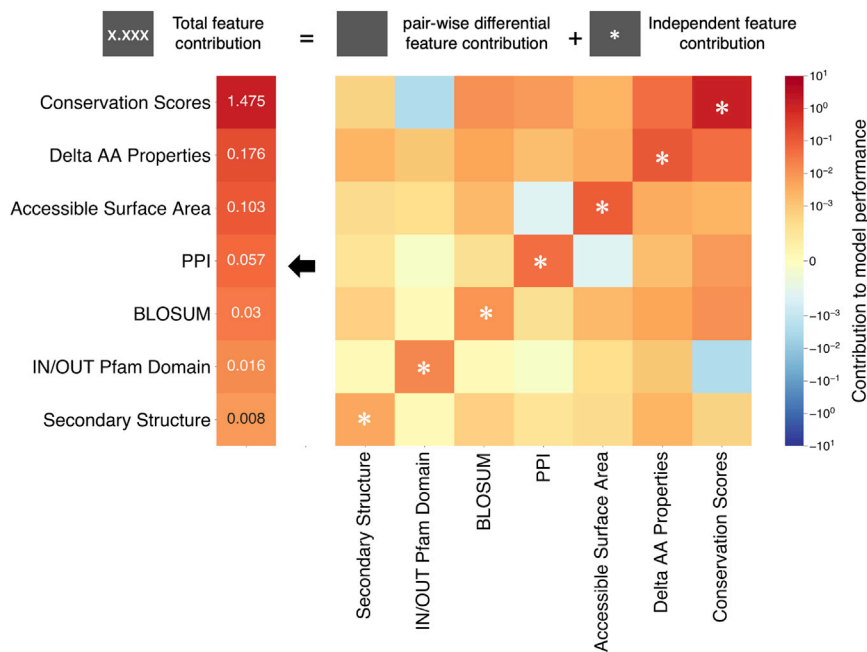
tial redundancy with one another was to be expected (see [Figure S9](#)). We also have observed that the combination of “conservation scores” with other feature groups such as “delta AA properties” and “PPI” provides a positive (synergistically favorable) pairwise feature contribution to model performance. Features contributing to each VARITY\_ER model prediction score were also identified, and the performance contributions of individual and grouped features was again analyzed with the weighted training examples, yielding similar results ([Figure S10](#) and [Figure S11](#)).

### Assessing VARITY with *de novo* variants in neurodevelopmental disorder studies

To assess VARITY, we examined missense variants from neurodevelopmental case/control studies, focusing on 253 genes enriched for *de novo* variation in neurodevelopmental disease patients.<sup>57,58</sup> For performance comparison, we considered all rare *de novo* variants in these genes that had not been used in training by VARITY and were not annotated by HGMD<sup>46</sup> (see [material and methods](#)). After retaining only variants that had MAF < 0.5% and that had been scored by all 23 of the methods to be compared, there were 188 variants in the case group and 27 in the control group.

We wished to judge performance using both receiver operating characteristics (ROC) and precision versus recall analysis. Precision versus recall analysis has the advantage of being more intuitively related to the needs of most users. However, precision varies as a function of the prior probability, and this prior is by default determined by the fraction of test examples with a Positive (putatively pathogenic) label. Moreover, this default prior may not have any relationship with the prior probability that is most relevant to a clinical geneticist using our output. To enable





**Figure 4. Feature group contribution to model performance for VARIETY\_R model**  
 The contribution of each feature group to VARIETY\_R model performance was averaged (weighted) across all training examples via the weight of each training example as optimized during hyperparameter tuning. The first column (left) indicates the total contribution to model performance of each feature group. For each feature group, the total contribution can be decomposed into the individual feature contribution (matrix cell with a star symbol on the corresponding row) and the differential contribution of that feature when it is combined with each other feature group (matrix cells without a star symbol on the corresponding row). Red and blue color indicates positive and negative contribution to model performance, respectively. A blue-colored cell for pairwise differential feature contribution indicates there is a certain amount of redundancy between two feature groups (e.g., between conservation scores and IN/OUT Pfam domain).

greater intuition and fairer comparisons, we therefore transformed each empirical precision-recall curve to the curve corresponding to a balanced (50% probability of pathogenicity) prior. Thus, our performance measures were as follows: area under the ROC curve (AUROC), area under the balanced precision recall curve (AUBPRC), and recall at 90% balanced precision (R90BP; see [material and methods](#) for details). For this *de novo* variant application, we used the case/control status of the patient(s) in which each variant appears as a proxy for pathogenicity.

VARIETY\_ER numerically outperformed all other methods in terms of AUBPRC. When performance was measured with the subset of variants that had been scored by all methods, AUBPRC performance improvement achieved statistical significance ( $p < 0.05$ ) relative to every method except MPC<sup>41</sup> ( $p = 0.125$ ) and M-CAP<sup>17</sup> ( $p = 0.099$ ). Moreover, when we used larger variant sets that had been scored by MPC<sup>41</sup> (285 variants in case group and 38 variants in control group) or M-CAP<sup>17</sup> (320 variants in case group and 45 variants in control group), VARIETY\_ER's performance improvement was statistically significant ( $p = 0.048$  and  $0.029$  for MPC<sup>41</sup> and M-CAP<sup>17</sup>, respectively). When the score threshold for every method was tuned to achieve 90% balanced precision, VARIETY\_ER recovered 10% more (presumed pathogenic) variants from the case group than the closest non-VARIETY method. VARIETY\_ER outperformed VARIETY\_R with a 2% greater AUBPRC ( $p = 0.122$ ), which may be explained by the fact that most *de novo* missense variants are extremely rare. However, VARIETY\_R also numerically outperformed all other methods (see [Figure 5](#) and [Table S5](#) for AUBPRC comparison; [Figure S12](#) and [Table S6](#) for AUROC comparison; [Table S7](#) for individual comparisons with MPC<sup>41</sup> and M-CAP<sup>17</sup>).

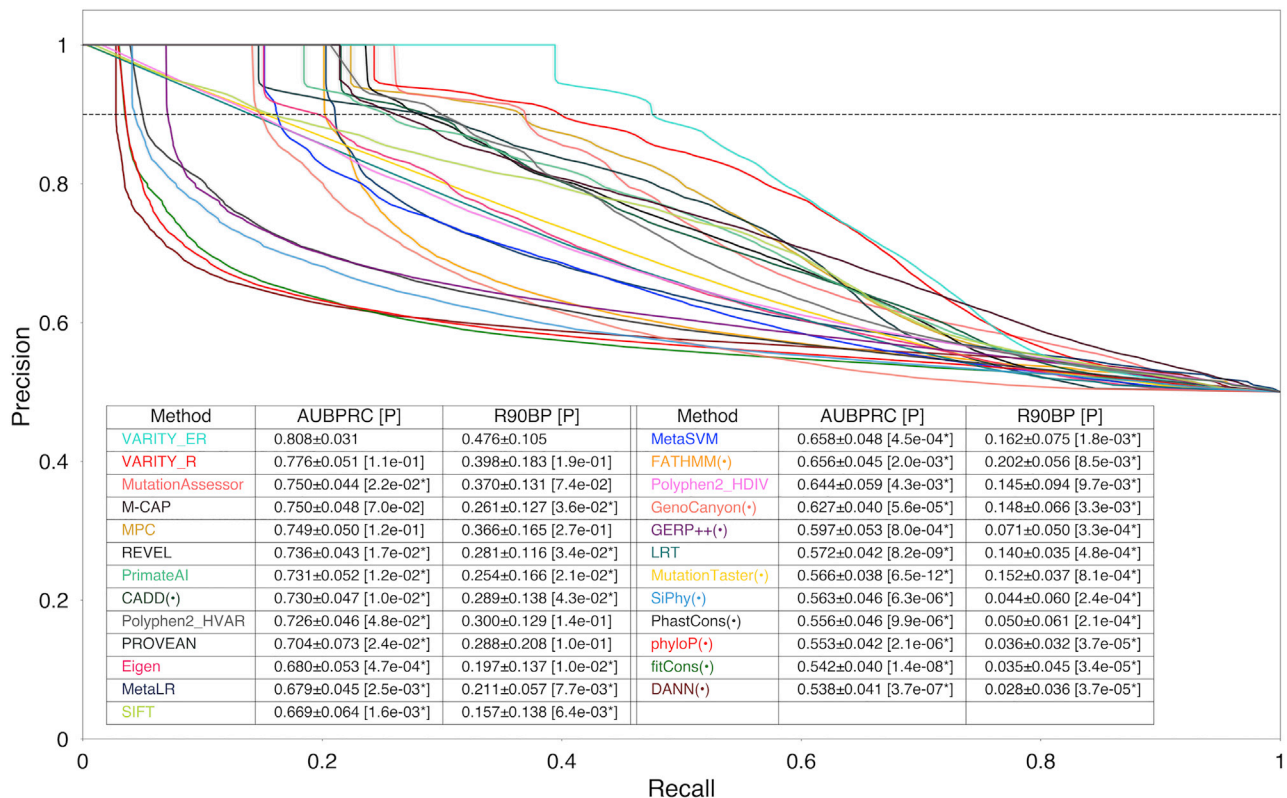
#### Assessing VARIETY with experimental variant effect maps

As another form of independent validation, we wished to assess agreement between impact scores and systematic experimental assessments of variant function. We therefore evaluated the ability of scores from each computational method to predict the quantitative scores from each of six recently published missense variant effect maps on human disease related genes.<sup>6–11</sup> Because five out of six maps (all except VKORC1 map) had been indirectly used in VARIETY training, we included leave-one-variant-out scores (VARIETY\_R\_LOO and VARIETY\_ER\_LOO) for comparison. VARIETY\_ER\_LOO showed best correspondence in terms of Pearson correlation coefficient (PCC) with variant effect map scores (PCC = 0.405), significantly outperforming all other non-VARIETY predictors ( $p < 0.05$ ; [Figure 6](#) and [Table S8](#) for PCC results; [Table S9](#) for Spearman's rank correlation coefficient [SRC] results).

#### More broadly assessing VARIETY performance via nested cross-validation

To further compare generalization performance of VARIETY models with previously developed computational predictors of missense variant effects, we used the high-quality core sets from ClinVar.<sup>44</sup> To avoid possible circularity from model training and hyperparameter tuning, we adopted a nested cross-validation strategy (see [material and methods](#)).

VARIETY\_R significantly outperformed all 23 methods on rare variants from ClinVar<sup>44</sup> (MAF < 0.5%; VARIETY\_R core sets) by all criteria (i.e., AUBPRC, R90BP, and AUROC;  $p < 0.05$  via one-sided paired t test), recovering 13% more pathogenic rare missense variants at 90% balanced precision than the next-best method, REVEL<sup>16</sup> ([Figure 7](#) and [Table S10](#) for AUBPRC comparison; [Figure S13](#) and [Table S11](#)



**Figure 5. Assessing VARITY balanced precision recall performance for *de novo* variants in neurodevelopmental disorder studies**

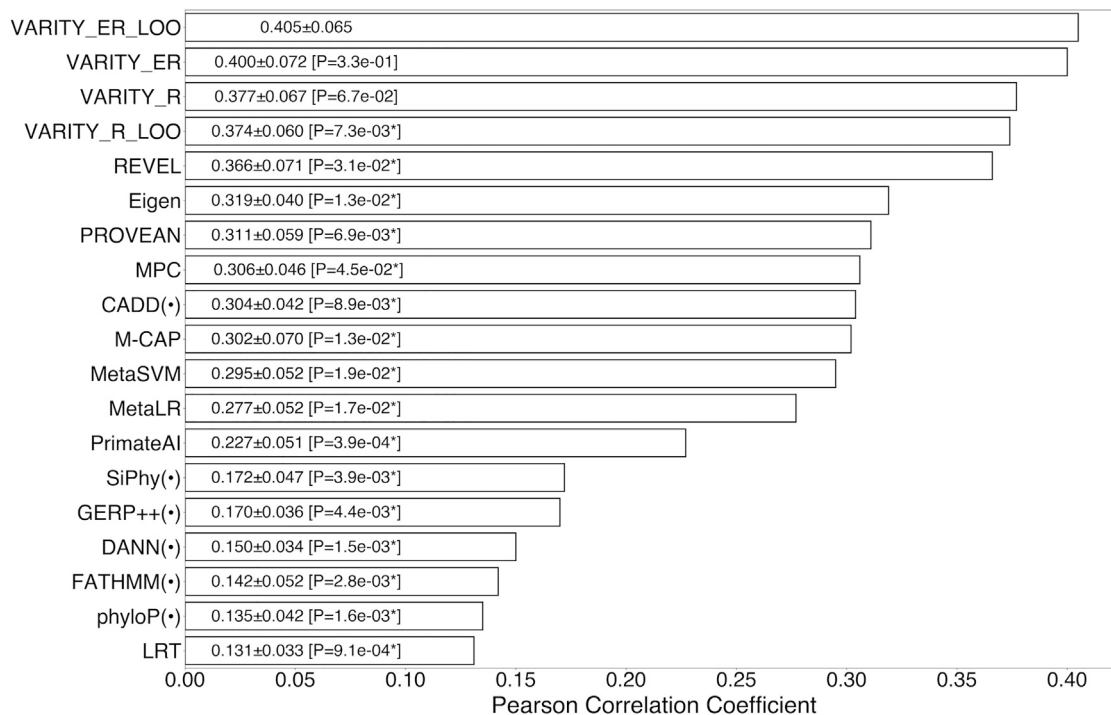
Here, we show balanced precision versus recall curves for VARITY\_ER, VARITY\_R, and 23 other variant pathogenicity predictors. Predictors designed specifically for nucleotide variants are indicated with a “(●).” Here, the test set was a total 215 rare *de novo* variants (188 positive and 27 negative examples, see [material and methods](#)). Recall was averaged over 2,000 bootstrapped test sets with standard error indicated by the surrounding gray region. As overall performance measures, AUBPRC and R90BP (the black dotted line) and their standard errors are shown. Statistical significance relative to VARITY\_ER applied a one-sided Z test based on 2,000 bootstrapped test sets ( $p$  values in brackets were indicated with a “\*” where  $p < 0.05$ ). Other test statistics, such as 95% confidence interval and effect size, can be found in [Table S5](#). When individual comparison used a larger variant set that had been scored by MPC<sup>41</sup> (285 variants in case group and 38 variants in control group) or M-CAP<sup>17</sup> (320 variants in case group and 45 variants in control group), VARITY\_ER’s improved AUBPRC performance was statistically significantly ( $p = 0.048$  and  $0.029$  for MPC<sup>41</sup> and M-CAP<sup>17</sup>, respectively; [Table S7](#)).

for ROC comparison). EVMutation<sup>27</sup> and DeepSequence<sup>42</sup> had few scores available for core set variants and, because we limited the overall comparison to variants assessed by all methods, their inclusion would have dramatically reduced the space of test variants. However, direct comparison of VARITY\_R with each method with the intersection of test variants scored by each respective method showed that VARITY\_R significantly outperformed both methods ([Table S12](#)). Methods designed to evaluate amino acid substitutions generally outperformed methods that were designed for nucleotide-level changes (see predictors marked with “(●)” in [Figure 7](#)). Indeed, among nucleotide-level methods, only CADD<sup>18</sup> (which may have benefited from amino-acid-level PolyPhen2<sup>19,20</sup> scores as training features) was able to outperform an amino-acid-level method.

To evaluate the extent to which VARITY’s success depended on the variant-weighting scheme, we retained a “VARITY\_R\_unweighted” model, giving equal weight to all core and add-on variants. This reduced AUBPRC performance from 92% to 90.9% ( $p = 6.5 \times 10^{-5}$ ; Student’s  $t$  test comparing the two sets of outer-loop validation perfor-

mance values). We expected that performance for the no-variant-weighting version of VARITY would fall further as more add-on training examples with uncertain quality are included. To evaluate this, we included additional gnomAD<sup>47</sup> variants that were originally excluded from the putatively-benign add-on training set because they had never been seen in a homozygous individual. After including these variants in training, the nested cross-validation AUBPRC performance of VARITY\_R\_unweighted was reduced further from 90.9% to 89.9% ( $p = 2.1 \times 10^{-4}$ ).

We also evaluated VARITY\_ER performance by using the core set of extremely rare variants from ClinVar<sup>44</sup> (MAF  $< 10^{-6}$ ; VARITY\_ER core sets), again with 10-fold nested cross-validation. VARITY\_ER outperformed all 23 methods by all criteria ( $p < 0.05$  via one-sided paired  $t$  test), recovering 12% more pathogenic rare missense variants at 90% balanced precision than the next-best method REVEL<sup>16</sup> ([Figure S14](#) and [Table S13](#) for AUBPRC comparison; [Figure S15](#) and [Table S14](#) for ROC comparison). Although EVMutation<sup>27</sup> and DeepSequence<sup>42</sup> were again excluded, VARITY\_ER outperformed EVMutation<sup>27</sup> in a



**Figure 6. Performance of VARITY\_R model and other predictors on variant effect maps**

Evaluation of VARITY models (VARITY\_R and VARITY\_ER) and 19 other computational predictors according to correlation (Pearson correlation coefficient or PCC) with variant function scores of rare (MAF < 0.5%) missense variants from six experimental variant effect maps.

SIFT,<sup>25,26</sup> PolyPhen2\_HVAR,<sup>19,20</sup> PolyPhen2\_HDIV,<sup>19,20</sup> and MutationAccessor<sup>33</sup> were not included because we did not have their scores for all six variant effect maps. Because some of the variant effect maps were indirectly used in VARITY training, we included VARITY\_R\_LOO and VARITY\_ER\_LOO for comparison. VARITY\_ER\_LOO had the overall best performance and was significantly better than all other models. Standard error of PCC was derived from the set of PCC values from each variant effect map. Statistical significance of each predictor's performance was evaluated relative to VARITY\_ER\_LOO via a one-sided paired t test with 5 degrees of freedom (p values in brackets were indicated with a "\*" where p < 0.05). For compactness, four methods that have correlation less than 0.1 are not shown (see Table S8). Other information, such as 95% confidence intervals and differences in PCC values, is in Table S8. The PCC comparison for each individual variant effect map is also in Table S8.

pairwise comparison based on extremely rare core set variants with available EVMutation<sup>27</sup> scores (Table S15). Only 16 extremely rare benign core set variants had DeepSequence<sup>42</sup> scores, leaving too few variants in each held-out subset during 10-fold nested cross-validation to accurately evaluate performance.

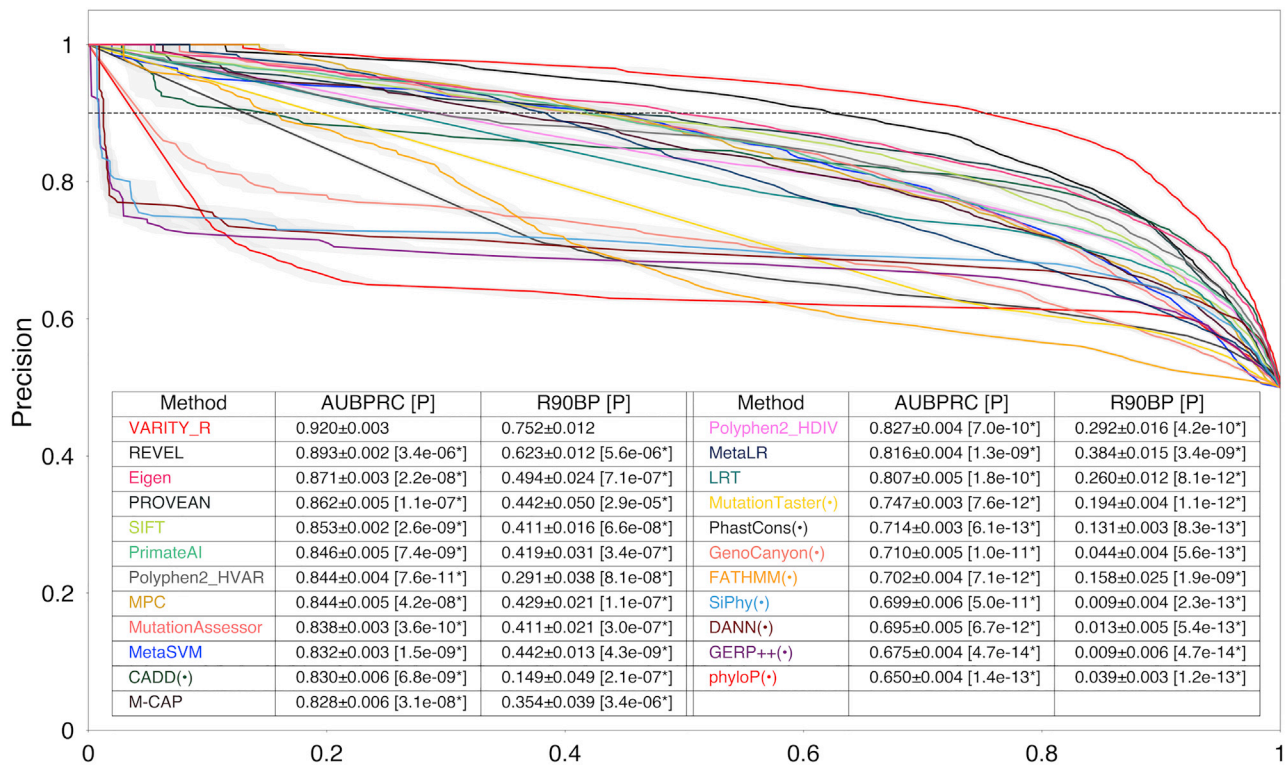
## Discussion

Evidence to evaluate variant pathogenicity is typically less available for rare as opposed to common variants. Detecting population-level correlation between a trait and a genotype has reduced power at lower minor allele frequencies and is futile for extremely rare variants. Experimental functional evidence is also less available for lower-allele-frequency variants. For these reasons, and because the majority of unique clinically observed variants are rare, there is a particularly strong motivation to improve computational methods for inferring the pathogenicity of rare variants.

Here, we developed predictors of variant pathogenicity that were specifically optimized for performance on rare and extremely rare missense variants. Given previous evidence that common variants are not representative of

rare variants in the context of training pathogenicity models,<sup>16</sup> perhaps the primary advantage of VARITY models is that rare variants have been given greater weight in training. Other key advantages are that VARITY makes use of large sets of potentially biased or error-prone training data while optimizing the potential for these add-on variant sets to increase performance. VARITY models benefitted not only from the process of filtering and weighting add-on examples but also from the use of features that have, either individually or in combination, not been used previously. To limit protein-specific bias and circularity, VARITY models also excluded features that are informative about protein identity or that may have been informed by previous pathogenicity annotations.

VARITY models performed well when evaluated via an independent set of *de novo* missense variants collected from neurodevelopmental case-control studies. Here, we only considered the subset of *de novo* missense variants that were in genes previously reported as being enriched for *de novo* missense and/or likely-gene-disruptive variants in neurodevelopmental disease cases. Because this certainly excluded some disease-relevant genes/variants,



**Figure 7. Comparing balanced precision recall performance of VARIETY\_R with other predictors in predicting a high-quality “core” variant set (MAF < 0.5%)**

We compare balanced precision versus recall performance for VARIETY\_R (with nested cross-validation) with 23 other variant pathogenicity predictors. For compactness, one predictor with AUBPRC < 0.6 is not shown (See Table S10). Predictors that had been designed specifically for nucleotide variants are indicated with a “(●).” The test set was 9,719 variants (5,912 positive and 3,807 negative examples) from the core set, after removing variants annotated by HGMD<sup>46</sup> and retaining only variants that had been scored by all methods. Recall was averaged over all ten outer-loop folds, and the standard error is indicated by the surrounding gray region. As overall performance measures, AUBPRC and R90BP (the black dotted line) and their standard errors are shown. Statistical significance of performance relative to VARIETY\_R used a one-sided paired t test with 9 degrees of freedom (p values in brackets were indicated with a “\*” where p < 0.05). Other test statistics, such as 95% confidence interval and effect size, are in Table S10.

it would be interesting in future studies to expand the gene/variant set based on more permissive identification of relevant genes/variants. VARIETY models might also be used for identification of candidate causal variants in other genes, e.g., in patients for whom a predicted-damaging variant has not been observed in an already-known disease gene.

More broadly, we used nested cross-validation to estimate the generalization performance of VARIETY models on the core set of ClinVar<sup>44</sup> variants. We note that the set of ClinVar<sup>44</sup> variants is influenced by historical ascertainment biases, in that they are enriched in disease genes discovered earlier, and some annotations were most likely influenced by early computational methods, such as BLOSUM<sup>15</sup> and SIFT.<sup>25,26</sup> Therefore, although ClinVar<sup>44</sup> variants enable useful measures of relative performance, it is difficult to estimate performance on an absolute scale for variants that will be observed in the future. Although the *de novo* variant evaluation set was free from these biases, it was necessarily limited in scope so that VARIETY’s generalization performance should in the future also be evaluated with additional independent test sets.

We provided information about which features and feature combinations were most useful in training VARIETY models and feature contributions for the prediction made for each variant. We also note that the quantitative VARIETY scores themselves can provide some intuition, e.g., there is a clear trend relating VARIETY\_R scores to the probability of pathogenicity (Figure S16).

The gradient boosted tree method has the advantage that training and testing examples for which some feature values are missing can be used directly without requiring imputation. Nevertheless, predictions may be less accurate for variants that are missing important features such as conservation-based scores and structure associated features. For example, if we remove all structural features from the core set, VARIETY\_R nested cross-validation performance (via AUBPRC) dropped from 92% to 91.5%.

For VARIETY training, we eliminated protein-identifying features to reduce protein-specific biases in our training set. However, we might want to augment VARIETY scores in the future by explicitly deriving a prior probability of missense variant pathogenicity (i.e., the probability of pathogenicity before considering the identity of the

missense variant) for each protein or protein region. However, because these priors may vary with the application setting, we suggest that the user should be explicitly informed about these priors and given the option to adjust them.

There are several avenues for future improvement of VARIETY models. More training examples could be added, e.g., adding putatively benign variants from primates as PrimateAI<sup>21</sup> has done. One might also develop specialized models for different classes of proteins or variants, e.g., those in predicted membrane-spanning domains. More quality-informative properties could be added to improve add-on set weighting. More features could also be added, e.g., based on knowledge of genetic or protein-protein interactions. We are also currently lacking features related to the typical mode of inheritance for a given gene and associated disease (e.g., dominant or recessive), the typical mechanistic class (e.g., gain or loss of function), or typical penetrance or variable expressivity of variants. Although adding mode-of-inheritance features would be complicated by the fact that this information is often unavailable and can differ for different variants within the same gene, it could improve performance in general.

To our knowledge, no previous computational variant effect prediction method has used the strategy of weighting variants used in training. However, it is not uncommon in machine learning to weight training examples. For example, the AdaBoost ensemble algorithm assigns higher weight to examples misclassified by previous classifiers in each learning iteration.<sup>62</sup> Where multiple annotators have provided possibly noisy labels, the weight and the true label of each example can be learned together with the model via a maximum-likelihood approach.<sup>63</sup> We also note that the modeling approach we describe is just one way to address the problem of model learning with noisy labels, which is the subject of intensive research especially within the deep learning field.<sup>64–68</sup> Given that variant effect interpretation is greatly limited by the availability of high-confidence unbiased training data, this field will be worth watching closely for future inspiration.

Other models tuned for other types of variation could employ this framework by simply changing the core and add-on training examples. For example, a computational model specialized for predicting the pathogenicity of variation in membrane-spanning domains could be trained by limiting the core set to variants from regions predicted to be in the membrane while still using all other variants as add-on training sets subject to differential weighting. Models for predicting deleteriousness of nucleotide-level variation in non-coding regions or spliced regions would require more extensive separate assembly of relevant features and both core and add-on variant sets but could then benefit from the algorithmic framework we describe here.

As the performance of computational variant effect predictors improves, an important unsolved problem is how to translate these gains to have a commensurate impact on clinical variant interpretation. The increasing quantity

of systematically collected large-scale functional evidence about missense variation has led to new recommendations for its use in the context of clinical interpretation.<sup>69</sup> We expect that a similarly quantitative evaluation of computational predictions will support a more nuanced use of computational evidence in clinical variant interpretation.

Users may search, browse, and download both VARIETY\_R and VARIETY\_ER scores (with associated feature contributions) for any of ~18,000 human proteins via the VARIETY web portal (see [web resources](#)).

### Data and code availability

All weighted core and add-on training data for the VARIETY\_R model can be found at [http://varity.varianteffect.org/downloads/VARIETY\\_R\\_training.csv](http://varity.varianteffect.org/downloads/VARIETY_R_training.csv) and, for the VARIETY\_ER model, at [http://varity.varianteffect.org/downloads/VARIETY\\_ER\\_training.csv](http://varity.varianteffect.org/downloads/VARIETY_ER_training.csv). Additional validation data (for *de novo* missense variants in neurodevelopmental disorder case/control studies and variants for six human-disease-related variant-effect maps) can be found at [http://varity.varianteffect.org/downloads/VARIETY\\_validation.zip](http://varity.varianteffect.org/downloads/VARIETY_validation.zip). All sources for features used in training are summarized in [Tables S1 and S2](#), and sources for labeling training variants can be found in [Table S3](#). All data are publicly available from the sources listed, except for (1) three MAVE datasets for *NCSI* (MIM: 603315; MaveDB: [urn:mavedb:00000065-a-1]), *GDII* (MIM: 300104; MaveDB: [urn:mavedb:00000066-a-1]), and *TECR* (MIM: 610057; MaveDB: [urn:mavedb:00000067-a-1]), which are currently unpublished but have been deposited to MaveDB<sup>14</sup> (and will be made available no later than 6 months after publication of this study), and (2) HGMD,<sup>46</sup> which was obtained under a commercial license that is generally available from QIAGEN/BioBase. All python code for training VARIETY models is available at <https://github.com/joewuca/varity>.

### Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.ajhg.2021.08.012>.

### Acknowledgments

We thank Thomas Hu for computing cluster support. This work was supported by a Canadian Institutes of Health Research Foundation grant (F.P.R.), by the National Human Genome Research Institute of the National Institutes of Health Center of Excellence in Genomic Science Initiative (NIH grants P50HG004233 and R01HG010461), by the Canada Excellence Research Chairs Program (F.P.R.), and by the One Brave Idea Initiative (jointly funded by the American Heart Association, Verily Life Sciences, and Astra-Zeneca).

### Declaration of interests

F.P.R. is a scientific advisor holding shares in Constantiam Biosciences and BioSymetrics and a Ranomics shareholder. S.S. is currently employed by Sanofi Pasteur (Canada). The authors declare no other competing interests.

Received: May 24, 2021

Accepted: August 18, 2021

Published: September 21, 2021

## Web resources

ClinVar, <https://www.ncbi.nlm.nih.gov/clinvar/>  
dbNSFP, <https://sites.google.com/site/jpopgen/dbNSFP>  
gnomAD, <https://gnomad.broadinstitute.org/>  
HGMD, <http://www.hgmd.cf.ac.uk/>  
HumsaVar, <https://www.uniprot.org/docs/humsavar>  
MaveDB, <https://www.mavedb.org/>  
VARITY, <http://varity.varianteffect.org/>

## References

1. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorf, L.A., Hunter, D.J., McCarthy, M.L., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.
2. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484.
3. Weile, J., and Roth, F.P. (2018). Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.* 137, 665–678.
4. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* 101, 315–325.
5. Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat. Methods* 11, 801–807.
6. 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.
7. 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.
8. 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.
9. 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, 874–882.
10. Sun, S., Weile, J., Verby, M., Wu, Y., Wang, Y., Cote, A.G., Fotiadou, I., Kitaygorodsky, J., Vidal, M., Rine, J., et al. (2020). A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Med.* 12, 13.
11. Chiasson, M.A., Rollins, N.J., Stephany, J.J., Sitko, K.A., Matreyek, K.A., Verby, M., Sun, S., Roth, F.P., DeSloover, D., Marks, D.S., et al. (2020). Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife* 9, 1–25.
12. 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.
13. Weile, J., Kishore, N., Sun, S., Maaieh, R., Verby, M., Li, R., Fotiadou, I., Kitaygorodsky, J., Wu, Y., Holenstein, A., et al. (2021). Shifting landscapes of human MTHFR missense-variant effects. *Am. J. Hum. Genet.* 108, 1283–1300.
14. Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T., Roth, F.P., Fowler, D.M., and Rubin, A.F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* 20, 223.
15. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 10915–10919.
16. 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.
17. 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.
18. 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.
19. 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.
20. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2, Chapter 7 (Curr. Protoc. Hum. Genet).
21. 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.
22. 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.
23. 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.
24. 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.
25. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081.
26. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9.
27. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135.
28. Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* 47, 276–283.
29. Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561.

30. 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.
31. 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.
32. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* *15*, 1034–1050.
33. 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.
34. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* *39*, e118.
35. 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.
36. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* *31*, 761–763.
37. 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.
38. Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K.H., and Zhao, H. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.* *5*, 10576.
39. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* *29*, 1504–1510.
40. Wong, K.C., and Zhang, Z. (2014). SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. *Bioinformatics* *30*, 1112–1119.
41. Samocha, K., Kosmicki, J., Karczewski, K., O'Donnell-Luria, A., Pierce-Hoffman, E., MacArthur, D., Neale, B., and Daly, M. (2017). Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* *10.1101/148353*.
42. 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*, 816–822.
43. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al.; ACMG Laboratory Quality Assurance Committee (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.
44. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* *46* (D1), D1062–D1067.
45. The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* *45* (D1), D158–D169.
46. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* *136*, 665–677.
47. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* *581*, 434–443.
48. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* *37*, 235–241.
49. Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* *372*, 774–797.
50. Buchan, D.W.A., and Jones, D.T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* *47* (W1), W402–W407.
51. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* *47* (D1), D427–D432.
52. Wu, Y., Weile, J., Cote, A.G., Sun, S., Knapp, J., Verby, M., and Roth, F.P. (2019). A web application and service for imputing and visualizing missense variant effect maps. *Bioinformatics* *35*, 3191–3193.
53. Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* *13-17-Aug*, 785–794.
54. 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.
55. Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Adv. Neural Inf. Process. Syst.* *24*, 2546–2554.
56. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D.D. (2015). Hyperopt: A Python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* *8*, 014008.
57. Turner, T.N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., F Stessman, H.A., Doebley, A.L., Bernier, R.A., Nickerson, D.A., and Eichler, E.E. (2017). denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* *45* (D1), D804–D811.
58. Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A., and Eichler, E.E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* *51*, 106–116.
59. Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* *2017-Decem*, 4766–4775.
60. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* *2*, 56–67.
61. 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.

62. Freund, Y., and Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* 55, 119–139.
63. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *J. Mach. Learn. Res.* 11, 1297–1322.
64. Vahdat, A. (2017). Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks. *Adv. Neural Inf. Process. Syst 2017-Decem*, 5597–5606.
65. Patrini, G., Rozza, A., Menon, A.K., Nock, R., and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition CVPR 2017 2017-Janua*, 2233–2241.
66. Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. (2018). Joint Optimization Framework for Learning with Noisy Labels. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR 2018*, 5552–5560.
67. Litany, O., and Freedman, D. (2018). SOSELETO: A Unified Approach to Transfer Learning and Training with Noisy Labels. *arXiv 1805.09622v2*. <https://arxiv.org/abs/1803.11364>.
68. Yao, J., Wang, J., Tsang, I.W., Zhang, Y., Sun, J., Zhang, C., and Zhang, R. (2018). Deep Learning from Noisy Image Labels with Quality Embedding. *IEEE Trans. Image Process.* 28, 1909–1922.
69. Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., Starita, L.M., et al.; Clinical Genome Resource Sequence Variant Interpretation Working Group (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* 12, 3.