Cold Spring Harbor Perspectives in Biology
www.cshperspectives.org

Variant Effect Prediction in the Age of Machine Learning

Yana Bromberg,[1,2] R. Prabakaran,[1,*] Anowarul Kabir,[3,*] and Amarda Shehu[3]

[1]Department of Biology; [2]Department of Computer Science, Emory University, Atlanta 30322, Georgia, USA
[3]Department of Computer Science, George Mason University, Fairfax 22030, Virginia, USA
Correspondence: yana.bromberg@emory.edu

Over the years, many computational methods have been created for the analysis of the impact of single amino acid substitutions resulting from single-nucleotide variants in genome coding regions. Historically, all methods have been supervised and thus limited by the inadequate sizes of experimentally curated data sets and by the lack of a standardized definition of variant effect. The emergence of unsupervised, deep learning (DL)-based methods raised an important question: Can machines learn the language of life from the unannotated protein sequence data well enough to identify significant errors in the protein "sentences"? Our analysis suggests that some unsupervised methods perform as well or better than existing supervised methods. Unsupervised methods are also faster and can, thus, be useful in large-scale variant evaluations. For all other methods, however, their performance varies by both evaluation metrics and by the type of variant effect being predicted. We also note that the evaluation of method performance is still lacking on less-studied, nonhuman proteins where unsupervised methods hold the most promise.

Prediction of genomic variant effect is arguably the holy grail of precision medicine, evolutionary analysis, and molecular function annotation. Over the last 20 years, machine learning techniques have captured the leading role in this field. This is primarily because of the enormous number of variables that go into establishing the effect of a specific variant. For example, evaluating the change in a protein function due to a single amino acid substitution, that is, the result of a missense single-nucleotide variant (SNV), may require evaluating all concomitant structural changes, proximity to active sites, and evolutionary history/conservation measures of the protein as a whole and at the specific position in question. Furthermore, the number of variants that have yet to be analyzed is large and growing. That is, every human genome differs from the reference by ~0.1% in SNVs, leading to roughly 3.5 million variants per individual (Pang et al. 2010; Pelak et al. 2010) and at least 84.7 million unique variants in the human population (1000 Genomes Project Consortium et al. 2015). Note that if we commit to the task of evaluating the variant effect on the abilities of other species, particularly bacteria, the number of variants needing annotation would grow exponentially.

[*]These authors contributed equally to this work.

Editors: Peter K. Koo, Christian Dallago, Ananthan Nambiar, and Kevin K. Yang
Additional Perspectives on Machine Learning for Protein Science and Engineering available at www.cshperspectives.org

Advanced Online Article. Cite this article as *Cold Spring Harb Perspect Biol* doi: 10.1101/cshperspect.a041467

A [NOT SO] PERFECT CASE FOR MACHINE LEARNING

Using machine learning methods for human variant analysis, however, has been historically complicated by the difficulty of defining measurable variant effect classes. For example, "Does this variant cause disease?" may be a yes/no question only for some cancer drivers and monogenic disorders; note that the variant effect may need to be further modified by the allelic dominance. Nevertheless, many methods have claimed the ability to predict variant pathogenicity, without explicitly defining the term's meaning. This reductionist paradigm that has long dominated the field of genetics (Gibson 2012; Katsanis 2016) answers a very different question than whether the complex interplay of all individual genome variants with the environment is likely to give rise to a particular disease.

Existing experimentally derived, quantified, and validated variant annotations that could be used for method training are few in the scientific databases and have, so far, been hard to extract from the literature. For example, the manual curation efforts to collect monogenic disease/phenotype-associated variants in OMIM (Hamosh et al. 2000) have only annotated ∼23,000 (0.027% of 84.7M) SNVs (Savojardo et al. 2021). Additional disease-associated and putatively neutral missense SNVs were collected in HumVar (Capriotti et al. 2006), from the work of SwissProt curators (Boeckmann et al. 2003; Bairoch et al. 2004), and in the more recent HuVarBase (Ganesan et al. 2019). The latter parses a large number of resources to extract 774,863 disease-annotated variants (718,590 are missense)—a "staggering" 0.92% of all known human SNVs.

Non-disease-relevant variant functional effect is even harder to come by. The Protein Mutant Database (PMD) (Nishikawa et al. 1994; Kawabata et al. 1999), a project started in 1989, required reading over 42,000 relevant scientific articles to extract the subjective, qualitatively binned measurements of the variant effect of over 200K variants in different species on protein structure, stability, and/or function; note that human variants make up less than a fifth of this data set. While recent small studies have been more proactive in distributing their data (see, e.g., the VariBench collection [Sarkar et al. 2020]), study reporting, data accessibility, and cross-study standardization, remain limited. Given the resulting disproportionately wide and short training data tables (i.e., relatively few variants and many features that can describe a variant), the field has gotten very creative in data collection for prediction method development/training.

NOT ALL VARIANT IMPACT IS THE SAME

Here lies the root of the confusion—understanding of exactly which variant effect is being predicted is limited across the nearly 200 currently available, supervised methods (Hu et al. 2019; Zhu et al. 2020). For example, SNAP (Bromberg and Rost 2007; Hecht et al. 2015) is trained using (all species) single amino acid substitutions collected from the PMD, where functional effect refers to the result of experimental evaluations of the activity of wild-type versus mutated proteins. As expected, SNAP aims to predict the variant functional effect. One version of PolyPhen-2 (Adzhubei et al. 2010) is trained on the HumVar data set, that is, pathogenic variants versus those with no such annotation. Curiously, PolyPhen is also meant to predict the possible effect of the variant on the function and structure of the affected protein. CADD (Kircher et al. 2014) contrasts observed coding and noncoding variants with simulated mutations, predicting variant functional effect and both disease causation and association. Furthermore, meta-methods (e.g., REVEL [Ioannidis et al. 2016]) rely on input from a variety of disparate techniques to summarize some fluid definition of variant effect. The recent advances in deep learning (DL) and unsupervised methods (e.g., DeepSequence [Riesselman et al. 2018], EVE [Frazer et al. 2021], and ESM1V [Meier et al. 2021]) have given rise to even less-explainable effect predictions. While these models are better suited to evaluate variants on a large scale, the specific type of effect that their variant scores capture is unclear. Numerous efforts have been undertaken to evaluate the performance of the various methods (e.g., The Critical Assessment of Genome Interpretation Consortium 2022), but these are also limited by the lack of correspondence between the questions being asked and the type of answers that these tools have been developed to provide.

Excluding the prediction of measurable structural or stability changes, there are three broad kinds of effects that are relevant: (evolutionary) fitness effect, pathogenicity (disease causation), and (molecular) function change. We argue that existing computational methods have largely failed to recognize the difference between these three types of effect.

IMPACT OF VARIANT EFFECT TYPE ON METHOD DEVELOPMENT

While closely related, the three types of variant effects are not identical and, in fact, are very different in non-edge cases. Evolutionary fitness, for example, is often evaluated in terms of population frequency or conservation of the mutated site across species; that is, variants are expected to be more or less common in human population or across species because of the impact they have on their carrier ability to survive and reproduce. Evolutionary history, however, does not guarantee current success. Thus, for individual human genomes, variant population frequency or site conservation alone is unlikely to lead to a precise conclusion about variant functional effect or involvement in causing a disease. Similarly, a slight change in the function of a given protein may be insufficient to make a fitness difference on the human population level or to bring on a clearly definable disease phenotype.

The relationship of evolutionary fitness with disease is complicated by the polygenic nature of most disorders, making it impossible to infer truly causative variants. However, even in the case of monogenic disorders, the selective disadvantage of causative variants is not always guaranteed, and population-specific frequency is paramount. For example, the sickle cell hemoglobin allele is exceedingly rare (∼0% minor allele frequency [MAF]) in Europe and the Americas, but more common (MAF > 0.5%) throughout most of the African continent, and pervasive (MAF > 9%) across the large area stretching from southern Ghana to northern Zambia (Piel et al. 2010). Thus, elucidating the specific relationship between disease and fitness requires a much deeper understanding of the nature of the disease and of the features of the affected populations.

Technological problems, for example, insufficient sequencing across populations, limited experimental resolution in establishing functional effect, and even statistical inference parameters, complicate matters further. For example, genome-wide association studies (GWAS), require that a given variant be present in some significant fraction of the population (e.g., 5%). This frequency, however, would be improbable, if not impossible, for a true disease-causing, selectively disadvantageous variant.

Furthermore, outside of human populations, the words fitness, pathogenicity, and population frequency are measured using different scales and, thus, methods pretrained with human data cannot be expected to produce similar results across organisms.

The lack of effect type differentiation in literature could be attributed to the variants historically analyzed in wet-lab experiments and used for, first, establishing the theoretical framework for effect identification and, later, for computational method development. Due to scientific interest and time/money constraints, scientists aimed to study variants likely to cause a disease; after all, why not start with a variant that is likely to cause diabetes or trigger cancer development? At the same time, due to experimental limitations, functional variant effect could only be measured reliably for high-effect variants with visible phenotypic displays, such as a drastic reduction in catalytic activity, reduced number/size of cell colonies, or even misshapen red blood cells. On the other hand, many experimentally annotated negative/no-effect findings were incidental and often not explicitly or rigorously evaluated.

This bias toward hunting for high-effect variants has been recently elucidated through the development of deep scanning mutagenesis (DMS) methods that quantitatively, for one type of function per experiment, evaluate many, if not all, possible variants per given protein (Araya and Fowler 2011). DMS experiments demonstrated that the scale of variant effects is continuous and, by sampling from only the high-effect variants, we forgo the understanding of what "effect" means (Miller et al. 2017, 2019; Pejaver et al. 2019). As such, even a perfectly trained supervised classifier would tend to underestimate the number of function effect-carrying variants in a given protein or gene, while method evaluations, of either supervised or unsupervised methods, may overestimate the correlation between the three types of variant effect.

In evaluating variant effect, one readily available representation of evolutionary fitness—conservation of the mutated site—has been exceedingly useful and used by many methods. The reasoning behind the use of this feature is clear—conserved positions are conserved for a reason and, thus, should not change. The use of conservation across methods has, however, ensured that most predictions are largely nonorthogonal, that is, they tend to predict the same conservation-evidenced outcomes. This conclusion is supported by the minimal variation in performance across methods and minimal performance improvement of meta-methods. The recent development of single-sequence-based, unsupervised, protein language models (pLMs) appears to bypass the explicit need for conservation information. These models may capture a different effect signal and, alone or in concert with other techniques, may be helpful in distinguishing variant effect types.

CAN THE TOOLS OF TODAY ANNOTATE VARIANT EFFECT?

In this work, we evaluate the advances in variant effect prediction across the various models and effect types and sizes. We first evaluate the contribution of variant frequency and position conservation to variant effect predictor scores. We suggest that these characteristics of variants are the best currently available, although indirect, representations of variant "fitness." We then use experimentally validated sets of variants to evaluate the method's ability to predict variant functional effects. Here, we assess both a selection of variants explored in the scientific literature (reported in the PMD) and an exhaustive collection of variants from DMS experiments. Finally, we assess the performance of methods in annotating variants as being pathogenic (i.e., likely disease-causing [reported in ClinVar; Landrum et al. 2018]).

We consider three types of methods (SOM methods): classical methods and supervised and unsupervised DL methods. Classical methods use computational (e.g., SIFT [Ng and Henikoff 2003; Hu and Ng 2013]) and machine learning–based (e.g., REVEL) techniques, where input features (e.g., conservation, structure, or other method predictions) are selected on the basis of biological relevance. Their training/development data classifies variants into two groups—effect or no effect—however defined. Supervised DL methods may similarly use selected biological variant or protein features (e.g., MetaRNN uses variant frequency), conservation scores, and output of other prediction methods, or they may rely simply on protein sequence and structure (e.g., Seq-UNet [Dunham et al. 2023]). They are also trained to reflect on variant effect (e.g., pathogenicity in the case of MetaRNN [Li et al. 2022]) and variant "rare-ness" (probability that the variant is rare), or the corresponding position-specific scoring matrix (PSSM) of amino acids (in the case of Seq-UNet). As opposed to classical methods, however, all methods in this category use DL architectures to achieve their task. Finally, unsupervised DL methods use sequence information alone without an attached variant classifier. Their outputs are then evaluated to explore the difference between wild-type and mutant sequences (SOM methods in Supplemental Material).

To evaluate the performance of all methods, we scaled each method's prediction scores into the [0, 1] range using min–max scaling across all variants in all data sets and made all scores unidirectional (1 = effect, 0 = no effect; SOM methods in Supplemental Material). Furthermore, we developed a standardized, population frequency-based means for score threshold selection. For each method, we identified a threshold to separate effect versus no-effect variants as the one below in which 95% of the variants common in the population (allele frequency ≥0.01) were observed (Zeng and Bromberg 2019). We argue that this approach allows for high precision in identifying variants of high effect, even if the recall of smaller effect variants is limited. Finally, to represent the conservation baseline, we used phastCons (Siepel et al. 2005), computed on multiple sequence alignments of 16 primate genomes to the human genome. Earlier work found that using primate genome alignments was more informative of variant effect than all vertebrate alignment (Sun and Yu 2019). Genome-alignment-based methods may not be as sensitive to protein structure/function changes as the more protein-focused methods. However, the conservation of genomic sites in protein-coding regions may be driven by non-protein-sequence-specific needs (e.g., splice sites, mRNA stability, tRNA binding, etc.). We thus argue that these estimates are better suited for baseline assessments of variant fitness and report phastCons performance for comparison with all data sets.

VARIANT EFFECT PREDICTORS AND LANGUAGE MODELS RECOGNIZE POPULATION FREQUENCY

Variants in the human genome are not evenly distributed across coding genes. In our analysis of the ALFA data (Phan et al. 2020), only slightly more than half (58%) of the 16,671 proteins considered in this study corresponded to genes that carried a common variant (allele frequency ≥0.01), while nearly all had a singleton (allele count = 1) or an ultra-rare variant (frequency <0.001). Rare variants were only present in roughly three-quarters of the proteins (0.01> allele frequency ≥0.001; 74%). Although the uneven distribution of variants in some of the proteins can be explained by insufficient data, disagreement between sequencing projects, which could be expected if individual project data were not representative, is rare. Where projects do disagree, it is often one, usually smaller project that assigns higher frequency, while all others do not. For example, the FLYWCH-type protein (NP_001294997) contains no common variants according to ALFA (sample size = 48,480) or any of the population-specific projects but has one SNP (rs61747748; ALFA MAF = 0.0014) labeled as common by the Simons genome diversity project (sample size = 12). Thus, we expect that common variants are indeed limited to a subset of genes. For these, the absence of common variants may identify evolutionary novelty and thus less time to perpetuate variants throughout the population. Alternatively, these genes can be essential and thus resist variance altogether. We note that in-depth evaluation of the reasons for this variant disparity across genes/proteins is beyond the scope of this work.

Here, we set out to evaluate the relationship of the variant frequency with the predicted variant effect. We considered the genes lacking common variants to be somewhat biologically unique and, possibly, misrepresented in training data of computational methods. We thus retained only the variants in those genes that carried both common and other variant types together and sampled 35,082 variants in 9,142 proteins (Supplemental Fig. S1; Supplemental Table S1).

The nature of training/development data used by many supervised methods had, due to experimental limitations, mostly labeled rare human pathogenic variants as deleterious (i.e., pathogenic and/or of large negative functional effect), while common variants were largely deemed neutral (i.e., not pathogenic nor bearing functional effect). For example, very common (i.e., >5% allele frequency) variants are defined as benign by ACMG–AMP experts (i.e., stand-alone evidence of benign effect [BA1]). We thus expected that classical methods for prediction of variant effect, represented here by CADD, REVEL, PolyPhen, and SIFT, as well as supervised DL methods (i.e., MetaRNN, MVP [Qi et al. 2021], Seq-UNet, VESPA [Marquet et al. 2022], and AlphaMissense [Cheng et al. 2023]), would identify less frequent variants as having a larger effect.

We observed the expected method behavior (Fig. 1A), although the difference between scores of common versus rare-type variants was not large for most methods. All classical methods performed roughly the same, but, curiously, supervised DL methods spanned the range of abilities in this analysis; they ranged from identifying almost no disparity in population frequency (Seq-UNet and MVP) to classic-like performance (VESPA) to nearly perfect frequency classification (MetaRNN). Note that while all methods were better at differentiating common from ultra-rare variants than from rare ones, MetaRNN was sig-

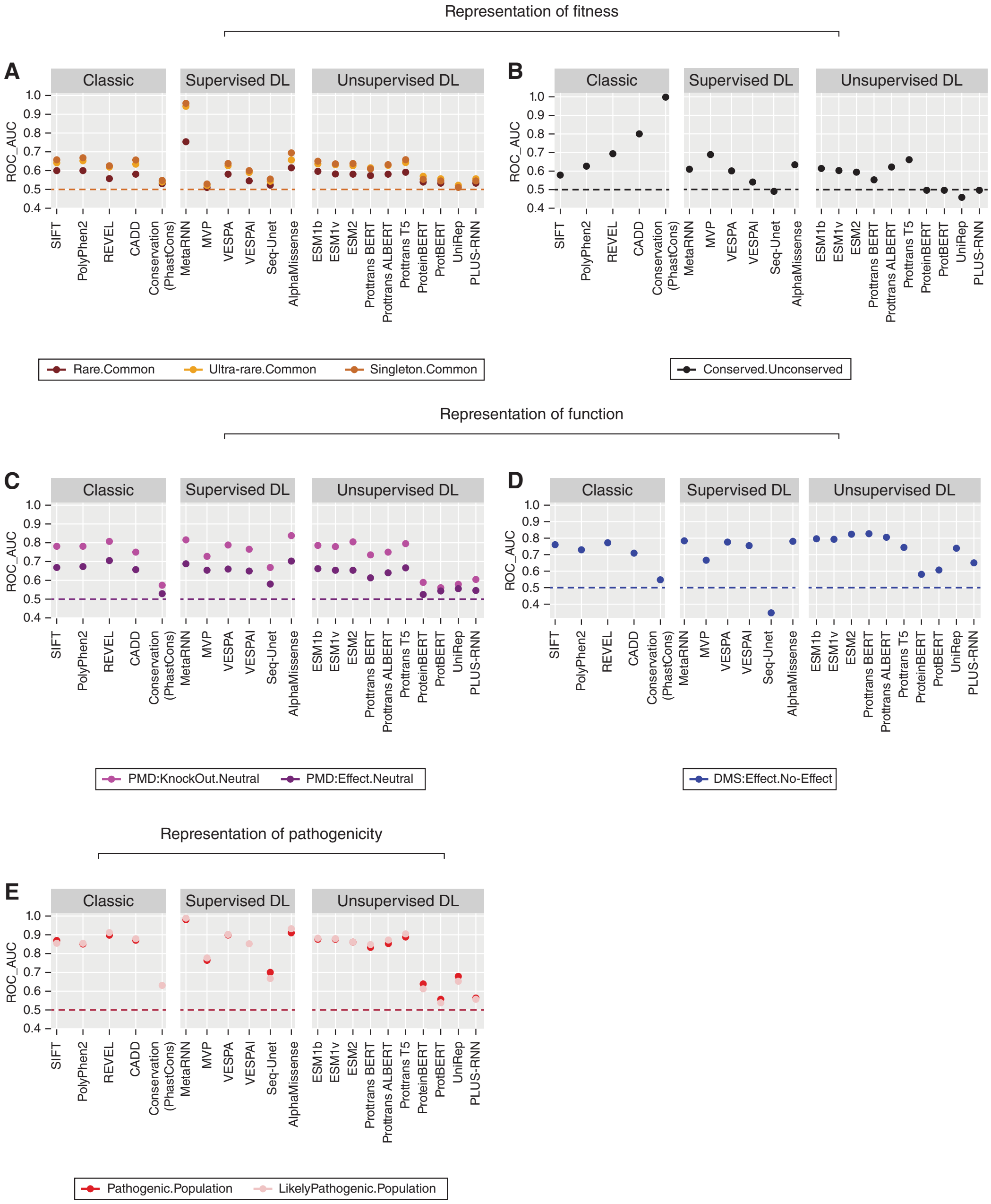


Figure 1. Evaluation of performance of various methods on data sets. Performance is reported as the ROC AUC for all data sets and methods used in this study. Higher ROC indicates better ability to differentiate (*A*) common variants from all other population frequency variants, (*B*) conserved from unconserved variants, (*C*) functionally neutral variants in the Protein Mutant Database (PMD) from variants assigned an effect, (*D*) effect/no-effect deep scanning mutagenesis (DMS) variants, and (*E*) pathogenic variants from putatively benign baseline. Dashed lines indicate the performance of an empirical random classifier.

nificantly more so, suggesting that the pathogenicity signal that it learned is primarily frequency-driven. Note that at our selected scoring threshold most methods included the majority of all variants—common or not—indicating the rarity of variants with high effect (threshold line is higher than the bulk of variant scores) (Fig. 2; Supplemental Fig. S2).

We expected that unsupervised models, trained on most available protein sequences, would capture global sequence signals rather than specific population frequencies. In fact, common human variants may be expected to have some level of evolutionary significance for phenotypically distinct human populations (Bromberg et al. 2013; Mahlich et al. 2017) but are unlikely to disturb the global protein language patterns. Furthermore, while rare and ultra-rare variants carry the bulk of evolutionarily deleterious variation, these drastically damaging changes make up only a small fraction of the massive total number of rare variants. That is, we expected to see little difference between the scores of common and rare variants. Indeed, for some models (e.g., proteinBERT [Brandes et al. 2022] or UniRep [Alley et al. 2019]), scores were not a major indicator of variant frequency. For others, however (e.g., ESM1b [Rives et al. 2021] and ProtTransT5 [Elnaggar et al. 2022]), we observed that scores were as different across variant frequency classes as for some classical methods (Fig. 1A). Note AlphaMissense was trained using frequency labels; thus, its improved performance in capturing frequency signals was expected if underwhelming.

Our findings thus indicate that, on average, a given variant's population frequency is only moderately indicative of its predicted effect, as captured by (most) methods regardless of predictor class.

VARIANT PREDICTIONS CORRELATE ACROSS MANY METHODS

In the presence of many methods for predicting variant effect, it is natural for scientists to evaluate individual variants using a number of tools to

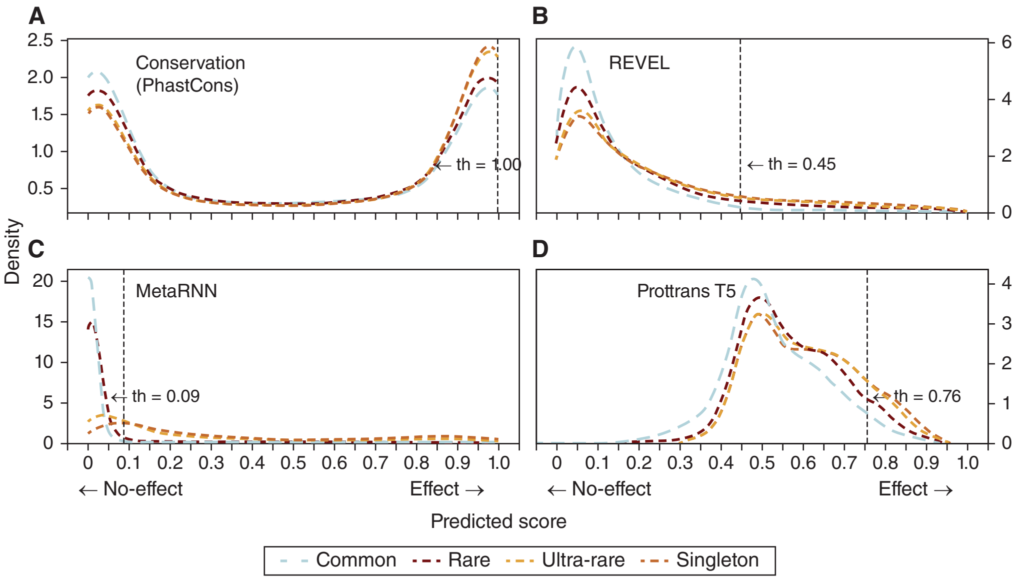
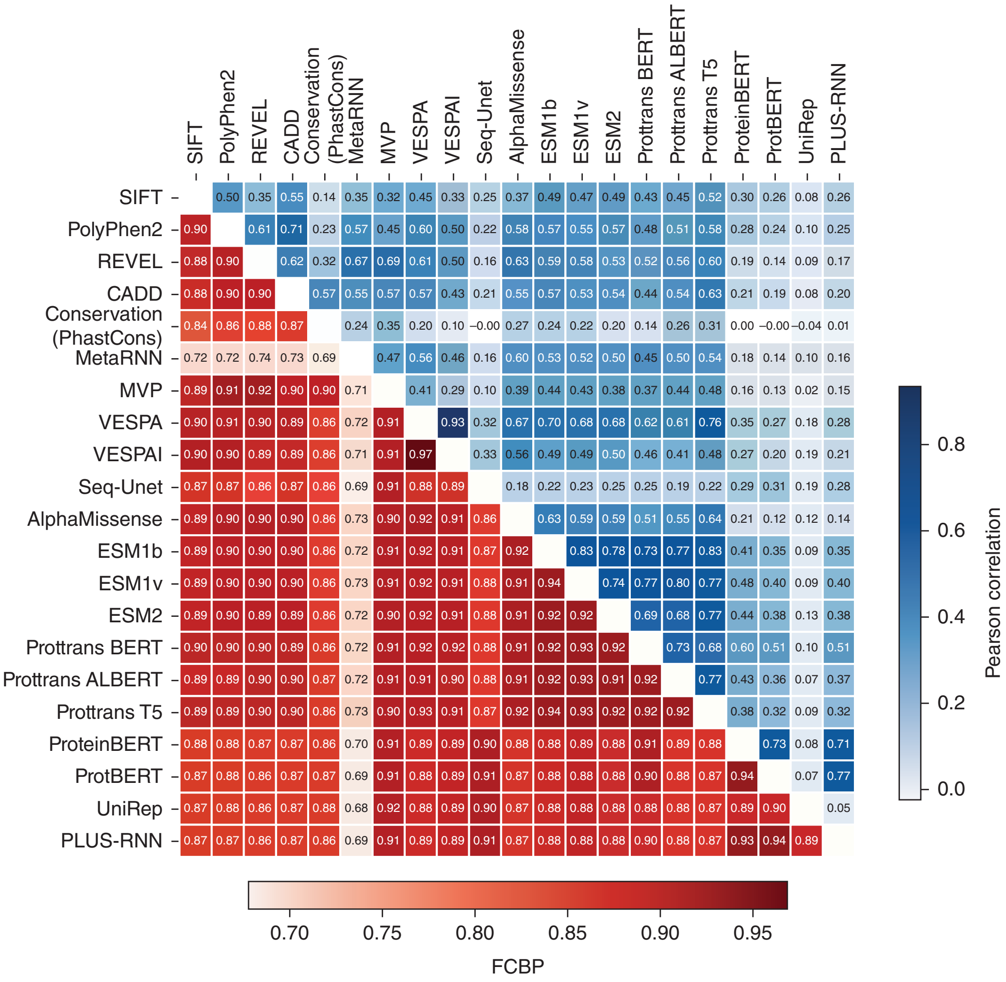


Figure 2. Distribution of scores for different variant population frequencies. Scores attained by variants in different population frequency classes (line colors) from (*A*) phastCons, our conservation score, (*B*) REVEL, an example of a classical method in our set, (*C*) MetaRNN an example of a supervised deep learning (DL), and (*D*) ProtTransT5, an example of an unsupervised predictor. Thresholds indicate scores below which 95% of common variants were observed. Representative methods attained the best ROC AUCs in Figure 1A; other methods are in Supplemental Figure S2.

establish agreement. The above findings, however, confirm that many methods are nonorthogonal and will often produce similar results—regardless of the actual effect that the variant may have. Indeed, it is informative to evaluate the performance of different methods on the set of variants where their predictions disagree (Bromberg and Rost 2007). We thus asked what is the relationship between method scores? (Fig. 3).

We observed that in binary variant assessment at the established threshold (Fig. 3, red), all methods, except MetaRNN, were in nearly perfect agreement. We also found that conservation (phastCons) scores correlated with method predictions, suggesting that conservation alone may have been sufficiently informative of these predictions.

The relationship between method scores, however, was somewhat more informative (Fig. 3, blue). As expected from binary comparisons, many predictor scores correlated to some extent, UniRep and Seq-UNet being the exceptions. However, conservation was no longer as well representative of other methods. There was a significant level of correlation between most supervised methods, whether DL-based or not. ESM



	SIFT	PolyPhen2	REVEL	CADD	Conservation (PhastCons)	MetaRNN	MVP	VESPA	VESPAl	Seq-Unet	AlphaMissense	ESM1b	ESM1v	ESM2	Prottrans BERT	Prottrans ALBERT	Prottrans T5	ProteinBERT	ProtBERT	UniRep	PLUS-RNN
SIFT		0.50	0.35	0.55	0.14	0.35	0.32	0.45	0.33	0.25	0.37	0.49	0.47	0.49	0.43	0.45	0.52	0.30	0.26	0.08	0.26
PolyPhen2	0.90		0.61	0.71	0.23	0.57	0.45	0.60	0.50	0.22	0.58	0.57	0.55	0.57	0.48	0.51	0.58	0.28	0.24	0.10	0.25
REVEL	0.88	0.90		0.62	0.32	0.67	0.69	0.61	0.50	0.16	0.63	0.59	0.58	0.53	0.52	0.56	0.60	0.19	0.14	0.09	0.17
CADD	0.88	0.90	0.90		0.57	0.55	0.57	0.57	0.43	0.21	0.55	0.57	0.53	0.54	0.44	0.54	0.63	0.21	0.19	0.08	0.20
Conservation (PhastCons)	0.84	0.86	0.88	0.87		0.24	0.35	0.20	0.10	−0.00	0.27	0.24	0.22	0.20	0.14	0.26	0.31	0.00	−0.00	−0.04	0.01
MetaRNN	0.72	0.72	0.74	0.73	0.69		0.47	0.56	0.46	0.16	0.60	0.53	0.52	0.50	0.45	0.50	0.54	0.18	0.14	0.10	0.16
MVP	0.89	0.91	0.92	0.90	0.90	0.71		0.41	0.29	0.10	0.39	0.44	0.43	0.38	0.37	0.44	0.48	0.16	0.13	0.02	0.15
VESPA	0.90	0.91	0.90	0.89	0.86	0.72	0.91		0.93	0.32	0.67	0.70	0.68	0.68	0.62	0.61	0.76	0.35	0.27	0.18	0.28
VESPAl	0.90	0.90	0.89	0.89	0.86	0.71	0.91	0.97		0.33	0.56	0.49	0.49	0.50	0.46	0.41	0.48	0.27	0.20	0.19	0.21
Seq-Unet	0.87	0.87	0.86	0.87	0.86	0.69	0.91	0.88	0.89		0.18	0.22	0.23	0.25	0.25	0.19	0.22	0.29	0.31	0.19	0.28
AlphaMissense	0.89	0.90	0.90	0.90	0.86	0.73	0.90	0.92	0.91	0.86		0.63	0.59	0.59	0.51	0.55	0.64	0.21	0.12	0.12	0.14
ESM1b	0.89	0.90	0.90	0.90	0.86	0.72	0.91	0.92	0.91	0.87	0.92		0.83	0.78	0.73	0.77	0.83	0.41	0.35	0.09	0.35
ESM1v	0.89	0.90	0.90	0.90	0.86	0.73	0.91	0.92	0.91	0.88	0.91	0.94		0.74	0.77	0.80	0.77	0.48	0.40	0.09	0.40
ESM2	0.89	0.90	0.89	0.89	0.86	0.72	0.90	0.92	0.91	0.88	0.91	0.92	0.92		0.69	0.68	0.77	0.44	0.38	0.13	0.38
Prottrans BERT	0.90	0.90	0.90	0.89	0.86	0.72	0.91	0.92	0.92	0.88	0.91	0.92	0.93	0.92		0.73	0.68	0.60	0.51	0.10	0.51
Prottrans ALBERT	0.89	0.89	0.90	0.90	0.87	0.72	0.91	0.91	0.90	0.88	0.91	0.92	0.93	0.91	0.92		0.77	0.43	0.36	0.07	0.37
Prottrans T5	0.89	0.89	0.90	0.90	0.86	0.73	0.90	0.93	0.91	0.87	0.92	0.94	0.93	0.92	0.92	0.92		0.38	0.32	0.09	0.32
ProteinBERT	0.88	0.88	0.87	0.87	0.86	0.70	0.91	0.89	0.89	0.90	0.88	0.88	0.89	0.88	0.91	0.89	0.88		0.73	0.08	0.71
ProtBERT	0.87	0.88	0.86	0.87	0.87	0.69	0.91	0.88	0.89	0.91	0.87	0.88	0.88	0.88	0.90	0.88	0.87	0.94		0.07	0.77
UniRep	0.87	0.88	0.86	0.87	0.88	0.68	0.92	0.88	0.89	0.90	0.87	0.88	0.88	0.88	0.88	0.88	0.87	0.89	0.90		0.05
PLUS-RNN	0.87	0.87	0.86	0.87	0.86	0.69	0.91	0.89	0.89	0.91	0.87	0.88	0.88	0.88	0.90	0.88	0.87	0.93	0.94	0.89	

Figure 3. Correlation of variant predictor scores. Correlation of predictor scores for the population frequency data set is reported as the Pearson correlation coefficient (blue) and the fraction of consensus binary predictions (FCBPs) (red). Higher scores (darker colors) indicate a better correlation of predictor outputs.

and ProtTrans scores were correlated among themselves and also with supervised methods. ProteinBert (Brandes et al. 2022), ProtBert (Rao et al. 2019), and PLUS-RNN (Min et al. 2021) agreed in their predictions, but only somewhat correlated with other unsupervised methods, suggesting that these pLMs captured a somewhat different signal.

Importantly, PLUS-RNN sequence representations were contextualized in protein structure, as were Seq-UNet predictions, allowing these methods to capture longer-range residue contacts. We thus suggest that Seq-UNet, PLUS-RNN and, by correlation, ProteinBert and ProtBert, predictions may be orthogonal to ESM and ProtTrans unsupervised DL methods. That is, the agreement of these methods on variant effect may be more informative than either one of the methods alone. We further explore this notion with functionally annotated PMD variants below.

Conservation Is Orthogonal to Frequency and Recognized by All Predictors

Many, if not most, of the classical variant effect prediction methods heavily rely on the conservation of variant position across protein homologs. This is warranted as significantly deleterious substitutions could be expected to be eliminated in evolution. In our work, this observation is also supported by the agreement between binary assessments of variants using phastCons versus other methods (Fig. 3, red). However, conservation scores alone are limited in the prediction of nuances of variant effect, as is evident from lower corresponding score correlations (Fig. 3, blue). There are multiple reasons for this observation. First, sufficiently descriptive estimates of per residue conservation are complicated and often limited to large gene/protein families (Triant and Pearson 2015; Malhis et al. 2019). Second, co-occurring mutations across multiple positions may dampen or increase individual variant effects (Holcomb et al. 2021). Finally, position conservation does not easily translate into quantitative descriptions of the severity of the effect (Miller et al. 2017). Here, the nature of a particular variant substitution may ameliorate the impact of affecting a conserved position or worsen the impact of tweaking an unconserved one. Thus, an aspartic to glutamic acid substitution in a conserved negatively charged site may be acceptable, while a serine to tryptophan change in a variable but buried position may be severely disruptive.

We used the population frequency variant set to evaluate how well variant effect predictors capture conservation. We labeled as "conserved" all variants with a phastCons score ≥0.5, and "unconserved" otherwise. We then asked whether variant conservation and population frequency were related terms. That is, we evaluated whether common variants are differently conserved than rare ones. In a discretized comparison, common variants were indeed less conserved than rare and ultra-rare ones; that is, common, rare, ultra-rare, and singleton phastCons score medians were = 0.43, 0.59, 0.73, 0.76, respectively (Fig. 2A). However, there was no significant correlation between conservation scores and frequency of variants in the population (Pearson correlation = −0.09, Spearman correlation = −0.12). These observations suggest that the signal describing the conservation of variant sites is orthogonal to that describing variant frequency.

We further observed that most effect prediction methods distinguished between variants in differentially conserved positions (Fig. 1B). However, classical method performance greatly varied, highlighting the differential emphasis on conservation (as reported by genome-based alignments) in evaluating variant effect—some surprisingly low (SIFT) and some high (CADD). Supervised DL methods also displayed significantly varied performance. As expected, MetaRNN conservation prediction performance was drastically lower than for prediction of variant frequency (Fig. 1A,B), while MVP's performance improved, likely due to the latter's reliance on conservation scores.

We then asked whether conservation can be predicted by methods that do not use it in training. In earlier work, Marquet et al. (2022) and Dunham et al. (2023) found this to be possible. For our data, pLMs appeared to be as good at differentiating variant conservation as they were for variant frequency, but their performance was not as good as that of some classical methods (Fig. 1B) that use conservation for making predictions. Notably, the supervised DL VESPA method performed worse than its baseline language model ProtTransT5, highlighting the fact that recognizing variant effect is not equivalent to recognizing conservation. Also note that ProtTransT5 was significantly better at this task than other pLMs, suggesting that some language models may produce more biologically interpretable embeddings than others.

If predictions of the methods in our study could be considered equivalent to experimental, in vitro or in vivo, analysis, these results would indicate that conservation plays a significant, if not all-encompassing, role in explaining variant effect (i.e., it contributes 0.1–0.3 of overall ROC AUC). However, given that method performance in predicting true variant effect is limited and exceedingly varied by the test/evaluation set used to establish performance metrics, our results suggest that the value of conservation alone is unlikely large, as confirmed by phastCons performance on function-relevant and pathogenic variants (Fig. 1C–E). That is, the signal of billions of years of evolution has to be seen through the prism of more information to be interpreted and applied to effect prediction.

Functional Effect Is a Combination of Many Factors Recognizable by Unsupervised Methods

As expected from the previously described bias in experimental studies, functionally significant changes affect conserved sites somewhat more frequently than unconserved ones (phastCons performance; Fig. 1C,D). Thus, given their ability to differentiate variants by conservation, most methods could be expected to perform at least as well or better in differentiating mutations of functional effect from those of no effect. Note that both effect and neutral (no-effect) variants were found across the full range of conservation scores, somewhat complicating the problem (Supplemental Fig. S3).

All methods were indeed significantly better at identifying knockout (large effects) than mild and moderate effect variants (Fig. 2C; Supplemental Fig. S3). Notably, unsupervised methods were as good as earlier techniques in differentiating the knockout/effect versus no-effect variants. However, all method score distributions were sufficiently overlapping as to often “misidentify” experimentally labeled neutral variants as having an effect (Supplemental Fig. S3). Note that experimentally establishing variants as neutral is a difficult task, with literature reports often disproven in later publications (Bromberg et al. 2013; Zeng and Bromberg 2019).

We found that, at the binary effect threshold, REVEL was excellent at labeling all effect/knockout variants (PMD set of human variants, F1 measure = 0.78; Table 1). Similar behavior was observed for MetaRNN (0.84). Note that MetaRNN was trained to predict pathogenic variants, this result suggesting that only variants of high functional effect in disease genes would be identified as impactful. Its high recall confirms that many of the MetaRNN predicted pathogenic variants were also of high functional effect. On the other hand, VESPA (F1 = 0.62) was trained using the PMD data and hence could be expected to perform well. However, our population-based threshold has somewhat altered its performance in favor of very high precision (90%), but lower recall (47%) predictions. At the default threshold (Supplemental Table S4), VESPA F1 was higher (0.66), while MetaRNN performance lower (0.81); for MVP, the difference between thresholds was even greater (F1 = 0.65 at our threshold, F1 = 0.84 at default threshold; Table 1; Supplemental Table S4). These differences highlight the issues of applying scoring thresholds to methods without optimization for specific tasks.

This method performance may also seem higher than expected in light of the small number of no-effect variants in our data. In fact, all methods mislabeled neutral variants to a certain extent (e.g., MetaRNN labeled 86% of all experimental neutrals as having an effect). VESPA, however, only tagged 16% of the neutral variants incorrectly (Table 1). This mislabeling by all supervised methods highlights the effect of the previously mentioned bias in selecting variants for evaluation—variants in disease genes are more likely to be experimentally evaluated and may be erroneously tagged as neutral on the basis of nonexhaustive experimentation. We note that

Table 1. Method performance at the frequency-established threshold

			PMD					Pathogenicity					DMS				
	Predictor	Thrsh	Prec	rec	F1	%at	%nat	Prec	rec	F1	%at	%nat	Prec	rec	F1	%at	%nat
Classic	CADD	0.48	0.87	0.40	0.55	0.34	0.18	0.90	0.55	0.68	0.36	0.09	0.17	**0.93**	0.29	0.75	0.72
	PolyPhen2	0.99	**0.89**	0.40	0.55	0.34	0.15	0.90	0.48	0.63	0.32	0.08	**0.26**	0.50	**0.35**	0.26	0.22
	REVEL	0.45	0.87	**0.70**	**0.78**	0.61	0.32	0.87	**0.89**	**0.88**	0.61	0.20	0.20	0.87	0.33	0.59	0.54
	SIFT	1.00	0.90	0.32	0.47	0.26	**0.11**	**0.91**	0.46	0.61	0.30	**0.06**	0.09	0.01	0.02	0.02	**0.02**
Supervised DL	MetaRNN	0.09	0.77	**0.94**	**0.84**	0.92	0.86	**0.93**	**0.99**	**0.96**	0.64	0.12	0.14	**0.98**	0.25	0.94	0.93
	MVP	0.92	0.86	0.52	0.65	0.46	0.26	0.90	0.70	0.79	0.54	0.18	0.21	0.58	0.31	0.38	0.35
	Seq-UNet	0.92	0.87	0.09	0.16	0.07	**0.04**	0.82	0.19	0.30	0.14	**0.06**	0.00	0.00	0.00	0.00	**0.00**
	VESPA	0.55	**0.90**	0.47	0.62	0.39	0.16	0.92	0.64	0.76	0.42	0.08	0.34	0.78	**0.48**	0.40	0.32
	VESPAl	0.58	**0.90**	0.40	0.56	0.34	0.13	**0.93**	0.49	0.65	0.32	**0.06**	**0.39**	0.53	0.45	0.24	0.18
	AlphaMissense	0.57	0.88	0.65	0.75	0.55	0.26	0.91	0.80	0.86	0.52	0.11	0.28	0.92	0.43	0.57	0.50
Unsupervised DL	ESM1b	0.65	0.89	0.47	0.62	0.40	0.18	0.90	0.69	0.78	0.46	0.11	0.05	0.04	0.04	0.32	0.38
	ESM1v	0.64	0.89	0.45	0.60	0.38	0.17	0.90	0.67	0.77	0.44	0.10	0.05	0.04	0.04	0.23	0.27
	ESM2	0.68	0.90	0.44	0.59	0.37	0.15	**0.91**	0.59	0.72	0.39	0.09	0.02	0.02	0.02	0.13	0.16
	ProtBERT	0.53	0.89	0.13	0.23	0.11	0.05	0.79	0.13	0.22	0.10	0.05	0.80	0.26	0.26	0.91	0.93
	ProteinBERT	0.43	**0.91**	0.13	0.22	0.11	**0.04**	0.88	0.25	0.39	0.17	0.05	**0.98**	**0.30**	**0.30**	0.99	0.99
	ProtTransALBERT	0.66	0.88	0.42	0.57	0.36	0.18	0.89	0.63	0.74	0.42	0.11	0.02	0.03	0.03	0.16	0.19
	ProtTransBERT	0.66	0.89	0.44	0.58	0.37	0.16	0.90	0.63	0.74	0.42	0.10	0.17	0.09	0.09	0.45	0.51
	ProtTransT5	0.76	0.88	**0.50**	**0.64**	0.42	0.20	0.90	**0.74**	**0.81**	0.49	0.12	0.00	0.01	0.01	0.02	**0.02**
	UniRep	0.07	0.85	0.13	0.23	0.12	0.07	0.81	0.10	0.17	0.07	**0.03**	0.20	0.27	0.27	0.09	0.07
	PLUS-RNN	0.41	0.90	0.16	0.27	0.13	0.05	0.82	0.19	0.31	0.14	0.06	0.91	0.28	0.28	0.97	0.99

(Prec) Precision, (rec) recall, (F1) F measure, (%at) percent of all variants predicted above threshold (positive), (%nat) percent of all negatives (neutrals/no effect) above threshold (false positives).

Best performance values in each class of tools and for each set are highlighted in bold. Note that all values are rounded to the second digit, leading to ~0 values in some entries.

these could potentially be successfully recovered as having an effect using computational analysis.

ESM and ProtTrans pLMs did as well as supervised methods in differentiating distributions of knockout/effect versus no-effect variants (Fig. 1C) and worse than supervised models at our selected cutoff (Table 1). Nevertheless, ProtTransT5, the best performer of all unsupervised models (F1 = 0.64; Table 1), mislabeled only 20% of the neutrals—a performance on par with best-supervised methods. We also note that at the standard (nonfrequency-optimized threshold; Supplemental Table S4), all ESM and ProtTrans models improved in performance as measured by the F1 measure (e.g., ProtTransT5 $F1_{nonoptimized} = 0.84$, as the cost of drastically reduced precision [prec = 0.78 vs. = 0.88, respectively]).

Given the bias in available experimental evaluation data toward variants of high effect, we further evaluated predictions of variant annotations extracted by DMS techniques. Specifically, we considered annotations of two proteins (PTEN and TPMT) (Notin et al. 2022; Supplemental Material). For this set of variants, supervised DL methods performed best (highest ROC AUC; Fig. 1D) across the scoring spectrum. Specifically, the best performers were VESPA (and VESPAl), both pLM (ProtTransT5)-based models, closely followed by AlphaMissense (Fig. 1D; Table 1), also pLM-based. Their improved performance (over unsupervised methods) suggests significant value in fine-tuning.

Given the above results, we suggest that while ESM and ProtTrans pLMs could possibly be used in identifying variant effect, large-scale analysis of variants benefits from more fine-tuned method application. Nevertheless, pLMs use may be particularly meaningful for nonhuman variants, where gene/protein large family alignments and experimental annotations are not readily available.

Prediction Methods Capture Different Signals

Given the results of our score correlation experiments, we asked whether combining methods may produce more precise classification of variants. We evaluated the precision of a jury-of-two method on PMD knockout versus no-effect variants, by only considering an agreement between the two methods as an effect prediction. In fact, asking two methods to agree significantly improved precision, albeit at the cost of recall (Fig. 4).

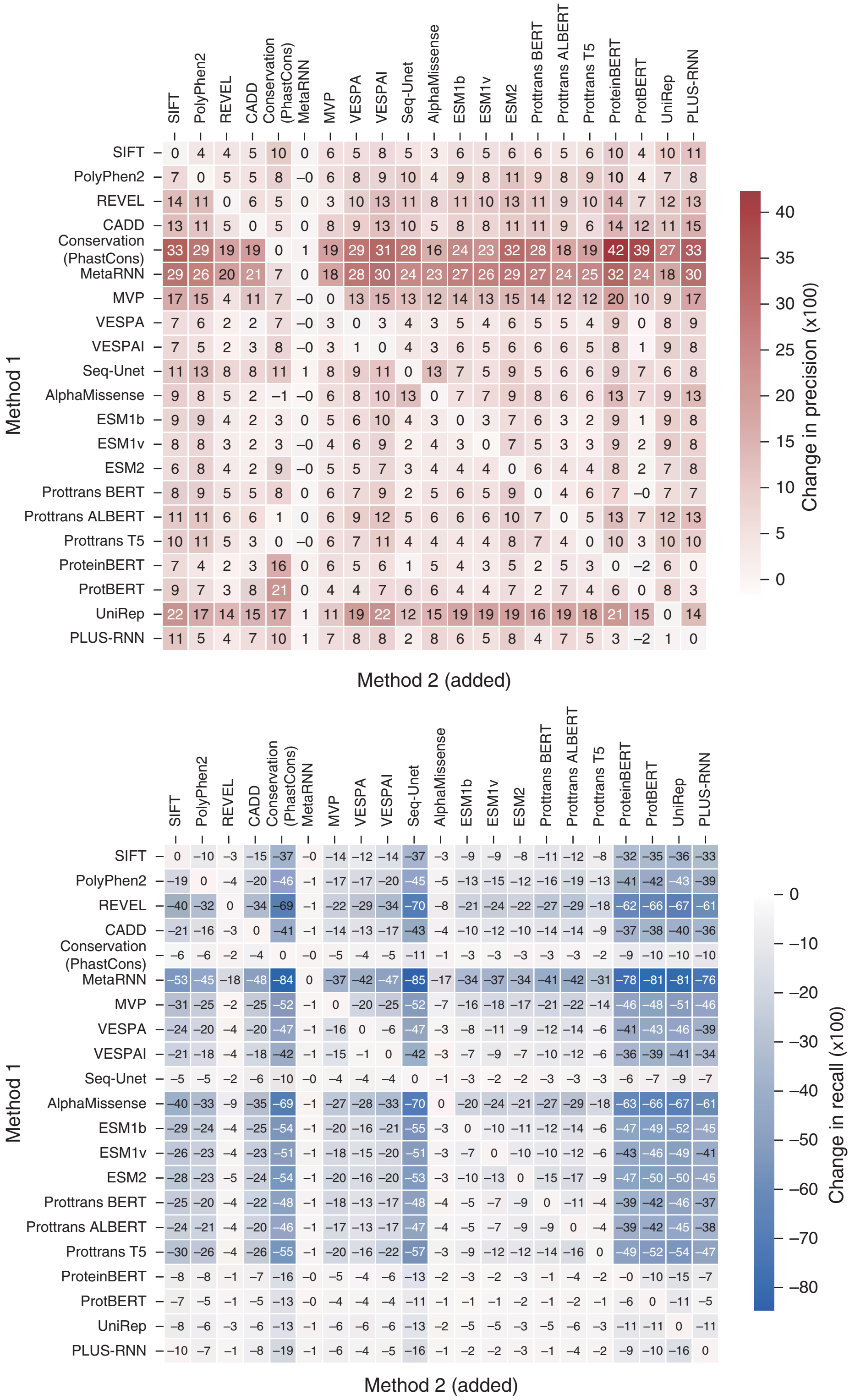
We observed this behavior for almost all combinations of methods. As expected, methods that performed worse on their own, got more of a boost. phastCons, for example, greatly benefitted from the addition of almost any other method's input—even without major cost to recall. Some interesting combinations were present (e.g., REVEL—an ensemble method including SIFT, PolyPhen, and CADD scores, still benefitted from the addition of either of these methods), albeit at a significant >15% cost to recall.

We did not expect much improvement from combining pLMs with correlated scores. However, adding ESM2 (Lin et al. 2023) to ProtTransT5, did result in an 8% gain in precision and a 12% loss in recall. This observation suggests that scores could be further fine-tuned to eke out only the (few) high-reliability variants in each set. Adding uncorrelated unsupervised model scores (PLUS-RNN or ProteinBERT) was somewhat beneficial for ProtTransALBERT precision (adding 13%), but less so for ProtTransT5 or ESM2 precision (adding 8%–10%), while drastically reducing recall for all (by 38%–47%).

These results suggest that relying on the differences in latent spaces described by individual unsupervised models does not necessarily improve variant effect capture. More analysis of data set selection and parameter optimization choices is necessary to define the unique characteristics of pLMs that care about individual "words" (residues) in protein "sentences" (sequences).

Unsupervised Methods Clearly Capture Signals of Pathogenicity

Identifying variant pathogenicity has long been the focus of human genetics and a major driver of research initiatives. Currently, two major resources provide information about designations of variant pathogenicity—ClinVar and ClinGen (Rehm et al. 2015). ClinVar focuses on variant pathogenicity, as recommended by ACMG–



	SIFT	PolyPhen2	REVEL	CADD	Conservation (PhastCons)	MetaRNN	MVP	VESPA	VESPAI	Seq-Unet	AlphaMissense	ESM1b	ESM1v	ESM2	Prottrans BERT	Prottrans ALBERT	Prottrans T5	ProteinBERT	ProtBERT	UniRep	PLUS-RNN
SIFT	0	4	4	5	10	0	6	5	8	5	3	6	5	6	6	5	6	10	4	10	11
PolyPhen2	7	0	5	5	8	−0	6	8	9	10	4	9	8	11	9	8	9	10	4	7	8
REVEL	14	11	0	6	5	0	3	10	13	11	8	11	10	13	11	9	10	14	7	12	13
CADD	13	11	5	0	5	0	8	9	13	10	5	8	8	11	11	9	6	14	12	11	15
Conservation (PhastCons)	33	29	19	19	0	1	19	29	31	28	16	24	23	32	28	18	19	42	39	27	33
MetaRNN	29	26	20	21	7	0	18	28	30	24	23	27	26	29	27	24	25	32	24	18	30
MVP	17	15	4	11	7	−0	0	13	15	13	12	14	13	15	14	12	12	20	10	9	17
VESPA	7	6	2	2	7	−0	3	0	3	4	3	5	4	6	5	5	4	9	0	8	9
VESPAI	7	5	2	3	8	−0	3	1	0	4	3	6	5	6	6	6	5	8	1	9	8
Seq-Unet	11	13	8	8	11	1	8	9	11	0	13	7	5	9	5	6	6	9	7	6	8
AlphaMissense	9	8	5	2	−1	−0	6	8	10	13	0	7	7	9	8	6	6	13	7	9	13
ESM1b	9	9	4	2	3	0	5	6	10	4	3	0	3	7	6	3	2	9	1	9	8
ESM1v	8	8	3	2	3	−0	4	6	9	2	4	3	0	7	5	3	3	9	2	9	8
ESM2	6	8	4	2	9	−0	5	5	7	3	4	4	4	0	6	4	4	8	2	7	8
Prottrans BERT	8	9	5	5	8	0	6	7	9	2	5	6	5	9	0	4	6	7	−0	7	7
Prottrans ALBERT	11	11	6	6	1	0	6	9	12	5	6	6	6	10	7	0	5	13	7	12	13
Prottrans T5	10	11	5	3	0	−0	6	7	11	4	4	4	4	8	7	4	0	10	3	10	10
ProteinBERT	7	4	2	3	16	0	6	5	6	1	5	4	3	5	2	5	3	0	−2	6	0
ProtBERT	9	7	3	8	21	0	4	4	7	6	6	4	4	7	2	7	4	6	0	8	3
UniRep	22	17	14	15	17	1	11	19	22	12	15	19	19	19	16	19	18	21	15	0	14
PLUS-RNN	11	5	4	7	10	1	7	8	8	2	8	6	5	8	4	7	5	3	−2	1	0

	SIFT	PolyPhen2	REVEL	CADD	Conservation (PhastCons)	MetaRNN	MVP	VESPA	VESPAI	Seq-Unet	AlphaMissense	ESM1b	ESM1v	ESM2	Prottrans BERT	Prottrans ALBERT	Prottrans T5	ProteinBERT	ProtBERT	UniRep	PLUS-RNN
SIFT	0	−10	−3	−15	−37	−0	−14	−12	−14	−37	−3	−9	−9	−8	−11	−12	−8	−32	−35	−36	−33
PolyPhen2	−19	0	−4	−20	−46	−1	−17	−17	−20	−45	−5	−13	−15	−12	−16	−19	−13	−41	−42	−43	−39
REVEL	−40	−32	0	−34	−69	−1	−22	−29	−34	−70	−8	−21	−24	−22	−27	−29	−18	−62	−66	−67	−61
CADD	−21	−16	−3	0	−41	−1	−14	−13	−17	−43	−4	−10	−12	−10	−14	−14	−9	−37	−38	−40	−36
Conservation (PhastCons)	−6	−6	−2	−4	0	−0	−5	−4	−5	−11	−1	−3	−3	−3	−3	−3	−2	−9	−10	−10	−10
MetaRNN	−53	−45	−18	−48	−84	0	−37	−42	−47	−85	−17	−34	−37	−34	−41	−42	−31	−78	−81	−81	−76
MVP	−31	−25	−2	−25	−52	−1	0	−20	−25	−52	−7	−16	−18	−17	−21	−22	−14	−46	−48	−51	−46
VESPA	−24	−20	−4	−20	−47	−1	−16	0	−6	−47	−3	−8	−11	−9	−12	−14	−6	−41	−43	−46	−39
VESPAI	−21	−18	−4	−18	−42	−1	−15	−1	0	−42	−3	−7	−9	−7	−10	−12	−6	−36	−39	−41	−34
Seq-Unet	−5	−5	−2	−6	−10	−0	−4	−4	−4	0	−1	−3	−2	−2	−3	−3	−3	−6	−7	−9	−7
AlphaMissense	−40	−33	−9	−35	−69	−1	−27	−28	−33	−70	0	−20	−24	−21	−27	−29	−18	−63	−66	−67	−61
ESM1b	−29	−24	−4	−25	−54	−1	−20	−16	−21	−55	−3	0	−10	−11	−12	−14	−6	−47	−49	−52	−45
ESM1v	−26	−23	−4	−23	−51	−1	−18	−15	−20	−51	−3	−7	0	−10	−10	−12	−6	−43	−46	−49	−41
ESM2	−28	−23	−5	−24	−54	−1	−20	−16	−20	−53	−3	−10	−13	0	−15	−17	−9	−47	−50	−50	−45
Prottrans BERT	−25	−20	−4	−22	−48	−1	−18	−13	−17	−48	−4	−5	−7	−9	0	−11	−4	−39	−42	−46	−37
Prottrans ALBERT	−24	−21	−4	−20	−46	−1	−17	−13	−17	−47	−4	−5	−7	−9	−9	0	−4	−39	−42	−45	−38
Prottrans T5	−30	−26	−4	−26	−55	−1	−20	−16	−22	−57	−3	−9	−12	−12	−14	−16	0	−49	−52	−54	−47
ProteinBERT	−8	−8	−1	−7	−16	−0	−5	−4	−6	−13	−2	−3	−2	−3	−1	−4	−2	−0	−10	−15	−7
ProtBERT	−7	−5	−1	−5	−13	−0	−4	−4	−4	−11	−1	−1	−1	−2	−1	−2	−1	−6	0	−11	−5
UniRep	−8	−6	−3	−6	−13	−1	−6	−6	−6	−13	−2	−5	−5	−3	−5	−6	−3	−11	−11	0	−11
PLUS-RNN	−10	−7	−1	−8	−19	−1	−6	−4	−5	−16	−1	−2	−2	−3	−1	−4	−2	−9	−10	−16	0

Figure 4. Improved precision at cost of recall by method jury. Most predictors' precision (red) benefits and recall (blue) suffers from the addition of another method in identifying severe (knockout) variants.

AMP (Richards et al. 2015) for variants interpreted for Mendelian conditions. ClinGen, on the other hand, is primarily concerned with establishing gene-disease involvement. Variant pathogenicity is then determined by evaluating genetic variants using the ACMG–AMP guidelines and gene-specific criteria developed by ClinGen expert panels. Note that in a sort of circular logic, disease genes are often defined on the basis of the pathogenic variants they carry.

For Mendelian/monogenic disorders, the process for establishing variant pathogenicity is a relatively well-defined, if laborious, task. For complex disorders, clinical observations and the experimental evidence of variant effect need to be overwhelming. As such, at the time of writing of this article (July 2023), 5312 variants in 2599 genes have been curated by ClinGen (ClinGen Statistics 2023), of which 2546 were designated as pathogenic or likely pathogenic. ClinVar, on the other hand, has collected 15,410 variants labeled by curators and nearly two million with some level of annotation (ClinVar Statistics 2023). Note that even given all precautions, a sufficient number of designated pathogenic variants have been observed in seemingly healthy individuals, suggesting the involvement of supporting and/or alternative molecular pathways (Shah et al. 2018). This biological incongruence is arguably even more pronounced in larger, literature-curated repositories of pathogenicity data (Cassa et al. 2013) (e.g., Human Genome Mutation Database [HGMD] [Stenson et al. 2020]). Nevertheless, in this work, we assumed that, regardless of the possible mislabeling of some variants, the full collection of ACMG–AMP guidelines-verified pathogenic variants is greatly enriched in disease-causing mutations. Note that to evaluate the methods' performance on a set that would not have likely overlapped with the methods' training data, we limited the extraction of pathogenic variants to those identified after 2022 (Supplemental Material).

Identifying putatively non-disease-causing variants for comparison to pathogenic ones was nearly impossible. Given the ACNG requirements for classifying variants as benign, we do not expect that these variants' characteristics (e.g., high population frequency, experimental data showing lack of functional effect, and nonsegregation with known disease) could in any way significantly overlap with those of the putatively pathogenic variants. Moreover, accepting multiple lines of computational evidence as a strong support for the likely benign-ness of variants is logically circuitous—train predictors to recognize benign variants and then label them as likely benign. Thus, in this work, we simply asked whether pathogenic variants can be recognized by the existing methods as different from the rest of the variants observed in the human population.

In our evaluation, methods differentiated likely pathogenic versus baseline variants with similar accuracy as pathogenic versus baseline variants (Fig. 1E; Supplemental Fig. S4). It thus stands to reason that curated likely pathogenic variants are indeed pathogenic according to the current criteria or that the predictor resolution is insufficient to tell the difference between the two. The first inference is more likely, given the inability of the variant and protein characteristic-naive, unsupervised methods to significantly better label pathogenic variants than likely pathogenic ones. This observation further suggests that the current process of pathogenic variant accumulation is either near perfect or heavily biased by the used experimental and clinical techniques.

All classical methods, however designed, explicitly select features of variants and their host protein sequences (e.g., conservation, structure, solvent accessibility, etc.) to attempt capture of variant effect. As the pathogenicity of a variant generally translates into a large effect on protein function, if not vice versa, these tools consistently did better in our hands-on differentiating pathogenic variants from putatively benign ones than functionally impactful variants from neutral ones (higher ROC AUC; Fig. 1C,E). Even simply using conservation was somewhat more informative for the pathogenicity set of variants than for the functional effect set. However, conservation alone was insufficient to precisely differentiate pathogenic variants, suggesting that they are not, contrary to expectations, confined to the strongly conserved sites.

Of all supervised methods, MVP and SeqUNet were the worst performers for the pathogenicity set, but even they attained an ROC AUC = ~0.7 (Fig. 1E). Note that the contribution of variant rarity in defining pathogenicity was illustrated by MetaRNN performance, that is, a method that considers population frequency explicitly and was thus able to differentiate baseline (more frequent) variants from pathogenic ones almost perfectly.

Of the unsupervised DL methods, ESM and ProtTrans were able to differentiate clinically significant variants from the general population better than simply using conservation. These models were also as good as or better than many of the supervised methods. At our selected thresholds, all ESM and ProtTrans models did well in recognizing pathogenicity (F1 ≥ 0.72, Table 1), and were better for this set than at predicting functional effect. This observation once again reaffirms that pathogenic variants are of high functional effect and are almost never common. However, as neither of the unsupervised methods captured variant population frequency well (Fig. 1A), the rarity of pathogenic variants is an unlikely cause of these models' pathogenicity classification abilities.

Variant Effects Are Correlated but Not Interchangeable for Effect Prediction

The ability of all methods in our study to differentiate variant effect across effect types is worth exploring further (Fig. 5; Supplemental Fig. S5). For example, all methods note that large-scale functional effect (purple lines) is almost as bad as pathogenicity (red lines). Furthermore, most methods frequently label variants found in the population, regardless of their frequency, as having no effect. In fact, variants that are annotated as having no functional effect (green line) are more frequently predicted to have an effect by all methods than ultra-rare variants (orange dashed line in Fig. 5; Supplemental Fig. S5).

Our graphs also tell us that optimal threshold selection, rarely considered by new methods in the field, is a difficult to capture. However, establishing this threshold is a necessary exercise for allowing the evaluation of individual variants. That is, a higher ROC AUC of the method is

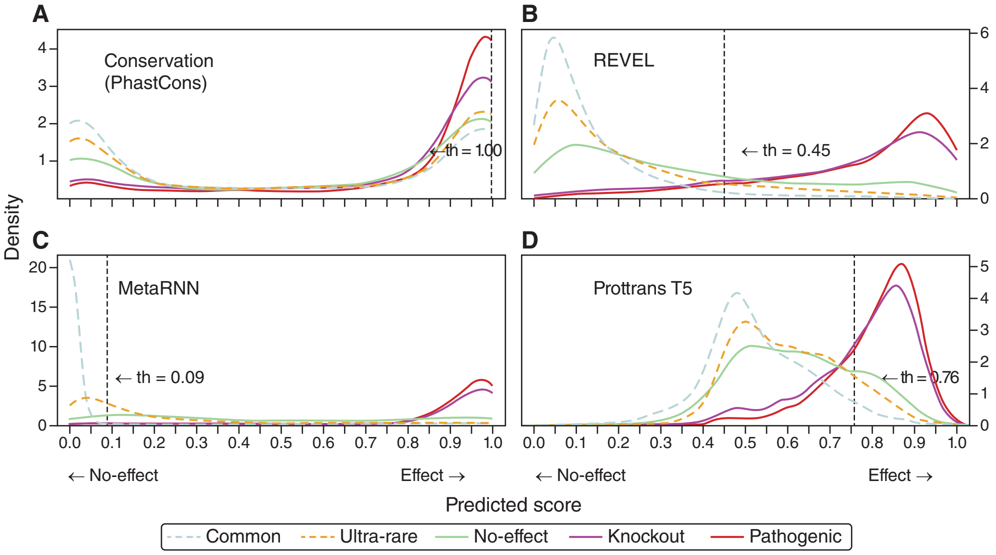


Figure 5. Distribution of scores for different variant classes. Scores attained by variants in different population frequency, functional effect, and pathogenicity classes (line colors) from (*A*) PhastCons, our conservation score, (*B*) REVEL, an example of a classical method in our set, (*C*) MetaRNN, an example of supervised deep learner, and (*D*) ProtTransT5, an example of an unsupervised predictor. Thresholds indicate scores below which 95% of common variants were observed. Representative methods were selected from the complete set (Supplemental Fig. S2) by best performance (ROC AUC) in differentiating population variant frequencies.

meaningless to a scientist looking for an assessment of the effect of their variant of interest. Furthermore, comparisons of method performance on different data sets and at differently selected cutoffs are bound to bring different performance results, making the selection of the best variant effect predictor nearly impossible (Fig. 1; Table 1; Supplemental Table S4).

Our results indicate that (some) unsupervised methods capture more than variant population frequency or variant site conservation. Instead, they seem to reflect functionally relevant features of variants, learning to extract information directly from the language of life. This ability allows for their correct labeling of pathogenic variants as well. Interpreting their assessments as binary classifications of a particular variant, however, requires a much deeper understanding of what makes an SNV unacceptable to the protein or organism it affects. It is also bound to broaden our horizons, allowing for the evaluation of variants in less well-annotated genes—a significant gain for exploration of, for example, the bacterial world.

CONCLUDING REMARKS

Multiple methods have been developed for annotation of the effect of missense variants. Appearance of unsupervised models has produced yet more of such methods. Our evaluation of method performance suggests that ESM and ProtTrans-based methods are the best performers in this space, exhibiting similar or better performance than specifically trained tools. However, neither of the existing supervised or unsupervised methods is able to evaluate all variant effect correctly.

To improve performance, better definitions of variant effect, as well as larger training sets for model fine-tuning are still necessary. The future "gold standard" predictor should indeed identify variant effect precisely. However, we hope that these models are also able to speed up discovery by labeling impactful variants in less studied spaces before experimental annotations catch up.

ACKNOWLEDGMENTS

We thank Drs. Michael Heinzinger (Technical University of Munich), Celine Marquet (Technical University of Munich), Ethan C. Alley (Alvea), and Mohammed AlQuraishi (Columbia University) for help with running and understanding their methods. A huge thanks goes to Christian Dallago (NVIDIA) for the invitation to participate in the special issue. A particular thanks is due the CSH Press Perspectives editor Barbara Acosta for her patience as this work was being delayed over and over again. Finally, we thank all researchers who deposit their data into public databases and those who maintain these databases. Y.B.'s and R.P.'s work on this manuscript was funded by the Emory Start-up Funding and NASA Astrobiology Institute Cycle 8 (80NSSC18M0093). All gene/protein lists, scores, and method evaluation measures are available from Bromberg et al. (2023).

REFERENCES

1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248–249. doi:10.1038/nmeth0410-248

Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. 2019. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* **16:** 1315–1322. doi:10.1038/s41592-019-0598-1

Araya CL, Fowler DM. 2011. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol* **29:** 435–442. doi:10.1016/j.tibtech.2011.04.003

Bairoch A, Boeckmann B, Ferro S, Gasteiger E. 2004. Swiss-Prot: juggling between evolution and stability. *Brief Bioinform* **5:** 39–55. doi:10.1093/bib/5.1.39

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31:** 365–370. doi:10.1093/nar/gkg095

Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38:** 2102–2110. doi:10.1093/bioinformatics/btac020

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

Bromberg Y, Kahn PC, Rost B. 2013. Neutral and weakly nonneutral sequence variants may define individuality. *Proc Natl Acad Sci* **110:** 14255–14260. doi:10.1073/pnas.1216613110

Bromberg Y, Kabir A, Ramakrishnan P, Shehu A. 2023. Variant prediction in the age of machine learning. figshare doi:10.6084/m9.figshare.c.6746316.v2

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

Cassa CA, Tong MY, Jordan DM. 2013. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum Mutat* **34:** 1216–1220. doi:10.1002/humu.22375

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381:** eadg7492. doi:10.1126/science.adg7492

ClinGen Statistics. 2023. https://search.clinicalgenome.org/kb/reports/stats

ClinVar Statistics. 2023. https://www.ncbi.nlm.nih.gov/clinvar/submitters

Dunham AS, Beltrao P, AlQuraishi M. 2023. High-throughput deep learning variant effect prediction with Sequence UNET. *Genome Biol* **24:** 110. doi:10.1186/s13059-023-02948-3

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. 2022. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* **44:** 7112–7127. doi:10.1109/TPAMI.2021.3095381

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

Ganesan K, Kulandaisamy A, Binny Priya S, Gromiha MM. 2019. Huvarbase: a human variant database with comprehensive information at gene and protein levels. *PLoS ONE* **14:** e0210475. doi:10.1371/journal.pone.0210475

Gibson G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet* **13:** 135–145. doi:10.1038/nrg3118

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* **15:** 57–61. doi:10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G

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

Holcomb D, Hamasaki-Katagiri N, Laurie K, Katneni U, Kames J, Alexaki A, Bar H, Kimchi-Sarfaty C. 2021. New approaches to predict the effect of co-occurring variants on protein characteristics. *Am J Hum Genet* **108:** 1502–1511. doi:10.1016/j.ajhg.2021.06.011

Hu J, Ng PC. 2013. SIFT indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS ONE* **8:** e77940. doi:10.1371/journal.pone.0077940

Hu Z, Yu C, Furutsuki M, Andreoletti G, Ly M, Hoskins R, Adhikari AN, Brenner SE. 2019. VIPdb, a genetic variant impact predictor database. *Hum Mutat* **40:** 1202–1214. doi:10.1002/humu.23858

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* **99:** 877–885. doi:10.1016/j.ajhg.2016.08.016

Katsanis N. 2016. The continuum of causality in human genetic disorders. *Genome Biol* **17:** 233. doi:10.1186/s13059-016-1107-9

Kawabata T, Ota M, Nishikawa K. 1999. The protein mutant database. *Nucleic Acids Res* **27:** 355–357. doi:10.1093/nar/27.1.355

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46:** 310–315. doi:10.1038/ng.2892

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46:** D1062–D1067. doi:10.1093/nar/gkx1153

Li C, Zhi D, Wang K, Liu X. 2022. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Med* **14:** 115. doi:10.1186/s13073-022-01120-z

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379:** 1123–1130. doi:10.1126/science.ade2574

Mahlich Y, Reeb J, Hecht M, Schelling M, De Beer TAP, Bromberg Y, Rost B. 2017. Common sequence variants affect molecular function more than rare variants? *Sci Rep* **7:** 1608. doi:10.1038/s41598-017-01054-2

Malhis N, Jones SJM, Gsponer J. 2019. Improved measures for evolutionary conservation that exploit taxonomy distances. *Nat Commun* **10:** 1556. doi:10.1038/s41467-019-09583-2

Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, Nechaev D, Rost B. 2022. Embeddings from protein language models predict conservation and variant effects. *Hum Genet* **141:** 1629–1647. doi:10.1007/s00439-021-02411-y

Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv Neural Inf Process Syst* 29287–29303.

Miller M, Bromberg Y, Swint-Kruse L. 2017. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci Rep* **7:** 41329. doi:10.1038/srep41329

Miller M, Vitale D, Kahn PC, Rost B, Bromberg Y. 2019. Funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res* **47:** e142. doi:10.1093/nar/gkz818

Min S, Park S, Kim S, Choi HS, Lee B, Yoon S. 2021. Pretraining of deep bidirectional protein sequence representations with structural information. *IEEE Access* **9:** 123912–123926. doi:10.1109/ACCESS.2021.3110269

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812–3814. doi:10.1093/nar/gkg509

Nishikawa K, Ishino S, Takenaka H, Norioka N, Hirai T, Yao T, Seto Y. 1994. Constructing a protein mutant database. *Protein Eng* **7:** 733. doi:10.1093/protein/7.5.733

Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, Gal Y. 2022. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11:** R52. doi:10.1186/gb-2010-11-5-r52

Pejaver V, Babbi G, Casadio R, Folkman L, Katsonis P, Kundu K, Lichtarge O, Martelli PL, Miller M, Moult J, et al. 2019. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. *Hum Mutat* **40:** 1495–1506. doi:10.1002/humu.23838

Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, et al. 2010. The characterization of twenty sequenced human genomes. *PLoS Genet* **6:** e1001111. doi:10.1371/journal.pgen.1001111

Phan L, Jin Y, Zhang H, Qiang W, Shekhtman E, Shao D, Revoe D, Villamarin R, Ivanchenko E, Kimura M. 2020. ALFA: Allele frequency aggregator. NCBI, U.S. NLM Gatew. http://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, Weatherall DJ, Hay SI. 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat Commun* **1:** 104. doi:10.1038/ncomms1104

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

Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, Abbeel P, Song YS. 2019. Evaluating protein transfer learning with TAPE. *Adv Neural Inf Process Syst* **32:** 9689–9701.

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, et al. 2015. Clingen—the clinical genome resource. *N Engl J Med* **372:** 2235–2242. doi:10.1056/NEJMsr1406261

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17:** 405–424. doi:10.1038/gim.2015.30

Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* **15:** 816–822. doi:10.1038/s41592-018-0138-4

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* **118:** e2016239118: doi:10.1073/pnas.2016239118

Sarkar A, Yang Y, Vihinen M. 2020. Variation benchmark datasets: update, criteria, quality and applications. *Database* **2020:** baz117. doi:10.1093/database/baz117

Savojardo C, Babbi G, Martelli PL, Casadio R. 2021. Mapping OMIM disease–related variations on protein domains reveals an association among variation type, Pfam models, and disease classes. *Front Mol Biosci* **8:** 617016. doi:10.3389/fmolb.2021.617016

Shah N, Hou YC, Yu HC, Sainger R, Caskey CT, Venter JC, Telenti A. 2018. Identification of misclassified ClinVar Variants via disease population prevalence. *Am J Hum Genet* **102:** 609–619. doi:10.1016/j.ajhg.2018.02.019

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050. doi:10.1101/gr.3715005

Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, et al. 2020. The Human Gene Mutation Database (HGMD): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* **139:** 1197–1207. doi:10.1007/s00439-020-02199-3

Sun H, Yu G. 2019. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci Rep* **9:** 1667. doi:10.1038/s41598-018-38189-9

The Critical Assessment of Genome Interpretation Consortium. 2022. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. arXiv doi:10.48550/arXiv.2205.05897

Triant DA, Pearson WR. 2015. Most partial domains in proteins are alignment and annotation artifacts. *Genome Biol* **16:** 99. doi:10.1186/s13059-015-0656-7

Zeng Z, Bromberg Y. 2019. Predicting functional effects of synonymous variants: a systematic review and perspectives. *Front Genet* **10:** 914. doi:10.3389/fgene.2019.00914

Zhu C, Miller M, Zeng Z, Wang Y, Mahlich Y, Aptekmann A, Bromberg Y. 2020. Computational approaches for unraveling the effects of variation in the human genome and microbiome. *Annu Rev Biomed Data Sci* **3:** 411–432. doi:10.1146/annurev-biodatasci-030320-041014

Cold Spring Harbor Perspectives in Biology

Variant Effect Prediction in the Age of Machine Learning

Yana Bromberg, R. Prabakaran, Anowarul Kabir and Amarda Shehu

Cold Spring Harb Perspect Biol published online April 15, 2024

Subject Collection Machine Learning for Protein Science and Engineering

Variant Effect Prediction in the Age of Machine Learning
Yana Bromberg, R. Prabakaran, Anowarul Kabir, et al.

Protein Design Using Structure-Prediction Networks: AlphaFold and RoseTTAFold as Protein Structure Foundation Models
Jue Wang, Joseph L. Watson and Sidney L. Lisanza

Petabase-Scale Homology Search for Structure Prediction
Sewon Lee, Gyuri Kim, Eli Levy Karin, et al.

Environmental Impacts of Machine Learning Applications in Protein Science
Loïc Lannelongue and Michael Inouye

Engineering Proteins Using Statistical Models of Coevolutionary Sequence Information
Jerry C. Dinan, James W. McCormick and Kimberly A. Reynolds

Building Representation Learning Models for Antibody Comprehension
Justin Barton, Aretas Gaspariunas, Jacob D. Galson, et al.

Is Novelty Predictable?
Clara Fannjiang and Jennifer Listgarten

Exploring the Protein Sequence Space with Global Generative Models
Sergio Romero-Romero, Sebastian Lindner and Noelia Ferruz

For additional articles in this collection, see http://cshperspectives.cshlp.org/cgi/collection/