Human Genetics (2025) 144:281–293
https://doi.org/10.1007/s00439-025-02732-2

ORIGINAL INVESTIGATION

Critical assessment of missense variant effect predictors on disease-relevant variant data

Ruchir Rastogi[1] · Ryan Chung[2] · Sindy Li[3] · Chang Li[4] · Kyoungyeul Lee[5] · Junwoo Woo[5] · Dong-Wook Kim[5] · Changwon Keum[5] · Giulia Babbi[6] · Pier Luigi Martelli[6] · Castrense Savojardo[6] · Rita Casadio[6] · Kirsley Chennen[7] · Thomas Weber[7] · Olivier Poch[7] · François Ancien[8,9] · Gabriel Cia[8,9] · Fabrizio Pucci[8,9] · Daniele Raimondi[10,19] · Wim Vranken[9,11] · Marianne Rooman[8,9] · Céline Marquet[12] · Tobias Olenyi[12] · Burkhard Rost[12] · Gaia Andreoletti[3,18] · Akash Kamandula[13] · Yisu Peng[13] · Constantina Bakolitsa[3] · Matthew Mort[14] · David N. Cooper[14] · Timothy Bergquist[15] · Vikas Pejaver[15,16] · Xiaoming Liu[4] · Predrag Radivojac[13] · Steven E. Brenner[2,3] · Nilah M. Ioannidis[1,2,17]

Received: 6 June 2024 / Accepted: 7 February 2025 / Published online: 21 March 2025

Abstract

Regular, systematic, and independent assessments of computational tools that are used to predict the pathogenicity of missense variants are necessary to evaluate their clinical and research utility and guide future improvements. The Critical Assessment of Genome Interpretation (CAGI) conducts the ongoing Annotate-All-Missense (Missense Marathon) challenge, in which missense variant effect predictors (also called variant impact predictors) are evaluated on missense variants added to disease-relevant databases following the prediction submission deadline. Here we assess predictors submitted to the CAGI 6 Annotate-All-Missense challenge, predictors commonly used in clinical genetics, and recently developed deep learning methods. We examine performance across a range of settings relevant for clinical and research applications, focusing on different subsets of the evaluation data as well as high-specificity and high-sensitivity regimes. Our evaluations reveal notable advances in current methods relative to older, well-cited tools in the field. While meta-predictors tend to outperform their constituent individual predictors, several newer individual predictors perform comparably to commonly used meta-predictors. Predictor performance varies between high-specificity and high-sensitivity regimes, highlighting that different methods may be optimal for different use cases. We also characterize two potential sources of bias. Predictors that incorporate allele frequency as a predictive feature tend to have reduced performance when distinguishing pathogenic variants from very rare benign variants, and predictors trained on pathogenicity labels from curated variant databases often inherit gene-level label imbalances. Our findings help illuminate the clinical and research utility of modern missense variant effect predictors and identify potential areas for future development.

Introduction

Predicting the significance of genetic variation is an ongoing challenge that is essential for determining genetic susceptibility to disease and identifying causal variants in rare disease diagnosis (Critical Assessment of Genome Interpretation Consortium 2024). Clinical sequencing laboratories often struggle with the interpretation of low-frequency, rare, and *de novo* variants seen in patients, classifying them as variants of uncertain significance (VUS) due to a lack of available evidence about their pathogenicity. Interpretation of missense variants is of particular interest due to their frequent occurrence and wide range of potential effects on protein function and clinical phenotypes, ranging from no effect to either an adaptive effect or a highly penetrant pathogenic loss or gain of function (Rost et al. 2016).

To address this challenge, many computational tools—collectively termed missense variant effect predictors or variant impact predictors—have been developed over the past three decades to predict the consequences of missense variants. The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have developed guidelines for classification of clinical variants as pathogenic or benign, which provide rules for integrating numerous lines of evidence, including predictions from computational tools as

Extended author information available on the last page of the article

supporting evidence (Richards et al. 2015). More recently, ClinGen provided updated recommendations, advising that some independently calibrated computational tools provide stronger levels of evidence for high-scoring variants (Pejaver et al. 2022; Stenton et al. 2024; Bergquist et al. 2024).

These computational tools leverage variant- and gene-level annotations, such as evolutionary conservation and protein structural properties, as predictive features (Hu et al. 2019; Katsonis et al. 2022). Variant effect predictors have varied goals: for example, some are trained to predict a variant's potential to disrupt protein molecular function, others attempt to infer effects on organismal fitness, while many, especially those trained on clinical databases, aim to predict variant effects on disease (pathogenicity). In this study, we evaluate all types of computational methods on their performance in pathogenicity prediction. Predictions from individual tools often disagree, which motivated the development of ensemble methods, or meta-predictors, trained to aggregate predictions from multiple tools. Meta-predictors tend to have improved performance over their component predictors, but rely on the continued development of individual predictors that incorporate information from new or complementary predictive features. Therefore, it is important to assess the performance of both meta-predictors and individual predictors on the task of missense variant pathogenicity prediction.

The goal of the CAGI Annotate-All-Missense challenge, also termed the Missense Marathon, is to conduct an ongoing assessment of missense pathogenicity predictors (both existing tools and those newly submitted to the CAGI challenge) using variants that have been classified as pathogenic or benign in clinical variant databases or identified as disease-causing variants since the close of the most recent challenge. Teams submitting to the challenge were asked to provide prediction scores for all possible missense single nucleotide variants (SNVs) in the human reference genome, based on dbNSFP v4 (Liu et al. 2020), similar to existing missense variant effect predictors with precomputed scores (Lin et al. 2024). A preliminary, limited assessment of missense predictors was previously performed as part of CAGI 5 (Critical Assessment of Genome Interpretation Consortium 2024). Here we perform a more extensive analysis of missense variant effect predictors for the challenge in CAGI 6, using variants with pathogenicity information made available between November 2021 and April 2023 in our evaluation set.

Results

Evaluation dataset

We curated a dataset of low-frequency (allele frequency < 0.05) missense variants classified in ClinVar (Landrum et al. 2018) as either Pathogenic or Benign with at least one star (excluding those with conflicting assertions) or listed as disease-causing (DM) entries in the Human Gene Mutation Database (HGMD) (Stenson et al. 2020). Our dataset was restricted to variants that were newly added to these databases after the close of the CAGI 6 Annotate-All-Missense challenge in October 2021. We consider the ClinVar and HGMD data both together and separately in the analyses below, to explore differences between the databases. Variants with pathogenicity information available in ClinVar, HGMD, or UniProt (McGarvey et al. 2019) prior to the close of the challenge were explicitly excluded. Common variants were removed, as they should be considered benign per the updated standalone BA1 rule in the ACMG/AMP guidelines (Ghosh et al. 2018). Additional details of dataset construction are provided in Methods. The resulting dataset contained 6,103 pathogenic and 4,353 benign variants from 2,115 genes, with an allele frequency distribution shown in Fig. S1.

Missense variant effect predictors

We evaluated the performance of 60 missense variant effect predictors, of which 12 were submitted by 6 teams to the CAGI 6 Annotate-All-Missense challenge. The additional tested methods include predictors commonly used by the clinical genetics community and recently developed deep learning methods for missense variant interpretation (listed in Table S1 and Methods). All assessed predictors were either released before the close of the challenge or did not train on variant pathogenicity data released after the challenge ended, to ensure no overlap with the evaluation set. Nonetheless, other more subtle forms of circularity may exist (e.g. Grimm et al. (2015)) and are discussed in more detail below. We also note that some predictors were trained or fine-tuned on variants from population databases such as gnomAD (Karczewski et al. 2020), which contain allele frequency information but not clinical classifications; therefore, we did not specifically exclude such variants from the evaluation set.

For presentation clarity, in most analyses, we show results for a select subset of 26 predictors: the top-performing model from each team from the CAGI 6 challenge, predictors widely used by the clinical genetics community, and recently developed methods that have garnered interest. We note that 5 of the 6 top-performing CAGI 6 team submissions are (nearly) identical to previously published methods—3Cnet (Won et al. 2021), MetaRNN (Li et al. 2022), MISTIC (Chennen et al. 2020), SNPs&GO (Calabrese et al. 2009), and VESPAl (Marquet et al. 2022). In figures, we label them with both their familiar method name and the submitting team identifier (e.g. 3Cnet/3billion). The sixth team submission, labeled

as DEOGEN2/(IB)2, uses DEOGEN2 (Raimondi et al. 2017) scores when available and rescaled PROVEAN (Choi et al. 2012) scores when not. Of the remaining 20 highlighted predictors, 15 are commonly used in clinical genetics and research applications: BayesDel (with and without allele frequency) (Feng 2017), CADD (Rentzsch et al. 2019), ClinPred (Alirezaie et al. 2018), Eigen (Ionita-Laza et al. 2016), FATHMM-XF (Rogers et al. 2018), MutationAssessor (Reva et al. 2011), MutPred2 (Pejaver et al. 2020), M-CAP (Jagadeesh et al. 2016), PolyPhen2 (Adzhubei et al. 2010), phyloP (Pollard et al. 2010), PROVEAN (Choi et al. 2012), REVEL (Ioannidis et al. 2016), SIFT4G (Vaser et al. 2016), and VEST4 (Carter et al. 2013). Among the 5 recently developed methods (AlphaMissense (Cheng et al. 2023), ESM-1b (Rives et al. 2021), EVE (Frazer et al. 2021), PrimateAI-3D (Gao et al. 2023), and VARITY (Wu et al. 2021)), all but VARITY are deep learning methods that are not supervised on known variant pathogenicity labels. Correlations between the predictions from these tools, as measured on our evaluation dataset, are shown in Fig. S2.

In addition to these 26 predictors shown in the figures, summary metrics for the full set of 60 tested predictors are provided in Table S1.

Full ROC curve performance

For each predictor, we first constructed its Receiver Operating Characteristic (ROC) curve based on the full evaluation dataset and computed the Area Under the ROC (AUROC) (Fig. 1 and Table S1). We show performance for meta-predictors and individual predictors separately in all figures, since we find that meta-predictors tend to achieve higher AUROCs by combining scores from individual predictors, consistent with previous observations. We also demarcate predictors that incorporate allele frequency as an explicit predictive feature with diamonds, due to limitations in their compatibility with other ACMG/AMP lines of evidence.

On the full evaluation dataset, meta-predictors that explicitly incorporate allele frequency as a predictive feature reach an AUROC of 0.92 (ClinPred, MetaRNN), while meta-predictors that do not explicitly incorporate allele frequency reach an AUROC of 0.88 (VARITY_R) (Fig. 1). Although we limited the evaluation dataset to low-frequency variants, the methods that explicitly use allele frequency likely benefit from the remaining allele frequency imbalance in the evaluation dataset, in which pathogenic variants have lower allele frequencies than benign variants (Fig. S1). We further explore the effect of allele frequency below. The individual predictor with the highest AUROC on the full evaluation dataset (AlphaMissense) also reaches an AUROC of

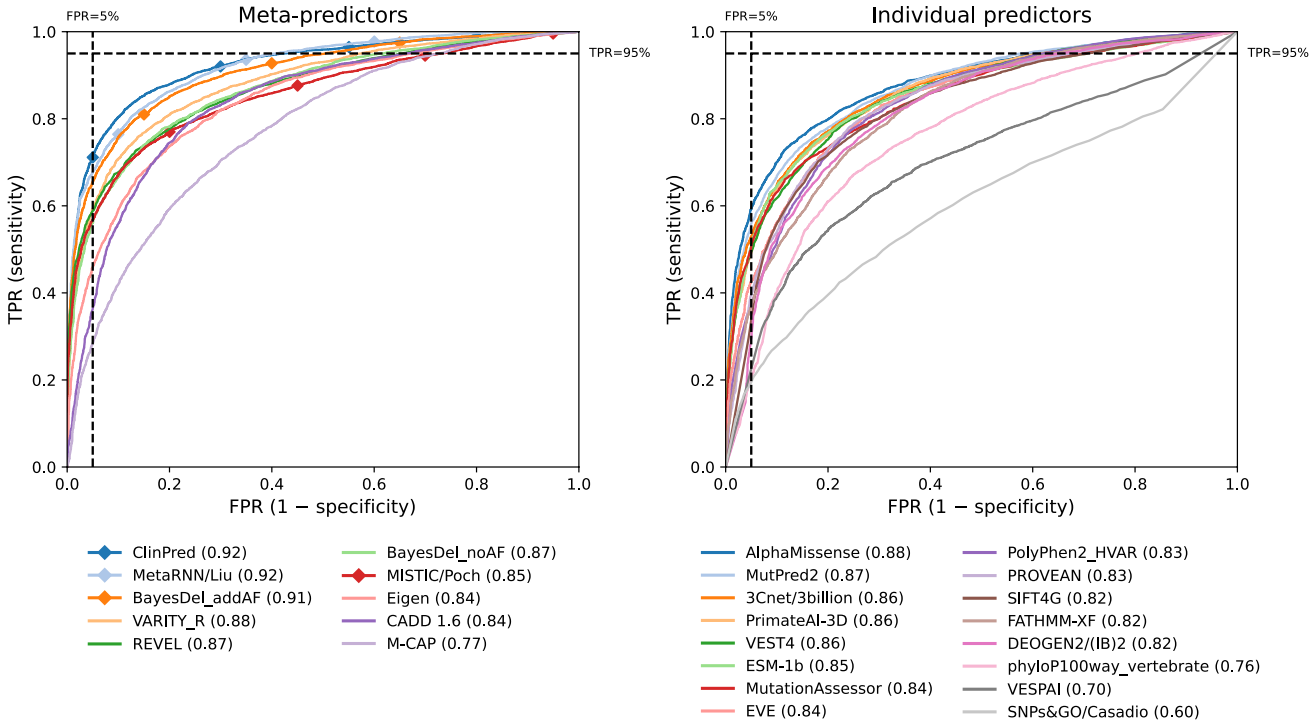

Fig. 1 Full ROC curve performance. We show the ROC curves and AUROCs for meta-predictors (left) and individual predictors (right) on the full evaluation dataset. Predictors marked by diamonds use allele frequency as a feature. The black dashed lines at 5% FPR and 95% TPR demarcate the boundaries of the high-specificity and high-sensitivity regions, respectively, which are enlarged in Fig. 2

0.88. In general, there are multiple predictors with AUROCs within a few percentage points of one another—including AlphaMissense, MutPred2, 3Cnet, PrimateAI-3D, and VEST4—indicating that a number of approaches all have strong performance. Among the deep learning methods that do not supervise on labeled pathogenic or disease-causing variants, those that model both protein structure and protein language (AlphaMissense and PrimateAI-3D) slightly outperform those that model only protein language (ESM-1b and EVE).

The above results were computed for each predictor on only the subset of variants from the evaluation dataset that were scored by that predictor, ignoring missing predictions. However, 8 out of 26 predictors do not report scores for at least 5% of the evaluation dataset (Fig. S3). Most notably, EVE does not supply predictions for 37% of the dataset. Therefore, in Fig. S4, we compare performance on the full dataset ($n = 10,456$) to performance on the smaller set of variants that are scored by all predictors ($n = 4,769$). 23 out of 26 predictors have higher performance on the latter set, suggesting that variants that are not scored by some predictors tend to be harder to predict. However, we find that the ordering of predictors by AUROC is largely similar for both sets of variants.

ClinVar and HGMD data subsets

Pathogenic variants in our evaluation dataset were sourced from two databases with very different curation strategies. ClinVar is a publicly accessible database with variant classifications primarily submitted by genetic testing laboratories, which apply ACMG/AMP guidelines for systematic classification of likely clinical relevance. HGMD is a licensed database that compiles disease-relevant variants from the primary literature, including basic research studies. Owing to its varied data sources, HGMD variant assertions are not subjected to a standardized, weighted evaluation of evidence like those in ClinVar.

To distinguish performance on the two databases, we constructed two subsets of our evaluation dataset: one containing pathogenic variants only from ClinVar, and the other containing pathogenic variants only from HGMD. In both cases, all benign variants were from ClinVar, since HGMD does not curate benign variants. Figure S5 shows the difference in performance on these two data subsets. All predictors have higher performance on the evaluation subset containing pathogenic variants from ClinVar, indicating a qualitative difference between pathogenic variants from the two databases. This difference is likely due in part to false positives among HGMD's DM assertions (McLaughlin et al. 2014; Sharo et al. 2023), though it may also reflect differences in variant predictability or complexity between the two databases. Notably, predictor rankings are largely similar in the two cases.

To further explore how model performance varies with the confidence level of variant classifications, we created a subset of our evaluation dataset containing only the ClinVar variants with two or more stars ($n = 47$); i.e., variants with multiple submitters and no conflicts. Our original dataset also included one-star, single-submitter classifications (with conflicting entries already excluded). On the higher-confidence, two-star subset, nearly all models have higher performance (Fig. S6), with some even achieving a perfect AUROC. This improvement may indicate that our one-star assessments underestimate true accuracy, or that two-star variants have greater bias or circularity concerns (such as those discussed below). However, the size of the two-star subset is too small to draw robust conclusions about performance or model rankings, or to perform many of our subsequent analyses.

High-specificity and high-sensitivity performance

The AUROC metric used above aggregates performance across all possible decision rules, or score thresholds, for separating predicted benign variants from predicted pathogenic variants. However, in practice, typically only a single decision rule is used for a particular application. While not all practitioners may choose the same decision rule, there are two general regimes in which computational predictors of missense variant pathogenicity are most likely to be used. First, for clinical variant interpretation, high confidence classifications of pathogenicity are required when reporting results to patients. Especially when reporting secondary genomic findings, which are putatively pathogenic variants of concern unrelated to the original reason for testing (Katz et al. 2020), false positives should be minimized to avoid overdiagnosis. Accordingly, practitioners will employ a decision rule with a low false positive rate (FPR), or equivalently, high specificity. To measure performance in this setting, we examined the high-specificity region (FPR $\leq$ 5%) of the ROC curves from Fig. 1 (Fig. 2a). (This particular 5% FPR threshold is arbitrary but represents a useful decision rule in this scenario.) Second, for exploratory analysis of whole-exome or whole-genome sequencing data, high sensitivity is often desired. For example, in a research environment, when analyzing data from a patient with an undiagnosed genetic disorder, computational predictors can be used to narrow down a list of VUS to those variants that should be prioritized in follow-up studies, ideally without mistaking the true pathogenic variant as benign. Accordingly, practitioners will employ a decision rule with a high true positive rate (TPR), or equivalently, high sensitivity (Rastogi et al. 2022). To measure performance in this setting, we examine the high-sensitivity region (TPR $\geq$ 95%)

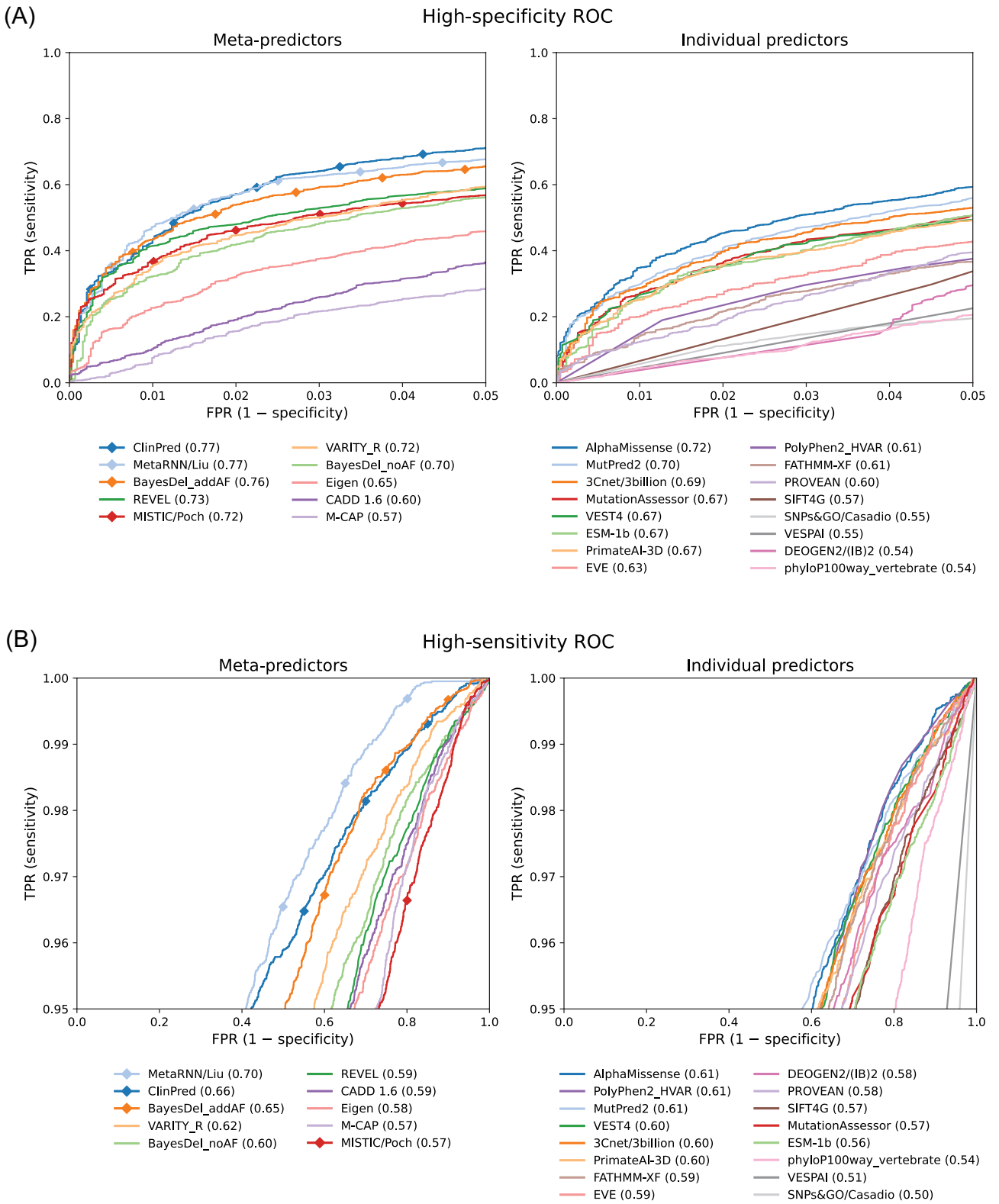

Fig. 2 Performance in high-specificity and high-sensitivity regimes. We show enlarged portions of the ROC curves from Fig. 1 to focus on (**A**) the high-specificity region (FPR $\lesssim 5\%$) and (**B**) the high-sensitivity region (TPR $\gtrsim 95\%$) for meta-predictors (left) and individual predictors (right). We also show the normalized area under the curve in these regions (normalized such that a perfect classifier gets a score of 1 and a random classifier gets a score of 0.5). Predictors marked by diamonds use allele frequency as a feature

of the ROC curves (Fig. 2b). For both the high-specificity (FPR ≤ 5%) and high-sensitivity (TPR ≥ 95%) regions of the ROC curves, we compute a normalized area under the curve in these regions (McClish 1989). Table S1 lists the full-curve AUROC, high-specificity AUROC, and high-sensitivity AUROC for all 60 predictors included in our evaluation.

Notably, the performance of some predictors varies substantially between the two classification regimes. MetaRNN, which uses allele frequency as a predictive feature, excels in the high-sensitivity region, particularly for true positive rates that approach 100%. On the other hand, MISTIC performs well in high-specificity regions but struggles in high-sensitivity regions, indicating its suitability for clinical variant classification rather than exploratory research. Among the individual predictors, PolyPhen2_HVAR has strong performance in the high-sensitivity region, but lower performance relative to other predictors in the high-specificity region, whereas MutationAssessor and ESM-1b both have lower performance in the high-sensitivity region. These findings underscore the notion that different methods may be better suited for different clinical or research applications.

Effect of allele frequency

As illustrated in Fig. S1, pathogenic variants tend to have lower allele frequencies than benign variants in our evaluation dataset. This trend is expected, as deleterious variants are more likely to be under negative selection and therefore less common in the population. However, for many clinical use cases, it is important to be able to distinguish very rare benign variants from pathogenic variants, for example in data from rare disease patients. Methods that utilize allele frequency as a predictive feature might struggle with variant classification in this setting. To evaluate the effect of allele frequency on performance, we binned the benign variants in our dataset by their allele frequencies and compared performance when differentiating benign variants in each bin from the full set of pathogenic variants (Fig. 3). Three of the four methods that utilize allele frequency as a predictive feature—ClinPred, MetaRNN, and BayesDel_addAF, which are also top-performing predictors in the above analyses—show a marked performance decrease on very rare benign variants. Despite this decline, MetaRNN and ClinPred still outperform most other predictors in distinguishing very rare benign variants from pathogenic variants, indicating that their predictions are not excessively reliant on allele frequency. VARITY_R also has notably high performance on very rare benign variants. The effect of including allele frequency as a predictive feature can also be illustrated by comparing the two versions of BayesDel with and without allele frequency. BayesDel_addAF (which performs 4% better than BayesDel_noAF on the full-dataset AUROC) has much higher performance in most of the benign allele frequency bins, but BayesDel_noAF outperforms BayesDel_addAF in the lowest allele frequency bin.

To minimize the effect of allele frequency on our performance metrics, we created a subset of the evaluation dataset in which allele frequencies were matched between pathogenic and benign variants. We then compared performance on the full dataset to the allele frequency matched dataset (Fig. S7). Methods that use allele frequency as a predictive feature have the largest drop in performance. Some methods that do not explicitly use

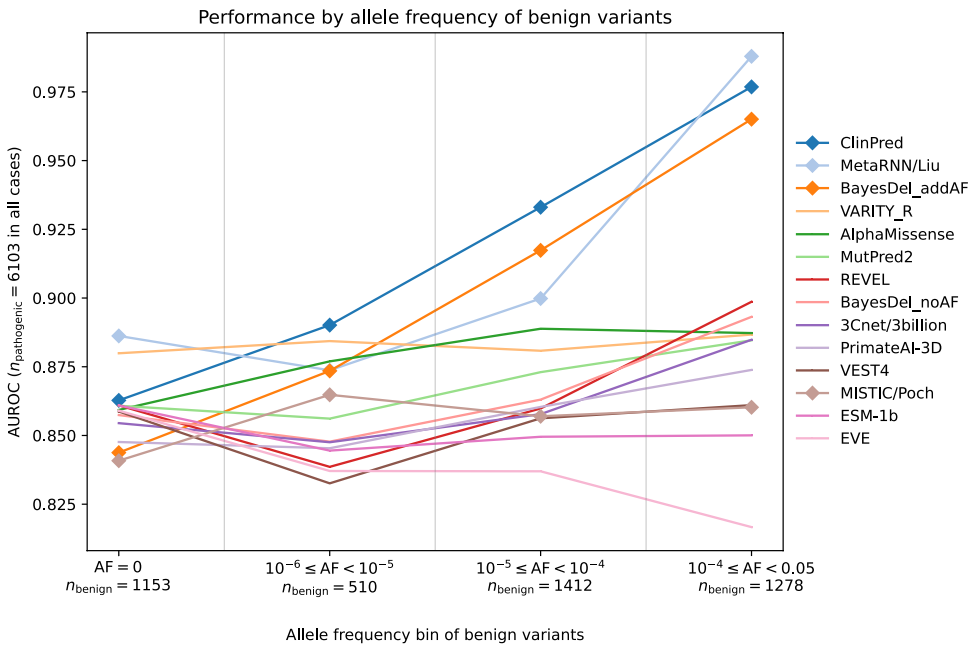

Fig. 3 Allele frequency bias. Top-performing predictors are evaluated for distinguishing benign variants in different allele frequency bins from pathogenic variants. All 6103 pathogenic variants were used in each evaluation, and benign variants were stratified by their allele frequencies obtained from the control cohort exomes in gnomAD v2.1.1 (Karczewski et al. 2020). Predictors marked by diamonds use allele frequency as a feature

allele frequency also have slightly lower performance on the allele frequency matched dataset, likely because allele frequency correlates with other features used by these tools (e.g. conservation scores).

Finally, while we used an upper allele frequency threshold of 0.05 to construct our dataset of low-frequency variants (following the BA1 rule), some clinical settings might use more stringent thresholds. To test whether our results are robust to the choice of allele frequency threshold, we varied this upper bound from 0.05 down to 0.001, which is commonly used in clinical settings (Fig. S8). We find that performance remains largely unchanged across these thresholds, though methods utilizing allele frequency as a feature start to show marginal performance degradation at the 0.001 threshold. These results indicate that the choice of allele frequency threshold does not significantly affect our conclusions.

Effect of gene label imbalance

Databases of disease-relevant variants, such as ClinVar and HGMD, have large imbalances in the ratio of pathogenic to benign variants per gene, which may reflect bias in which variants have been studied rather than the true fitness landscape for those genes (Grimm et al. 2015). To gauge the degree of label imbalance in our evaluation dataset, we tested a simple baseline model, similar to the one outlined in Cheng et al. (2023). The baseline model assigns the same score to all variants in a gene, equal to the fraction of high-confidence missense variants from ClinVar and HGMD that were available before the cutoff date for our evaluation dataset and were labeled as pathogenic or disease-causing in those databases. On our evaluation dataset, this simple model achieves an AUROC of 0.74 (Fig. S9), rivaling the performance of the best conservation score, phyloP100way_vertebrate.

To minimize the effect of gene label imbalance on our performance metrics, we created a subset of the evaluation dataset containing an equal number of pathogenic and benign variants per gene. We then compared performance on the original dataset to the gene label-balanced dataset (Fig. 4). Many of the tested predictors, particularly many of the meta-predictors, have lower performance on the label-balanced dataset. However, predictors that do not train on labeled pathogenic or disease-causing variants (including but not limited to AlphaMissense, PrimateAI-3D, CADD, Eigen, EVE, ESM-1b, phyloP, and VESPAl) do not show a degradation in performance. The largest increase in performance on the gene label-balanced dataset is observed for SNPs&GO.

Effect of prior pathogenicity probability on evidence thresholds

A recently developed calibration method adopts a principled probabilistic approach to determine, for any given predictor, the thresholds at which its scores meet ACMG/AMP evidence strengths (supporting to very strong) for both pathogenicity and benignity (Tavtigian et al. 2018; Pejaver et al. 2022) using an estimated prior probability of pathogenicity (Zeiberg et al. 2020). For different applications, particularly in research settings, a variety of prior probabilities may be relevant. Therefore, we applied this calibration method to all tested predictors at five different prior probabilities of pathogenicity (0.02, 0.04, 0.06, 0.08, and 0.10) using our evaluation dataset. Figures S10 and S11 display the dependence of the resulting score thresholds on the prior probability for meta-predictors and individual predictors, respectively. In all cases, lower prior pathogenicity probabilities lead to reduced evidence strengths for both benignity and pathogenicity. We note that due to the previously observed effect of allele frequency on the performance of tools that explicitly include allele frequency as a predictive feature, in future studies such tools should be calibrated separately on variants within each allele frequency bin.

Discussion

We present the first full assessment of the ongoing CAGI Annotate-All-Missense challenge, evaluating the ability of computational variant effect prediction tools to classify missense variants as pathogenic or benign under a variety of evaluation conditions. In general, we find strong performance of many predictors on an evaluation dataset of missense variants that were classified in ClinVar or added as disease-causing in HGMD after the close of the CAGI 6 Annotate-All-Missense challenge in October 2021.

Rather than using a single overall performance metric, it is important to evaluate missense variant effect predictors in a variety of settings that are relevant to different clinical or research applications. We examined performance in high-specificity and high-sensitivity settings separately, since high specificity is most relevant to clinical variant classification and high sensitivity is most relevant to exploratory analysis of whole-genome or whole-exome sequencing data in a research setting. We also examined performance on subsets of the evaluation dataset that were either matched for pathogenic and benign allele-frequency distributions or that included only very rare benign variants. These evaluation settings are important for applications that already use allele frequency as a separate criterion for establishing benignity (e.g. BA1 (Ghosh et al. 2018)) or that aim to classify missense variants within a pool of very rare variants

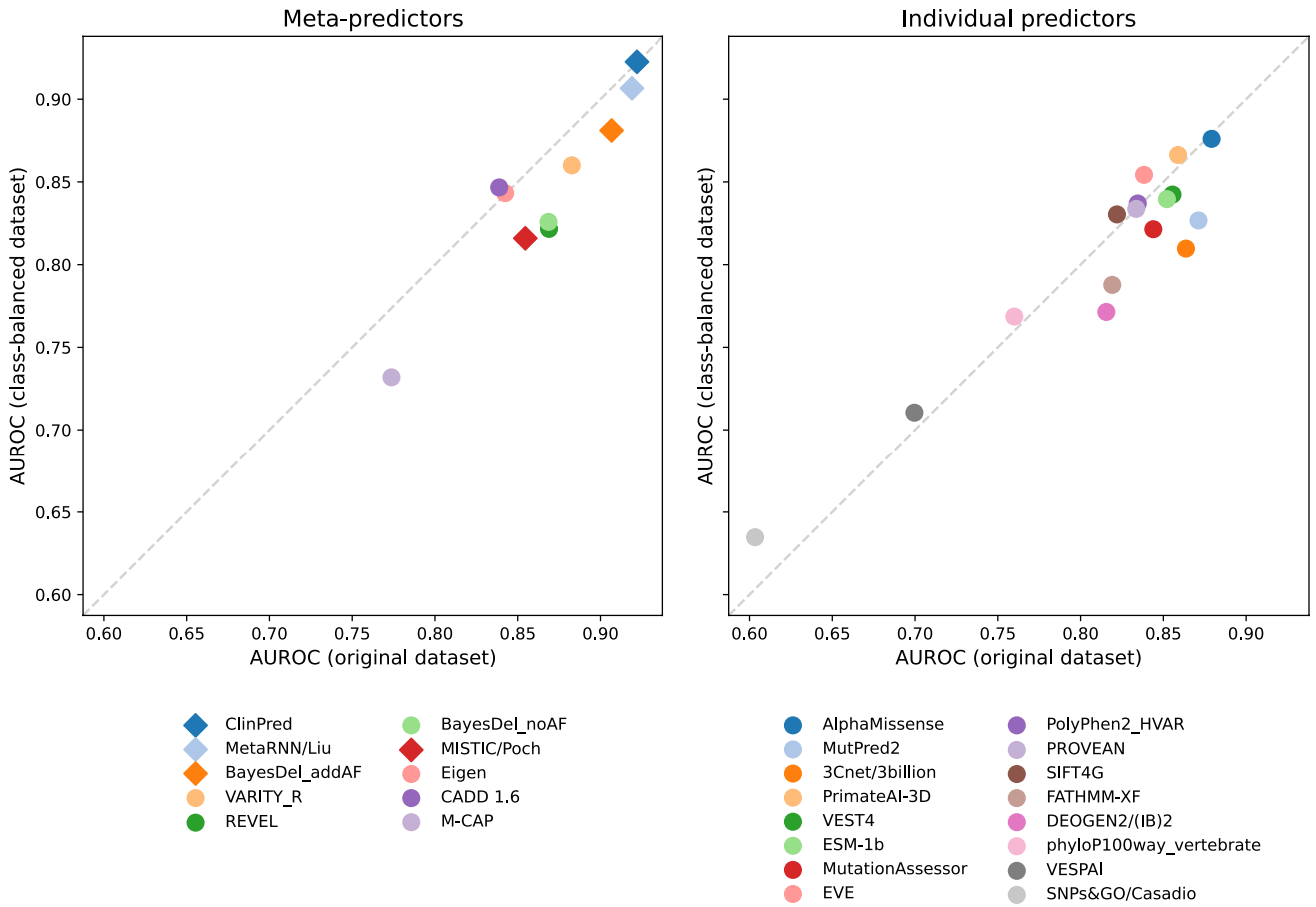

Fig. 4 Gene label balancing. We constructed a gene label-balanced subset of our evaluation dataset containing an equal number of pathogenic and benign variants per gene. This label-balanced dataset consists of 2140 variants from 504 genes. Performance on the label-balanced dataset (y-axis) is compared to performance on the full dataset from Fig. 1 (x-axis) for meta-predictors (left) and individual predictors (right). Predictors marked by diamonds use allele frequency as a feature

from rare disease patients. In general, we find that predictors with strong performance tend to perform well across multiple settings, but that the specific predictor rankings differ between settings, suggesting that different predictors may be best suited to different clinical or research applications. We recommend that practitioners consider the most relevant evaluation data subsets or the most relevant portions of ROC curves for their application and choose methods based on performance in those settings, rather than examining only the full AUROC on the full evaluation dataset.

We also evaluated performance on a subset of the evaluation dataset with equal numbers of pathogenic and benign variants per gene, due to substantial imbalance in the class labels of the available variants for many genes in the full dataset. This type of imbalance is present within clinical variant databases and leads predictors trained on variants from such databases to learn gene-level properties in addition to variant-level properties for variant classification, which tends to result in reduced performance on the gene label-balanced subset. However, interpreting performance on this subset is complicated by the fact that some gene label imbalance is reflective of true biology, such as a different tolerance to mutation for different genes, while some is due to current practices in clinical testing or bias in the amount of attention and research effort devoted to particular genes and diseases. Separating the different factors contributing to gene label imbalance is an ongoing challenge in the evaluation of missense variant effect predictors. As above, we note that while the specific predictor rankings differ on this evaluation subset, predictors with strong performance in other settings also tend to perform well in the gene label-balanced setting.

For this assessment, we used recently classified or disease-relevant variants from clinical variant databases due to their relevance for evaluating clinical utility; however, we note that this source of evaluation data has several limitations in addition to the gene label imbalance discussed above. Importantly, there are likely to be some errors in the labels provided by these databases, which limit the maximum achievable performance in our evaluation. Although

we attempted to reduce such errors by using high-confidence labels from each database—variants with Benign or Pathogenic labels (excluding Likely Benign and Likely Pathogenic), non-conflicting interpretations, and either at least 1-star ratings from ClinVar or disease-causing (DM) label from HGMD—some incorrectly labeled variants likely remain. To account for possible systematic differences between the two databases, we report performance on the ClinVar and HGMD subsets of pathogenic variants separately, in addition to performance on the full evaluation dataset. The substantially lower performance that we observe for all predictors on the HGMD disease-causing variants likely reflects differences in annotation practices between the two databases. These differences may contribute to the higher false positive rate previously reported for HGMD (McLaughlin et al. 2014; Sharo et al. 2023) but may also result in ClinVar containing variants that are inherently easier to classify. Nonetheless, predictor rankings are largely similar when evaluated on pathogenic variants from the two databases separately.

In addition, many missense variant effect predictors were trained using data from clinical variant databases, and any overlap between these training variants and the variants in the evaluation dataset would inflate the performance estimates for such predictors. To avoid overlap, we specifically excluded variants from our evaluation dataset that had pathogenicity information available in ClinVar, HGMD, or UniProt, which are the most commonly used databases for training predictors, prior to the close of the CAGI 6 challenge. However, it is possible that some evaluation set variants had been studied in the literature or included in other specialized databases prior to this date, where pathogenicity labels could have been available for training. Another limitation of using clinical variant databases for evaluation is the potential for circularity if predictions from any of the tested tools were considered when making pathogenicity classifications for the variants in the evaluation dataset. Based on the Richards et al. 2015 (Richards et al. 2015) ACMG/AMP guidelines, computational predictions had primarily been used only as supporting evidence for classification, but recent clinical recommendations (Pejaver et al. 2022) identified thresholds at which predictions from certain tools provide stronger levels of evidence. Therefore, it is possible that computational predictions had some influence on the most recently classified variants, which would result in inflated performance estimates for those and related tools. To enable continued unbiased assessments of missense variant effect predictors in the future, it will be essential for clinical variant databases to document the lines of evidence used for each variant classification.

Overall, our results indicate that currently available tools for missense variant effect prediction provide a powerful line of evidence for classifying missense variants of uncertain significance. While we find that meta-predictors tend to outperform their constituent individual predictors, a number of individual predictors have performance close to that of commonly used meta-predictors, particularly meta-predictors that do not explicitly include allele frequency as a predictive feature. We note continued progress in the field relative to the oldest and most cited tools, as well as recent advancement in developing individual predictors that are not trained on variants from clinical variant databases, making them less susceptible to biases in the collection and interpretation of variant data. Several such predictors achieve strong performance in our assessment, including predictors that use only unsupervised or self-supervised training schemes. These types of predictors are promising candidates to be incorporated into future meta-predictors and combined with other complementary information related to variant pathogenicity. This ongoing CAGI challenge will continue to evaluate such developments and to assess state-of-the-art methods as the field progresses.

Methods

Evaluation dataset construction

We created an evaluation dataset by incorporating variants from both the April 4, 2023 version of ClinVar (Landrum et al. 2018), which contains both pathogenic and benign variants, and the 2023.1 Professional version of the Human Gene Mutation Database (HGMD) (Stenson et al. 2020), which contains only pathogenic variants.

We assigned molecular effects to all variants using SnpEff (Cingolani et al. 2012), which was configured with the Ensembl 105 gene set, and retained all single-nucleotide variants that were annotated as missense in at least one affected transcript, excluding variants that were assigned a higher impact annotation (`HIGH` impact or `splice_region_variant`) in another transcript.

To ensure that all predictors were tested on variants that they had not previously seen during training, we removed (1) all variants, except those of uncertain significance, present in the November 7, 2021 version of ClinVar, (2) all `DM` variants present in the 2021.4 version of HGMD, (3) all variants, expect those of uncertain significance, present in the 2021.4 version of the UniProt Humsavar database (McGarvey et al. 2019), and (4) variants in the AlphaMissense validation set (used for early stopping). These cutoff dates were chosen based on the CAGI 6 Annotate-All-Missense challenge, which closed on October 11, 2021. Motivated by PM5 (Richards et al. 2015), we also excluded variants affecting the same codon as any of the aforementioned removed variants to minimize data leakage (Fig. S12).

Among the remaining variants, we only retained those with high-confidence pathogenicity classifications in ClinVar (either `Benign` or `Pathogenic` with 1 star or above, except those with conflicting interpretations) and high-confidence disease-causing (`DM`) HGMD variants. We further removed variants that can be inferred to be benign by their allele frequency in gnomAD exomes v2.1.1 (Karczewski et al. 2020), as per the revised BA1 criterion (Ghosh et al. 2018). Specifically, we removed variants with a control global allele frequency > 0.05 or control continental allele frequency > 0.05 with at least 2000 observed alleles in any of the five major continental populations: African/African American (AFR), Latino/Admixed American (AMR), East Asian (EAS), South Asian (SAS), and non-Finnish European (NFE). Furthermore, we discarded all variants that were not present in a Mendelian disease gene. For the purposes of this study, we consider Mendelian disease genes ($n = 3465$) to be those with at least one high-confidence (as described above) pathogenic variant of any mutation class in the April 4, 2023 version of ClinVar. Lastly, we excluded variants that predictors submitting to the CAGI 6 Annotate-All-Missense challenge were not asked to score. Our final dataset contains 6,103 pathogenic and 4,353 benign variants.

Procuring predictions

We evaluated 60 missense variant effect predictors, none of which were trained on clinical pathogenicity data released after the CAGI 6 Annotate-All-Missense challenge deadline.

CAGI 6 Annotate-All-Missense submissions. Six teams submitted a total of twelve models to the challenge. Submitters were asked to provide a prediction score for a pre-specified list of missense variants throughout the genome (based on dbNSFP v4 (Liu et al. 2020)), of which our evaluation dataset is a subset. All team identities and model details were hidden until the conclusion of the analysis.

Predictors available in dbNSFP. We obtained predictions for 40 tools from the dbNSFP v4.2a database (released on April 6, 2021) (Liu et al. 2020). Version 4.2a of the database was chosen as the last release before the CAGI 6 Annotate-All-Missense challenge deadline. For each predictor, we extracted the rank score of all dataset variants using SnpSift (Cingolani et al. 2012). The rank score of a variant is its percentile, scaled between 0 and 1, among all variants in dbNSFP, with higher rank scores corresponding to more deleterious predictions. If a variant was assigned multiple rank scores (e.g. if the method makes separate predictions for each affected transcript), we took the highest rank score.

VARITY. Predictions from the VARITY class of models (Wu et al. 2021) (VARITY_R, VARITY_R_LOO, VARITY_ER, VARITY_ER_LOO), were added to dbNSFP v4.4a (released on May 6, 2023 after the challenge deadline). However, the VARITY models were trained on ClinVar and HGMD data released prior to the challenge deadline (F. Roth and J. Wu, personal communication, July 12, 2023). The same procedure described above was used to extract VARITY scores from dbNSFP v4.4a.

MutPred2. Predictions from MutPred2 (Pejaver et al. 2020) on the evaluation dataset were provided by the original authors (V. Pejaver, personal communication, August 25, 2023). Scores were provided per affected isoform, and the most pathogenic score was chosen.

PrimateAI-3D. Predictions from PrimateAI-3D (Sundaram et al. 2018) for most human missense variants were provided by Illumina. If a variant mapped to multiple genes, the most pathogenic score was chosen.

ESM-1b. ESM-1b (Rives et al. 2021) variant effect scores were computed for all possible amino acid changes in most proteins in the human proteome by Brandes et al. (2023). SnpSift was used to annotate each dataset variant with affected Ensembl transcripts and corresponding amino acid changes. Because ESM-1b scores are indexed by UniProt identifiers, we used the UniProt ID mapping service (https://www.uniprot.org/id-mapping) to convert Ensembl transcripts to UniProt IDs. If a variant mapped to multiple proteins, the most pathogenic score was chosen.

EVE. EVE (Frazer et al. 2021) provides variant effect predictions in the form of VCF files for 2951 human proteins. The VCF files were downloaded on May 23, 2023 from https://evemodel.org/download/bulk. If a variant mapped to multiple proteins, the most pathogenic score was chosen.

AlphaMissense. We downloaded AlphaMissense (Cheng et al. 2023) scores for variants in canonical isoforms and non-canonical isoforms on September 19, 2023 from https://console.cloud.google.com/storage/browser/dm_alphamissense. If a variant mapped to multiple isoforms, the most pathogenic score was chosen.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-025-02732-2.

Acknowledgements The CAGI Annotate-All-Missense challenge was originally proposed by Sean Mooney.

Author Contributions R.R. and R.Ch. performed the analyses. R.R., R.Ch., S.L., P.R., S.E.B., and N.M.I. designed the assessment and interpreted results. S.E.B. and G.A. designed and developed the original challenge, respectively. P.R. organized the challenge for CAGI 6. G.A., A.K., Y.P., and C.B. provided technical support for the challenge. C.L. and X.L. provided dbNSFP variants for the challenge. C.L., K.L., J.W., D.W.K., C.K., G.B., P.L.M., C.S., R.Ca., K.C., T.W., O.P., F.A., G.C., F.P., D.R., W.V., M.R., C.M., T.O., B.R., and X.L. participated as predictors in the challenge. M.M. and D.N.C. curated HGMD data for the assessment. T.B. and V.P. provided MutPred2 predictions and guidance on the evidence threshold calibration analysis. R.R., R.Ch., P.R., S.E.B., and N.M.I. wrote the manuscript with feedback from all authors.

Funding This work was supported in part by the U.S. National Institutes of Health (NIH) awards U24HG007346 (S.E.B.), U41HG007346

(S.E.B.), R13HG006650 (S.E.B.), and U01HG012022 (P.R.). N.M.I. is a Chan Zuckerberg Biohub Investigator. We also thank the Belgian Fund for Scientific Research (F.R.S.-FNRS) and the Research Foundation Flanders (FWO) for financial support.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

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. Nature Methods 7(4):248–249

Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD (2018) ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. The American Journal of Human Genetics 103(4):474–483

Ancien F, Pucci F, Godfroid M, Rooman M (2018) Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Scientific Reports 8(1):4480

Bergquist T, Stenton SL, Nadeau EA, Byrne AB, Greenblatt MS, Harrison SM, Tavtigian SV, O'Donnell-Luria A, Biesecker LG, Radivojac P, et al. (2025) Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria. Genetics in Medicine

Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V (2023) Genome-wide prediction of disease variant effects with a deep protein language model. Nature Genetics 55(9):1512–1522

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Human Mutation 30(8):1237–1244

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying mendelian disease genes with the variant effect scoring tool. BMC Genomics 14(3):1–16

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(6664):7492

Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, Laporte J, Poch O (2020) MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One 15(7):0236962

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7(10):46688

Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. Genome Research 19(9):1553–1561

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X (2012) Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics 3:35

Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6(2):80–92

Critical Assessment of Genome Interpretation Consortium (2024) CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. Genome Biology 25(1):53

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Computational Biology 6(12):1001025

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X (2014) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics 24(8):2125–2137

Feng B-J (2017) PERCH: a unified framework for disease gene prioritization. Human Mutation 38(3):243–251

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(7883):91–95

Freedman D, Diaconis P (1981) On the histogram as a density estimator: L_2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57(4):453–476

Gao H, Hamp T, Ede J, Schraiber JG, McRae J, Singer-Berk M, Yang Y, Dietrich AS, Fiziev PP, Kuderna LF et al (2023) The landscape of tolerated genetic variation in humans and primates. Science 380(6648):8153

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X (2009) Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25(12):54–62

Ghosh R, Harrison SM, Rehm HL, Plon SE, Biesecker LG (2018) ClinGen Sequence Variant Interpretation Working Group: Updated recommendation for the benign stand-alone ACMG/AMP criterion. Human Mutation 39(11):1525–1530

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation 36(5):513–523

Hu Z, Yu C, Furutsuki M, Andreoletti G, Ly M, Hoskins R, Adhikari AN, Brenner SE (2019) VIPdb, a genetic Variant Impact Predictor Database. Human Mutation 40(9):1202–1214

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. The American Journal of Human Genetics 99(4):877–885

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics 48(2):214–220

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics 48(12):1581–1586

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP et al (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581(7809):434–443

Katsonis P, Wilhelm K, Williams A, Lichtarge O (2022) Genome interpretation using in silico predictors of variant impact. Human Genetics 141(10):1549–1577

Katz AE, Nussbaum RL, Solomon BD, Rehm HL, Williams MS, Biesecker LG (2020) Management of secondary genomic findings. The American Journal of Human Genetics 107(1):3–14

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4:1073–1081

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 Research 46(D1):1062–1067

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25(21):2744–2750

Li C, Zhi D, Wang K, Liu X (2022) MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Medicine 14(1):115

Lin YJ, Menon AS, Hu Z, Brenner SE (2024) Variant Impact Predictor database (VIPdb), version 2: trends from three decades of genetic variant impact predictors. Hum Genomics 18(1):90

Liu X, Li C, Mou C, Dong Y, Tu Y (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Medicine 12(1):1–8

Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Scientific Reports 5:10576

Malhis N, Jacobson M, Jones SJM, Gsponer J (2020) LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Research 48(W1):154–161

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. Human Genetics 141(10):1629–1647

McClish DK (1989) Analyzing a portion of the ROC curve. Medical Decision Making 9(3):190–195

McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C (2019) UniProt Consortium: UniProt genomic mapping for deciphering functional effects of missense variants. Human Mutation 40(6):694–705

McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, Kohane IS, Krier J, Lane WJ, Lautenbach D, Lebo MS, Machini K, MacRae CA et al (2014) A systematic approach to the reporting of medically relevant findings from whole genome sequencing. BMC Medical Genetics 15:1–12

McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genetics 5(5):1000471

Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, Mort M, Cooper DN, Sebat J, Iakoucheva LM et al (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature Communications 11(1):5918

Pejaver V, Byrne AB, Feng B-J, Pagel KA, Mooney SD, Karchin R, O'Donnell-Luria A, Harrison SM, Tavtigian SV, Greenblatt MS et al (2022) Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. The American Journal of Human Genetics 109(12):2163–2177

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20(1):110–121

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. Nature Communications 12(1):510

Quang D, Chen Y, Xie X (2014) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31(5):761–763

Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, Rooman M, Vranken W (2017) DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Research 45(W1):201–206

Rastogi R, Stenson PD, Cooper DN, Bejerano G (2022) X-CAP improves pathogenicity prediction of stopgain variants. Genome Medicine 14(1):1–11

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research 47(D1):886–894

Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Research 39(17):118–118

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. Genetics in Medicine 17(5):405–423

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. Proceedings of the National Academy of Sciences 118(15):2016239118

Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 34(3):511–513

Rost B, Radivojac P, Bromberg Y (2016) Protein function in precision medicine: deep understanding with machine learning. FEBS Letters 590(15):2327–2341

Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ (2017) Regional missense constraint improves variant deleteriousness prediction. bioRxiv, 148353

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods 7:575–576

Sharo AG, Zou Y, Adhikari AN, Brenner SE (2023) ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. Genome Medicine 15(1):51

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation 34(1):57–65

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31(10):1536–1543

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15(8):1034–1050

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. Human Genetics 139:1197–1207

Stenton SL, Pejaver V, Bergquist T, Biesecker LG, Byrne AB, Nadeau EA, Greenblatt MS, Harrison SM, Tavtigian SV, Radivojac P et al (2024) Assessment of the evidence yield for the calibrated PP3/BP4 computational recommendations. Genetics in Medicine 26(11):101213

Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J et al (2018) Predicting the clinical impact of human mutation with deep neural networks. Nature Genetics 50(8):1161–1170

Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, Biesecker LG (2018) ClinGen Sequence Variant Interpretation Working Group: Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genetics in Medicine 20(9):1054–1060

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC (2016) SIFT missense predictions for genomes. Nature Protocols 11(1):1–9

Won D-G, Kim D-W, Woo J, Lee K (2021) 3Cnet: pathogenicity prediction of human variants using multitask learning with evolutionary constraints. Bioinformatics 37(24):4626–4634

Wu Y, Liu H, Li R, Sun S, Weile J, Roth FP (2021) Improved pathogenicity prediction for rare human missense variants. The American Journal of Human Genetics 108(10):1891–1906

Zeiberg D, Jain S, Radivojac P (2020) Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6729–6736

Authors and Affiliations

Ruchir Rastogi[1] · Ryan Chung[2] · Sindy Li[3] · Chang Li[4] · Kyoungyeul Lee[5] · Junwoo Woo[5] · Dong-Wook Kim[5] · Changwon Keum[5] · Giulia Babbi[6] · Pier Luigi Martelli[6] · Castrense Savojardo[6] · Rita Casadio[6] · Kirsley Chennen[7] · Thomas Weber[7] · Olivier Poch[7] · François Ancien[8,9] · Gabriel Cia[8,9] · Fabrizio Pucci[8,9] · Daniele Raimondi[10,19] · Wim Vranken[9,11] · Marianne Rooman[8,9] · Céline Marquet[12] · Tobias Olenyi[12] · Burkhard Rost[12] · Gaia Andreoletti[3,18] · Akash Kamandula[13] · Yisu Peng[13] · Constantina Bakolitsa[3] · Matthew Mort[14] · David N. Cooper[14] · Timothy Bergquist[15] · Vikas Pejaver[15,16] · Xiaoming Liu[4] · Predrag Radivojac[13] · Steven E. Brenner[2,3] · Nilah M. Ioannidis[1,2,17]

✉ Ruchir Rastogi
ruchir_rastogi@berkeley.edu

✉ Steven E. Brenner
brenner@compbio.berkeley.edu

✉ Nilah M. Ioannidis
nilah@berkeley.edu

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA

[2] Center for Computational Biology, University of California, Berkeley, CA, USA

[3] Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

[4] USF Genomics, College of Public Health, University of South Florida, Tampa, FL, USA

[5] 3billion Inc., Seoul, South Korea

[6] Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[7] University of Strasbourg, Strasbourg, France

[8] Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium

[9] Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels, Belgium

[10] ESAT-STADIUS, KU Leuven, Leuven, Belgium

[11] Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium

[12] Department of Informatics, Bioinformatics and Computational Biology, Technical University of Munich, Munich, Germany

[13] Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA

[14] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK

[15] Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[16] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[17] Chan Zuckerberg Biohub, San Francisco, CA, USA

[18] Present Address: Sage Bionetworks, Seattle, WA, USA

[19] Institut de Génétique Moléculaire de Montpellier, Université de Montpellier, Montpellier, France