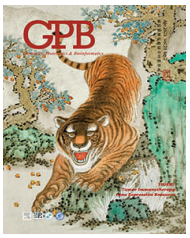
Genomics Proteomics Bioinformatics 21 (2023) 414–426

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



APPLICATION NOTE

mvPPT: A Highly Efficient and Sensitive Pathogenicity Prediction Tool for Missense Variants

Shi-Yuan Tong [1], Ke Fan [1], Zai-Wei Zhou [2], Lin-Yun Liu [1], Shu-Qing Zhang [1], Yinghui Fu [1], Guang-Zhong Wang [3], Ying Zhu [4,*], Yong-Chun Yu [1,*]

[1] *Jing'an District Central Hospital of Shanghai, State Key Laboratory of Medical Neurobiology, MOE Frontiers Center for Brain Science, Institutes of Brain Science, Fudan University, Shanghai 200032, China*
[2] *Shanghai Xunyin Biotechnology Co., Ltd., Shanghai 201802, China*
[3] *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[4] *Huashan Hospital, State Key Laboratory of Medical Neurobiology, MOE Frontiers Center for Brain Science, Institutes of Brain Science, Fudan University, Shanghai 200032, China*

Received 14 September 2021; revised 19 May 2022; accepted 29 July 2022
Available online 5 August 2022

Handled by Jingfa Xiao

KEYWORDS

Machine learning;
Missense variant;
Genomics;
Computational biology;
Pathogenicity prediction

Abstract Next-generation sequencing technologies both boost the discovery of variants in the human genome and exacerbate the challenges of pathogenic variant identification. In this study, we developed **Pathogenicity Prediction** Tool for **missense variants** (mvPPT), a highly sensitive and accurate missense variant classifier based on gradient boosting. mvPPT adopts high-confidence training sets with a wide spectrum of variant profiles, and extracts three categories of features, including scores from existing prediction tools, frequencies (allele frequencies, amino acid frequencies, and genotype frequencies), and genomic context. Compared with established predictors, mvPPT achieves superior performance in all test sets, regardless of data source. In addition, our study also provides guidance for training set and feature selection strategies, as well as reveals highly relevant features, which may further provide biological insights into variant pathogenicity. mvPPT is freely available at http://www.mvppt.club/.

Introduction

Whole-exome sequencing (WES) and whole-genome sequencing (WGS) enable large-scale parallel assessment of genetic variants and have been increasingly adopted in clinical diagnosis, which makes interpreting the effect of the identified

* Corresponding authors.
E-mail: ying_zhu@fudan.edu.cn (Zhu Y), ycyu@fudan.edu.cn (Yu YC).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.07.005

variants a serious challenge [1,2]. Unlike synonymous variants and loss-of-function variants for which the impact on the protein can be relatively easy to predict, missense variants that often lead to inconclusive genomic outcomes remain a major challenge in pathogenicity interpretation. Compared with the reference genome, a human exome on average contains around 20,000 single nucleotide variants (SNVs) and approximately half are missense variants [3,4]. Nevertheless, the effects of most missense variants on proteins are unclear, as experimental validation of large numbers of variants is limited by efficiency and cost. To address these limitations, many computational tools have been developed to predict the potential impact of variants [5–21]. Early prediction models compute pathogenicity scores based on a single property of variants, such as evolutionary conservation [8,10,15] and protein structure/function [16,17]. And recent ensemble methods achieve higher classification accuracy by integrating information from individual predictors [5–7,9,11–13,20]. Although these existing tools have made significant contributions to the prediction of the hazard of genetic variants, the sensitivity of prediction still needs to be improved when assessing the pathogenicity of massive variants in clinical scenarios.

While the existing tools provide positive predictive power, their prediction results are often inconsistent with each other [6,18]. It is believed that the predictive power of current ensemble methods is hampered by the lack of appropriate training data and incomplete features [7,8,22]. For training set inclusion, the widely-adopted strategies to create training sets include using variants from disease databases only [7,9] or using variants from both disease and population databases to balance the ratio of benign and pathogenic variants in the dataset [5,6]. However, there is no conclusion on which strategy results in the best performance. Specifically, existing ensemble tools mostly train machine learning models on variants with known labels in disease databases such as ClinVar [23] and/or Human Gene Mutation Database (HGMD) [24]. However, variants in disease databases may cluster around well-described disease genes, *i.e.*, the more a gene has been studied, the more variants on this gene are likely to be discovered. Unfortunately, it is not clear yet whether the clustering of variants on certain genes introduces bias in the prediction of variant pathogenicity. Moreover, since each resource maintains different variant inclusion criteria, vast variants in ClinVar and HGMD databases were labeled with conflict or even opposite clinical significance [25,26], which might attenuate the prediction accuracy of the computational tools. To further expand the training data, some existing tools include sequence variants from population databases, such as Exome Sequencing Project (ESP) [6,11,27]. Most of these tools consider the sequence variants in general populations above a certain allele frequency (AF) as benign; however, how to choose a proper AF threshold for defining neutral training variants remains a question.

The common features adopted by most of the present ensemble models are scores computed by individual predictors based on amino acid or nucleotide conservation, and biochemical properties of the amino acid substitutions [5,6,9,12,13]. While these scores are proven to be highly relevant to variant deleteriousness, other features linked to variant pathogenicity have been shown strong correlations with human diseases. For example, AF has been used as an important criterion in deleterious selection in practice for a long time but was rarely considered in an ensemble model. ClinPred adds AF as input features for the first time and is shown to be more effective than many other ensemble machines [7]. Similarly, genotype frequency (GF) and amino acid frequency (AAF) contain hints of natural selection, which may provide extra information for pathogenicity inference. Additionally, recent studies have shown that intolerance to variation is a strong predictor of human disease relevance, emphasizing the role of genomic context in variant pathogenicity prediction [28–30].

As the algorithm is the "brain" of a machine learning model, the efficiency of a model is largely dependent on algorithm selection. Varieties of machine learning approaches, such as logistic regression [9,13], support vector machine (SVM) [9], random forest [6,13], and boosting algorithms [5,7] have been implemented in variant classification. In general, tree-based approaches achieve higher accuracy and precision according to prior studies [5–7,19]; however, few studies have systematically evaluated the effects of different algorithms on pathogenic variant prioritization. LightGBM is a gradient boosting framework that uses tree-based learning algorithms [31]. Unlike random forests where the component trees are trained independently, in gradient boosting, trees are built in a stepwise manner, where each successive tree is optimized on the residuals of the prediction of the preceding tree. In a previous study, it has been demonstrated that compared with other gradient boosting frameworks such as XGBoost and Catboost, LightGBM converges on a solution that generalizes better [32].

Considering the aforementioned observations, we introduce Pathogenicity Prediction Tool for missense variants (mvPPT), a novel gradient boosting machine for missense variant pathogenicity prediction. By selecting 62 features (including scores from individual predictors, AF/GF/AAF, and genomic context information) and adopting high-confidence variant training sets, mvPPT demonstrates a best-to-date performance in variant pathogenicity prediction, paving its way in molecular diagnosis and clinical scenario applications. mvPPT and pre-computed scores of missense variants in the human exome can be accessed through http://www.mvppt.club/.

Method

Web resources

Web resources of all databases and software used in this study are listed in Table S1.

Missense variant annotation

Variants were annotated by the latest version of ANNOVAR software (version 2019Oct24) [33], with gene-based annotation set to ensGene (assembly version hg19). Variants whose functional consequences were marked as nonsynonymous SNVs were selected. Further removing loss-of-function (stop gain or stop loss) variants ensured that our model was trained and evaluated nearly exclusively on missense variants.

Training set

Training set variants were collected from disease databases: ClinVar (2020.7), HGMD (Pro version 2020.3), and UniProt

(2020.6) [34], as well as a population database from Genome Aggregation Database (gnomAD) genomes (version 2.1.1) [28]. Each variant in ClinVar has a review status tag reporting the level of review supporting the assertion of clinical significance, and a clinical significance tag labeling variants as pathogenic, likely pathogenic, uncertain significance, likely benign, and benign. To select variants with reliable tags, we kept variants with review status of "criteria provided" from submitters and "reviewed by expert panel". The variants were further filtered according to their significance tag: variants that were categorized as (1) benign or likely benign and (2) pathogenic or likely pathogenic were selected as negative (benign) and positive (pathogenic) labels, respectively. The variants in HGMD were labeled by seven different tags, including disease-causing mutation (DM), disease-causing mutation? (DM?), disease-associated polymorphism (DP), disease-associated polymorphism with supporting functional evidence (DFP), *in vitro*/laboratory or *in vivo* functional polymorphism (FP), polymorphic or rare variants reported in the literature (FTV), and retired entry (R). The variants with the DM and DM? labels were reported to be disease-causing in the original literature report. The question mark denotes that a degree of doubt has been found regarding pathogenicity. We only kept the variants with the "DM" label in this study. For UniProt, there are three variant labels: Disease, Polymorphism, and Unclassified. These labels were curated from literature reports. We kept the variants labeled with "Disease" and "Polymorphism" in this study. All variants with conflict labels in different databases were excluded. Population variants were obtained from gnomAD genomes (version 2.1.1), which combines variation data from 15,708 individuals. To avoid any bias, the population variants were further filtered to remove any variants in disease databases. Ten-fold cross-validation was implemented through the Python package scikit-learn (version 0.23.2).

Cross-validation in algorithm selection and feature selection

To avoid overfitting, we designed the ten-fold cross-validation procedure as follows: (1) we divided the variants from disease and 1000 Genomes Project (1KGP) [35] databases into ten subsets; (2) in each round, we selected nine subsets of variants from disease databases to generate the training set, and combined the remaining one subset of variants from disease databases with one subset of 1KGP variants to form the test set.

Cross-validation in training set analysis

The same cross-validation procedure was conducted as above, except that in each round, the training set was generated by using variants from disease databases only, or by adding variants from gnomAD as benign variants. More specifically, we first divided the variants from disease databases into ten subsets. In each round, we selected 90% of variants from the disease databases and combined them with variants in gnomAD genomes passing different AF thresholds (all, removing singleton, AF > 0.0001, AF > 0.001, and AF > 0.01) to generate the training sets for comparison; the remaining 10% of variants from disease databases were then combined with 10% variants from 1KGP to form the test set.

Test set

An independent test set was generated by combining variants from Vereniging Klinisch Genetische Laboratoriumdiagnostiek (VKGL, 2020.9) [36], VariSNP (2017.2.16) [37], Database of Curated Mutations (DoCM, version 3.2) [38], Database of Pathogenic Variants (DPV, 2020.12.29) [39], Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome (CAPICE, version 4) test [19], and MetaLR/SVM_Test [9]. To guarantee a second independent test set with variants that have never been used in any tools' training set, we used PubMed to search for papers reporting new genetic disease-causing genes. In total, we found five papers covering seven genes, and none of these genes include pathogenic variants in the training or the test sets in our study [40–44]. All missense variants on these genes reported in the literature were collected. Population variants were obtained from 1KGP. For each simulated exome, we randomly selected 1000 neutral variants from 1KGP without replacement and added one disease-causing variant. The random seed was set to 1. To validate the robustness of our model and avoid overfitting, variants that were used in any training set or our features' training data were discarded from all test sets, and only variants with comprehensive scores required by all comparator models were included.

Features

mvPPT adopted 62 features from three categories: (A) pathogenicity likelihood scores assessed by different component tools, including Sorting Tolerant From Intolerant (SIFT) [10], MutationAssessor [15], Protein Variation Effect Analyzer (PROVEAN) [45], GERP++RS [46], phyloP [47], phastCons [48], and SIte-specific PHYlogenetic analysis (SiPhy) [49]. (B) AFs, GFs, and AAFs of variants estimated from 125,748 exomes in gnomAD (version 2.1.1); and (C) genomic context of the variant, *i.e.*, region/gene-based information from Gene Variation Intolerance Rank (GeVIR) [29], VIRLoF [29], oe_mis_upper (from gnomAD), Haploinsufficiency Predictions (HIP) [50], Constrained Coding Regions (CCRs) [51], Interpro domain [52], and amino acid sequences before and after mutation. To avoid overfitting, the seven tools we used in category A did not generate scores based on machine learning algorithms. We annotated datasets with ANNOVAR using the database for nonsynonymous SNPs' functional predictions (dbNSFP, v.4.1a) [53,54] to generate some of the required prediction scores from different component tools, including Interpro domain, MutationAssessor, phyloP, GERP++RS, phastCons, PROVEAN, and SiPhy. Mutations located in the Interpro domains were recorded as 1 and the rest were recorded as 0. AFs, GFs, and AAFs of each variant in different populations were obtained from the gnomAD exomes. AFs, AAFs, homozygous frequencies (HomFs), and heterozygous frequencies (HetFs) were assigned 0, and wild-type frequencies (WtFs) were assigned 1, if the variant was not present in the database. The GeVIR, VIRLoF, oe_mis_upper, HIP, and CCRs scores were downloaded from their respective websites. One-hot encoding has been applied to amino acid sequence, representing each amino acid with a binary vector of length 20 with a single non-zero value. All the features were selected to provide complementary information, and they

either did not require training or their training data are publicly available to allow exclusion from our data.

Outlier detection and Gene Ontology enrichment analysis

The interquartile range (IQR) was used to identify outliers. The IQR criterion is summarized as follows:
1. Compute the first and third quartiles, Q_{1j} and Q_{3j}, for each peptide j, and then its IQR: $IQR_j = Q_{3j} - Q_{1j}$.
2. For each peptide j, observation y_{ij} is flagged as an outlier if $y_{ij} < Q_{1j} - k \times IQR_j$ or $y_{ij} > Q_{3j} + k \times IQR_j$, where $k = 1.5$.

Gene Ontology (GO) enrichment analysis was performed with the R package *clusterProfiler* [55–58].

Metrics for performance evaluation

We used 11 different metrics to evaluate the performance of the prediction tools. A detailed description of the metrics is provided in Table S2.

mvPPT training

mvPPT was trained using the Python package LightGBM (version 2.3.1) [31], and parameters were tuned by Bayesian optimization (version 1.2.0). The random status was set as 1 throughout the model training process. For Bayesian optimization process, the number of iterations was set as 100 (n_iter = 100) and the number of steps of random exploration was set as 15 (init_points = 15). The ranges of the hyperparameters in the LightGBM for Bayesian optimization were set as follows: num_leaves (24, 45), feature_fraction (0.1, 0.9), bagging_fraction (0.8, 1), max_depth (5, 8.99), lambda_l1 (0, 5), lambda_l2 (0, 3), min_split_gain (0.001, 0.1), and min_child_weight (5, 50). After the parameter optimization process, the final used values of the parameters were as follows: num_leaves = 45, min_child_weight = 6.163, learning_rate = 0.01, bagging_fraction = 0.870, feature_fraction = 0.632, lambda_l1 = 0. 921, lambda_l2 = 0.193, min_gain_to_split = 0.039, and max_depth = 9.

Scores from existing tools

The scores for REVEL [6], ClinPred [7], PrimateAI [8], MetaSVM/MetaLR [9], VEST4 [11], MVP [12], PolyPhen-2 [16], and FATHMM-XF [20] were obtained from dbNSFP v4.1a. The scores for Mendelian Clinically Applicable Pathogenicity (M-CAP) [5], MISsense deleTeriousness predICtor (MISTIC) [13], CAPICE [19], Combined Annotation Dependent Depletion (CADD) [21], and ReVe [22] were downloaded from their respective websites.

Statistical analysis

Wilcoxon matched-pairs signed-rank test was conducted using the stats module in SciPy Python package (version 1.5.4). Adjusted P value in GO enrichment analysis was calculated by the R package clusterProfiler. All the metrics in this study were calculated based on the scikit-learn Python package.

Results

The prediction model was refined with various algorithm and feature selection

The performance of a machine learning model is mainly determined by the algorithm, the features, and the training set used. Therefore, we designed mvPPT by careful selection of the algorithm, features, and training set (**Figure 1**).

We first benchmarked the performance of ten commonly used algorithms, including SVM, naive Bayes, logistic regression, decision tree, random forest, extra forest, gradient boosting machine (GBM), AdaBoost, LightGBM, and bagging on the data test set mentioned above. The performance of each algorithm was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). We found that the performance of the ensemble learning (random forest, extra forest, GBM, AdaBoost, LightGBM, and bagging) achieved the higher AUROC and AUPRC (**Figure 2**). Among them, LightGBM has the highest AUROC (0.970 ± 0.001) and AUPRC (0.952 ± 0.002) (Figure 2).

To obtain a proper feature space, we extracted three types of features, including prediction scores (category A), frequency features (category B), and region-based features (category C) (Figure 1; **Table 1**). We evaluated the performance of models with the following combinations of feature categories: A, A + B, A + C, and A + B + C (Table 1). Overall, adding category B or category C greatly boosted the performance, compared with the model including only category A, which is the case of most previous ensemble machines. Specifically, the model including A + B + C achieved the highest performance (**Figure 3**A and B). As pathogenic variants in databases may be enriched with low-frequency variants, we also assessed the performance of different models on rare variants. When all the variants in the test set were rare (AF = 0, based on gnomAD exomes), the main contributions came from category A and category C, and adding category B showed few but slightly positive impacts (Figure 3C and D). Altogether, both category B and category C improved forecasting accuracy and were thus included in mvPPT.

As population variants tend to have higher AFs than pathogenic variants, modes including AFs as features may perform better in test sets including population variants. Therefore, we excluded 1KGP variants from the test set and reconducted the comparison models that displayed similar performance on test sets with or without population variants (Figure S1).

Training data prefiltration improves the model performance

We observed that genomic locations of variants recorded in disease databases are likely to be biased by interests of the research field, *i.e.*, variants in the databases are likely to be enriched in "hotspots" of the human genome. To evaluate the enrichment pattern of variants from different databases on genome, we calculated the ratio between the number of missense variants in each gene and the length of the gene's protein-coding sequence (VPR). We found that VPR in disease

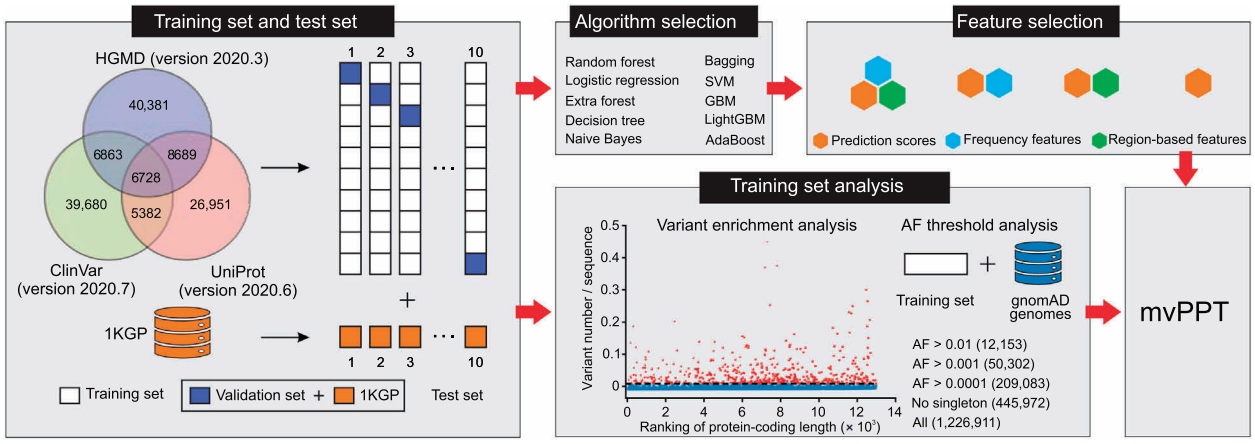


Figure 1 mvPPT workflow

The detailed mvPPT construction process is shown. High-confidence variant sets were extracted from ClinVar, HGMD, and UniProt. Models were trained using LightGBM with parameters tuned by Bayesian global optimization. Ten-fold cross-validation is carried out to verify the effectiveness of the prediction model. mvPPT is built after algorithm selection, feature selection, and training set analysis. mvPPT, Pathogenicity Prediction Tool for missense variants; HGMD, Human Gene Mutation Database; 1KGP, 1000 Genomes Project; SVM, support vector machine; GBM, gradient boosting machine; gnomAD, Genome Aggregation Database; AF, allele frequency.

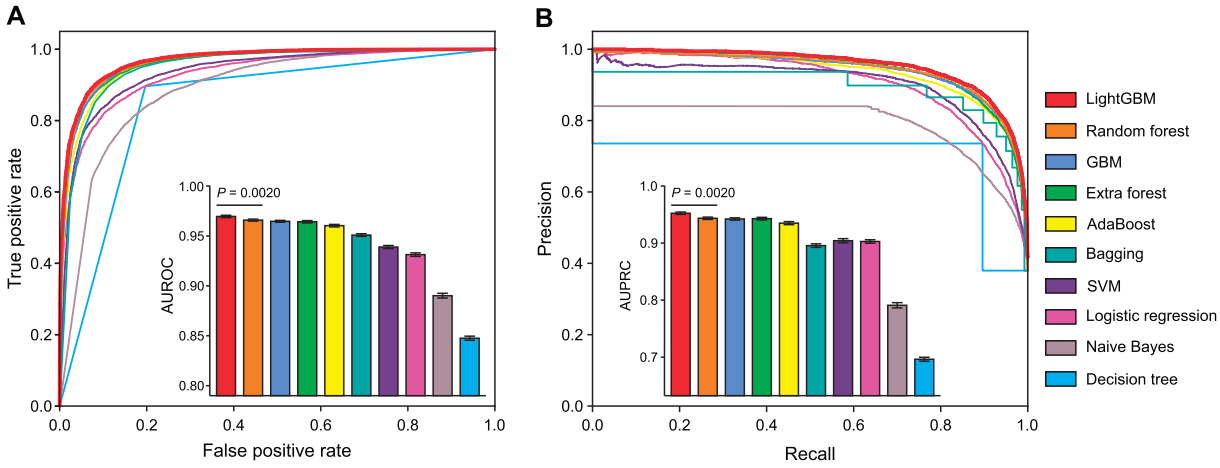


Figure 2 Performance comparison of different algorithms

Performance comparison of models using different algorithms. **A.** The ROC for models trained on different algorithms. **B.** The PRC for models trained on different algorithms. Below: barplot quantifications of the AUROC and the AUPRC values by using ten-fold cross-validation. *P* value was calculated by Wilcoxon matched-pairs signed-rank test. ROC, operating characteristic curve; PRC, precision-recall curve; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.

databases is much variable than that in gnomAD, with the coefficient of variation (CV) of 0.815% for gnomAD and 2.35% for disease databases (**Figure 4**A–C). Outlier detection based on IQR detected 1.12% of the genes as outliers (VPR > mean + 1.5 × IQR) in gnomAD, but 13.48% of the genes as outliers in disease databases (Figure 4D; Table S3). Likewise, when plotting the number of variants against the length of the coding sequence, we observed a significant positive correlation for variants in gnomAD, but not for variants in the pathogenic databases (Figure 4E and F). GO enrichment analysis revealed that outlier genes in gnomAD are enriched in pathways associated with immune response, which are known to be hotspots of positive selection (Figure 4G). In contrast, top enriched pathways of outlier genes in pathogenic databases include gland development (adjusted P = 1.33E−12), regulation of body fluid levels (adjusted P = 8.61E−12), and response to an inorganic substance (adjusted P = 9.70E−11) (Figure 4H), reflecting that different variant enrichment patterns are there in the disease and population databases.

To further test if aggregation of genetic variants on genes impairs the model performance, we down-sampled pathogenic variants on genes with large numbers of pathogenic variants. Specifically, on genes with VPR greater than a set threshold (0.008, 0.015, 0.040, 0.065, and 0.090), we randomly selected a fixed number of variants and combined them with variants from other genes to form the down-sampled training sets (Figure S2A). The down-sampled training sets were then fed to

Table 1 Features associated with missense variants mined in this work

Category	Feature	Definition
A	MutationAssessor, SIFT, PROVEAN, GERP + + RS, phyloP100way_vertebrate, phyloP30way_mammalian, phyloP17way_primate, phastCons100way_vertebrate, phastCons30way_mammalian, phastCons17way_primate, SiPhy_29way_logOdds	Pathogenicity likelihood scores generated by 7 non-machine learning-based component tools
B	gnomAD_WtFs, gnomAD_AFs, gnomAD_HetFs, gnomAD_HomFs, gnomAD_AAFs	AAFs, AFs, and GFs of each variant estimated from 125,748 exomes in gnomAD (version 2.1.1)
C	GeVIR, VIRLoF, HIP, CCRs, Interpro_domain, oe_mis_upper, RefAA_A, RefAA_C, RefAA_D, RefAA_E, RefAA_F, RefAA_G, RefAA_H, RefAA_I, RefAA_K, RefAA_L, RefAA_M, RefAA_N, RefAA_P, RefAA_Q, RefAA_R, RefAA_S, RefAA_T, RefAA_V, RefAA_W, RefAA_Y, AltAA_A, AltAA_C, AltAA_D, AltAA_E, AltAA_F, AltAA_G, AltAA_H, AltAA_I, AltAA_K, AltAA_L, AltAA_M, AltAA_N, AltAA_P, AltAA_Q, AltAA_R, AltAA_S, AltAA_T, AltAA_V, AltAA_W, AltAA_Y	Region/gene-based scores

Note: gnomAD_WtFs, wild-type frequencies estimated from all exomes in gnomAD; gnomAD_AFs, allele frequencies estimated from all exomes in gnomAD; gnomAD_HetFs, heterozygous frequencies estimated from all exomes in gnomAD; gnomAD_HomFs, homozygous frequencies estimated from all exomes in gnomAD; gnomAD_AAFs, amino acid frequencies estimated from all exomes in gnomAD; RefAA, reference amino acid; AltAA, alternate amino acid; AAF, amino acid frequency; AF, allele frequency; GF, genotype frequency.

models using different feature combinations (A + B and A + B + C). To avoid a similar variant enrichment pattern in the test set, we randomly selected one pathogenic variant from each gene to form a test set. Ten test sets were created in each round of the ten-fold cross-validation. We found that the predictive power of the models reduced with down-sampling (Figure S2B and C), which is possible because the number of variants available for learning is largely reduced with down-sampling. Adding category C slowed down the reduction. However, overall, down-sampling appears to attenuate the performance of the model, and thus is disfavored (Figure S2B and C).

Next, we compared the performance of models trained on six training sets generated by different strategies. In our assessment, we found that adding variants from the population database skewed the distribution of benign variant AFs toward zero, making it similar to that of pathogenic variants (Figure S3). Including neutral variants from the population database have a positive impact on the model, with adding variants without singletons displaying the highest performance (Figure 4I and J). In contrast, adding the full variants from gnomAD jeopardizes the model performance, possibly due to the inclusion of variants in the population database that are not true benign variants (Figure 4I and J). To further investigate whether this improvement is due to the correlation of AFs between training and test sets, we divided the original test sets into bins based on AFs (based on gnomAD exomes) and tested the performance of the models on variants within a specific AF range. We found that including population variants removing singletons slightly enhances the performance of the model in most of the bins, especially the low-AF bins, where most of the pathogenic variants are located, likely because it expands the training set while excluding unreliable (singleton) samples (Table S4). In contrast, the performance of the model drops with AF cutoff increases, probably due to the lack of rare benign variants in the training set.

Based on the aforementioned analyses, mvPPT was finally trained using LightGBM (tuned by Bayesian optimization) on variants from three disease databases and gnomAD with singletons removed, using all features from categories A, B, and C. The correlation among the individual features and relative importance of these features are shown in Figure S4 and the description of the training set is shown in Table S5.

mvPPT outperforms existing prediction tools

For assessment, we collected variants from VariSNP, VKGL, DPV, DoCM, MetaLR/SVM_Test, and CAPICE_Test, to generate an independent test set. Variants that were used in any training set or in our features' training data were discarded from the test set, and only variants with comprehensive scores required by all comparators were included. In total, there are 175,144 variants in the test set with 168,222 benign variants and 6922 pathogenic variants (Table S6).

Using the new test set, the performance of mvPPT was benchmarked against 15 prediction tools that are widely used and readily implemented, including MVP, CAPICE, FATHMM-XF, REVEL, M-CAP (version 1.4), ClinPred, ReVe, PrimateAI, MetaSVM, MetaLR, MISTIC, CADD (version 1.4), PolyPhen-2 HDIV, PolyPhen-2 HVAR, and VEST (version 4). Among all these tools, mvPPT has the highest

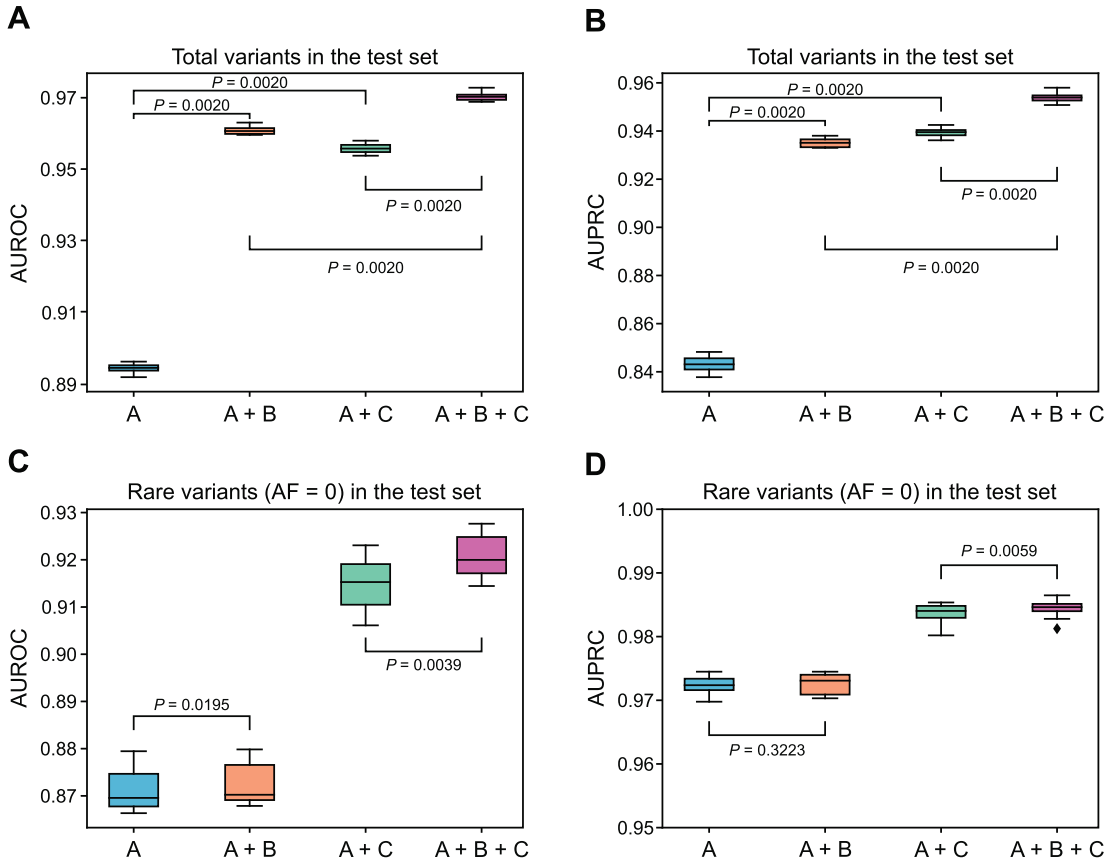


Figure 3 The prediction power of features by category

A. AUROC obtained from models trained on each combination of categories. **B.** AUPRC obtained from models trained on each combination of categories. **C.** AUROC evaluating the performance of models trained on different combinations of categories when all the variants in the test set were rare (AF = 0, based on gnomAD exomes). **D.** AUPRC evaluating the performance of models trained on different combinations of categories when all the variants in the test set were rare (AF = 0, based on gnomAD exomes). AUROC and AUPRC from ten-fold cross-validation were plotted. *P* value was calculated by Wilcoxon matched-pairs signed-rank test.

AUROC of 0.960 and the highest AUPRC of 0.791 (**Figure 5**). MISTIC has the second-best overall performance, with AUROC of 0.920 and AUPRC of 0.565. Besides, we also calculated other metrics include accuracy, precision [also known as positive predictive value (PPV)], sensitivity [also known as true positive rate (TPR)], F1 score, log loss, Matthews correlation coefficient (MCC), true negative rate (TNR; also known as specificity), false positive rate (FPR), and diagnostic odd ratio (DOR) (**Table 2**). M-CAP has the highest sensitivity of 0.953 (second-best: mvPPT, 0.888), but this comes at the cost of a low precision of 0.078 (best: mvPPT, 0.323). PrimateAI has the highest specificity of 0.925 (second-best: mvPPT, 0.923) and lowest FPR of 0.075 (second-best: mvPPT, 0.077). Here, mvPPT has the highest accuracy of 0.922, the highest F1 score of 0.473, the highest DOR of 95.102, the highest MCC value of 0.508, the highest precision of 0.323, and the lowest log loss of 2.697 (Table 2).

To further evaluate the robustness of mvPPT, we proceeded to random sampling. We repeated the random sampling for 20 rounds. In each round, 20% of the variants in the independent test set were sampled. The results showed that mvPPT displayed the highest efficiency and robustness (Figure S5). Since mvPPT included AFs as features, we then tested the performance of our model when lacking AF information. We compared the predictive power of mvPPT with existing methods on variants with different AF levels (based on gnomAD exomes). As shown in Figure S6, mvPPT performed the best on variants with different AF levels.

For further assessment, we assembled a test set with pathogenic variants from DoCM, a highly curated database of known, disease-causing mutations in cancer-derived from literature, and benign variants randomly selected from VariSNP and VKGL. mvPPT again achieved the best performance in this test set (Figure S7).

Performance of mvPPT on pathogenic variants within novel disease-causing genes

To further evaluate the performance of our predictor on variants in new disease-causing genes (*i.e.*, genes which were reported as causative genes of Mendelian diseases for the first time but have not been included in disease database yet), we collected seven disease-causing genes from five recent publications [40–44], and 62 missense variants were retained with complete scores on all comparators (Table S7). We simulated 62 exomes of Mendelian diseases, by selecting one

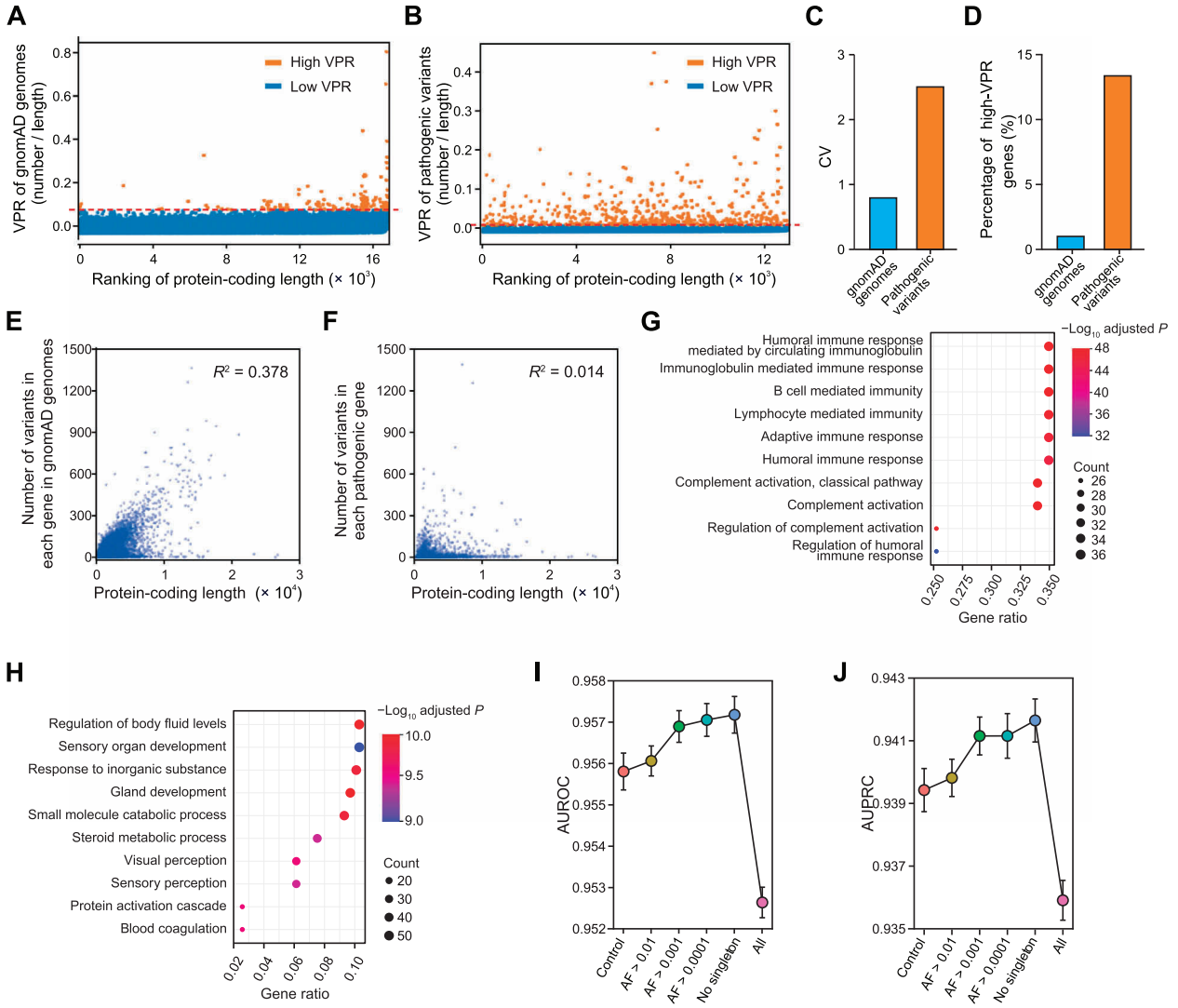


Figure 4 The enrichment pattern of variants from different databases and AFs selected

The VPR was calculated. **A.** After being centered, VPR in gnomAD genomes was plotted against the ranking of protein-coding sequence length. **B.** After being centered, VPR in pathogenic databases was plotted against the ranking of protein-coding sequence length. **C.** The CV for VPR in gnomAD genomes and pathogenic databases. **D.** The percentage of outlier genes related to gnomAD genomes and pathogenic databases. **E.** Variant number profile across all genes in gnomAD genomes. **F.** Variant number profile across all genes in pathogenic databases. R^2 corresponds to the coefficient of determination. **G.** Bubble chart showing GO enrichment analysis of variants in gnomAD genomes. **H.** Bubble chart showing GO enrichment analysis of variants in pathogenic databases. **I.** AUROC obtained from models trained on training sets combining gnomAD variants selected according to different AF thresholds. **J.** AUPRC obtained from models trained on training sets combining gnomAD variants selected according to different AF thresholds. "Control" represents the model without incorporating gnomAD variants, "No singleton" represents the model incorporating gnomAD variants with singletons removed, and "All" represents the model with all gnomAD variants incorporated. VPR, ratio between the number of missense variants in each gene and the length of the gene's protein-coding sequence; CV, coefficient of variation; GO, Gene Ontology.

disease-causing variant and randomly selecting 1000 neutral variants from 1KGP. For each simulated exome, we calculated the percentage of predicted pathogenic variants obtained by different predictors, according to the authors' recommended threshold (**Figure 6**A; Table S8). The ranking of pathogenic variants among all variants in each simulated exome is presented in Figure 6B. Among all predictors, PrimateAI generated the shortest list of pathogenic variants, followed by MISTIC and mvPPT. However, only 41 and 44 of the 62 variants were predicted as pathogenic by PrimateAI and MISTIC, respectively (Table S8). CADD identified 100% of these 62 variants as pathogenic, but may cause a plenty of false positives (PPV = 0.615 ± 0.002) (Figure 6A; Table S8). Instead, mvPPT performed relatively well in both sensitivity (60/62) and PPV (0.137 ± 0.001) (Figure 6A; Table S8). To further evaluate the ability of each predictor in prioritizing the pathogenic variants, we computed the ranking of pathogenic variants among all variants in simulated exomes (Figure 6B). Pathogenic variants showed the best ranking in mvPPT (29 ± 9), significantly better than the rest of the tools (second-best: ClinPred 75 ± 11, P = 2.12E−08, Wilcoxon matched-pairs signed-rank test), further demonstrating the

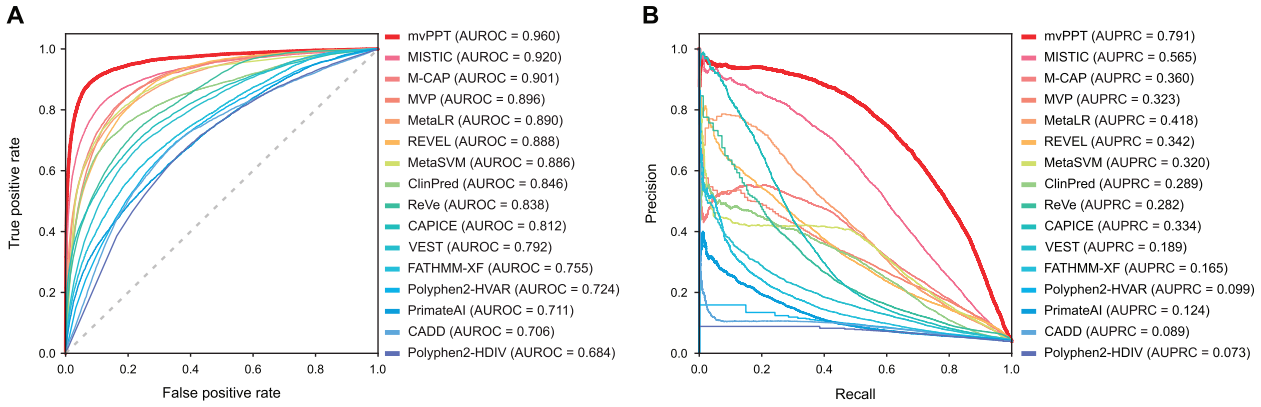


Figure 5 Performance comparison of mvPPT and existing prediction tools

A. ROCs of mvPPT and 15 established prediction methods when tested on an independent test set. **B.** PRCs of mvPPT and 15 established prediction methods.

Table 2 Overview of the performance of mvPPT in comparison to other tools in an independent test set

Predictor	AUROC	AUPRC	Accuracy	Precision	Sensitivity	F1 score	Log loss	MCC	TNR	FPR	DOR
mvPPT	**0.960**	**0.719**	**0.922**	**0.323**	0.888	**0.473**	**2.697**	**0.508**	0.923	0.077	**95.102**
MISTIC	0.920	0.565	0.863	0.203	0.839	0.327	4.718	0.371	0.864	0.136	33.238
M-CAP	0.901	0.360	0.551	0.078	**0.953**	0.144	15.507	0.190	0.534	0.466	23.381
MVP	0.896	0.323	0.773	0.134	0.869	0.232	7.853	0.285	0.769	0.231	22.063
MetaLR	0.890	0.418	0.830	0.160	0.782	0.266	5.889	0.303	0.831	0.169	17.726
REVEL	0.888	0.342	0.849	0.171	0.733	0.278	5.204	0.305	0.854	0.146	16.087
MetaSVM	0.886	0.320	0.847	0.175	0.772	0.286	5.271	0.320	0.851	0.149	19.266
ClinPred	0.846	0.289	0.683	0.096	0.829	0.171	10.942	0.208	0.677	0.323	10.157
ReVe	0.838	0.282	0.608	0.080	0.856	0.147	13.545	0.179	0.598	0.402	8.817
CAPICE	0.812	0.334	0.733	0.101	0.731	0.178	9.224	0.200	0.733	0.267	7.457
VEST	0.792	0.189	0.634	0.079	0.780	0.144	12.638	0.163	0.628	0.372	5.997
FATHMM-XF	0.755	0.165	0.586	0.069	0.761	0.127	14.285	0.134	0.579	0.421	4.388
Polyphen2_HVAR	0.724	0.099	0.687	0.079	0.651	0.141	10.809	0.141	0.689	0.311	4.128
PrimateAI	0.711	0.124	0.901	0.145	0.308	0.197	3.426	0.164	**0.925**	**0.075**	5.505
CADD	0.706	0.087	0.417	0.054	0.835	0.102	20.153	0.093	0.399	0.601	3.354
Polyphen2_HDIV	0.684	0.073	0.545	0.061	0.737	0.114	15.711	0.107	0.537	0.463	3.250

Note: mvPPT, Pathogenicity Prediction Tool for missense variants; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; MCC, Matthews correlation coefficient; TNR, true negative rate; FPR, false positive rate; DOR, diagnostic odd ratio. The best scores in each column are bolded.

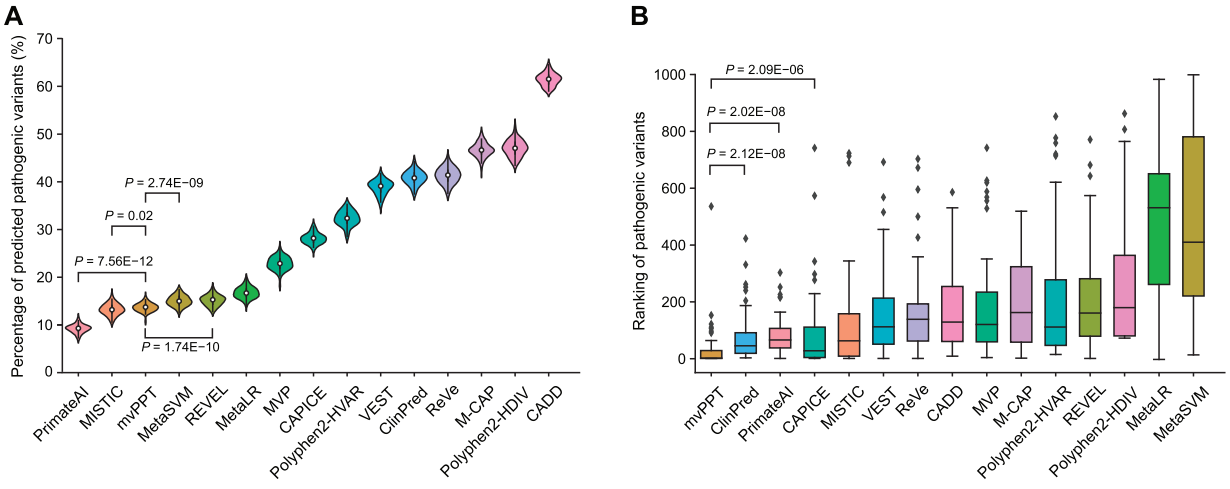


Figure 6 Evaluation of the different prediction tools using simulated disease exomes

A. Distribution of the percentage of predicted pathogenic variants in the simulated disease exomes. **B.** Ranking of the pathogenic variants in the simulated disease exomes.

advantages of mvPPT on detecting pathogenic variants within novel disease-causing genes.

Discussion

Missense variants as the most common category of SNVs have important implications for human genetic diseases. Although a variety of variant pathogenicity assessment tools have been established and have made important contributions to genetic variant evaluation, there is still room for improvement in prediction accuracy and precision, which is of great importance for the explanation of the tremendous number of genetic variants. In this study, we present a comprehensive prediction tool, mvPPT, and demonstrate that the performance of mvPPT is superior to other existing comparators in AUROC, AUPRC, F1 score, and many other metrics. We found that the improvement in prediction probably resulted from the careful selection of the algorithms, the features, and the training sets.

Boosting algorithms are widely adopted by many ensemble models of variant classification to improve the accuracy of prediction. mvPPT adopted a recently developed LightGBM algorithm proven to outperform the existing boosting frameworks on both efficiency and accuracy. Other than the boosting algorithms, we also tried a deep neural network framework and found that our model gave higher accuracy and precision than 10-layer, 15-layer, and 20-layer fully connected neural networks (data not shown). Compared with traditional machine learning algorithms, deep learning performs better as the scale of data increases [59]. We expect more accurate prediction models based on deep neural networks in the near future, with the accumulation of larger amounts of high-confidence training data. Recently developed deep learning models such as primateAI, SpliceAI [60], and EVE [61] provided new perspectives of variant pathogenicity prediction, which use surrounding DNA sequences as input, without requiring explicit features. These approaches could be possibly further improved with larger training data, as well as incorporating sequence conservation, constraints, and protein structure information.

In this study, mvPPT adopted 62 features belonging to three categories. Other than commonly used features extracted from previous predictors (category A), mvPPT included two categories of features associated with allele/genotype/amino acid frequencies (category B) and genetic constraint of adjunction regions (category C). Our benchmarking studies revealed that category B and category C contributed significantly to the prediction. Among them, features in category C contributed the most as a whole, and adding category B further promotes the performance (Figure S4). The high predictive power of features in categories B and C could be explained by the fact that natural selection constantly eliminates deleterious variants during evolution, and thus deleterious variants tend to locate in more conserved and intraspecies constrained genomic regions with lower AF/GF/AAF in human populations compared with neutral mutations. Considering that the pathogenic variants in disease databases are likely to have small AF and GF due to existing pathogenic variant selection criteria, we further tested our model on rare variants with AF = 0 and confirmed that the model works well even AF is not functioning as a feature in this case. Other than the features we used, protein structural changes corresponding to changes in amino acid sequence may also be critical predictors of variant pathogenicity. Newly developed protein structure prediction tools, such as AlphaFold2 [62,63], have made it possible to include protein structure information in future tools.

In addition to the algorithm and feature selection, the improvement in the performance of mvPPT also came from a cautious selection of the training set. While most of the prevailing variant classifiers trained their models on single databases, previous studies have uncovered considerable disagreement among databases [7,9]. Some tools also enrolled variants in general population databases (*e.g.*, Exome Aggregation Consortium [64]) as benign variants, to increase the size of the training set. However, this setting may add noise to the training data, as not all variants in population databases are truly benign. In this study, we found that adding full sets of variants from gnomAD attenuates the predictive performance of our model. On the contrary, applying appropriate filters to population variants assures high-quality training data, and improves the predictive ability of our model. The possible explanation of this observation is that the singletons are more likely to be contaminated with false benign variants or less confident variants compared with non-singletons. Therefore, we examined the labels of singletons from gnomAD that we added to the training set in ClinVar. Among gnomAD with labels in ClinVar, we found that 8.1% of them were benign (labeled as "benign", "likely benign", and "benign/likely benign"), 78.5% were labeled as "uncertain", 7.4% were labeled as "conflicting interpretations of pathogenicity", and 6.0% were pathogenic (labeled as "pathogenic", "likely pathogenic", and "pathogenic/likely pathogenic"). In contrast, among the overlapping non-singleton missense variants, 47.8% were benign, 34.8% were labeled as "uncertain", 15.3% were labeled as "conflicting interpretations of pathogenicity", and 2.1% were pathogenic. Overall, our observation highlights the importance of maintaining a balance between size and purity of the training set, as well as provides practical guidance of training set selection.

As reported by previous studies, we observed that the pathogenic variants in disease databases are enriched in certain genes. However, our down-sampling experiments suggest that the aggregation of variants has few effects on pathogenicity prediction. This can be possibly explained by the fact that most features of variants are independent of genes. Furthermore, we found that adding region/gene-based information (category C) slowed down the effects of down-sampling, suggesting that incorporating genomic context further lessens the impacts on the variant aggregation (Figure S2B and C).

In conclusion, we developed an ensemble classifier, mvPPT, for predicting the pathogenicity of missense variants, and demonstrated that mvPPT achieved superior performance compared with other established prediction tools. Particularly, in clinical data, mvPPT showed the highest accuracy and robustness in classifying variants associated with both Mendelian diseases and cancer. Therefore, mvPPT promises to facilitate a better clinical interpretation of missense variants with uncertain significance. For convenient use, we built a searchable website and all pre-computed mvPPT scores are available at http://www.mvppt.club/.

Code availability

The mvPPT scores for potential missense variants in the human genome are available at http://www.mvppt.club/. The

mvPPT codes are available at https://ngdc.cncb.ac.cn/biocode/tools/BT007292 as well as at https://github.com/tongshiyuan/mvPPT for noncommercial purposes.

Competing interests

Zai-Wei Zhou is a current employee of Shanghai Xunyin Biotechnology Co., Ltd. All the other authors have declared no competing interests.

CRediT authorship contribution statement

Shi-Yuan Tong: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Visualization. **Ke Fan:** Validation, Formal analysis, Investigation, Writing – original draft. **Zai-Wei Zhou:** Methodology, Investigation, Writing – original draft. **Lin-Yun Liu:** Validation, Writing – original draft. **Shu-Qing Zhang:** Validation, Data curation, Writing – original draft. **Yinghui Fu:** Validation, Data curation, Writing – original draft. **Guang-Zhong Wang:** Writing – original draft, Data curation. **Ying Zhu:** Conceptualization, Formal analysis, Resources, Writing – review & editing, Supervision, Project administration. **Yong-Chun Yu:** Conceptualization, Resources, Writing – review & editing, Visualization, Supervision, Project administration. All authors have read and approved the final manuscript.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2021ZD0202500), the Shanghai Natural Science Foundation, China (Grant No. 20ZR1403800), the National Natural Science Foundation of China (Grant Nos. 31900476, 82071259, 31930044, and 31725012), the Shanghai Municipal Science and Technology Major Project (Grant No. 2018SHZDZX01) and ZJ Lab, the Shanghai Center for Brain Science and Brain-Inspired Technology, China, the Foundation of Shanghai Municipal Education Commission, China (Grant No. 2019-01-07-00-07-E00062), and the Collaborative Innovation Program of Shanghai Municipal Health Commission, China (Grant No. 2020CXJQ01). We are thankful to Dr. Yvette Chin for English language editing.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.07.005.

ORCID

ORCID 0000-0002-7899-5728 (Shi-Yuan Tong)
ORCID 0000-0003-1488-1200 (Ke Fan)
ORCID 0000-0002-3955-9018 (Zai-Wei Zhou)
ORCID 0000-0002-6902-759X (Lin-Yun Liu)
ORCID 0000-0001-7216-9284 (Shu-Qing Zhang)
ORCID 0000-0003-4748-4498 (Yinghui Fu)
ORCID 0000-0001-6432-8310 (Guang-Zhong Wang)
ORCID 0000-0002-6594-3734 (Ying Zhu)
ORCID 0000-0002-7456-7451 (Yong-Chun Yu)

References

[1] Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA 2014;312:1880–7.
[2] Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. N Engl J Med 2013;369:1502–11.
[3] Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum Genomics 2014;8:11.
[4] Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, et al. Genetic variation in an individual human exome. PLoS Genet 2008;4:e1000160.
[5] Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 2016;48:1581–6.
[6] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 2016;99:877–85.
[7] Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. Am J Hum Genet 2018;103:474–83.
[8] Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018;50:1161–70.
[9] Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 2015;24:2125–37.
[10] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
[11] Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 2013;14:S3.
[12] Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 2021;12:510.
[13] Chennen K, Weber T, Lornage X, Kress A, Bohm J, Thompson J, et al. MISTIC: a prediction tool to reveal disease-relevant deleterious missense variants. PLoS One 2020;15:e0236962.
[14] Ip E, Chapman G, Winlaw D, Dunwoodie SL, Giannoulatou E. VPOT: a customizable variant prioritization ordering tool for annotated variants. Genomics Proteomics Bioinformatics 2019;17:540–5.
[15] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118.
[16] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[17] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014;11:361–2.
[18] Li Q, Liu X, Gibbs RA, Boerwinkle E, Polychronakos C, Qu HQ. Gene-specific function prediction for non-synonymous mutations in monogenic diabetes genes. PLoS One 2014;9:e104452.
[19] Li S, van der Velde KJ, de Ridder D, van Dijk ADJ, Soudis D, Zwerwer LR, et al. CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. Genome Med 2020;12:75.
[20] Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic

point mutations via extended features. Bioinformatics 2018;34:511–3.
[21] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47:D886–94.
[22] Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res 2018;46:7793–804.
[23] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–7.
[24] Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017;136:665–77.
[25] Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol 2013;425:4047–63.
[26] Salnikova LE, Kolobkov DS, Sviridova DA, Abilev SK. An overview of germline variations in genes of primary immunodeficiences through integrative analysis of ClinVar, HGMD and dbSNP databases. Hum Genet 2021;140:1379–93.
[27] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64–9.
[28] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–43.
[29] Abramovs N, Brass A, Tassabehji M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. Nat Genet 2020;52:35–9.
[30] Vitsios D, Dhindsa RS, Middleton L, Gussow AB, Petrovski S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nat Commun 2021;12:1504.
[31] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 31st Annual Conference on Neural Information Processing Systems 2017:3149–57.
[32] Anghel A, Papandreou N, Parnell T, De Palma A, Pozidis H. Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv 2018;1809.04559.
[33] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
[34] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.
[35] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
[36] Fokkema IFAC, van der Velde KJ, Slofstra MK, Ruivenkamp CAL, Vogel MJ, Pfundt R, et al. Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. Hum Mutat 2019;40:2230–8.
[37] Schaafsma GC, Vihinen M. VariSNP, a benchmark database for variations from dbSNP. Hum Mutat 2015;36:161–6.
[38] Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods 2016;13:806–7.
[39] Suzuki H, Kurosawa K, Fukuda K, Ijima K, Sumazaki R, Saito S, et al. Japanese pathogenic variant database: DPV. Transl Sci Rare Dis 2018;3:133–7.
[40] Fliedner A, Kirchner P, Wiesener A, van de Beek I, Waisfisz Q, van Haelst M, et al. Variants in *SCAF4* cause a neurodevelopmental disorder and are associated with impaired mRNA processing. Am J Hum Genet 2020;107:544–54.
[41] Palencia-Campos A, Aoto PC, Machal EMF, Rivera-Barahona A, Soto-Bielicka P, Bertinetti D, et al. Germline and mosaic variants in *PRKACA* and *PRKACB* cause a multiple congenital malformation syndrome. Am J Hum Genet 2020;107:977–88.
[42] Tsai MH, Muir AM, Wang WJ, Kang YN, Yang KC, Chao NH, et al. Pathogenic variants in *CEP85L* cause sporadic and familial posterior predominant lissencephaly. Neuron 2020;106:237–45.
[43] Hadjadj J, Castro CN, Tusseau M, Stolzenberg MC, Mazerolles F, Aladjidi N, et al. Early-onset autoimmunity associated with *SOCS1* haploinsufficiency. Nat Commun 2020;11:5341.
[44] Lessel D, Zeitler DM, Reijnders MRF, Kazantsev A, Nia FH, Bartholomaus A, et al. Germline *AGO2* mutations impair RNA interference and human neurological development. Nat Commun 2020;11:5797.
[45] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One 2012;7:e46688.
[46] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 2010;6:e1001025.
[47] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110–21.
[48] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–50.
[49] Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 2009;25:i54–62.
[50] Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 2010;6:e1001154.
[51] Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet 2019;51:88–95.
[52] Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;30:1236–40.
[53] Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
[54] Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020;12:103.
[55] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[56] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 2019;47:D419–26.
[57] Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021;49:D325–34.
[58] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[59] Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep learning and its applications in biomedicine. Genomics Proteomics Bioinformatics 2018;16:17–32.
[60] Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell 2019;176:535–48.

[61] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature 2021;599:91–5.
[62] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[63] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50:D439–44.
[64] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91.