Supplementary Material

PhD-SNP^g: updating a webserver and lightweight tool for scoring nucleotide variants.

Emidio Capriotti[1]* and Piero Fariselli[2]*

[1] BioFolD Unit, Department Pharmacy and Biotechnology (FaBiT),
University of Bologna, Via F. Selmi 3, Bologna, 40126, Italy.

[2] Department of Medical Sciences, University of Torino,
Via Santena 19, 10126, Torino, Italy.

*Correspondence should be addressed: emidio.capriotti@unibo.it, piero.fariselli@unito.it

1. Collection of the variant's datasets

The datasets used for training and testing PhD-SNP^g have been extracted from two releases of the Clinvar database (1) (December 2020 and 2022). From the first version of Clinvar, the Single Nucleotide Variants (SNVs) and InDels annotated as "*Benign/Likely Benign*" and *"Pathogenic/Likely Pathogenic"* were collected in an initial dataset (Clinvar122020). A second dataset was generated by collecting the newly annotated variants from December 2020 and December 2022 (NewClinvar122022). The composition of Clinvar122020 and NewClinvar122022 datasets in terms of SNVs and InDels is summarized in Tables S1-S2 and Figure S1. For balancing the composition of *Pathogenic* and *Benign* variants in the training and testing sets, a downsampling procedure of the most abundant class was adopted. Following this approach, from the initial version of Clinvar (December 2020) were collected the Clinvar122020-SNV and Clinvar122020-InDel datasets, which consist of 103,916 SNVs and 34,290 InDels respectively. From the newly annotated set of variants (NewClinvar122022) were generated the datasets NewClinvar122022-SNV and NewClinvar122022-InDel which includes of 42,594 SNVs and 9,046 InDels. For each one of the four datasets, 5 replicates with an equally distributed number of *Pathogenic* and *Benign* variants were generated by sampling the most abundant class.

2. Input features

PhD-SNP^g takes in input sequence and conservation-based features from the UCSC (University of California, Santa Cruz) repository (http://hgdownload.cse.ucsc.edu/).

PhD-SNP^g predicts the impact of SNVs and InDels considering as basic input data is a 35-element vector, 25 elements encode for the sequence of a 5-nucleotide window centered around the mutated position (2 nucleotides in each direction). Each position in the window is represented by a 5-element vector for the 5 possible nucleotides (A,C,G,T,N). The element corresponding to the nucleotide in the sequence is set to +1 and the remaining elements are 0. The 5-element vector encoding for the mutated position in the center of the window is built by settings: -1 to the element associated with the reference nucleotide, +1 to the element associated with the mutant nucleotide, and 0 to the remaining elements.

The 35-element vector is completed by two 5-element vectors corresponding to the PhyloP conservation scores (2) in the 5-nucleotide window. PhyloP conservation scores in each vector are derived from 100way (PhyloP100) and 470way (PhyloP470) UCSC alignments.

The effect of an InDel is predicted by computing the impact of the closest SNV to the mutated loci generated by the deletion/insertion of nucleotides. Thus, the input data for an InDel consists of a 38-element vector including the 35-element vector described above and 3 features representing the size and the location of the InDel. In detail, the size of the InDel is encoded with two values which encode for the length of the reference and the alternative alleles, while the location of the InDel is represented by a boolean variable indicating if the variant occurs in a coding or noncoding region. Predictions can be performed, providing input genomics coordinates from hg38 human assembly. Genomics coordinates based on hg19 human assembly are internally converted through UCSC *liftOver* program. The representation of the PhD-SNP^g input features is reported in Fig. 1C in the main manuscript.

3. Training and testing procedure

The performance of PhD-SNP^g has been assessed by a 10-fold cross-validation procedure in which all the variants corresponding to each chromosome were kept in the same subset to reduce possible overfitting. We evaluated possible gene sharing among different chromosomes and found only a few genes common between X and Y chromosomes. For this reason, the variants from chromosomes X and Y are kept in the same fold. This procedure also allows keeping the variants belonging to the same gene in the same subset, assigning them either to the testing or training set. Furthermore, the cross-validation splitting has been evaluated 5 times by bootstrapping the cross-validation sets.

More in detail, the performance of PhD-SNP^g in predicting the impact of SNVs was tested by a 10-fold cross-validation procedure on the Clinvar122020-SNV dataset and by a blind test on the NewClinvar122022-SNV dataset. Given the composition of the datasets described above (section 1), the performance of PhD-SNP^g in predicting the pathogenicity of the InDels was scored by a 10-fold cross-validation procedure on the NewClinvar12022-InDel dataset, which includes and high number of annotated variants, while the Clinvar122020 was used as blind set.

When we tested the method on the blind NewClinvar122022-SNV and NewClinvar12022-InDel test sets, we used the same strategy to evaluate the variants in the cross-validation procedure. We predict variants of a given chromosome using the model fitted during the cross-validation phase, which did not contain variants from the chromosome to test in its training set. A representation of this procedure is shown in Fig. S4.

The performance of PhD-SNP^g have been calculated on six classes of variants including "*exonic*", "*intronic*", "*splicing*", "*noncoding RNA*", and "*other*". The classification of the variants is obtained using ANNOVAR (3).

4. Method optimization

The Gradient Boosting algorithm from the *scikit-learn* package (4) (http://scikit-learn.org/) was not optimized, but the best hyper-parameters were kept fixed at those obtained in the previous PhD-SNP^g version. In detail, the method was trained by considering a tree depth of 7 and 500 estimators (5).

5. Comparison with CADD and FATHMM

One of the main aims for the development of PhD-SNP^g is the creation of a benchmark tool for testing new algorithms for SNVs prioritization. For this reason, we provided as Supplementary File the results of all the 10-fold cross-validation tests and blind test on datasets generated from Clinvar122020 and NewClinvar122022.

In this paper we compared PhD-SNP^g with CADD (6) and FATHMM (7,8), although it was not possible to compare them on the same bases, because the cross-validation predictions for CADD and FATHMM are not available. Moreover, some SNVs included in our dataset can overlap with the training set of CADD. For example, comparing the datasets used for training and testing the previous version of CADD and PhD-SNP^g algorithms, we estimated that a minimum of ~24% of the variants are in common.

The results of our tests on NewClinvar122022-SNV and Clinvar120220-InDel are summarized in Tables 1 and 2 respectively. The performance achieved by PhD-SNP^g with the 10-fold cross-validation on the Clinvar122020 and NewClinvar122022-InDel are summarized in Tables S7-S8. The relative ROC curves are shown in Figure 2 of the main manuscript and Figure S5.

6. Evaluation measures for binary classifiers

For the predictions of PhD-SNP^g and FATHMM, the binary classification (*Pathogenic/Benign*) is made at the output threshold of 0.5. Thus, if the probability of *Pathogenic* classification is

>0.5 the mutation is predicted to be *Pathogenic*. For CADD, a Phred-like score threshold of 20 was used to calculate the performance.

In all the performance measures - assuming that positives indicate *Pathogenic* and negatives indicate *Benign* - TP (true positives) are correctly predicted Pathogenic Single Nucleotide Variants (SNVs), TN (true negatives) are correctly predicted *Benign* variants, FP (false positives) *Benign* SNVs annotated as *Pathogenic*, and FN (false negatives) are *Pathogenic* variants predicted to be *Benign*.

Predictor performance was evaluated using the following metrics: true positive and negative rates (*TPR, TNR*), positive and negative predicted values (*PPV, NPV*), *F1* score and overall accuracy (Q_2)

$$\begin{aligned} &Pathogenic: \; PPV = \frac{TP}{TP+FP} \quad TPR = \frac{TP}{TP+FN} \\ &Benign: \; NPV = \frac{TN}{TN+FN} \quad TNR = \frac{TN}{TN+FP} \\ &F1 = \frac{2TP}{2TP+FP+FN} \quad Q_2 = \frac{TP+TN}{TP+FP+TN+FN} \end{aligned} \qquad \text{[Eq. 1]}$$

We computed the Matthew’s correlation coefficient *MCC* (Eq. 2) as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad \text{[Eq. 2]}$$

We also calculated the area under the receiver operating characteristic (ROC) curve (AUC), by plotting the True Positive Rate as a function of the False Positive Rate at different probability thresholds of annotating a variant as *Pathogenic* or *Benign*. PhD-SNP^g calculates the False Discovery Rate (FDR) as a function of the returned output (s_0).

$$Pathogenic: \; FDR(s > s_0) = \frac{FP}{FP+TP} \quad Benign: \; FDR(s < s_0) = \frac{FN}{FN+TN} \qquad \text{[Eq. 3]}$$

In this analysis, the calibration of the predictor was scored by calculating the Bier score (9). The calibration curve shows whether the predicted probabilities agree with the observed probabilities. If the calibration curve lies on the diagonal, the predictor is perfectly calibrated, and it requires no further investigation. The deviation from the diagonal indicates the miscalibration. Brier score ranges from zero to one (one being totally uncalibrated, zero being perfect calibration).

REFERENCES

1. Landrum,M.J., Chitipiralla,S., Brown,G.R., Chen,C., Gu,B., Hart,J., Hoffman,D., Jang,W., Kaur,K., Liu,C., *et al.* (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res*, **48**, D835–D844.
https://doi.org/10.1093/nar/gkz972
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6943040

2. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110–121.
https://doi.org/10.1101/gr.097857.109
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2798823

3. Yang,H. and Wang,K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc, 10, 1556–1566.
https://doi.org/10.1038/nprot.2015.105
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4718734

4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* (2011) Scikit-learn: Machine Learning in Python. *JMLR*, **12**, 2825–2830

5. Capriotti,E. and Fariselli,P. (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res*, **45**, W247–W252.
https://doi.org/10.1093/nar/gkx369
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570245

6. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, **47**, D886–D894.
https://doi.org/10.1093/nar/gky1016
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323892

7. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of noncoding and coding sequence variation. Bioinformatics, 31, 1536–43.
https://doi.org/10.1093/bioinformatics/btv009
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426838

8. Ferlaino,M., Rogers,M.F., Shihab,H.A., Mort,M., Cooper,D.N., Gaunt,T.R. and Campbell,C. (2017) An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. BMC Bioinformatics, 18, 442.
https://doi.org/10.1186/s12859-017-1862-y
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5955213

9. Brier,G.W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**, 1–3.

Supplementary Tables

Dataset	Effect	All	Coding	Noncoding
Clinvar122020-SNV	*Benign*	103243	66119	37124
	Pathogenic	51958	45697	6261
	Total	155201	111816	43385
NewClinvar122022-SNV	*Benign*	83417	20500	62917
	Pathogenic	21299	18965	2334
	Total	104716	39465	65251

Table S1. SNVs from December 2020 Clinvar version (Clinvar122020-SNV) and new ones in December 2022 Clinvar (NewClinvar122022-SNV).

Dataset	Effect	All	Coding	Noncoding
Clinvar122020-InDel	*Benign*	4523	1118	3405
	Pathogenic	37421	36221	1200
	Total	41944	37339	4605
NewClinvar122022-InDel	*Benign*	17145	786	16359
	Pathogenic	24382	23460	922
	Total	41527	24246	17281

Table S2. InDels from December 2020 Clinvar version (Clinvar122020-InDel) and new ones in December 2022 Clinvar (NewClinvar122022-InDel).

Dataset	Effect	Insertions	Deletions	Ins+Dels
Clinvar122020-InDel	*Benign*	1937	2456	130
	Pathogenic	11190	24926	1305
	Total	13127	27382	1435
NewClinvar122022-InDel	*Benign*	8145	8908	92
	Pathogenic	7481	16348	553
	Total	15626	25256	645

Table S3. Types of variants in the Clinvar122020-InDel and NewClinvar122022-InDel datasets. Ins+Dels: Insertion and deletion.

Conservation	Subset	Pathogenic			Benign		
		Mean	**Median**	**Std**	**Mean**	**Median**	**Std**
PhyloP7	*all*	0.741	0.871	0.423	-0.128	-0.027	0.889
	coding	0.729	0.871	0.430	0.029	0.028	0.825
	noncoding	0.836	0.917	0.351	-0.280	-0.154	0.921
PhyloP100	*all*	5.458	6.091	3.184	0.173	-0.018	2.263
	coding	5.367	5.965	3.206	0.647	0.249	2.827
	noncoding	6.143	7.085	2.925	-0.285	-0.163	1.384
PhyloP470	*all*	6.555	7.622	4.333	-0.382	-0.147	4.319
	coding	6.429	7.601	4.340	-0.283	0.073	5.736
	noncoding	7.502	7.843	4.159	-0.477	-0.270	2.201

Table S4. Mean, median and standard deviation (std) of the PhyloP conservation scores (PhyloP7, PhyloP100 and PhyloP470) for *Pathogenic* and *Benign* Single Nucleotide Variants (SNVs) from Clinvar122022-SNV dataset and its subsets of coding and SNVs.

Conservation	Subset	Pathogenic			Benign		
		Mean	**Median**	**Std**	**Mean**	**Median**	**Std**
PhyloP7	*all*	0.526	0.871	0.629	-0.047	0.020	0.772
	coding	0.529	0.871	0.625	0.113	0.054	0.772
	noncoding	0.460	0.871	0.729	-0.062	0.017	0.770
PhyloP100	*all*	3.220	2.302	3.468	-0.008	-0.037	1.557
	coding	3.222	2.300	3.451	0.682	0.219	2.430
	noncoding	3.170	2.482	3.914	-0.074	-0.058	1.427
PhyloP470	*all*	4.024	3.483	4.927	-0.137	-0.114	2.478
	coding	4.028	3.493	4.916	0.452	0.103	4.848
	noncoding	3.921	2.987	5.239	-0.194	-0.134	2.105

Table S5. Mean, median and standard deviation (std) of the PhyloP conservation scores (PhyloP7, PhyloP100 and PhyloP470) for *Pathogenic* and *Benign* InDels from Clinvar122022-InDels dataset and its subsets of coding and InDels.

Dataset	PhyloP	Th	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	D_{KS}
SNVs	PhyloP7	0.768	0.751	0.703	0.778	0.799	0.729	0.505	0.763	0.770	0.587
	PhyloP100	2.393	0.785	0.817	0.767	0.752	0.804	0.570	0.777	0.865	0.672
	PhyloP470	4.646	0.773	0.845	0.738	0.700	0.819	0.551	0.755	0.834	0.641
InDels	PhyloP7	0.824	0.696	0.786	0.666	0.606	0.739	0.398	0.666	0.697	0.468
	PhyloP100	1.298	0.714	0.831	0.673	0.596	0.779	0.440	0.675	0.768	0.502
	PhyloP470	1.956	0.710	0.835	0.668	0.585	0.780	0.434	0.669	0.749	0.509

Table S6. Performance of simple predictor based on the PhyloP score of the mutated loci on the testing datasets of SNVs (NewClinvar122022-SNV) and InDels (Clinvar122020-InDel). Average results of the 5 bootstrap tests (10-fold) optimizing the threshold (Th) on the training set and applying it on the testing set. Q_2, TNR, NPV, TPR, PPV, MCC, F1, AUC and D_{KS} are defined in the section above.

Method	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
PhD-SNP^g	*all*	0.890	0.889	0.891	0.891	0.890	0.781	0.890	0.954	0.082	100.0
	coding	0.887	0.882	0.896	0.893	0.879	0.775	0.886	0.953	0.083	84.2
	noncoding	0.905	0.937	0.863	0.880	0.945	0.812	0.911	0.961	0.072	15.8
CADD	*all*	0.910	0.852	0.964	0.968	0.867	0.825	0.915	0.976	NA	99.7
	coding	0.902	0.831	0.973	0.976	0.848	0.814	0.907	0.976	NA	84.2
	noncoding	0.952	0.979	0.918	0.929	0.983	0.904	0.955	0.982	NA	15.5
FATHMM-MKL	*all*	0.767	0.626	0.871	0.908	0.708	0.556	0.796	0.879	0.173	99.5
	coding	0.740	0.582	0.864	0.905	0.675	0.512	0.773	0.861	0.193	84.0
	noncoding	0.911	0.898	0.901	0.921	0.919	0.819	0.920	0.963	0.070	15.6

Table S7. Performance of PhD-SNP^g, CADD and FATHMM-MKL on the training dataset of SNVs (Clinvar122020-SNV). Average results of the 5 bootstrap tests (10-fold) performed on the both datasets. Q_2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above. For CADD, a Phred score threshold of 20 was considered for binary classification.

Method	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
PhD-SNP^g	*all*	0.976	0.966	0.986	0.986	0.966	0.952	0.976	0.992	0.022	100.0
	coding	0.977	0.416	0.760	0.995	0.981	0.552	0.988	0.919	0.020	48.7
	noncoding	0.975	0.983	0.990	0.829	0.748	0.774	0.787	0.945	0.023	51.3
CADD	*all*	0.951	0.990	0.919	0.912	0.989	0.906	0.949	0.983	NA	99.6
	coding	0.919	0.775	0.257	0.924	0.992	0.417	0.957	0.919	NA	48.4
	noncoding	0.982	0.997	0.984	0.717	0.941	0.813	0.814	0.855	NA	51.3
FATHMM-indel	*all*	0.909	0.919	0.904	0.898	0.914	0.818	0.906	0.967	0.068	97.1
	coding	0.900	0.748	0.192	0.905	0.992	0.346	0.946	0.902	0.077	46.7
	noncoding	0.918	0.924	0.989	0.778	0.321	0.467	0.454	0.908	0.060	50.4

Table S8. Performance of PhD-SNP^g, CADD and FATHMM-indel on the training dataset of InDels (NewClinvar122022-InDels). Average results of the 5 bootstrap tests (10-fold) performed on the both

datasets. Q_2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above. For CADD, a Phred score threshold of 20 was considered for binary classification.

PhD-SNP^g	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
2017	*all*	0.884	0.881	0.886	0.887	0.882	0.768	0.884	0.951	0.085	100.0
	coding	0.881	0.873	0.891	0.889	0.871	0.763	0.880	0.949	0.087	84.2
	noncoding	0.900	0.929	0.859	0.876	0.938	0.801	0.906	0.958	0.077	15.8
2023	*all*	0.890	0.889	0.891	0.891	0.890	0.781	0.890	0.954	0.082	100.0
	coding	0.887	0.882	0.896	0.893	0.879	0.775	0.886	0.953	0.083	84.2
	noncoding	0.905	0.937	0.863	0.880	0.945	0.812	0.911	0.961	0.072	15.8

Table S9. Performance of the old (2017) and new (2023) versions of PhD-SNP^g on the Clinvar122020-SNV dataset. Average performance on the 5 bootstrap tests (10-fold) performed on both datasets. Q_2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above.

Dataset	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
2017	*all*	0.893	0.879	0.904	0.907	0.882	0.786	0.894	0.957	0.079	100.0
	coding	0.887	0.865	0.905	0.910	0.871	0.775	0.890	0.953	0.083	84.0
	noncoding	0.923	0.953	0.901	0.891	0.949	0.847	0.919	0.968	0.060	16.0
2023	*all*	0.898	0.883	0.909	0.912	0.887	0.796	0.899	0.959	0.076	100.0
	coding	0.892	0.869	0.910	0.914	0.876	0.784	0.894	0.956	0.080	84.0
	noncoding	0.930	0.958	0.909	0.900	0.954	0.861	0.926	0.970	0.055	16.0

Table S10. Performance of the old (2017) and new (2023) versions of PhD-SNP^g on the NewClinvar122022-SNV dataset. Average performance on the 5 bootstrap tests (10-fold) performed on both datasets. Q2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above.

Dataset	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
All	*all*	0.907	0.830	0.981	0.984	0.853	0.824	0.914	0.966	0.081	100.0
	coding	0.877	0.357	0.933	0.994	0.873	0.532	0.930	0.880	0.107	56.6
	noncoding	0.946	0.954	0.987	0.865	0.642	0.718	0.737	0.962	0.048	43.4
Deletions	*all*	0.915	0.827	0.978	0.986	0.877	0.834	0.928	0.968	0.074	61.4
	coding	0.896	0.354	0.937	0.996	0.894	0.539	0.942	0.887	0.090	36.9
	noncoding	0.943	0.952	0.983	0.883	0.721	0.767	0.794	0.969	0.050	24.5
Insertion	*all*	0.914	0.865	0.987	0.984	0.838	0.837	0.905	0.966	0.075	36.7
	coding	0.875	0.418	0.934	0.992	0.869	0.574	0.927	0.865	0.109	18.3
	noncoding	0.953	0.959	0.992	0.812	0.433	0.572	0.563	0.941	0.041	18.4

Table S11. Performance of PhD-SNP^g on the Clinvar122020-InDels dataset and Insertion and Deletions subsets. Average performance on the 5 bootstrap tests (10-fold) performed on the Clinvar122020-InDel datasets. Q_2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above.

SNVs	Subset	Q_2	MCC	F1	AUC
Clinvar122020	*all*	0.891±0.004	0.781±0.007	0.890±0.007	0.955±0.003
	coding	0.888±0.004	0.776±0.008	0.886±0.007	0.954±0.003
	noncoding	0.904±0.008	0.808±0.016	0.909±0.012	0.960±0.006
NewClinvar122022	*all*	0.898±0.004	0.796±0.008	0.899±0.006	0.960±0.002
	coding	0.892±0.004	0.784±0.007	0.894±0.006	0.956±0.002
	noncoding	0.930±0.007	0.860±0.014	0.925±0.011	0.970±0.005

Table S12. Mean and standard deviation of the overall accuracy (Q_2), Matthews Correlation Coefficient (MC), F1 and area under the receiver operating characteristic curve (AUC) in the prediction of pathogenic SNVs by using a 10-fold cross-validation procedure.

InDels	Subset	Q_2	MCC	F1	AUC
Clinvar122020	*all*	0.976±0.003	0.952±0.006	0.976±0.003	0.993±0.003
	coding	0.976±0.004	0.551±0.081	0.988±0.002	0.921±0.017
	noncoding	0.976±0.005	0.778±0.021	0.789±0.018	0.956±0.032
NewClinvar122022	*all*	0.907±0.005	0.824±0.010	0.913±0.007	0.966±0.006
	coding	0.876±0.009	0.531±0.030	0.928±0.007	0.880±0.028
	noncoding	0.947±0.006	0.711±0.045	0.729±0.046	0.962±0.012

Table S13. Mean and standard deviation of the overall accuracy (Q_2), Matthews Correlation Coefficient (MC), F1 and area under the receiver operating characteristic curve (AUC) in the prediction of pathogenic InDels by using a 10-fold cross-validation procedure.

SNVs	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
Clinvar122020	*all*	0.890	0.889	0.891	0.891	0.890	0.781	0.890	0.954	0.082	100.0
	coding	0.887	0.882	0.896	0.893	0.879	0.775	0.886	0.953	0.083	84.2
	10	0.883	0.885	0.932	0.881	0.804	0.751	0.841	0.949	0.086	39.7
	5	0.882	0.885	0.940	0.874	0.772	0.736	0.820	0.947	0.088	27.2
NewClinvar122022	*all*	0.898	0.883	0.909	0.912	0.887	0.796	0.899	0.959	0.076	100.0
	coding	0.892	0.869	0.910	0.914	0.876	0.784	0.894	0.956	0.080	84.0
	10	0.890	0.874	0.918	0.908	0.860	0.780	0.883	0.956	0.081	50.9
	5	0.888	0.874	0.920	0.906	0.854	0.776	0.879	0.955	0.082	35.8

Table S14. Performance PhD-SNP^g on the dowsampled subsets of SNVs from Clinvar122020 and NewClinvar122022. The subsets 5 and 10 are obtained, selecting a maximum of coding 5 or 10 variants for each gene. Average performance on the 5 bootstrap tests (10-fold) performed on both datasets. Q2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above.

InDels	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
NewClinvar122022	*all*	0.976	0.966	0.986	0.986	0.967	0.953	0.977	0.990	0.029	100.0
	coding	0.977	0.416	0.760	0.995	0.981	0.552	0.988	0.919	0.020	48.7
	10	0.965	0.398	0.788	0.995	0.970	0.546	0.982	0.917	0.030	26.3
	5	0.961	0.404	0.813	0.994	0.965	0.556	0.979	0.915	0.034	19.3
Clinvar122020	*all*	0.907	0.830	0.981	0.984	0.853	0.824	0.914	0.966	0.081	100.0
	coding	0.877	0.357	0.933	0.994	0.873	0.532	0.930	0.880	0.107	56.6
	10	0.837	0.351	0.950	0.994	0.826	0.517	0.902	0.890	0.142	40.4
	5	0.816	0.342	0.951	0.993	0.802	0.503	0.888	0.885	0.160	34.1

Table S15. Performance PhD-SNP^g on the dowsampled subsets of InDels from Clinvar122020 and NewClinvar122022. The subsets 5 and 10 are obtained, selecting a maximum of 5 or 10 coding variants for each gene. Average performance on the 5 bootstrap tests (10-fold) performed on both datasets. Q2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in the section above.

Variants	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
SNVs	*all*	0.890	0.889	0.891	0.891	0.890	0.781	0.890	0.954	0.082	100.0
(Clinvar122020)	*exonic*	0.887	0.882	0.895	0.892	0.879	0.774	0.885	0.952	0.084	85.2
	intronic	0.839	0.949	0.854	0.496	0.757	0.521	0.599	0.852	0.121	4.8
	splicing	0.979	0.422	0.068	0.981	0.998	0.163	0.989	0.851	0.019	6.7
	ncRNA	0.867	0.930	0.881	0.732	0.830	0.686	0.778	0.906	0.102	0.4
	UTR	0.914	0.937	0.973	0.375	0.198	0.231	0.259	0.752	0.067	2.5
	other	0.663	0.968	0.639	0.239	0.842	0.315	0.371	0.792	0.248	0.1
InDels	*all*	0.976	0.966	0.986	0.986	0.967	0.953	0.977	0.990	0.029	100.0
(NewClinvar122022)	*exonic*	0.973	0.428	0.687	0.993	0.980	0.530	0.986	0.923	0.023	50.6
	intronic	0.993	0.995	0.997	0.568	0.397	0.471	0.467	0.934	0.007	42.9
	splicing	0.913	0.940	0.812	0.900	0.970	0.811	0.934	0.972	0.079	1.0
	ncRNA	0.903	0.929	0.969	0.200	0.091	0.089	0.125	0.405	0.091	1.6
	UTR	0.837	0.841	0.991	0.582	0.060	0.147	0.109	0.847	0.142	2.5
	other	0.916	0.997	0.919	0.033	0.450	0.103	0.061	0.329	0.082	1.4

Table S16. Performance of PhD-SNP^g on the training set of SNVs (ClinVar122020) and InDels (NewClinVar122022) classified according to their location. The variants are classified in *exonic, intronic, splicing, noncoding RNA and UTR* using ANNOVAR. Average performance on the 5 bootstrap tests (10-fold) performed on both datasets. Q2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in Supplementary Materials.

Supplementary Figures

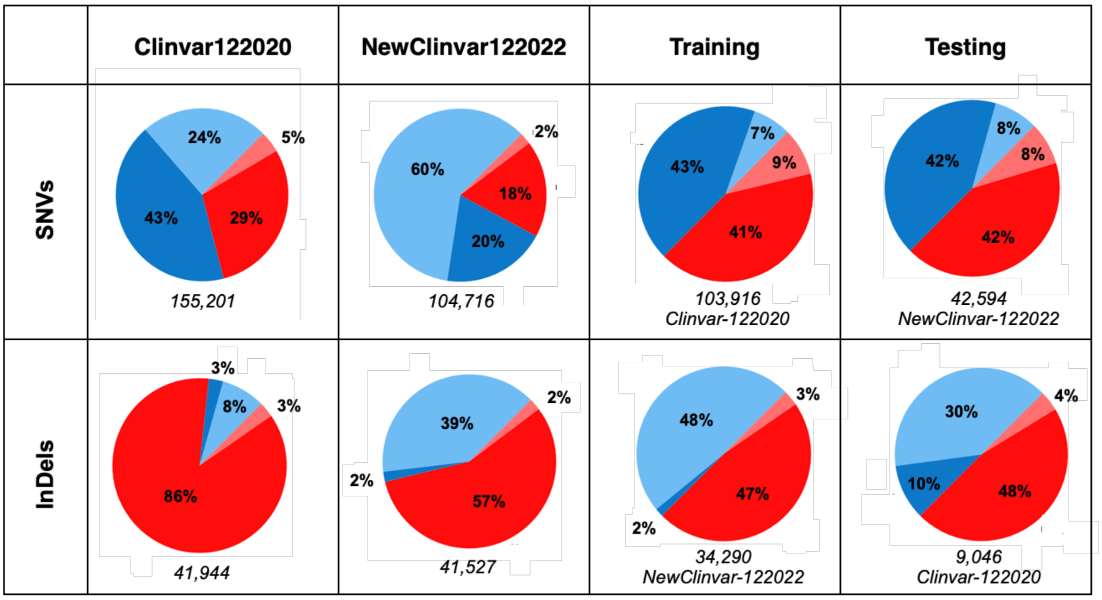

Figure S1. Pie chart of the four datasets used in this work. *Benign* variants in coding and noncoding regions are indicated in dark and light blue respectively, while *Pathogenic* variants are reported in dark and light red respectively. The most abundant The most abundant class of variants is downsampled randomly.

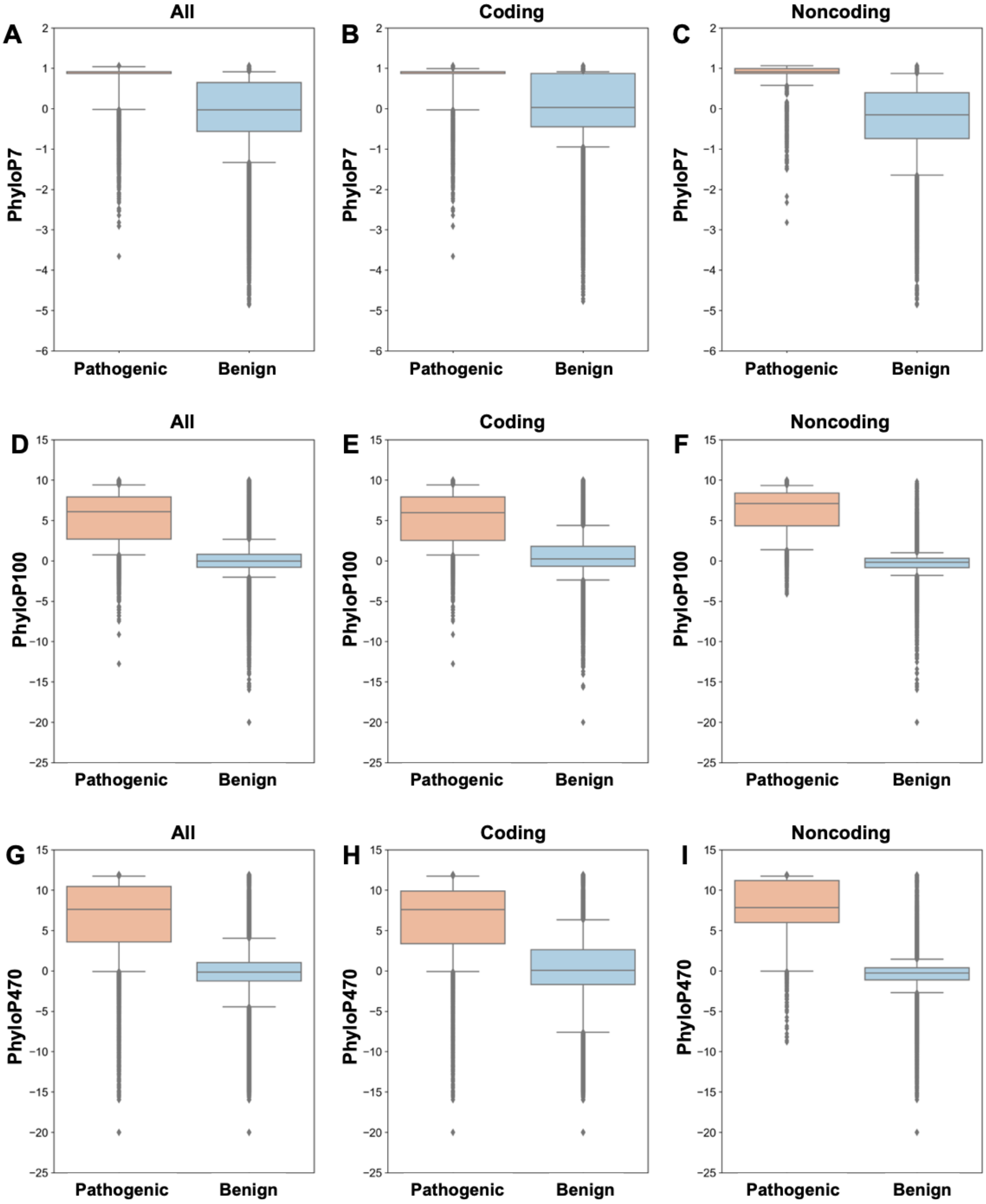

Figure S2. Distribution of the PhyloP conservation scores (PhyloP7, PhyloP100 and PhyloP470) for *Pathogenic* and *Benign* Single Nucleotide Variants (SNVs) from Clinvar122022-SNV dataset and its subsets of coding and SNVs.

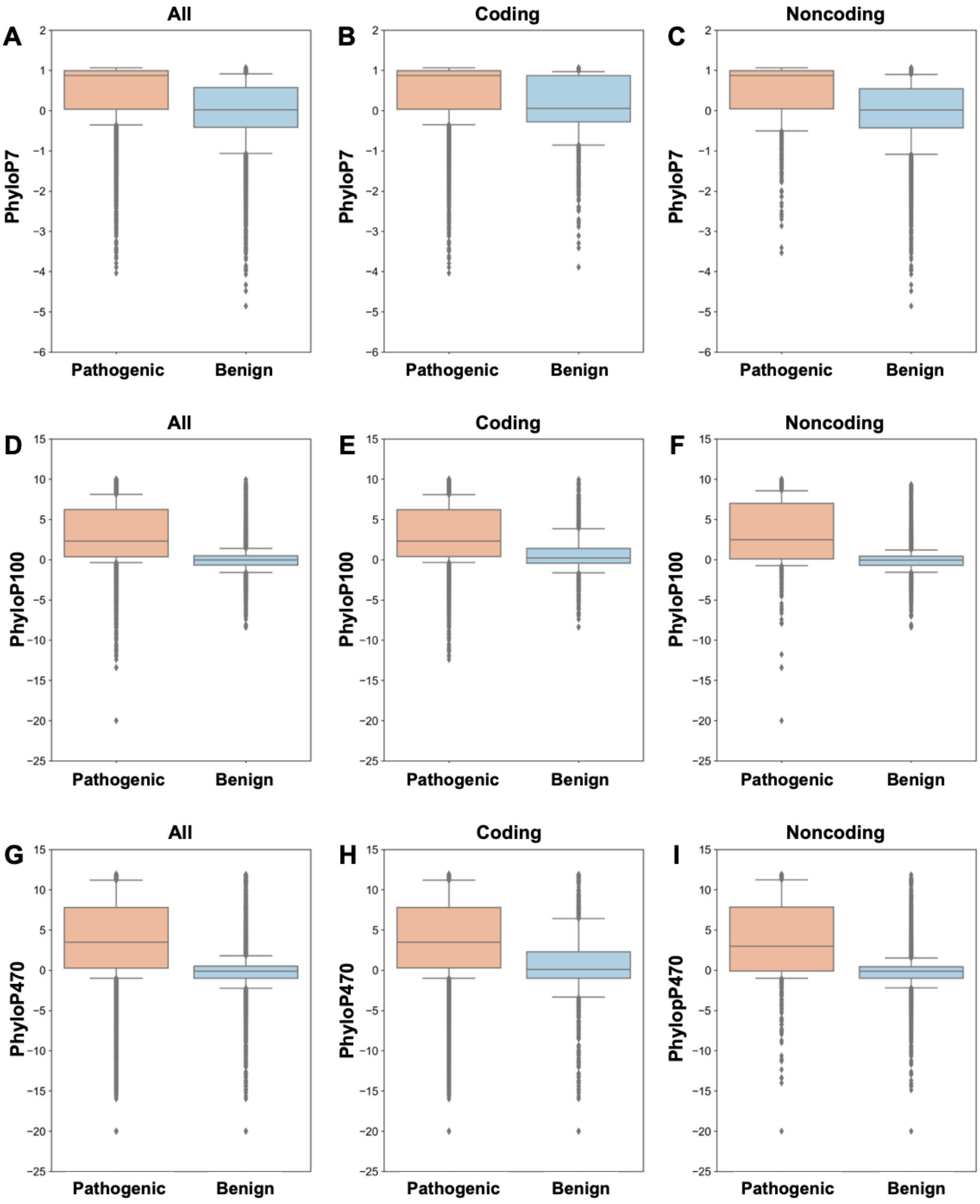

Figure S3. Distribution of the PhyloP conservation scores (PhyloP7, PhyloP100 and PhyloP470) for *Pathogenic* and *Benign* InDels from Clinvar122022-InDels dataset and its subsets of coding and InDels.

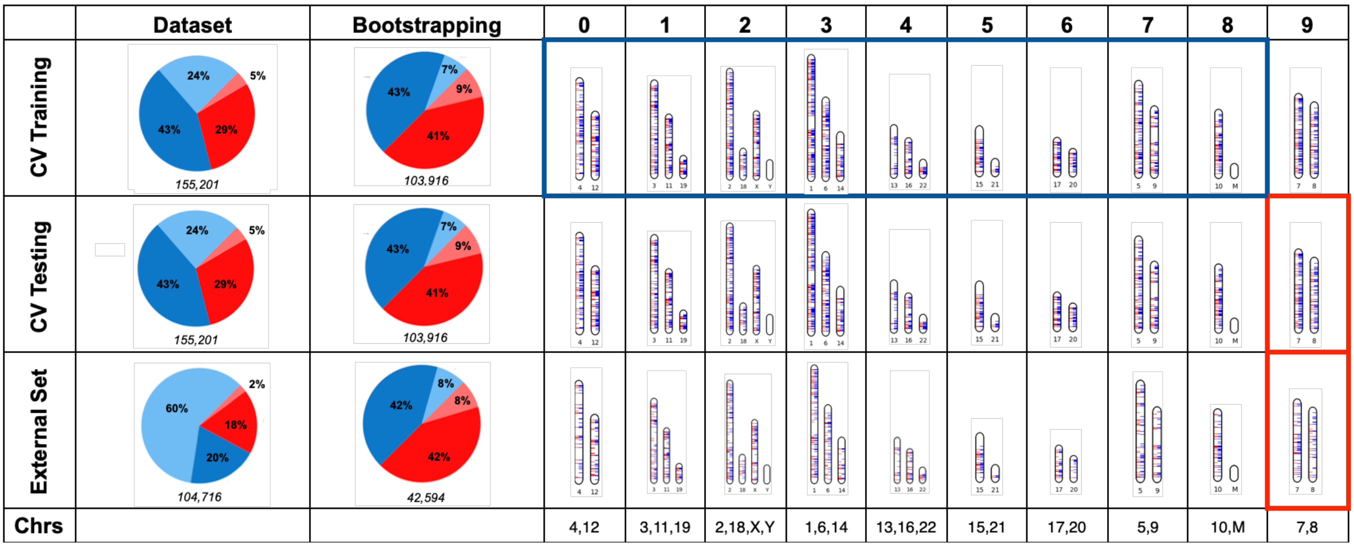

Figure S4. Representation of the training and testing procedure on balanced datasets of *Pathogenic* (coding and noncoding in dark and light red, respectively) and *Benign* (coding and noncoding in dark and light blue, respectively) variants. A bootstrapping procedure is used for downsampling the most abundant class for training and testing sets. Both datasets are divided into 10 parts, keeping all the variants from the same chromosome in the same subset. Variants in X and Y chromosomes are kept together. To avoid prediction bias, the 10-fold cross-validation procedure is performed, excluding a subset of variants from a group of chromosomes not considered in the training step. In the presented example of 10-fold cross-validation, the prediction of subset 9, which includes only variants from chromosomes 7 and 8, is obtained by using a model generated from the subsets 0-8, which do not include variants in chromosomes 7 and 8. Thus, all the returned 10-fold cross-validation predictions are from variants in *“never seen”* chromosomes. Furthermore, the model obtained on the subset 0-8 of the *Training* dataset is used for predicting the effect of the variants in subset 9 of the *Testing* sets. With this approach, all the variants in the *Testing* set are never used in training and are predicted with models trained on variants from other chromosomes.

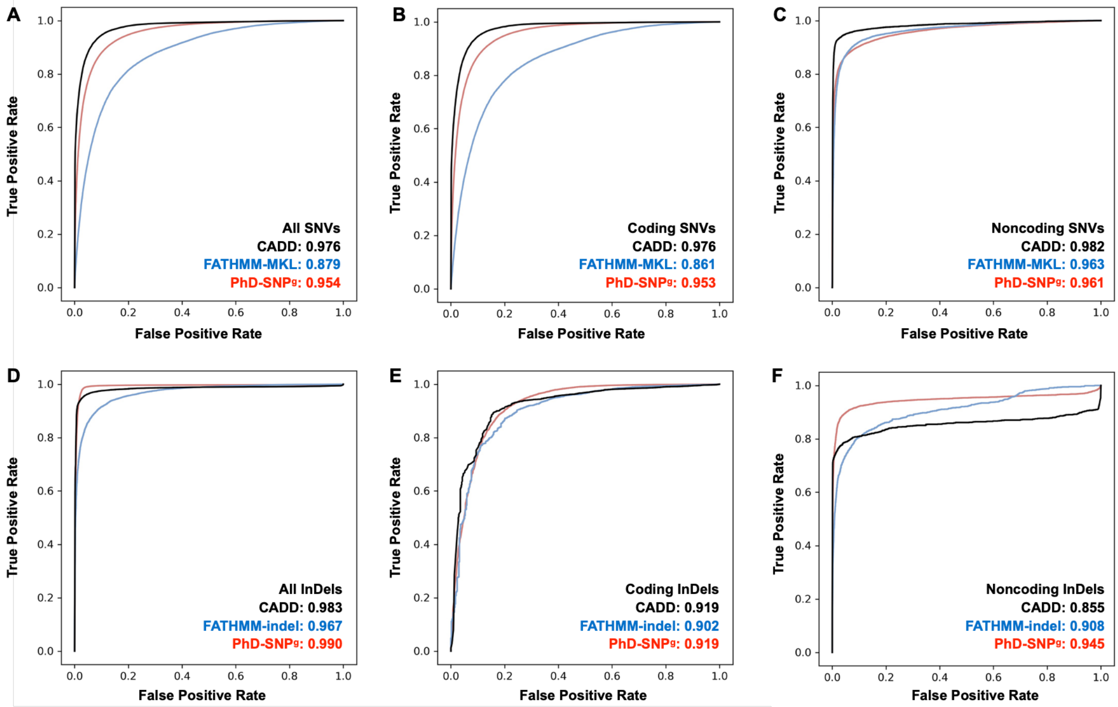

Figure S5. Comparison of the Area Under Receiver Operator Characteristic Curve (ROC) for CADD (black), PhD-SNP^g (red) and FATHMM-MKL/indel (blue). The ROC curves are calculated on the datasets Clinvar122020-SNV (A) and NewClinvar122022-InDels (D) and their subset of coding and noncoding SNVs (B, C) and InDels (E, F).