SUPPLEMENTARY MATERIALS

Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants

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

[1] BioFold Unit, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Via F. Selmi 3, 40126 Bologna, Italy.
[2] Department of Medical Sciences, University of Torino, Via Santena 19, 10126, Torino, Italy.

Contacts: emidio.capriotti@unibo.it, piero.fariselli@unito.it

Performance measures for binary classifiers

For each prediction, the binary classification (*Pathogenic/Benign*) is made at the threshold *t*. Thus, if a selected score for *Pathogenic* classification is >*t* the variant is predicted to be *Pathogenic*. For single feature-based predictors the classification threshold is optimized on the *CommonClinvar* dataset (Table S2) while for the gradient boosting algorithm and REVEL (Ioannidis *et al.* 2016) the output threshold is set to 0.5. For CADD (Rentzsch *et al.* 2019) a raw score threshold of 3.1 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)

$$PPV = \frac{TP}{TP+FP} \quad TPR = \frac{TP}{TP+FN}$$
$$NPV = \frac{TN}{TN+FN} \quad TNR = \frac{TN}{TN+FP} \qquad \text{[Eq. 1]}$$
$$F1 = \frac{2TP}{2TP+FP+FN} \quad Q_2 = \frac{TP+TN}{TP+FP+TN+FN}$$

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 and the Area and the Precision Recall Curve (AUP) at different probability thresholds of annotating a variant as *Pathogenic* or *Benign*. Sensitivity = Recall = TPR, Precision = PPV.

For each method, the reported scoring measures are obtained averaging the performance on ten randomly selected sets from *CommonClinvar* and *NewClinvar* datasets. The selection procedure is performed for generating balanced datasets of *Pathogenic* and *Benign* variants downscaling the most abundant class of variants.

Supplementary Tables

Dataset	Annotation	Proteins	Mutations
CommonClinvar	All	7,582	36,751
	Benign	6,444	19,659 (53.5%)
	Pathogenic	3,117	17,092 (46.5%)
NewClinvar	All	1,855	5,172
	Benign	935	2,247 (43.4%)
	Pathogenic	1,119	2,925 (56.6%)

Table S1. Composition of the datasets extracted from *Clinvar* database (https://www.ncbi.nlm.nih.gov/clinvar/),

Dataset	Features	D_{KS}	*Pathogenic*			*Benign*		
			Mean	*Median*	*Std*	*Mean*	*Median*	*Std*
CommonClinvar	*PC*	0.529	0.957	1.000	0.185	0.562	0.893	0.466
	PP	0.573	6.523	7.521	2.794	2.120	1.374	2.789
	f_{ref}	0.579	0.956	1.000	0.108	0.739	0.830	0.270
	f_{alt}	0.598	0.008	0.000	0.036	0.147	0.047	0.226
	f_{wt}	0.545	0.797	0.945	0.261	0.414	0.362	0.270
	f_{mut}	0.609	0.011	0.000	0.043	0.092	0.037	0.143
NewClinvar	*PC*	0.473	0.951	1.000	0.198	0.598	0.966	0.460
	PP	0.548	6.461	7.509	2.861	2.301	1.521	2.782
	f_{ref}	0.563	0.953	1.000	0.109	0.734	0.836	0.283
	f_{alt}	0.590	0.008	0.000	0.037	0.159	0.047	0.244
	f_{wt}	0.546	0.802	0.943	0.257	0.420	0.378	0.273
	f_{mut}	0.632	0.010	0.000	0.038	0.096	0.037	0.148

Table S2. Statistics of the distribution of six conservation features for the subset of *Pathogenic* and *Benign* variants in the *CommonClinvar* and *NewClinvar* datasets. *PC*: *PhastCons100way* score. PP: *PhyloP100way* score, f_{ref} and f_{alt}: frequencies of the reference and alternate alleles in the *multiz100way* genomic alignment, f_{wt} and f_{mut} : frequencies of the wild-type and mutant residues from a multiple protein sequence alignment. D_{KS} is the distance between two cumulative distributions calculated through the Kolmogorov-Smirnov test.

Feature	Threshold	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC	AUP
PC	1.000	0.764	0.655	0.838	0.874	0.717	0.541	0.787	0.782	0.836
PP	4.704	0.786	0.815	0.771	0.758	0.804	0.574	0.780	0.856	0.844
f_{ref}	0.977	0.789	0.825	0.770	0.754	0.812	0.580	0.781	0.844	0.848
f_{alt}	0.000	0.798	0.758	0.824	0.838	0.776	0.598	0.806	0.832	0.865
f_{wt}	0.702	0.773	0.822	0.748	0.723	0.802	0.548	0.761	0.842	0.831
f_{mut}	0.005	0.805	0.807	0.803	0.802	0.807	0.609	0.804	0.851	0.850

Table S3. Performance of the basic predictors based on a single feature on the *CommonClinvar* dataset Prediction threshold are optimized maximizing both the True Positive Rate (TPR) and the True Negative Rate (TNR) dataset. Q2: Overall Accuracy, TNR: True negative rate, NPV: Negative predicted value, TPR: True Positive Rate, PPV: Positive Predicted Value, MCC: Matthews Correlation Coefficient, F1: harmonic mean of precision and sensitivity, AUC: Area Under the Receiver Operator Characteristic Curve, AUP Area under the Precision Recall Curve. All the performance measures are defined above.

Method	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC	AUP
CADD	0.841	0.819	0.857	0.864	0.826	0.683	0.845	0.910	0.904
REVEL	0.902	0.933	0.879	0.871	0.929	0.806	0.899	0.961	0.960
ProtProf	0.833	0.868	0.811	0.798	0.858	0.667	0.827	0.906	0.901
DNAProf	0.821	0.791	0.842	0.851	0.803	0.644	0.827	0.888	0.879
PPScore	0.792	0.798	0.788	0.785	0.796	0.583	0.790	0.868	0.859

Table S4: Prediction in cross-validation the *CommonClinvar* dataset. Q2: Overall Accuracy, TNR: True negative rate, NPV: Negative predicted value, TPR: True Positive Rate, PPV: Positive Predicted Value, MCC: Matthews Correlation Coefficient, F1: harmonic mean of precision and sensitivity, AUC: Area Under the Receiver Operator Characteristic Curve, AUP Area under the Precision Recall Curve. For CADD a raw score classification threshold of 3.1 was considered. All the performance measures are defined above.

Features	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC	AUP
PPScore - N_g	0.768	0.772	0.765	0.763	0.770	0.536	0.767	0.847	0.835
PPScore	0.771	0.776	0.769	0.767	0.774	0.543	0.770	0.855	0.846
DNAProf - N_g	0.794	0.739	0.830	0.849	0.765	0.592	0.805	0.868	0.855
DNAProf	0.812	0.780	0.834	0.845	0.794	0.626	0.818	0.881	0.873
ProtProf – N_p	0.827	0.829	0.825	0.824	0.829	0.654	0.827	0.895	0.888
ProtProf	0.831	0.865	0.809	0.796	0.855	0.662	0.824	0.910	0.905

Table S5. Testing prediction of the three basic methods (*PPScore, DNAProf and ProtProf*) excluding *Ng a*nd *Np* from the features. Q2: Overall Accuracy, TNR: True negative rate, NPV: Negative predicted value, TPR: True Positive Rate, PPV: Positive Predicted Value, MCC: Matthews Correlation Coefficient, F1: harmonic mean of precision and sensitivity, AUC: Area Under the Receiver Operator Characteristic Curve, AUP Area under the Precision Recall Curve. All the performance measures defined above are calculated on the *NewClinvar* dataset.

Frequency	Subset	D_{KS}	Pathogenic		Benign	
			Mean	*Std*	*Mean*	*Std*
f_{wt}	*Consensus*	0.697	0.858	0.220	0.367	0.249
	NotConsensus	0.133	0.657	0.285	0.614	0.266
f_{mut}	*Consensus*	0.798	0.005	0.030	0.114	0.158
	NotConsensus	0.152	0.022	0.054	0.028	0.062

Table S6. Comparison of the distributions of the frequencies of wild-type and mutant residues on the subset of *NewClinvar* for which the predictions of *PPScore, DNAProf* and *ProtProf* are in agreement (*Consensus*) or in disagreement (*NotConsensus*). D_{KT} is the Kolmogorov-Smirnov distance between the cumulative distributions of the frequencies for *Pathogenic* and *Benign* variants.

Frequency	Subset	D_{KS}	Pathogenic		Benign	
			Mean	*Std*	*Mean*	*Std*
f_{wt}	*Consensus*	0.697	0.849	0.225	0.386	0.264
	NotConsensus	0.133	0.692	0.290	0.525	0.273
f_{mut}	*Consensus*	0.798	0.006	0.030	0.108	0.154
	NotConsensus	0.152	0.020	0.053	0.061	0.124

Table S7. Comparison of the distributions of the frequencies of wild-type and mutant residues on the subset of *NewClinvar* for which the predictions of REVEL, CADD and *ProtProf* are in agreement (*Consensus*) or in disagreement (*NotConsensus*). D_{KT} is the Kolmogorov-Smirnov distance between the cumulative distributions of the frequencies for *Pathogenic* and *Benign* variants.

Supplementary figures

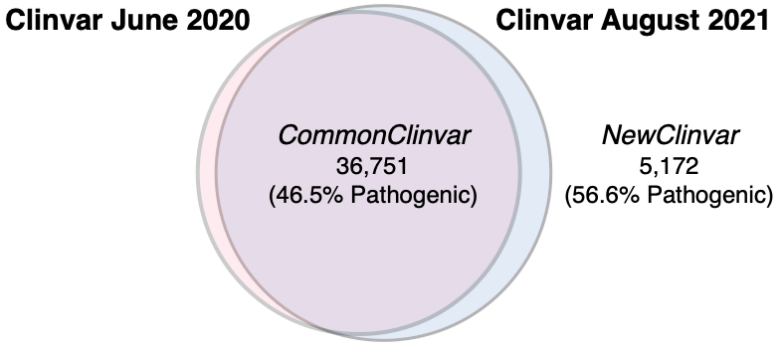

Figure S1. Venn diagram showing the intersection and difference between the two versions (August 2021 and June 2020) of Clinvar database (https://www.ncbi.nlm.nih.gov/clinvar/) used for generating the *CommonClinvar* and *NewClinvar* datasets

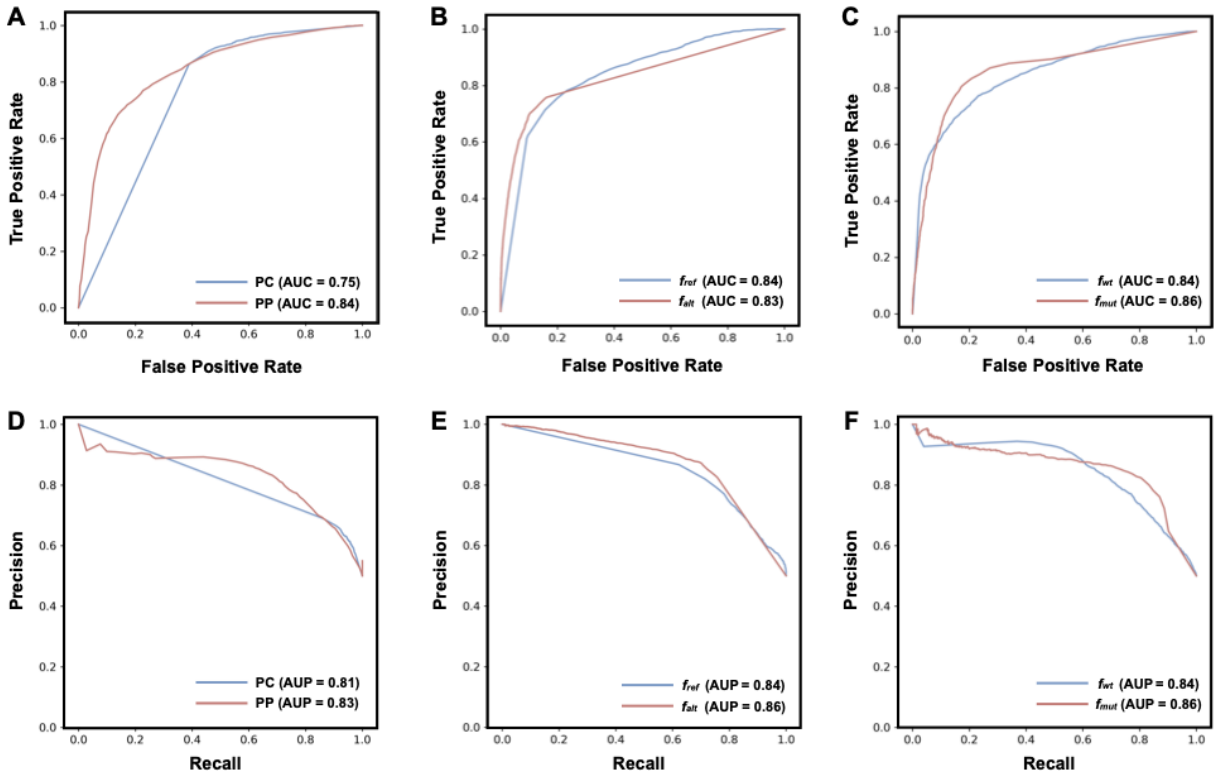

Figure S2. Receiver Operating Characteristic (A,B,C) and Precision Recall (D,E,F) curves for single feature predictors on the *NewClinvar* dataset. *PC*: *PhastCons100way* score. PP: *PhyloP100way* score, f_{ref} and f_{alt}: frequencies of the reference and alternate alleles in the *multiz100way* genomic alignment, f_{wt} and f_{mut} : frequencies of the wild-type and mutant residues from a multiple protein sequence alignment.

REFERENCES

Ioannidis NM, Rothstein JH, Pejaver V, et al (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* 99:877–885.
Rentzsch P, Witten D, Cooper GM, et al (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47:D886–D894.