Supplementary Materials for: **Evaluating ML-Based Continuous Predictors for Missense Variant Pathogenicity**

Author: Aitana Diaz Vasquez

Supplementary Tables

Predictor	Type of prediction	Methodology	Features	Training data	Score range	Cut-off value
CADD	Quantitative	Logistic regression	MSA, sequence, function, other predictors	Simulated variants from the ancestral genome	CADD raw: -28.378 to 25.512; PHRED: 1–99	Not defined, PHRED 10–20 recommended
MetaRNN	Quantitative	Recurrent neural network	MSA, other predictors	ClinVar, gnomAD	0 (benign) to 1 (pathogenic)	>0.5
Envision	Quantitative	Gradient boosting regression	MSA, sequence, structure	DMS/MAVE datasets	0 (pathogenic) to 1 (benign)	Not defined, <0.5 recommended
QAFI	Quantitative	Multiple linear regression models	Sequence, structure	DMS data	0 (pathogenic) to 1 (benign)	<0.81
EVE	Quantitative (binarizable)	Variational autoencoder (unsupervised)	MSA	MSA only	0 (benign) to 1 (pathogenic)	Not defined, <0.5 recommended
AlphaMissense	Binary	Protein Language Modeling	MSA, structure	MSA, population frequency data	0 (benign) to 1 (pathogenic)	>0.5
BayesDel	Binary	Naive Bayes classifier	MSA, other predictors	ClinVar, UniProt	-1.29334 (benign) to 0.75731 (pathogenic)	>0.0692655
REVEL	Binary	Supervised random forest	MSA, other predictors	HGMD	0 (benign) to 1 (pathogenic)	Not defined, <0.5 recommended
VEST4	Binary	Supervised random forest	MSA, sequence, function, structure	HGMD	0 (benign) to 1 (pathogenic)	Not defined

Table S1. Overview of selected predictive tools for variant pathogenicity classification.
Prediction type (binary or quantitative), core methodology, input features, training data sources, native score range, and recommended decision thresholds for each tool. Data were extracted from each tool's primary publication and the Atlas of Variant Effects resource.

Metric	Formula	Description
Sensitivity (Recall)	$\frac{TP}{TP+FN}$	Correctly predicted pathogenic variants divided by total pathogenic cases.
Specificity	$\frac{TN}{TN+FP}$	Correctly predicted benign variants divided by total benign cases.
PPV (Precision)	$\frac{TP}{TP+FP}$	True pathogenic predictions among all pathogenic predictions. Indicates how likely a pathogenic prediction is a TP.
NPV	$\frac{TN}{TN+FN}$	True benign predictions among all benign predictions. Indicates how likely a benign prediction is a TN.
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Ratio of correct predictions to total predictions made.
MCC	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$	Balanced measure of correlation between predictions and true cases. Ranges from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating random predictions.

Table S2. Performance metrics for binary classification models.
Definitions, mathematical formulae (in terms of TP, TN, FP, and FN), and interpretation for each metric—sensitivity (recall), specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and Matthews correlation coefficient (MCC)—used to evaluate pathogenicity classifiers. Metrics and descriptions were adapted from standard ML references.

Predictor	ClinVar	Humsavar
CADD	0.0	0.0
MetaRNN	71.4	7.9
Envision	6.8	0.8
QAFI	8.9	1.0
EVE	31.9	3.6
AlphaMissense	6.1	0.7
BayesDel	0.1	0.0
REVEL	7.6	0.8
VEST4	2.1	0.2

Table S3. Percentage of missing predictions per tool.
Fractions of ClinVar and Humsavar missense variants for which each predictor did not return a score, expressed as a percentage. High rates of missing data (e.g., MetaRNN in ClinVar) indicate potential coverage gaps that bias comparative performance analyses.

Predictor	Optimal Threshold
CADD	0.338
MetaRNN	0.548
Envision	0.332
QAFI	0.321
EVE	0.340

Table S4. Optimal decision thresholds for calibrated continuous predictors.
Youden's J derived cut-off probability for each continuous predictor after Platt scaling. The formula used was $J = Se + Sp - 1$. These thresholds were used to discretize calibrated scores into binary calls (pathogenic vs. benign) for downstream clinical metrics.

Predictor	AUROC	Brier	τ	Acc	Se	Sp	PPV	NPV	MCC
CADD	0.882	0.127	0.338	0.805	0.827	0.795	0.644	0.911	0.588
MetaRNN	0.964	0.064	0.548	0.917	0.892	0.929	0.850	0.950	0.810
Envision	0.824	0.152	0.332	0.760	0.743	0.768	0.598	0.865	0.487
QAFI	0.893	0.118	0.321	0.817	0.813	0.819	0.679	0.903	0.606
EVE	0.885	0.132	0.340	0.802	0.832	0.784	0.698	0.886	0.600
AlphaMissense	0.915	0.107	0.500	0.854	0.744	0.905	0.783	0.885	0.659
BayesDel	0.954	0.175	0.500	0.746	0.974	0.644	0.551	0.982	0.574
REVEL	0.941	0.104	0.500	0.868	0.888	0.860	0.744	0.944	0.717
VEST4	0.906	0.159	0.500	0.772	0.914	0.708	0.583	0.949	0.575

Table S5. Clinical performance metrics for calibrated continuous and native binary predictors.
Classification metrics are computed at the optimal thresholds (continuous) or default settings (binary) on the ClinVar & Humsavar dataset. Metrics include AUROC, Brier score, accuracy (Acc), sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and Matthews correlation coefficient (MCC).

Predictor	% Classified
CADD	79.86
MetaRNN	18.17
Envision	12.06
QAFI	55.40
EVE	36.87
AlphaMissense	63.36
BayesDel	73.67
REVEL	68.54
VEST4	77.32

Table S6. Coverage of ACMG evidence categories by predictor.
Percentage of variants that each tool classified into any defined ACMG computational evidence tier (i.e. not falling into the indeterminate zone). Higher values indicate broader variant coverage.

Predictor	Category	N	Precision
CADD	BP4_SUP	9884	0.911
CADD	PP3_MOD	8306	0.801
CADD	PP3_SUP	7091	0.584
CADD	BP4_MOD	10927	0.973
CADD	BP4_STR	1459	0.988
MetaRNN	PP3_MOD	2776	0.908
MetaRNN	BP4_MOD	1541	0.975
MetaRNN	BP4_SUP	1046	0.964
MetaRNN	BP4_VS	2638	0.987
MetaRNN	PP3_SUP	523	0.419
MetaRNN	BP4_STR	44	0.955
Envision	PP3_SUP	3730	0.724
Envision	PP3_MOD	1960	0.943
QAFI	PP3_MOD	6126	0.891
QAFI	PP3_SUP	2991	0.733
QAFI	BP4_SUP	14715	0.946
QAFI	BP4_MOD	2296	0.990
EVE	PP3_MOD	3196	0.932
EVE	PP3_SUP	4666	0.758
EVE	BP4_SUP	6326	0.939
EVE	BP4_MOD	3201	0.978
AlphaMissense	BP4_SUP	8347	0.932
AlphaMissense	PP3_MOD	6980	0.894
AlphaMissense	PP3_SUP	2653	0.754
AlphaMissense	BP4_MOD	11905	0.981
BayesDel	BP4_MOD	10617	0.991
BayesDel	PP3_SUP	4016	0.585
BayesDel	PP3_STR	3339	0.984
BayesDel	PP3_MOD	7322	0.907
BayesDel	BP4_SUP	9451	0.976
REVEL	BP4_SUP	6177	0.946
REVEL	PP3_MOD	5958	0.880
REVEL	PP3_SUP	3102	0.623
REVEL	PP3_STR	3674	0.979
REVEL	BP4_MOD	12828	0.981
REVEL	BP4_STR	573	0.993
REVEL	BP4_VS	15	1.000
VEST4	BP4_SUP	5403	0.940
VEST4	PP3_STR	3943	0.971
VEST4	PP3_MOD	6667	0.819
VEST4	BP4_MOD	16852	0.955
VEST4	PP3_SUP	3603	0.529

Table S7. Precision and sample size for each evidence category.
Number of variants (N) and precision (proportion of correct calls) achieved by each predictor in the supporting, moderate, strong, and very strong categories for benign (BP4) and pathogenic (PP3) evidence tiers.

	Benign (BP4)				Pathogenic (PP3)			
Method	**Very Strong**	**Strong**	**Moderate**	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very Strong**
CADD	–	≤ 0.15	0.15, 17.3	17.3, 22.7	25.3, 28.1	≥ 28.1	–	–
MetaRNN	≤ 0.025	0.025, 0.026	0.026, 0.1	0.1, 0.248	0.661, 0.816	≥ 0.816	–	–
Envision	–	–	–	–	–0.772, –0.673	$\geq$ –0.673	–	–
QAFI	–	–	$\leq$ –1.177	–1.177, –0.946	–0.7, –0.636	$\geq$ –0.636	–	–
EVE	–	–	≤ 0.097	0.097, 0.194	0.695, 0.861	≥ 0.861	–	–
AlphaMissense	–	–	≤ 0.09	0.09, 0.146	0.755, 0.887	≥ 0.887	–	–
BayesDel	–	–	$\leq$ –0.36	–0.36, –0.18	0.13, 0.27	0.27, 0.5	≥ 0.5	–
REVEL	≤ 0.003	0.003, 0.016	0.016, 0.183	0.183, 0.29	0.644, 0.773	0.773, 0.932	≥ 0.932	–
VEST4	–	–	≤ 0.302	0.302, 0.449	0.764, 0.861	0.861, 0.965	≥ 0.965	–

Table S8. Score ranges that define ACMG evidence tiers for each predictor.
Calibrated probability intervals corresponding to each evidence strength—very strong, strong, moderate, supporting—for benign (BP4) and pathogenic (PP3) categories, derived via the Pejaver odds-to-evidence mapping.

Supplementary Figures

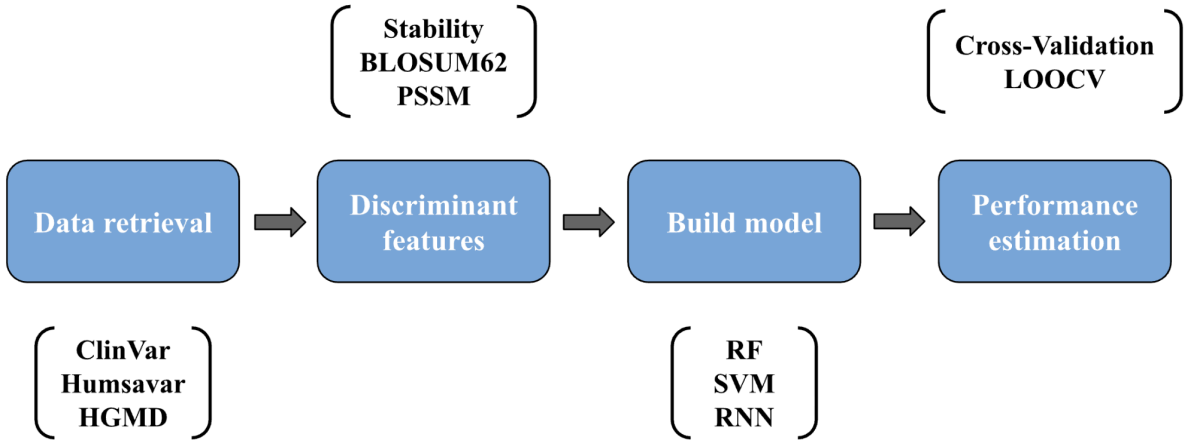

Figure S1. Machine learning pipeline for variant pathogenicity prediction.
Steps involved in developing an ML-based predictor for variant classification. The first step defines the classification problem. The next two steps focus on solving the classification problem. The final step validates the model. Detailed information is given on top or below each box.

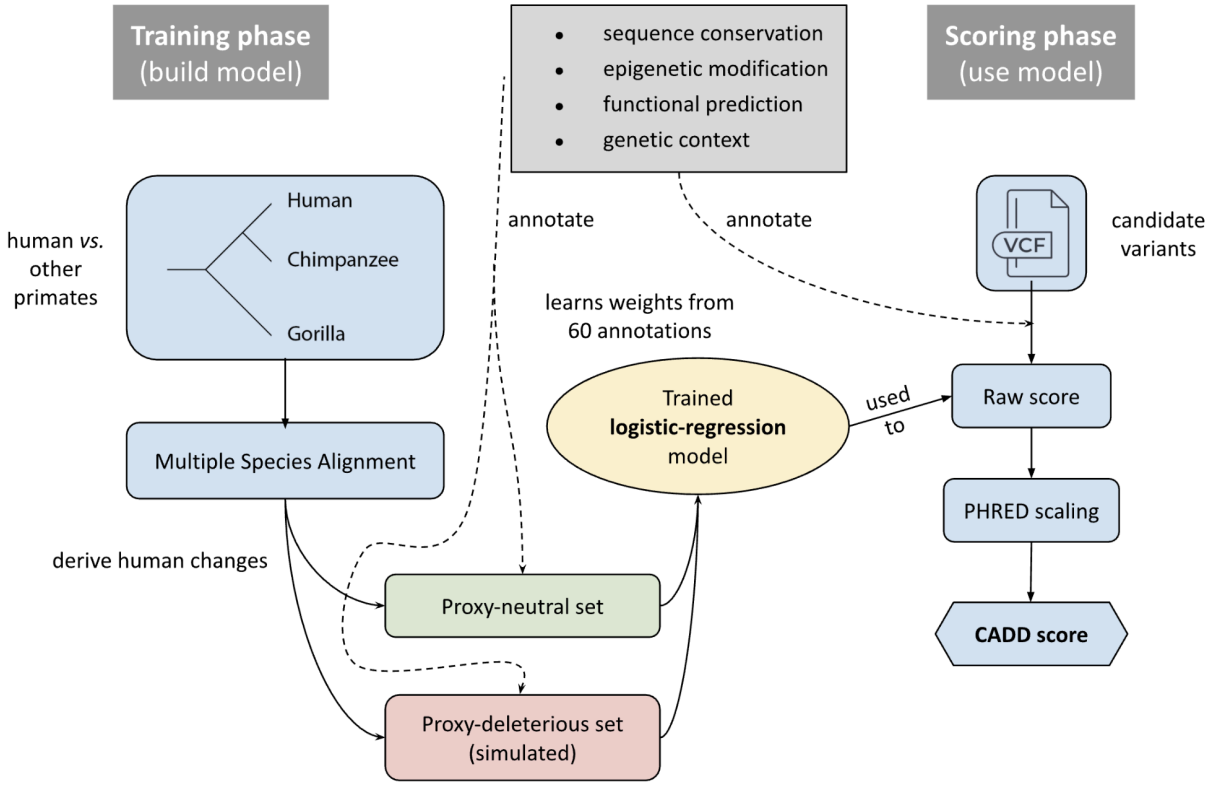

Figure S2. CADD workflow for genome-wide variant deleteriousness scoring.
Schematic of the Combined Annotation-Dependent Depletion (CADD) method.
***Training phase (left)**: fixed differences between humans and other primates are extracted from a multiple species alignment to form a proxy-neutral set, while matched simulated changes form a proxy-deleterious set. Both groups are annotated with 60 sequence, epigenetic, functional, and context-based features and used to fit a logistic regression model.*
***Scoring phase (right)**: user variants are annotated with the same features and passed through the trained model to obtain a raw log-odds score, which is then PHRED-scaled; the resulting CADD score ranks each variant by how deleterious-like it appears relative to all possible human SNVs.*

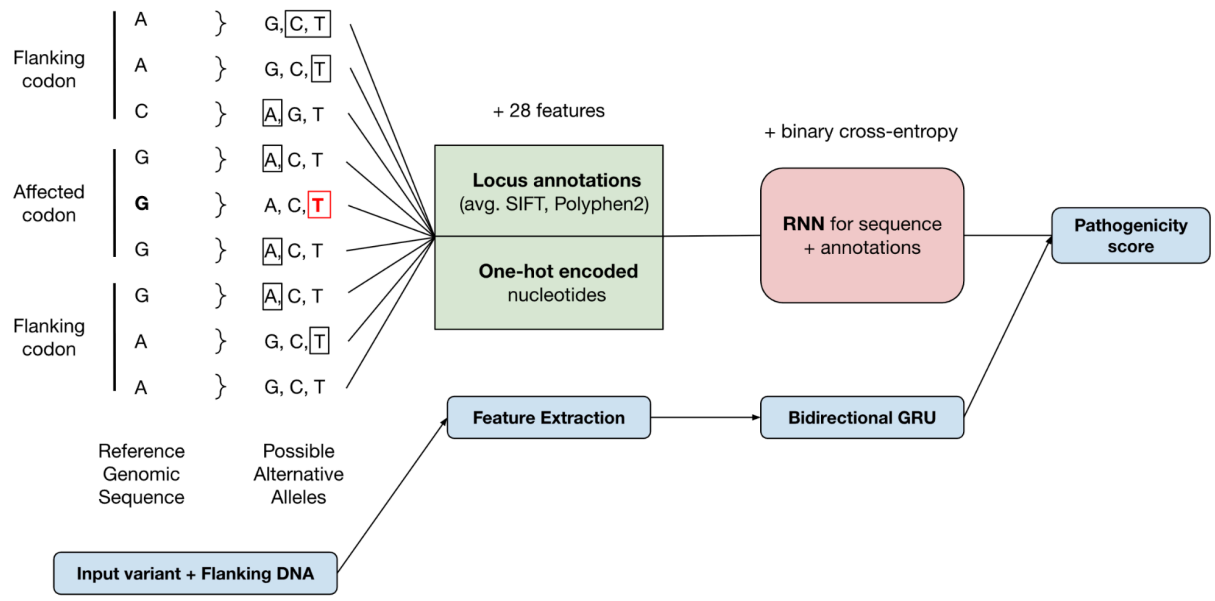

Figure S3. MetaRNN workflow for missense-variant pathogenicity prediction
The model extracts the ± 1 codon window, one-hot encodes each base, and appends 28 locus features (mean SIFT, PolyPhen2, allele frequency, etc.). A bidirectional GRU reads the window left→right and right→left, merges the hidden states, and returns a pathogenicity score. Weights are learned by minimizing binary cross-entropy, which penalizes deviation from benign/pathogenic labels. Rectangles in the "Possible Alternative Alleles" column show one-hot slots; the red T highlights the actual mutant allele (G → T) being evaluated.

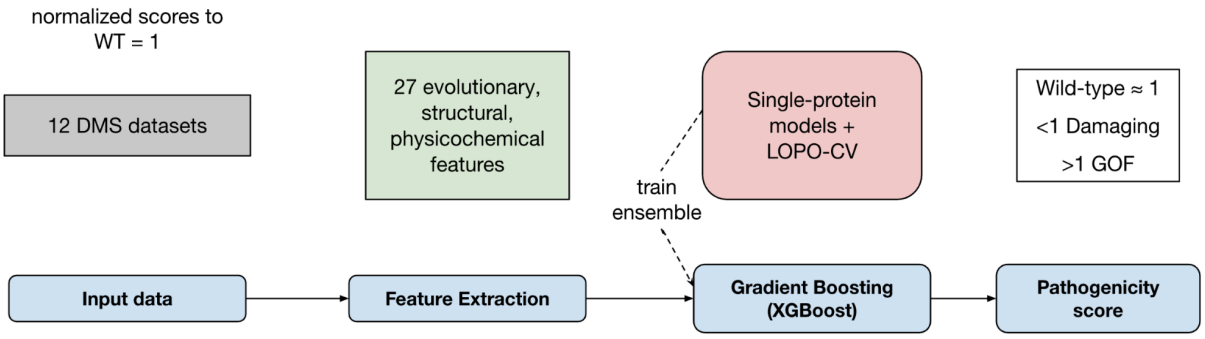

Figure S4. Envision workflow for protein-function impact prediction.
Variant effect scores from 12 deep-mutational-scanning (DMS) datasets are first normalized to wild type = 1. For every variant, 27 evolutionary, structural, and physicochemical features are extracted.
Single-protein gradient-boosting models are trained with leave-one-protein-out (LOPO) cross-validation, then ensembled via XGBoost to output a quantitative score: ≈ 1 wild-type-like, <1 damaging (lower = more severe), >1 gain-of-function (GOF).

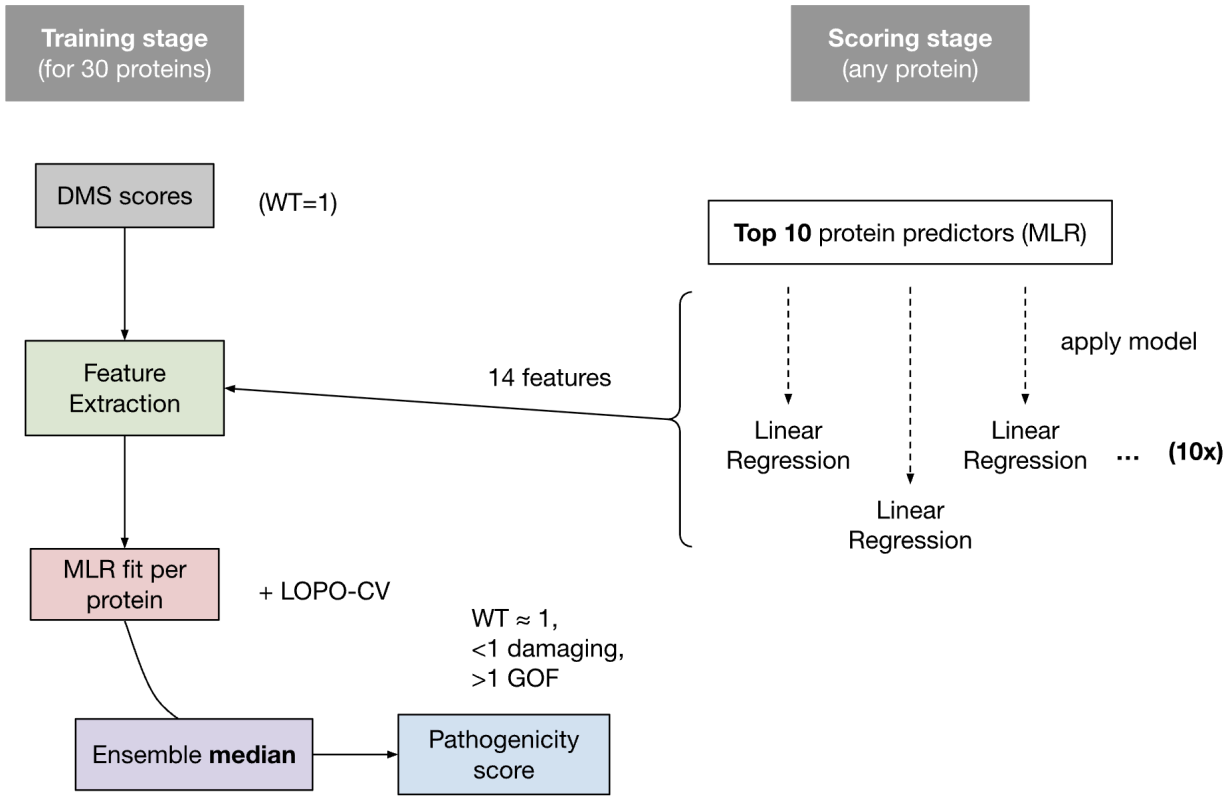

Figure S5. QAFI workflow for quantitative functional-impact prediction.
Deep-mutational-scanning data from 30 proteins are normalized to wild-type = 1 and used to train protein-specific multiple-linear-regression (MLR) models on 14 sequence- and structure-based features. Models are validated with leave-one-protein-out (LOPO) cross-validation.
During scoring, the same 14 features are computed for any query missense variant, passed through the top 10 closest protein predictors, and their outputs are aggregated by the median to yield the final QAFI score: ≈ 1 wild-type-like, < 1 damaging (lower = more severe), > 1 gain-of-function (GOF).

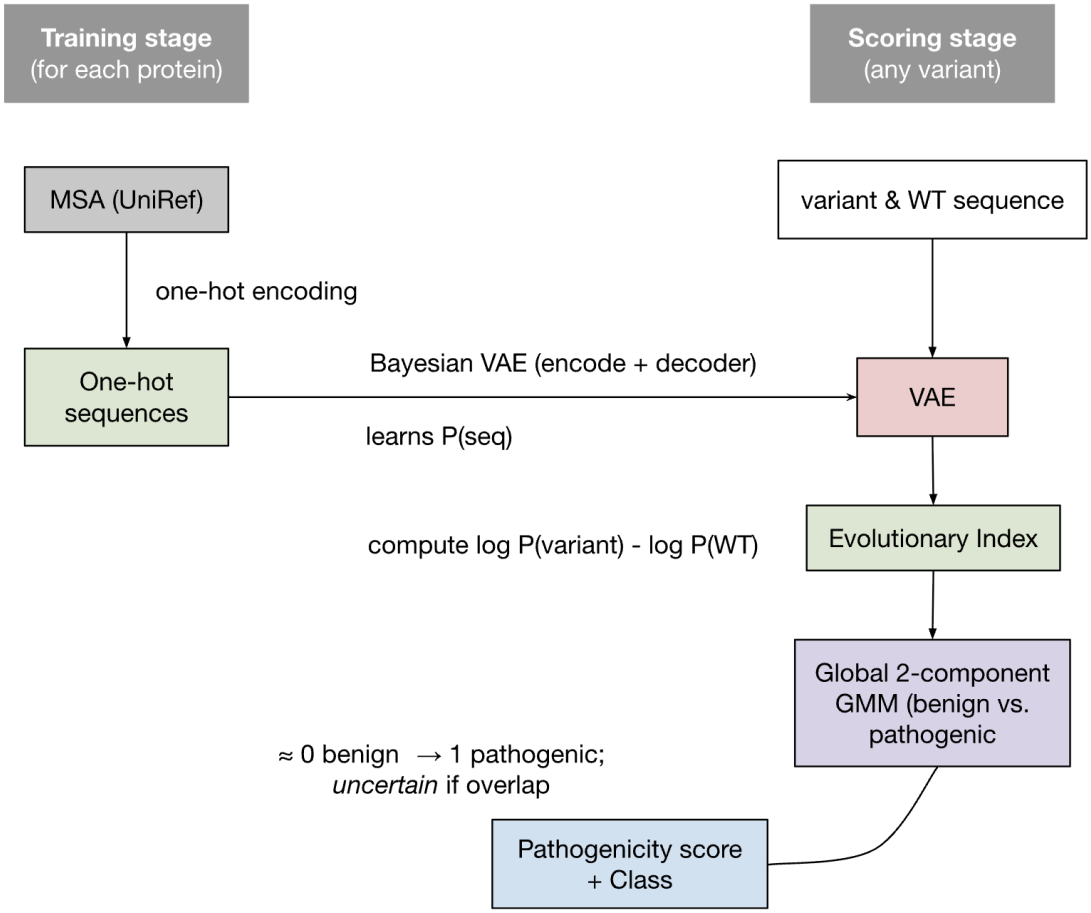

Figure S6. EVE workflow for unsupervised missense-variant pathogenicity prediction.
An evolutionary multiple-sequence alignment (MSA) is one-hot encoded and fed into a Bayesian variational auto-encoder (VAE) that learns the probability of any sequence for that protein. At scoring time, the trained VAE evaluates both the wild-type and the mutated sequence; the log-likelihood difference (evolutionary index) quantifies how much the variant deviates from evolutionary constraints.
Evolutionary indices from thousands of proteins are then calibrated by a 2 component Gaussian-mixture model (GMM; benign vs. pathogenic), converting each index into a continuous score, with variants falling in the mixture overlap flagged as uncertain.

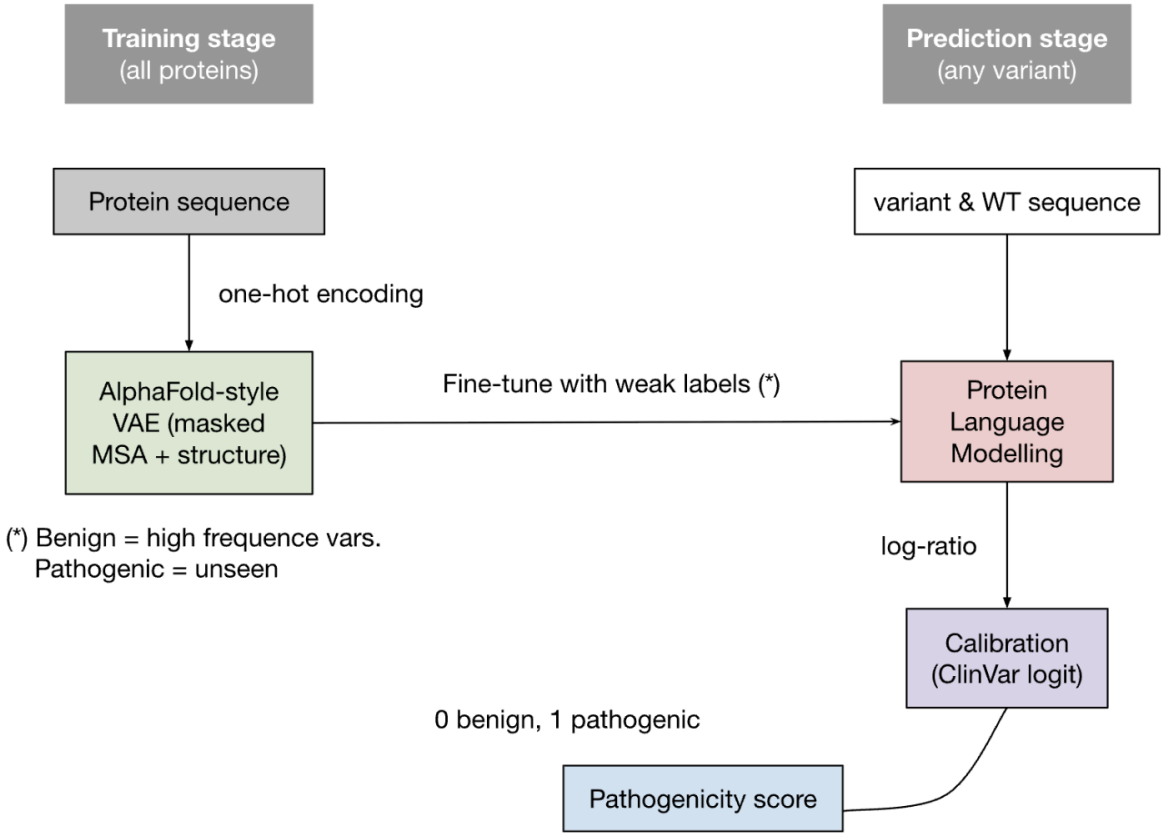

Figure S7. AlphaMissense workflow for missense-variant pathogenicity prediction.
An AlphaFold-style protein language model is first pre-trained on UniRef sequences and 3D structures, then weak-label fine-tuned: common human/primate variants serve as proxy-benign examples, while never-observed variants act as proxy-pathogenic. For any query missense, the model compares the log-likelihood of the mutant residue to that of the wild type; this log-likelihood ratio is a raw pathogenicity signal. A final ClinVar-based logistic calibration converts the ratio into an AlphaMissense score on a 0–1 scale.

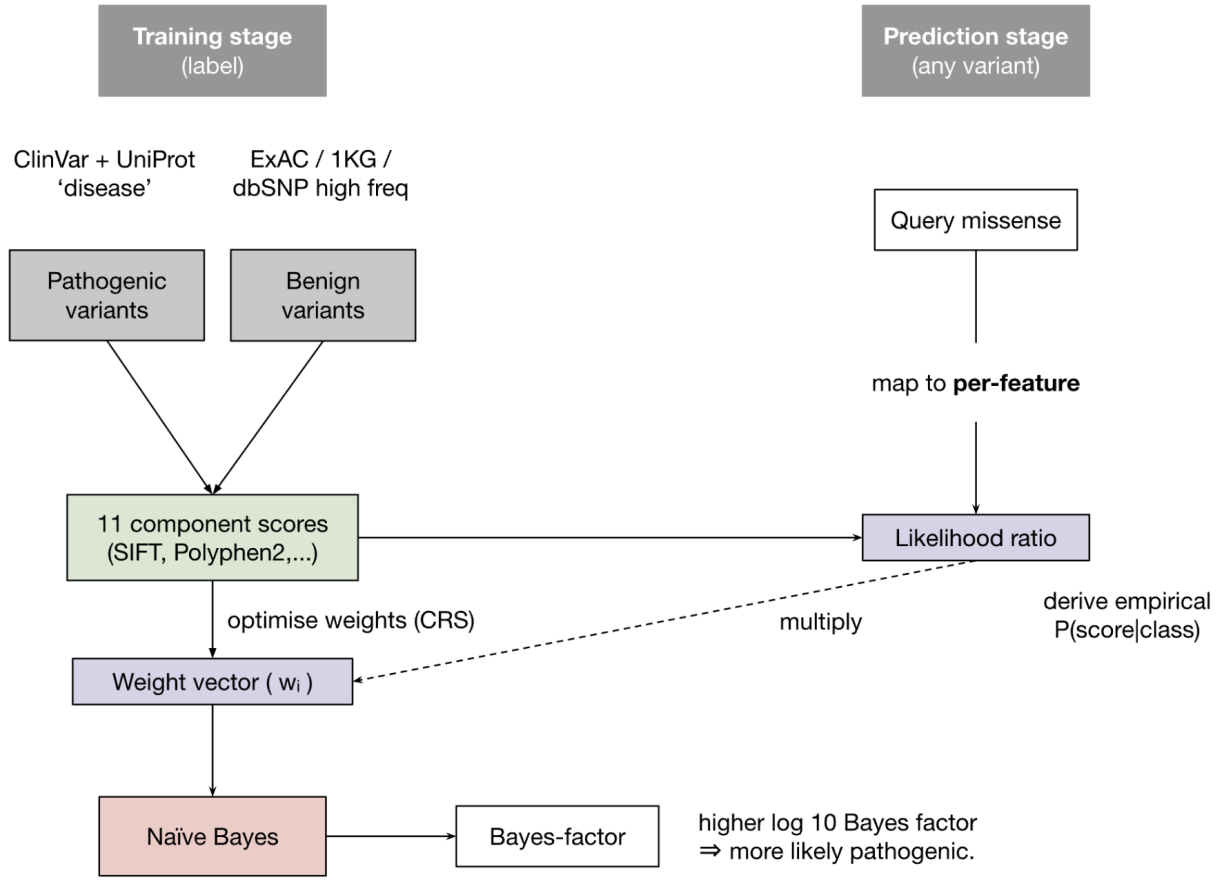

Figure S8. BayesDel workflow for likelihood-ratio-based missense-variant pathogenicity scoring.
ClinVar pathogenic variants and high-frequency population variants (benign) define two training sets. For each of 11 component predictors (PolyPhen-2, SIFT, CADD, etc.) BayesDel derives empirical score distributions P(score|pathogenic) and P(score|benign), then forms a per-feature likelihood ratio (LR). A controlled random search step learns feature weights w_i that maximize AUC on the training data.
At prediction time the same 11 scores are fetched for a query missense; each is converted to its LR, and the LRs are exponentiated by w_i and multiplied to yield a Bayes factor. The reported BayesDel score is log10 (Bayes factor): higher values represent stronger evidence of pathogenicity.

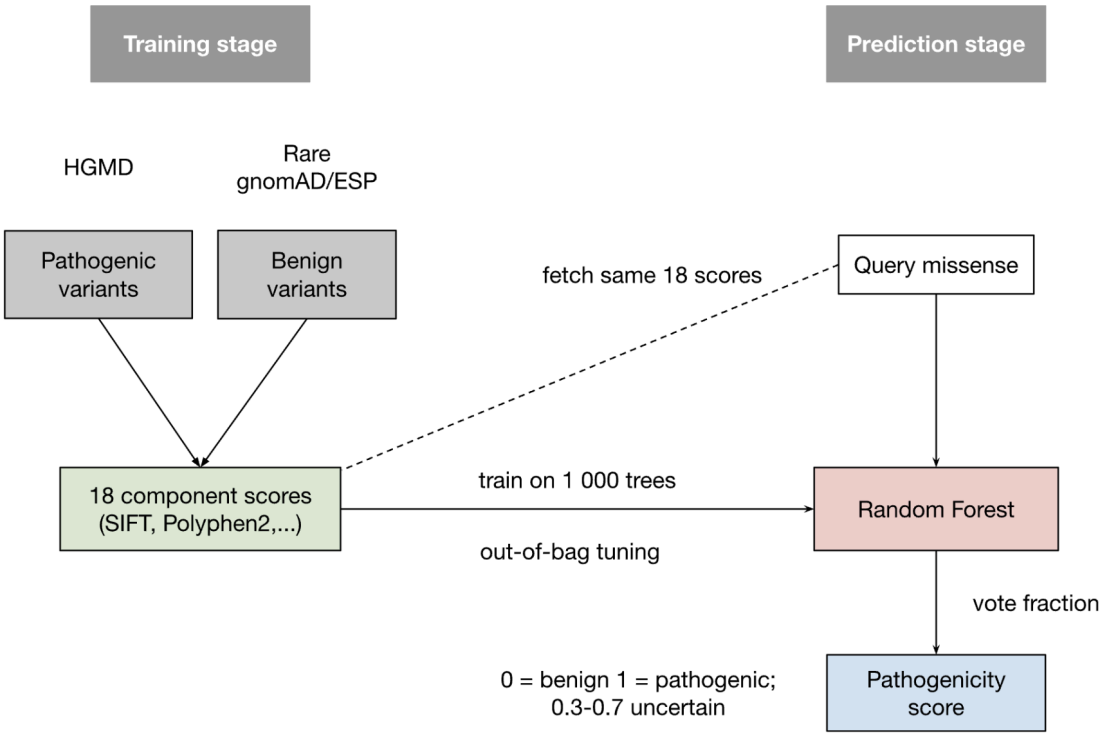

Figure S9. REVEL workflow for ensemble missense-variant pathogenicity prediction.
Pathogenic missense variants from HGMD and rare neutral variants from population databases supply the training labels. For each variant, REVEL retrieves 18 pre-existing scores (eight conservation, ten functional; e.g., SIFT, PolyPhen-2, MutPred, GERP++). A 1,000-tree random forest learns to distinguish the two classes using out-of-bag error for tuning. At inference, the same 18 scores are fetched for any query variant; the fraction of trees voting "pathogenic" is reported as the REVEL score (0 = benign, 1 = pathogenic).

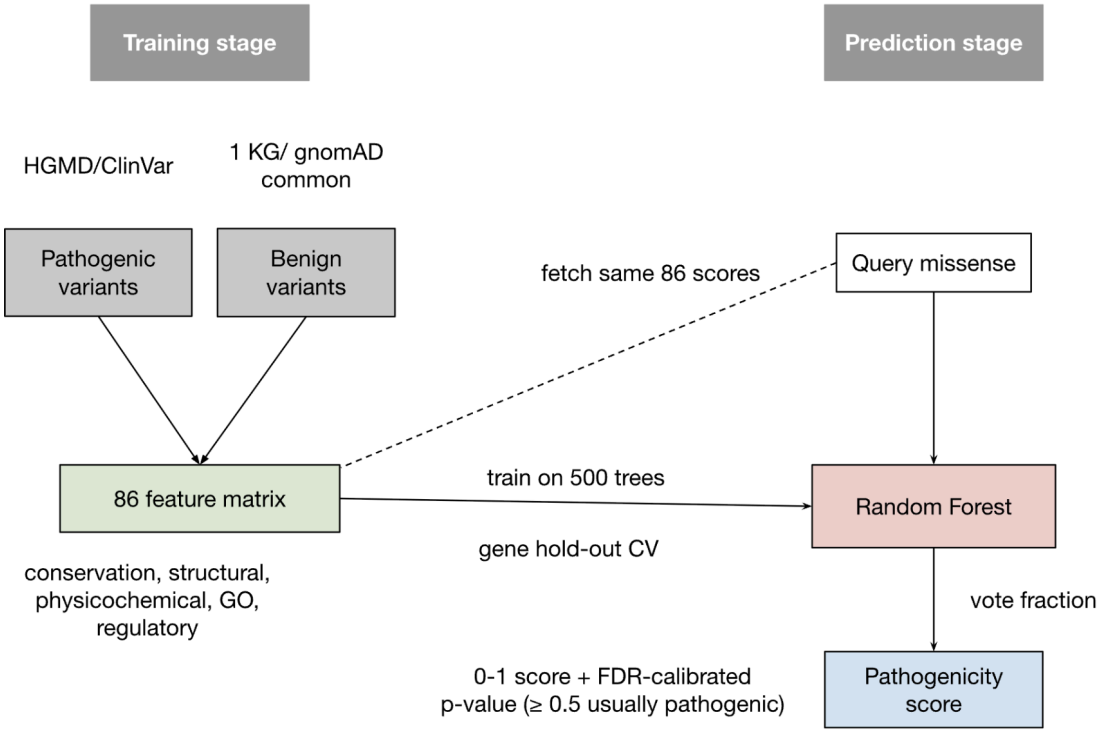

Figure S10. VEST4 workflow for supervised missense-variant pathogenicity prediction.
Pathogenic variants from HGMD/ClinVar and common benign variants from 1000 Genomes + gnomAD supply the training labels. Each variant is summarised by 86 sequence, structural, physicochemical, regulatory, and Gene-Ontology features. A 500-tree random-forest classifier is trained with "gene hold-out" cross-validation to avoid gene-specific overfitting. At inference, a query missense is annotated with the same 86 features, fed into the trained forest, and the fraction of trees voting "pathogenic" becomes the VEST4 score (0 = benign, 1 = pathogenic); an associated FDR-calibrated p-value is also reported, with scores ≥ 0.5 typically flagged as likely pathogenic.

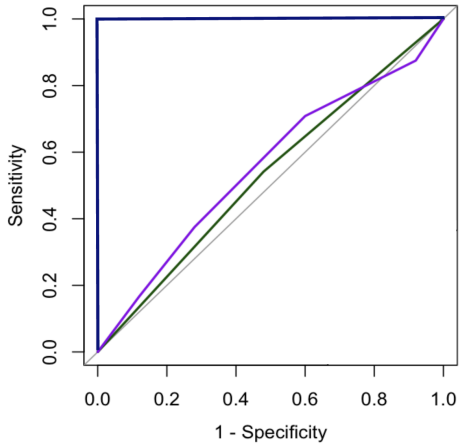

Figure S11. Area under the Receiver Operating Characteristic (ROC) curve (AUC).
The area under the purple and green curves represents the AUC for the respective models. The x and y axes show the false positive rate (FPR, or 1-specificity) and true positive rate (TPR, or sensitivity), respectively. The dark blue line represents the ROC curve of a perfect model, while the grey line corresponds to a random model.

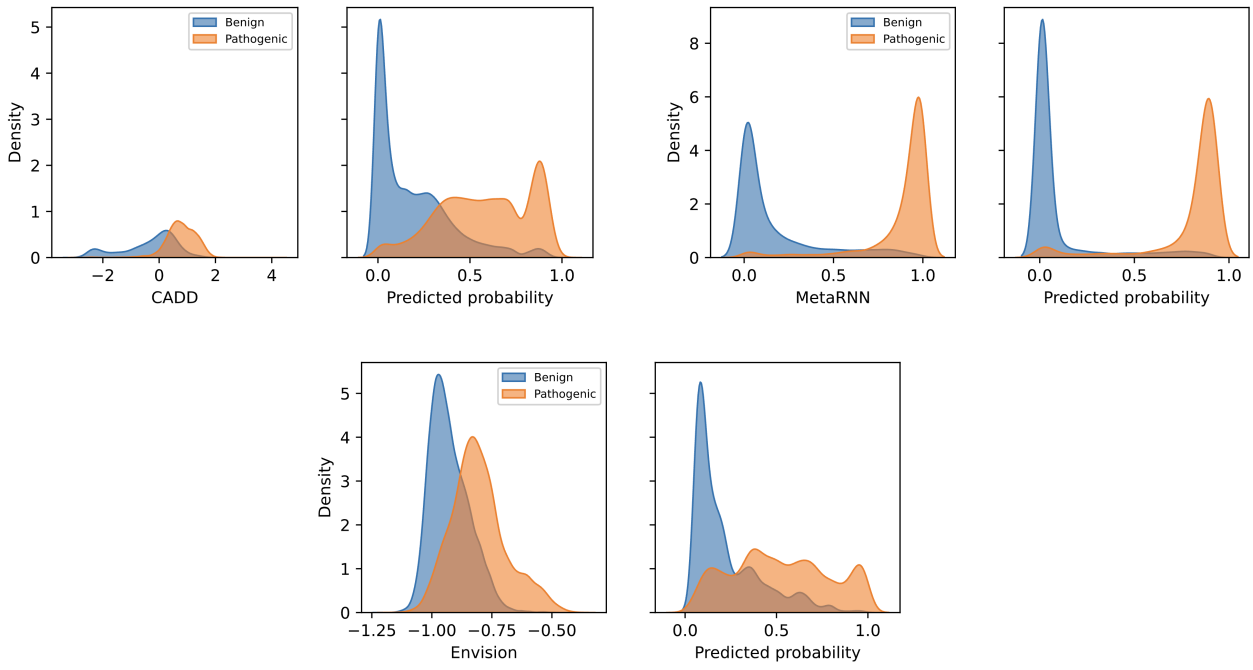

Figure S12. KDE plots of raw and calibrated score distributions for CADD, MetaRNN, and Envision.
Kernel density estimates of predictor scores for benign (blue) and pathogenic (orange) variants before (left panels) and after (right panels) Platt scaling. KDE illustrates how calibration shifts and sharpens the class-specific distributions. Detailed threshold values used are provided in Supplementary Table S4.

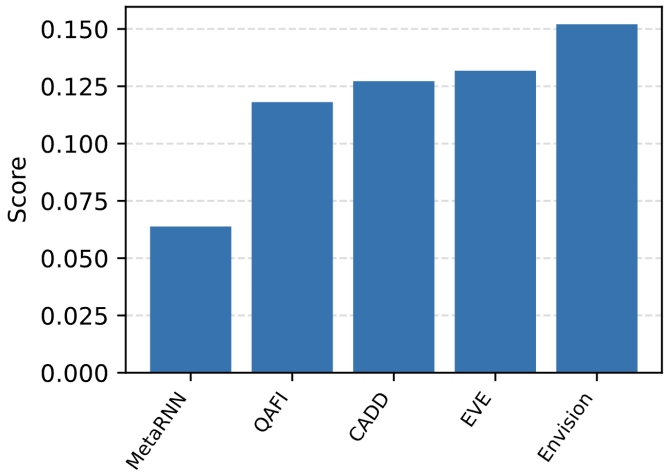

Figure S13. Brier scores of calibrated continuous predictors.
Barplot showing the Brier score for each continuous predictor after Platt scaling. Lower values indicate better overall probability calibration and accuracy. Continuous tools are ordered left to right by increasing the Brier score.