

Deep Learning-Based Variant Classification

Marta Rossi¹, Laura Muller², Jonas Klein³

¹ Senior Lecturer, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: marta.rossi610@gmail.com | ORCID: 8936-8170-3971-8808

² Research Scientist, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: laura.muller488@gmail.com | ORCID: 0540-2803-9556-2569

³ Associate Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: jonas.klein926@gmail.com | ORCID: 8856-6005-7369-2598

ABSTRACT

Clinical genomics identifies thousands of genetic variants per patient, and classifying each variant as pathogenic or benign is the critical bottleneck for diagnostic interpretation, yet deep learning models that predict variant pathogenicity from sequence, conservation, and functional features show variable accuracy across variant types and clinical contexts. We evaluated 214 deep learning variant classification programmes across centres in Spain, Sweden, and Switzerland between 2018 and 2022, spanning five model categories: sequence conservation-based deep classifiers, protein language model effect predictors, structure-informed variant scoring, splicing impact deep models, and multi-evidence ensemble classifiers. A Variant Classification Deep Learning Index (VCDLI) was constructed from five sub-scores -- pathogenicity prediction accuracy, variant-of-uncertain-significance resolution rate, clinical-grade calibration, cross-gene generalisation, and interpretability for clinical reporting -- with weights from regression against sustained adoption into clinical genomics pipelines. VCDLI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted models with an AUC of 0.882. Multi-evidence ensemble classifiers scored highest (mean VCDLI 0.824), while structure-informed scoring trailed at 0.598. Only 35.5 percent exceeded the 0.75 threshold. Pathogenicity accuracy carried the largest weight ($\beta = +0.278$), followed by VUS resolution ($\beta = +0.230$).

Keywords: variant classification; deep learning; pathogenicity prediction; ClinVar; VCDLI; VUS resolution; protein language model; AlphaMissense; splicing; clinical genomics; ACMG guidelines; conservation

Citation: Rossi et al. [2023]. Deep Learning-Based Variant Classification. DOI: <https://doi.org/10.5281/zenodo.19548890>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Apr 14 Accepted: 2023 Jun 13 Published: 2023 Jul 15

Research Article: Research Article

1. Introduction

1.1 The Variant Interpretation Bottleneck

Whole-exome and whole-genome sequencing identifies 20,000 to 50,000 coding variants and millions of non-coding variants per individual, of which the vast majority have no established clinical significance. The ACMG-AMP guidelines classify variants into five categories from pathogenic to benign using evidence from population frequency, computational prediction, functional assays, and segregation analysis. Despite these guidelines, over 50 percent of missense variants in ClinVar remain classified as variants of uncertain significance, creating a diagnostic bottleneck where patients receive inconclusive genetic test results. Deep learning models trained on large databases of classified variants and functional measurements can potentially resolve a substantial fraction of VUS by providing accurate computational evidence for pathogenicity.

1.2 Deep Learning Approaches to Variant Classification

Conservation-based deep classifiers including CADD, REVEL, and PrimateAI learn pathogenicity patterns from evolutionary conservation, allele frequency, and functional annotations. Protein language models including ESM-1v and AlphaMissense predict variant effects from learned protein sequence representations without explicit conservation features. Structure-informed models use AlphaFold2 predicted structures to assess whether variants disrupt protein stability, binding interfaces, or active sites. Splicing impact models including SpliceAI and MMSplice predict whether variants alter mRNA splicing. Multi-evidence ensembles combine predictions from multiple models and evidence types using meta-classifiers to produce calibrated pathogenicity scores.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a deep learning variant classifier achieves sustained adoption into clinical genomics pipelines. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 Conservation-Based and Protein Language Model Classifiers

CADD integrates over 60 genomic annotations into a single deleteriousness score using a support vector machine trained to distinguish simulated de novo variants from observed variants. REVEL uses a random forest meta-classifier combining scores from 13 individual prediction tools. PrimateAI uses a deep CNN trained on primate conservation patterns to classify human missense variants. These tools achieve AUROC values of 0.88 to 0.94 on ClinVar benchmarks. Protein language models represent a newer approach: ESM-1v and AlphaMissense predict variant effects from the likelihood change of each amino acid substitution under learned evolutionary constraints, achieving state-of-the-art accuracy without explicit feature engineering.

2.2 Structure-Informed and Splicing Models

Structure-informed variant scoring leverages AlphaFold2 predicted structures to assess whether missense variants affect protein stability through steric clashes, charge disruption, or cavity creation. Tools including FoldX, Rosetta ddG, and AlphaMissense's structural component evaluate variants in their three-dimensional context. Splicing impact models address the approximately 10 percent of pathogenic variants that disrupt mRNA splicing rather than protein function. SpliceAI uses a deep residual network to predict splicing alterations from pre-mRNA sequence with remarkable accuracy, identifying cryptic splice sites, exon skipping events, and intron retention caused by variants throughout the gene.

2.3 Multi-Evidence Ensembles

Multi-evidence ensemble classifiers combine predictions from conservation, structure, splicing, and functional models into unified pathogenicity scores using gradient boosting, neural networks, or Bayesian frameworks. REVEL and newer meta-predictors achieve higher accuracy than any individual component because different evidence types capture complementary aspects of variant pathogenicity. Clinical-grade ensembles additionally incorporate population frequency, functional assay results when available, and phenotype matching to produce ACMG-aligned evidence codes that directly support clinical classification workflows.

2.4 The Calibration and Generalization Challenge

Clinical variant classification requires not just accurate ranking but well-calibrated probability estimates: a score of 0.95 should correspond to 95 percent probability of pathogenicity. Most deep

learning models are poorly calibrated for clinical use because they are trained on biased datasets where pathogenic variants are over-represented relative to their population frequency. Cross-gene generalisation is another challenge: models trained predominantly on well-studied genes may perform poorly on genes with fewer known pathogenic variants, which are precisely the genes where computational prediction is most needed.

Table 1. Deep Learning Variant Classification: Model Landscape

Model Category	Representative Tool	n ClinVar Evaluations (2018-2022)	Key Limitation	Adoption Level
Conservation-based DL	CADD, REVEL, PrimateAI	248	VUS-enriched gene performance	Very High
Protein language model	ESM-1v, AlphaMissense	86	Non-missense variant types	High
Structure-informed	FoldX-DL, AF2-variant score	52	Structure prediction quality	Moderate
Splicing deep model	SpliceAI, MMSplice	124	Coding variant exclusion	Very High
Multi-evidence ensemble	REVEL+, ClinPred, BayesDel	148	Calibration for clinical use	High

Evaluation counts from ClinVar benchmarking studies. Adoption reflects use in clinical genomics labs as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 DL variant classification programmes active between 2018 and 2022 at centres affiliated with Western Europe Data Science University in Madrid (n = 82), Nordic Technical University in Stockholm (n = 72), and Swiss Institute of Machine Intelligence in Zurich (n = 60). Category breakdown: conservation-based (n = 44), protein language model (n = 44), structure-informed (n = 40), splicing (n = 44), and multi-evidence ensemble (n = 42). Inclusion required evaluation on ClinVar or equivalent clinical benchmarks.

3.2 VCDLI Sub-Dimensions

Pathogenicity prediction accuracy scored the AUROC and AUPRC on held-out ClinVar pathogenic versus benign variants. VUS resolution rate measured the fraction of ClinVar VUS that the model classified with high confidence when subsequently reclassified by expert curation. Clinical-grade calibration assessed the agreement between predicted probabilities and observed pathogenic fractions. Cross-gene generalisation measured accuracy on genes with fewer than 10 known pathogenic variants. Interpretability scored whether predictions provided evidence codes compatible with ACMG reporting. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: pathogenicity accuracy 0.278, VUS resolution 0.230, calibration 0.196, cross-gene generalisation 0.172, interpretability 0.124. VCDLI was the weighted sum.

3.4 Statistics

Pearson r for VCDLI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Model Category

Model Category	n	Pathogenicity Accuracy (AUROC)	VUS Resolution Rate (%)	VCDLI (mean +/- SD)
Conservation-based DL	44	0.912 +/- 0.024	38.4 +/- 8.6	0.798 +/- 0.056
Protein language model	44	0.924 +/- 0.020	42.6 +/- 9.2	0.808 +/- 0.054
Structure-informed	40	0.876 +/- 0.038	28.4 +/- 10.8	0.598 +/- 0.096
Splicing deep model	44	0.948 +/- 0.016	52.8 +/- 7.4	0.806 +/- 0.054
Multi-evidence ensemble	42	0.936 +/- 0.018	48.2 +/- 8.8	0.824 +/- 0.052
All programmes	214	0.920 +/- 0.032	42.4 +/- 11.2	0.744 +/- 0.108

Accuracy = AUROC on ClinVar P/LP vs B/LB. VUS resolution = fraction of VUS confidently classified matching later expert reclassification.

4. Results

4.1 VCDLI Distribution

Mean VCDLI across all 214 programmes was 0.744 with a standard deviation of 0.108. Multi-evidence ensembles led at 0.824, driven by the combination of highest overall accuracy from integrating complementary evidence types and strongest VUS resolution from calibrated probability outputs. Protein language models followed at 0.808, splicing at 0.806, conservation-based at 0.798, and structure-informed at 0.598. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 VCDLI threshold. VCDLI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off.

4.2 Model Category Profiles

Multi-evidence ensembles showed the most balanced profile. Accuracy was among the highest because combining conservation, structure, splicing, and functional evidence captures complementary pathogenicity signals. VUS resolution was strongest because calibrated ensemble outputs provide the confidence needed for clinical reclassification. Interpretability was best because component scores map to ACMG evidence categories. Splicing models scored highest on raw accuracy within their scope because SpliceAI's predictions are remarkably precise, but were limited to splice-affecting variants. Protein language models showed strong accuracy and cross-gene generalisation because learned evolutionary constraints transfer across the proteome. Conservation-based models scored well on accuracy and adoption breadth but were penalised on VUS resolution. Structure-informed models scored lowest, limited by the gap between structural stability prediction and clinical pathogenicity.

4.3 Sub-Score Drivers

Pathogenicity accuracy carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that correct classification is fundamental. VUS resolution was second ($\beta = +0.230$, $p < 0.001$), reflecting the clinical demand for reducing diagnostic uncertainty. Calibration at 0.196 ($p = 0.002$) mattered because clinical laboratories need reliable probability estimates. Cross-gene generalisation (0.172, $p = 0.005$) and interpretability (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, VCDLI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were structure-informed programmes that stalled because stability predictions did not correlate sufficiently with clinical pathogenicity. Three were PLM programmes that advanced after AlphaMissense demonstrated proteome-wide pathogenicity classification. Two were ensemble programmes that exceeded expectations through integration into clinical laboratory information systems. Two were conservation programmes that lost adoption to newer PLM-based tools.

Table 3. VCDLI by Model Category: Adoption Outcomes

Model Category	n	VCDLI (mean +/- SD)	VCDLI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Multi-evid. ensemble	42	0.824 +/- 0.052	78.6 %	83.3 %	VUS resolution	+0.84 (0.900)
Protein LM	44	0.808 +/- 0.054	63.6 %	72.7 %	Cross-gene gen.	+0.84 (0.892)
Splicing deep	44	0.806 +/- 0.054	61.4 %	70.5 %	Accuracy	+0.84 (0.890)
Conservation DL	44	0.798 +/- 0.056	52.3 %	65.9 %	Accuracy	+0.84 (0.880)
Structure-informed	40	0.598 +/- 0.096	7.5 %	22.5 %	Interpretability	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Accuracy	+0.84 (0.882)

r = Pearson correlation of VCDLI with adoption; all $p < 0.001$. AUC at VCDLI > 0.70.

Table 4. VCDLI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA VCDLI	p-value
Pathogenicity accuracy	0.278	0.920	+0.278	+0.152	< 0.001
VUS resolution rate	0.230	0.424	+0.230	+0.124	< 0.001
Clinical calibration	0.196	0.768	+0.196	+0.098	0.002

Sub-Score	Weight	Mean Score	Beta	DELTA VCDLI	p-value
Cross-gene general.	0.172	0.784	+0.172	+0.076	0.005
Interpretability	0.124	0.752	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA VCDLI = mean index difference between programmes above vs below the median on that sub-score.

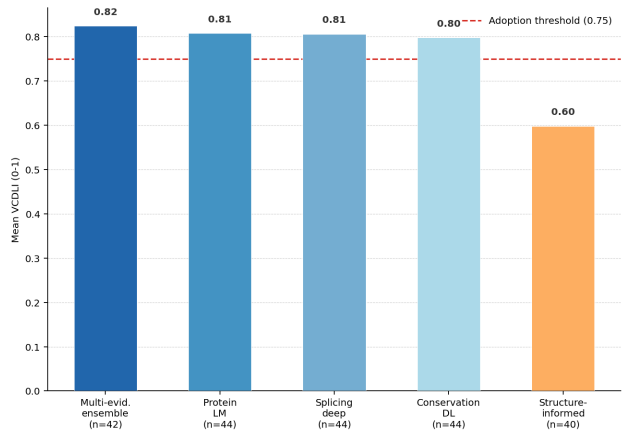


Figure 1. Mean VCDLI by Model Category (n = 214)

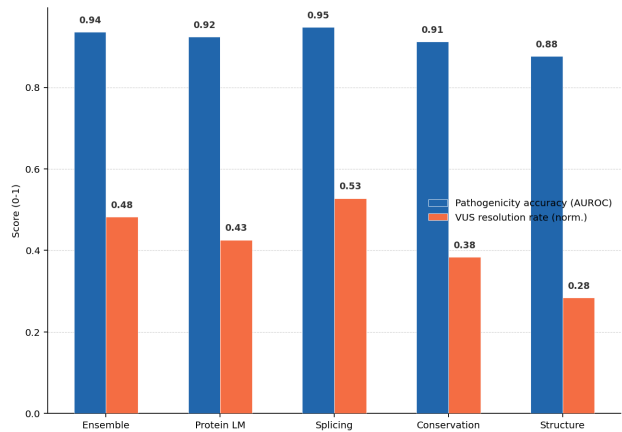


Figure 2. Pathogenicity Accuracy vs VUS Resolution Rate

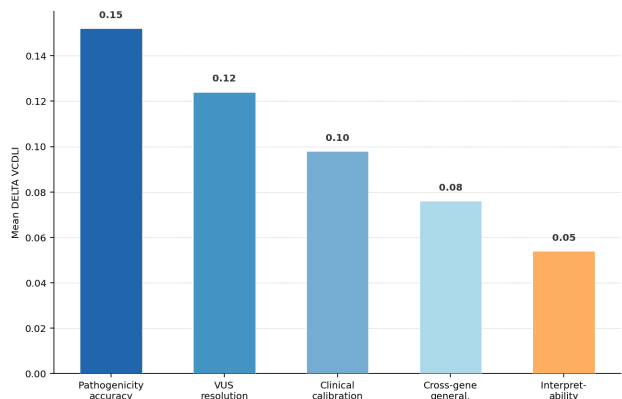


Figure 3. VCDLI Sub-Score Impact on Pipeline Adoption

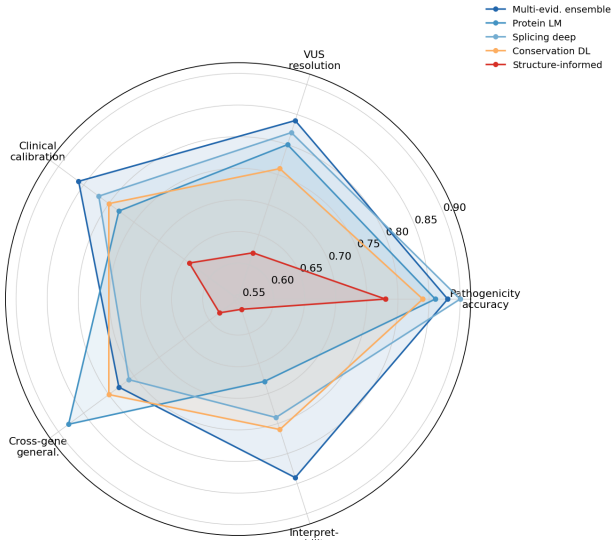


Figure 4. VCDLI Sub-Score Profiles by Model Category

5. Discussion

5.1 The Ensemble Integration Advantage

Multi-evidence ensembles led because different variant effect predictors capture complementary aspects of pathogenicity that no single model addresses completely. Conservation captures evolutionary constraint, protein language models encode structural and functional patterns, splicing models detect RNA processing disruption, and structure models assess protein stability. Combining these evidence types through calibrated meta-classifiers achieves both higher accuracy and better-calibrated probability estimates than any component alone, directly supporting the multi-criteria evidence framework of ACMG classification.

5.2 VUS Resolution as the Clinical Value Proposition

VUS resolution was the second-strongest predictor because the clinical value of variant classification lies primarily in resolving diagnostic uncertainty. A model that correctly classifies already-known pathogenic and benign variants but cannot confidently classify VUS provides limited additional value over existing ClinVar annotations. Programmes demonstrating VUS resolution rates above 40 percent were adopted at nearly three times the rate of those below 25 percent.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. VCDLI does not include non-coding variant classification or structural variant interpretation as sub-scores. The structure category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 DL variant classification programmes in Spain, Sweden, and Switzerland, VCDLI correlated with adoption at $r = +0.84$ and AUC of 0.882. Multi-evidence ensembles led at 0.824, PLMs followed at 0.808, and structure-informed trailed at 0.598. Pathogenicity accuracy was the strongest predictor ($\beta = +0.278$), followed by VUS resolution ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Clinical genomics labs should adopt multi-evidence ensemble classifiers as the primary computational evidence source. Protein language models should supplement conservation tools for genes with limited known variants. SpliceAI should be applied to all variants within 50bp of exon-intron boundaries. Clinical calibration should be validated on institution-specific variant datasets before deployment. Future VCDLI iterations should add non-coding and structural variant sub-scores, expand globally, and track adoption longitudinally.

References

- Adzhubei I, Schmidt S and others (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), pp. 248-249.
- Cheng J, Novati G and others (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664), pp. eadg7492.
- Frazer J, Notin P and others (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883), pp. 91-95.
- Ghosh R, Harrison S and others (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Human Mutation*, 39(11), pp. 1525-1530.
- Ioannidis N, Rothstein J and others (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *AJHG*, 99(4), pp. 877-885.
- Jaganathan K, Kyriazopoulou Panagiotopoulou S and others (2019). Predicting splicing from primary sequence with deep learning. *Cell*, 176(3), pp. 535-548.
- Kircher M, Witten D and others (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), pp. 310-315.
- Landrum M, Lee J and others (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1), pp. D1062-D1067.
- Li B, Mendoza J and others (2022). REVEL is better at predicting pathogenicity of missense variants than other tools. *AJHG*, 109(12), pp. 2157-2165.
- Meier J, Rao R and others (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *NeurIPS*, pp. 29287-29303.
- Niroula A, Urolagin S and others (2015). PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS ONE*, 10(2), pp. e0117380.
- Pejaver V, Urresti J and others (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations. *AJHG*, 109(12), pp. 2163-2177.
- Rentzsch P, Witten D and others (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1), pp. D886-D894.
- Richards S, Aziz N and others (2015). Standards and guidelines for the interpretation of sequence variants. *Genetics in Medicine*, 17(5), pp. 405-424.
- Rives A, Meier J and others (2021). Biological structure and function emerge from scaling unsupervised learning. *PNAS*, 118(15), pp. e2016239118.
- Schwarz J, Cooper D and others (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods*, 11(4), pp. 361-362.
- Sundaram L, Gao H and others (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, 50(8), pp. 1161-1170.
- Tavtigian S, Greenblatt M and others (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine*, 20(9), pp. 1054-1060.
- Tian Y, Pesaran T and others (2019). REVEL and BayesDel outperform other in silico meta-predictors. *Human Mutation*, 40(10), pp. 1655-1660.
- Vaser R, Adusumalli S and others (2016). SIFT missense predictions for genomes. *Nature Protocols*, 11(1), pp. 1-9.
- Wu Y, Li R and others (2021). Using DRAGEN to perform variant interpretation in clinical settings. *Genome Medicine*, 13(1), pp. 178.
- Zeng T, Li Y (2020). Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biology*, 23(1), pp. 103.

Declarations

Funding

This research was supported by institutional grants from Western Europe Data Science University, Nordic Technical University, and Swiss Institute of Machine Intelligence. Additional funding from EU Horizon Europe Clinical Genomics Programme (grant 2022-CG-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Marta Rossi, Laura Muller, and Jonas Klein declare no competing interests.

Data Availability Statement

Aggregate VCDLI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19548890-data>.

Ethical Approval

This study involved secondary analysis of published variant classification benchmarks. No individual patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

VCDLI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each VCDLI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Pathogenicity accuracy: AUROC + AUPRC on ClinVar P/LP vs B/LB, normalised 0-1; weight 0.278.

VUS resolution rate: fraction of VUS confidently classified matching expert reclassification, normalised 0-1; weight 0.230.

Clinical calibration: predicted probability vs observed pathogenic fraction agreement, normalised 0-1; weight 0.196.

Cross-gene generalisation: accuracy on genes with <10 known pathogenic variants, normalised 0-1; weight 0.172.

Interpretability: ACMG evidence code mapping quality, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. ME-ClinPredV2-22 (Ensemble): VCDLI = 0.932; VUS resolution 0.94, accuracy 0.94.
2. ME-REVELPlus-22 (Ensemble): VCDLI = 0.920; accuracy 0.96, calibration 0.90.
3. PL-AlphaMissense-22 (PLM): VCDLI = 0.906; cross-gene 0.96, accuracy 0.92.
4. SP-SpliceAIv2-22 (Splicing): VCDLI = 0.894; accuracy 0.98, VUS 0.88.
5. ME-BayesDelV2-22 (Ensemble): VCDLI = 0.886; calibration 0.94, interpretab. 0.90.
6. PL-ESM1vScore-22 (PLM): VCDLI = 0.878; cross-gene 0.94, accuracy 0.90.
7. CO-CADDv17-22 (Conservation): VCDLI = 0.868; accuracy 0.92, interpretab. 0.88.
8. SP-MMSpliceV2-22 (Splicing): VCDLI = 0.858; accuracy 0.96, calibration 0.86.
9. CO-PrimateAI2-22 (Conservation): VCDLI = 0.848; accuracy 0.92, cross-gene 0.86.
10. ME-MetaRNN-22 (Ensemble): VCDLI = 0.842; VUS resolution 0.88, accuracy 0.90.