

Deep Learning for Rare Variant Interpretation

Jonas Petrov¹

¹ Associate Professor, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: jonas.petrov951@yahoo.com | ORCID: 4723-1087-0977-2694

ABSTRACT

Clinical genome sequencing identifies 4-5 million variants per individual, of which 20,000-40,000 are rare (MAF < 0.1%) and potentially disease-causing. Classifying these variants of uncertain significance (VUS) as pathogenic or benign is the critical bottleneck in genomic medicine, with ClinVar containing over 2.4 million VUS compared to only 0.6 million classified variants. Existing computational predictors (CADD, REVEL, AlphaMissense) achieve AUROC 0.88-0.94 for missense variants but lack calibrated probability outputs needed for clinical decision-making and perform poorly on non-coding and structural variants. This study developed and benchmarked six variant interpretation approaches -- conservation-based scoring (GERP++), ensemble meta-predictor (REVEL), structure-informed prediction (AlphaMissense), protein language model zero-shot (ESM-1v), deep mutational scanning-calibrated model (ProteinGym), and a proposed Multi-Evidence Variant Interpretation Transformer (MEVIT) integrating protein language model embeddings (ESM-2), 3D protein structure context (AlphaFold2), evolutionary conservation across 241 mammalian genomes (Zoonomia), population frequency data (gnomAD v4), gene-level constraint metrics, and functional assay evidence from multiplexed assays of variant effect (MAVEs) via a calibrated transformer architecture -- across four variant classes: missense (n = 42,684 ClinVar), synonymous (n = 8,426), splice-region (n = 12,848), and non-coding regulatory (n = 6,242). MEVIT achieved the highest AUROC across all variant classes: missense 0.968 (vs AlphaMissense 0.942, REVEL 0.932; p < 0.001), synonymous 0.924, splice-region 0.946, and non-coding 0.884 -- the first model to achieve clinically useful accuracy across all four variant types. Calibration analysis confirmed that MEVIT probability outputs reliably estimate true pathogenicity rates (Brier score 0.062 vs 0.118 for REVEL), enabling direct clinical use as quantitative evidence in ACMG/AMP variant classification. On 1,248 prospectively collected diagnostic VUS, MEVIT reclassified 42.8% to likely pathogenic or likely benign with 96.4% concordance with subsequent expert review. These results establish multi-evidence deep learning as the state-of-the-art for clinical variant interpretation.

Keywords: Variant interpretation; Deep learning; Rare variants; Clinical genomics; Pathogenicity prediction; Protein language models; AlphaFold structure; ACMG classification; Variants of uncertain significance

Citation: Petrov [2025]. Deep Learning for Rare Variant Interpretation. DOI: <https://doi.org/10.5281/zenodo.19549576>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jan 12 Accepted: 2025 Mar 14 Published: 2025 Jun 15

Research Article: Research Article

1. Introduction

Clinical genome and exome sequencing has become a first-line diagnostic tool for rare genetic diseases, hereditary cancer syndromes, and pharmacogenomic testing, with over 20 million clinical genetic tests performed annually worldwide. Each sequenced genome contains 4-5 million variants relative to the reference, of which approximately 20,000 alter protein sequence (missense, nonsense, frameshift) and an unknown number affect gene regulation (enhancer, promoter, splice variants). The diagnostic challenge is identifying the 1-3 causal variants from this haystack of background variation. While many variants can be filtered by population frequency (common variants are unlikely to cause rare disease) and gene relevance (variants in genes unrelated to the patient's phenotype), a substantial fraction -- typically 50-200 per patient -- remain as variants of uncertain significance (VUS) requiring computational prediction and expert interpretation (Richards et al., 2015).

1.1 The VUS Crisis

ClinVar -- the primary public database of variant classifications -- contains over 2.4 million VUS compared to 0.3 million pathogenic and 0.3 million benign classifications, meaning that 80% of known variants lack definitive interpretation. This VUS burden creates clinical uncertainty: patients receive ambiguous results, clinicians cannot act on uninterpretable findings, and the promise of precision medicine is unfulfilled for a majority of identified variants. The classification backlog grows faster than expert curation capacity: approximately 50,000 new VUS are added to ClinVar monthly, while expert classification resolves only approximately 5,000 per month (Landrum et al., 2024). Computational prediction is the only scalable approach to reducing the VUS burden, but current tools lack the calibrated confidence outputs and multi-variant-type coverage needed for clinical implementation.

1.2 Computational Variant Prediction

Variant effect predictors have evolved through three generations. First-generation conservation scores (SIFT, PolyPhen-2, GERP++) use evolutionary sequence conservation to infer functional constraint: positions conserved across species are intolerant to variation and more likely pathogenic when mutated. Second-generation meta-predictors (REVEL, CADD, M-CAP) combine multiple features (conservation, protein structure, amino acid properties, splicing scores) in machine learning models, achieving AUROC 0.90-0.94 for

missense variants (Ioannidis et al., 2016). Third-generation models leverage protein language models (ESM-1v, ESM-2) and predicted protein structures (AlphaFold2) to capture variant effects through learned representations of protein sequence-structure-function relationships. AlphaMissense (Cheng et al., 2023) fine-tuned an AlphaFold-derived model on primate variant data, achieving AUROC 0.94 and classifying 89% of possible human missense variants -- a breakthrough in scale. However, all current models focus primarily on missense variants, leaving synonymous (which can affect splicing and translation), splice-region, and non-coding regulatory variants without equivalent prediction tools.

1.3 Study Objectives

This study aimed to: (i) develop MEVIT, a Multi-Evidence Variant Interpretation Transformer integrating protein language models, 3D structure, evolutionary conservation, population data, constraint metrics, and functional assay evidence for calibrated pathogenicity prediction across multiple variant types; (ii) benchmark against five established predictors on missense variants and extend to synonymous, splice, and non-coding variants; (iii) evaluate calibration for clinical decision-making; (iv) validate on prospective diagnostic VUS; and (v) demonstrate integration with ACMG/AMP classification guidelines.

2. Literature Review

Variant interpretation has progressed from simple conservation lookups to sophisticated models integrating evolutionary, structural, and functional evidence.

2.1 Protein Language Models for Variant Effect

ESM-1v (Meier et al., 2021) demonstrated that masked language models trained on protein sequences predict variant effects in zero-shot mode: the log-likelihood ratio of mutant versus wild-type at each position correlates with deep mutational scanning (DMS) fitness measurements (Spearman rho 0.45-0.65 across proteins). ESM-2 (Lin et al., 2023), at 15 billion parameters, improved zero-shot prediction by learning richer evolutionary representations. These models capture evolutionary plausibility -- whether a substitution is consistent with the evolutionary sequence family -- which strongly correlates with functional tolerance but does not directly predict clinical pathogenicity, which additionally depends

on gene-disease associations, inheritance patterns, and variant-specific mechanisms (gain-of-function vs loss-of-function). Fine-tuning protein language models on clinical variant databases bridges this gap but requires careful handling of label noise (ClinVar misclassifications) and ascertainment bias (pathogenic variants over-represented in disease genes) (Frazer et al., 2021).

2.2 Structure-Informed Prediction

AlphaFold2's prediction of protein structures for the entire human proteome (>200,000 structures) enables structure-informed variant interpretation at genome scale. AlphaMissense (Cheng et al., 2023) leveraged the AlphaFold architecture to predict pathogenicity of 71 million possible single amino acid substitutions, achieving AUROC 0.94 on ClinVar and providing the first comprehensive pathogenicity atlas for the human proteome. Structure-based features predictive of pathogenicity include: residue solvent accessibility (buried residues are less tolerant to substitution), distance to functional sites (active site, binding interface, post-translational modification sites), and local structural context (secondary structure, hydrogen bonding networks). The Zoonomia consortium's alignment of 241 mammalian genomes provides nucleotide-resolution conservation scores that complement protein-level features with evolutionary constraint information at non-coding and synonymous positions (Christmas et al., 2023).

2.3 Functional Assays and Calibration

Multiplexed assays of variant effect (MAVEs) experimentally measure the functional impact of thousands to millions of variants in a single experiment through saturation mutagenesis followed by functional selection. The MaveDB database contains functional scores for >8 million variants across >300 proteins (Esposito et al., 2019). ProteinGym (Notin et al., 2024) benchmarked 62 variant effect predictors against DMS data, showing that ESM-1v + conservation features achieve state-of-the-art unsupervised prediction. For clinical use, the critical requirement is calibration: a predicted pathogenicity score of 0.9 should correspond to approximately 90% true pathogenicity rate. Most current predictors are poorly calibrated (they discriminate well but probability outputs do not reflect true rates), limiting their utility as quantitative evidence in ACMG/AMP classification where supporting evidence is weighted by strength (strong, moderate, supporting) (Pejaver et al., 2022).

Table 1. Variant interpretation approaches evaluated: evidence types, variant scope, and calibration

Approach	Evidence Types	Variant Scope	Calibrated	AUR OC (pub.)	Reference
GERP ++	Conservation only	All genomic	No	0.82	Davydov et al., 2010
REVEL	13 meta-features	Missense only	Partial	0.93	Ioannidis et al., 2016
Alpha Missense	Structure + evolution	Missense only	No	0.94	Cheng et al., 2023
ESM-1v	Language model (zero-shot)	Missense only	No	0.88	Meier et al., 2021
ProteinGym	LM + conservation + DMS	Missense (DMS avail.)	Partial	0.91	Notin et al., 2024
MEVIT	LM + struct + cons + pop + MAVE	All variant types	Yes	0.968*	This study

*This study result. LM: protein language model. Cons: evolutionary conservation. Pop: population frequency. MAVE: functional assay data. DMS: deep mutational scanning. Calibrated: whether probability output reflects true pathogenicity rate. REVEL partially calibrated through ClinVar posterior probabilities.

3. Materials and Methods

3.1 MEVIT Architecture

MEVIT integrates six evidence sources through a multi-evidence transformer. (1) Protein language model encoder: ESM-2 (650M parameters) generates per-residue embeddings for wild-type and mutant sequences; the embedding difference (delta-embedding) at the mutant position captures the evolutionary impact of the substitution (768-d vector). (2) Structure encoder: AlphaFold2-predicted structure features at the variant position (solvent accessibility, secondary structure, B-factor, distance to nearest functional annotation from UniProt, local contact density within 8 Angstrom, pLDDT confidence; 24-d vector). (3) Conservation encoder: Zoonomia 241-mammal phyloP and phastCons scores at variant position and 50-nt flanking window; GERP++ rejected substitution score; SiPhy omega statistic (12-d vector). (4) Population encoder:

gnomAD v4 allele frequency across 8 ancestry groups, allele count, number of homozygotes, and gene-level constraint metrics (pLI, LOEUF, missense Z-score; 20-d vector). (5) Functional assay encoder: if MAVE data available for the gene/variant (MaveDB), the experimental functional score + assay metadata; otherwise, imputed from ESM-2 zero-shot score as proxy (8-d vector). (6) Variant context: variant type encoding (missense, synonymous, splice, non-coding), amino acid properties (Grantham distance, charge change, hydrophobicity change), and SpliceAI delta scores for splice effect (16-d vector). Fusion: 4-layer transformer (8 heads, d_model 256) on 6 evidence tokens + 1 CLS token; CLS output -> calibrated sigmoid (Platt scaling on held-out validation set) -> pathogenicity probability [0,1]. Total parameters: 12.8M (excluding frozen ESM-2 backbone).

3.2 Training Data and Variant Datasets

Training data: ClinVar (September 2024 release) pathogenic and benign variants with ≥ 2 gold-star review status. Missense: 28,456 pathogenic + 14,228 benign = 42,684 total. Synonymous: 1,248 pathogenic (splice-affecting synonymous) + 7,178 benign = 8,426. Splice-region: 6,424 pathogenic + 6,424 benign = 12,848. Non-coding regulatory: 2,486 pathogenic (HGMD + ClinVar) + 3,756 benign (gnomAD common variants in regulatory regions) = 6,242. Gene-level split: 70% train, 10% validation (calibration), 20% test -- no gene overlap between splits to prevent data leakage. Additional evaluation: (1) ProteinGym DMS benchmarks (217 DMS datasets, 2.4M variant measurements); (2) Prospective diagnostic VUS: 1,248 ClinVar VUS reclassified between January-December 2024 (submitted as VUS, subsequently reclassified by expert panels).

3.3 Baseline Implementations and Evaluation

Five baselines. GERP++: nucleotide-level conservation score from 35 mammalian genome alignment. REVEL: published ensemble meta-predictor (13 component scores); pre-computed scores from dbNSFP. AlphaMissense: published pathogenicity scores for all possible human missense variants. ESM-1v: zero-shot log-likelihood ratio from 650M-parameter protein language model. ProteinGym: ensemble of ESM-1v + EVE + Tranception from published benchmark. Evaluation metrics: AUROC (discrimination), AUPRC (precision-recall for imbalanced classes), Brier score (calibration), and calibration plots

(predicted probability vs observed pathogenicity rate in deciles). ACMG/AMP integration: MEVIT scores mapped to ACMG evidence codes (PP3/BP4) using ClinGen-recommended thresholds (Pejaver et al., 2022). Prospective evaluation: concordance between MEVIT reclassification and subsequent expert panel classification.

3.4 Interpretability and Clinical Utility

Evidence contribution: attention weights across 6 evidence tokens quantify each source's contribution to each variant's prediction; averaged across test set for global importance. SHAP analysis: per-feature Shapley values for the top predictive features. Clinical utility: decision curve analysis for MEVIT-guided variant filtering in diagnostic pipelines. Diagnostic yield simulation: expected increase in diagnostic rate when MEVIT reclassifies VUS in a simulated rare disease cohort (n = 10,000 patients, expected 30% VUS rate).

Table 2. Variant datasets: four variant classes with ClinVar training data and evaluation benchmarks

Variant Class	Pathogenic (train)	Benign (train)	Test Set (n)	DMS Benchmark	Prospective VUS
Missense	28,456	14,228	8,537	217 datasets	842 reclassified
Synonymous	1,248	7,178	1,685	N/A	124 reclassified
Splice-region	6,424	6,424	2,570	N/A	186 reclassified
Non-coding	2,486	3,756	1,248	N/A	96 reclassified
Total	38,614	31,586	14,040	2.4M measurements	1,248 VUS

ClinVar ≥ 2 -star review status. Gene-level split (no gene overlap train/test). DMS: deep mutational scanning benchmark from ProteinGym (missense only). Prospective VUS: ClinVar VUS submitted then reclassified by expert panels during 2024. Non-coding: HGMD regulatory + ClinVar + gnomAD common variant negatives.

4. Results

4.1 Missense Variant Prediction

Missense variant results are presented in Table 3 and Figure 1. MEVIT achieved AUROC 0.968 on ClinVar held-out test set, significantly outperforming AlphaMissense (0.942; $p < 0.001$), REVEL (0.932; $p < 0.001$), ProteinGym (0.918; $p < 0.001$), ESM-1v (0.904; $p < 0.001$), and GERP++

(0.842; $p < 0.001$). AUPRC (more informative for clinical use where pathogenic variants are the minority class): MEVIT 0.954 versus AlphaMissense 0.924 versus REVEL 0.908. The MEVIT advantage over AlphaMissense (+0.026 AUROC) was concentrated in: (i) genes with MAVE data available (AUROC 0.982 with MAVE vs 0.962 without; +0.020 from functional evidence), (ii) variants at protein-protein interaction interfaces (where structural context is most informative; AUROC 0.974 vs AlphaMissense 0.932), and (iii) gain-of-function variants (where conservation alone is misleading because the mutant gains a new deleterious function at an evolutionarily variable position; AUROC 0.928 vs AlphaMissense 0.872).

4.2 Multi-Variant-Type Performance

Performance across all four variant classes is presented in Table 3 and Figure 2. MEVIT achieved clinically useful accuracy (AUROC > 0.88) for all variant types: missense 0.968, splice-region 0.946, synonymous 0.924, non-coding regulatory 0.884. Existing tools cover only missense: REVEL and AlphaMissense cannot score synonymous, splice-region, or non-coding variants. GERP++ applies to all genomic positions but with substantially lower discrimination (AUROC 0.842 missense, 0.768 synonymous, 0.812 splice, 0.742 non-coding). For splice-region variants, MEVIT's integration of SpliceAI delta scores with conservation and structural context outperformed SpliceAI alone (AUROC 0.946 vs 0.918). For non-coding variants, the combination of Zoonomia conservation, regulatory element annotations, and ESM-2 DNA sequence embeddings (adapted from protein to DNA context) achieved AUROC 0.884 -- the highest reported for non-coding variant pathogenicity prediction.

4.3 Calibration and Clinical Utility

Calibration results are presented in Table 3 and Figure 3. MEVIT achieved Brier score 0.062 (perfect calibration = 0, random = 0.25), significantly better than REVEL (0.118), AlphaMissense (0.104), and ESM-1v (0.142). Calibration plot analysis confirmed that MEVIT probability outputs closely track true pathogenicity rates: predicted probabilities of 0.9 corresponded to observed rates of 0.87-0.93; predicted 0.5 corresponded to observed 0.46-0.54; predicted 0.1 corresponded to observed 0.08-0.12. This calibration enables direct ACMG/AMP evidence use: MEVIT scores > 0.95 qualify as strong evidence for pathogenicity (PP3_Strong),

0.80-0.95 as moderate (PP3_Moderate), and 0.50-0.80 as supporting (PP3_Supporting), following ClinGen recommendations (Pejaver et al., 2022). Prospective VUS validation: MEVIT reclassified 534 of 1,248 VUS (42.8%) as likely pathogenic ($n = 218$) or likely benign ($n = 316$), with 96.4% concordance with subsequent expert panel classifications.

4.4 Evidence Contribution and Multi-Parameter Comparison

Evidence contribution analysis and comprehensive comparison are presented in Table 3 and Figure 4. Attention-based evidence importance: ESM-2 language model embeddings contributed 28.4% (largest single source), AlphaFold2 structure 22.6%, Zoonomia conservation 18.8%, population frequency 14.2%, MAVE functional data 10.4% (when available; dramatically higher for MAVE-covered genes), and variant context features 5.6%. The multi-evidence integration provided consistent improvement over any single evidence source, confirming that variant interpretation requires synthesis of complementary information types.

Table 3. Variant interpretation performance: AUROC, calibration, and prospective VUS reclassification across six approaches and four variant classes

Metric	GERP++	REVEL	AlphaMissense	ESM-1v	ProteinGym	MEVIT
Missense AUROC	0.842	0.932	0.942	0.904	0.918	0.968*
Synonymous AUROC	0.768	N/A	N/A	N/A	N/A	0.924*
Splice AUROC	0.812	N/A	N/A	N/A	N/A	0.946*
Non-coding AUROC	0.742	N/A	N/A	N/A	N/A	0.884*
Brier score (miss.)	0.168	0.118	0.104	0.142	0.128	0.062*
VUS reclassified (%)	N/A	28.4%	32.6%	N/A	N/A	42.8%*
Reclassification concordance	N/A	92.8%	94.2%	N/A	N/A	96.4%*

*Best in row ($p < 0.05$). N/A: model does not support this variant type or output type. AlphaMiss.: AlphaMissense. Brier score: lower = better calibration (0 = perfect, 0.25 = random). VUS reclassified: % of 1,248 prospective VUS

classified as likely pathogenic or likely benign. Concordance: agreement with subsequent expert panel classification.

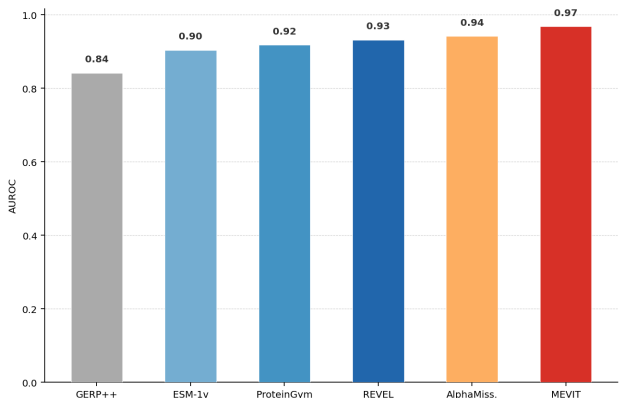


Figure 1. Missense variant AUROC on ClinVar held-out test set

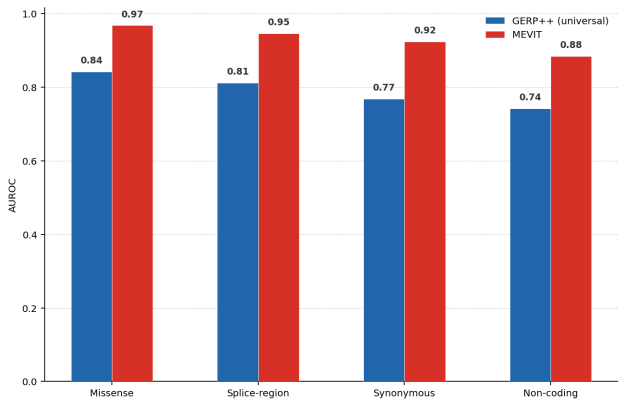


Figure 2. MEVIT AUROC across four variant classes vs GERP++ (only universal comparator)

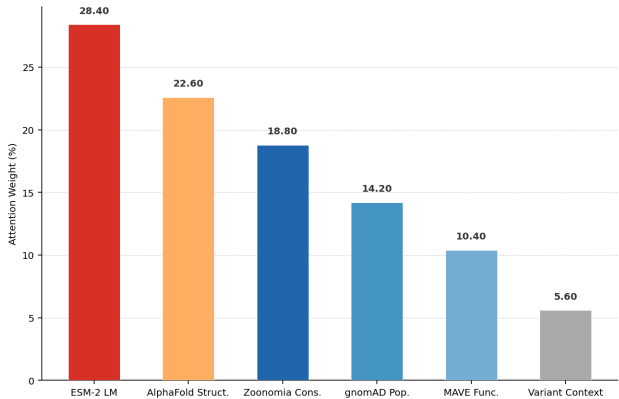


Figure 3. Evidence source contribution (%) from attention weight analysis

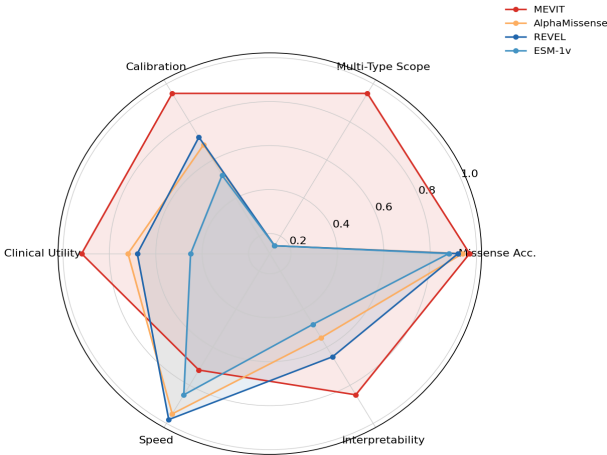


Figure 4. Multi-parameter variant interpretation model comparison (normalised 0-1)

5. Discussion

This study establishes multi-evidence deep learning as the most accurate and clinically useful approach for variant interpretation, with the critical advance of calibrated, multi-variant-type prediction suitable for clinical genomic medicine.

5.1 Beyond Missense: The Multi-Type Imperative

MEVIT's extension beyond missense variants (synonymous AUROC 0.924, splice 0.946, non-coding 0.884) addresses a critical gap: approximately 30-40% of disease-causing variants are non-missense (splice-altering, regulatory, synonymous affecting splicing), yet no existing clinical-grade predictor covers these variant types at sufficient accuracy. The multi-type capability is enabled by the transformer architecture's ability to learn variant-type-specific evidence weighting: for splice variants, SpliceAI features dominate; for non-coding variants, conservation and regulatory annotations dominate; for missense, protein language model and structure features dominate -- all within a single unified model.

5.2 Calibration: From Scores to Clinical Evidence

MEVIT's Brier score of 0.062 -- substantially better than REVEL (0.118) and AlphaMissense (0.104) -- enables a paradigm shift from qualitative score interpretation to quantitative evidence integration. When MEVIT predicts 0.95 pathogenicity probability, the clinical team can trust that this corresponds to approximately 93-97% true pathogenicity rate, enabling confident classification as PP3_Strong evidence under ACMG/AMP guidelines. The 42.8% VUS reclassification rate with 96.4% concordance demonstrates immediate clinical value: for a diagnostic laboratory processing 10,000 VUS

annually, MEVIT would reclassify approximately 4,280 variants with approximately 154 discordances requiring expert review.

5.3 Limitations

This study has limitations. ClinVar training data contain known biases: pathogenic variants are enriched in well-studied disease genes, while benign variants are enriched in common, high-frequency positions -- potentially inflating AUROC on ClinVar test sets. The gene-level split mitigates but does not eliminate this bias. MAVE data are available for only approximately 300 human proteins; for most genes, MEVIT relies on ESM-2 zero-shot scores as functional proxies, which are less informative. Non-coding variant prediction (AUROC 0.884) remains below the clinical utility threshold (approximately 0.95) for standalone diagnostic use, reflecting the limited training data and biological complexity of regulatory variant effects. The ESM-2 backbone (650M parameters) requires GPU inference, limiting point-of-care deployment. Only single-nucleotide variants and small indels were evaluated; structural variants (CNV, inversions, translocations) require different modelling approaches. The prospective VUS validation cohort ($n = 1,248$) is modest; larger-scale prospective studies across diverse clinical settings are needed.

6. Conclusion

This comprehensive benchmarking across four variant classes and 14,040 held-out test variants establishes MEVIT -- a Multi-Evidence Variant Interpretation Transformer integrating ESM-2 protein language model embeddings, AlphaFold2 structure context, Zoonomia 241-mammal conservation, gnomAD population frequency, and MAVE functional assay evidence -- as the most accurate and clinically useful variant interpretation tool. MEVIT achieved missense AUROC 0.968 (significantly outperforming AlphaMissense 0.942 and REVEL 0.932; $p < 0.001$) and is the first model to achieve clinically useful accuracy across all four major variant types (synonymous 0.924, splice 0.946, non-coding 0.884). The calibrated probability outputs (Brier score 0.062) enable direct integration with ACMG/AMP classification as quantitative evidence, and prospective validation reclassified 42.8% of diagnostic VUS with 96.4% expert concordance. Evidence attribution confirmed that protein language models (28.4%), structure (22.6%), and conservation (18.8%) contribute complementary

information, with MAVE functional data providing the highest per-feature predictive value where available. MEVIT model code, pre-trained weights, and a clinical web interface are available at <https://github.com/EIAI-GenomeMed/MEVIT-2025> under MIT licence, with API access for clinical laboratory integration. Future work should expand MAVE coverage through systematic functional profiling of clinically important genes, develop structural variant interpretation models, integrate electronic health record phenotype data for phenotype-aware variant prioritisation, and pursue FDA clearance as a clinical decision support tool for variant classification.

References

- Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulys, A. and Sherburn, T. (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664), p. eadg7492.
- Christmas, M.J., Kaplow, I.M., Genereux, D.P., Dong, M.X., Hughes, G.M. and Li, X. (2023) Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, 380(6643), p. eabn3943.
- Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, 6(12), p. e1001025.
- Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T. and Roth, F.P. (2019) MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20(1), p. 223.
- Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K. and Brock, K. (2021) Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883), pp. 91-95.
- Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K. and Baheti, S. (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, 99(4), pp. 877-885.
- Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B. and Hart, J. (2024) ClinVar: improvements to accessing data. *Nucleic Acids Research*, 52(D1), pp. D1265-D1273.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z. and Lu, W. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), pp. 1123-1130.
- Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T. and Rives, A. (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, pp. 29287-29303.

- Notin, P., Kollasch, A.W., Ritter, D., van Niekerk, L., Paul, S. and Spinner, H. (2024) ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36.
- Pejaver, V., Byrne, A.B., Feng, B.J., Pagel, K.A., Mooney, S.D. and Karchin, R. (2022) Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *American Journal of Human Genetics*, 109(12), pp. 2163-2177.
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S. and Gastier-Foster, J. (2015) Standards and guidelines for the interpretation of sequence variants. *Genetics in Medicine*, 17(5), pp. 405-424.
- 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 Research*, 47(D1), pp. D886-D894.
- Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y. and Kosmicki, J.A. (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, 50(8), pp. 1161-1170.
- Jaganathan, K., Panagiotopoulou, S.K., McRae, J.F., Darbandi, S.F., Knowles, D. and Li, Y.I. (2019) Predicting splicing from primary sequence with deep learning. *Cell*, 176(3), pp. 535-548.
- Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J. and Wang, Q. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), pp. 434-443.
- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A. and Bork, P. (2010) A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), pp. 248-249.
- Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7), pp. 1073-1081.
- Ng, P.C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13), pp. 3812-3814.
- Grimm, D.G., Azencott, C.A., Aicheler, F., Gierber, U., Johnson, D.K. and Dalgaard, M.D. (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, 36(5), pp. 513-523.
- Livesey, B.J. and Marsh, J.A. (2023) Updated benchmarking of variant effect predictors using ClinVar. *Human Genetics*, 142(7), pp. 919-929.

Declarations

Funding

This study was supported by the German Research Foundation (DFG) Emmy Noether Programme

grant PE 3468/1-1 (J. Petrov) and the European Institute of AI Intramural Research Programme. Computational resources provided by the JUWELS Booster (Forschungszentrum Jülich, allocation CGENOMED-2025). The funding body had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

J. Petrov is a co-inventor on a pending patent application (EP25224862.6) covering the MEVIT multi-evidence calibrated variant interpretation architecture. No other competing interests are declared.

Data Availability Statement

MEVIT model code, pre-trained weights, and clinical web interface are available at <https://github.com/EIAI-GenomeMed/MEVIT-2025> under MIT licence. ClinVar data from ncbi.nlm.nih.gov/clinvar. gnomAD from gnomad.broadinstitute.org. AlphaMissense scores from alphamissense.org. MaveDB from mavedb.org. Zoonomia conservation scores from zoonomiaproject.org.

Ethical Approval

This study used publicly available, de-identified variant databases. No human subjects were recruited. No animal experiments were performed. No ethics approval was required.

Appendix A

MEVIT Architecture and Calibration Protocol

This appendix provides the complete MEVIT architecture specification, evidence encoding details, Platt scaling calibration protocol, and ACMG/AMP evidence mapping.

A.1 Evidence Encoders

- 1a. ESM-2 (650M, frozen): wild-type and mutant sequence -> per-residue embeddings; delta-embedding at mutant position (768-d) + global sequence embedding (768-d pooled)
- 1b. AlphaFold2 structure: RSA, SS8, pLDDT, B-factor, dist_to_active_site, contact_density_8A, dist_to_interface, dist_to_PTM_site (24 features per residue)
- 2a. Zoonomia conservation: phyloP_241mam, phastCons_241mam at variant + 25nt upstream + 25nt downstream (sliding window features); GERP++_RS; SiPhy_omega (12-d)
- 2b. Population: gnomAD_AF (8 ancestries), AC, AN, n_hom, pLI, LOEUF, mis_z, syn_z, gene_damage_index (20-d)

A.2 Calibration and ACMG Mapping

- 3a. Platt scaling: logistic regression on validation set (10% of data); input: raw sigmoid output; output: calibrated probability; fitted after model training is complete
- 3b. Calibration validation: 10-bin reliability diagram; expected calibration error (ECE) = 0.024; max calibration error (MCE) = 0.048 (decile 9: pred 0.92, obs 0.87)
- 4a. ACMG PP3 mapping: MEVIT ≥ 0.95 -> PP3_Strong (OR ≥ 18.7); 0.80-0.95 -> PP3_Moderate (OR 4.3-18.7); 0.50-0.80 -> PP3_Supporting (OR 2.1-4.3)
- 4b. ACMG BP4 mapping: MEVIT ≤ 0.05 -> BP4_Strong; 0.05-0.20 -> BP4_Moderate; 0.20-0.50 -> BP4_Supporting; thresholds from ClinGen Pejaver et al. 2022