Additional file 1: Supplementary Information for "From Text to Translation: Using Language Models to Prioritize Variants for Clinical Review"

Supplementary Methods

Data cleaning and processing

With the dataset being parsed from the raw XML file, we then applied several text processing methods to prepare our dataset for model training and downstream evaluation tasks. Since ClinVar submission summary text is a specific domain of text that contains notations (i.e., HGVS nomenclature) and acronyms (i.e., ACMG evidence types) that only appear in this type of clinical text data, we develop text processing methods tailored to this dataset. The following describes our text processing steps in detail.

Training SentenceClassifier

In order to remove variant or submission classification labels and their proxies in submission summaries via a robust approach, we train a sentence classifier for labeling a sentence as `evidence` or `conclusion`. This classifier aims to label sentences as `evidence` or `conclusion`. We define a sentence as `conclusion` if it represents a decision made by the submitter or the testing lab, or institution, asserting a classification result for a specific case upon submission to ClinVar. References to classification results from other sources, such as other testing labs or publications, do not qualify as `conclusion` under this definition.

Rule-based conclusion vs. evidence labeling

To construct a dataset for training the classifier, we initially implemented a rule-based labeling method so that we could efficiently extract sentences labeled as conclusions from the ClinVar dataset. This rule-based methodology provides a more fine-grained approach to process data and is widely used in the field of machine learning and LLMs 31–34. In our case, our rules are various types of sentence patterns that we use for conclusion sentence matching. Analyzing the submission summaries from multiple submitters with various classification labels enables us to identify a unique list of keywords associated predominantly with conclusion sentences.

Using these keywords, we extracted a preliminary set of sentences, labeling them as `conclusion`. The remaining sentences were labeled as `evidence`. To enhance the reliability of this rule-based labeled dataset, we conducted manual reviews and corrections, resulting in a balanced dataset consisting of 2,500 evidence and 2,500 conclusion examples. Furthermore, to ensure data quality, a set of 100 randomly selected examples underwent a rigorous review by a domain expert.

Rule-Based Labeling: Conclusion Phrases and Keywords

- "In summary"
- "In conclusion", "To summarize", "To conclude"
- "Therefore,", "is therefore predicted to be"
- "Taken together", "Taking together", "In brief"
- "this alteration remains unclear", "Considering all the evidence"
- "Based on the available evidence", "Due to insufficient evidences and the lack of functional studies"
- "After careful consideration", "Upon review of the evidence"
- "Based on available information", "Based on the results"
- "it has been classified as", "Based on the available information"
- "based on the above information", "with clinical assertions as classified by the original submitter"
- "Based on insufficient or conflicting evidence", "the clinical significance of this alteration remains unclear"
- "Since supporting evidence", "For these reasons"
- "based on the currently available information", "We consider it to be"
- "As such", "Due to these contrasting evidences and the lack of functional studies"
- "The score for this variant resulted in a classification of", "Taking together", "we classify this variant as"
- "we classify the", "we classify it as", "Considering that this is a"
- "there is insufficient evidence to classify", "we interpret"
- "Considering available [...]"
- "Due to limited information", "Due to the potential impact of"

- "Based on the evidence outlined above", "Variant of Uncertain Significance due to insufficient evidence:"
- "Since supporting an evidence is limited at this time", "the clinical significance of this variant is"
- "Based on the classification scheme", "Given all the evidence"
- "this collective evidence supports the classification of", "leading us to conclude that"

Examples of labeled data for SentenceClassifier

As we train a classifier for identifying sentences including classification labels in text, we first have a labeled dataset that includes balanced numbers of labeled examples for `description`, `evidence`, and `conclusion`.

Example Labeled Data

Description-Labeled Data

- This sequence change replaces leucine with glutamine at codon 55 of the SRD5A2 protein (p.Leu55Gln).
- This sequence change replaces arginine with proline at codon 297 of the SKI3 protein (p.Arg297Pro).
- This sequence change replaces glutamine with proline at codon 332 of the NEF1 protein (p.Gln332Pro).
- This sequence change replaces arginine with histidine at codon 249 of the LDLR protein (p.Arg249His).
- The normal sequence, with the base that is deleted in brackets, is GAGAAGC[T]. This variant results from a T to A substitution at nucleotide position 34.
- The c.58860¿A (p.G19540R) alteration is located in exon 24 (CD5.24) of the HCFC1 gene.

Evidence-Labeled Data

- This variant is present in population databases (rs201097255, gnomAD 0.06%).
- This population frequency is higher than expected for a pathogenic variant in MSH2 causing Lynch syndrome (BS1).
- ClinVar contains an entry for this variant (Variation ID: 17355).
- (I) 0304 - Variant is present in gnomAD (v2) <0.01 for a recessive condition (93 heterozygotes, 0 homozygotes).
- BARD1 His686Arg occurs at a position that is not conserved and is located in the BRCT 2 domain (UniProt).
- The K61E variant was not observed in approximately 6,500 individuals of European and African American ancestry in the NHLBI Exome Sequencing Project, indicating it is not a common benign variant in these populations.

Conclusion-Labeled Data

- The co-occurring 3'-UTR variant is located three base pairs upstream of the polyadenylation signal of PHEX, thus it remains unclear whether it is just a marker for this pathogenic duplication, or can be also detrimental in isolation.
- Therefore, this collective evidence supports the classification of the c.416G>A (p.Ser139Asn) as a recessive Likely Pathogenic variant for Nonsyndromic hearing loss and deafness.
- Thus, the clinical significance of the p.Phe17754Ser variant cannot be determined with certainty.
- Based on the collective evidence, the p.Arg947Pro variant is classified as a variant of uncertain significance for autosomal dominant pseudohypoaldosteronism type 1.
- In summary, the clinical significance of the p.Arg343Gln variant is uncertain.
- Due to these contrasting evidences and the lack of functional studies, the clinical significance of the p.Glu886Ala change remains unknown at this time.

Examples of highly similar submission summaries

With a threshold of **0.95**, the following submission summaries are identified as highly similar by MinHash 15

Example summaries

- This missense variant replaces methionine with isoleucine at codon 141 of the MSH2 protein. Computational prediction tool is inconclusive regarding the impact of this variant on protein structure and function. Internally defined REVEL score threshold: 0.5 < inconclusive < 0.7 (PMID: 27666373)...

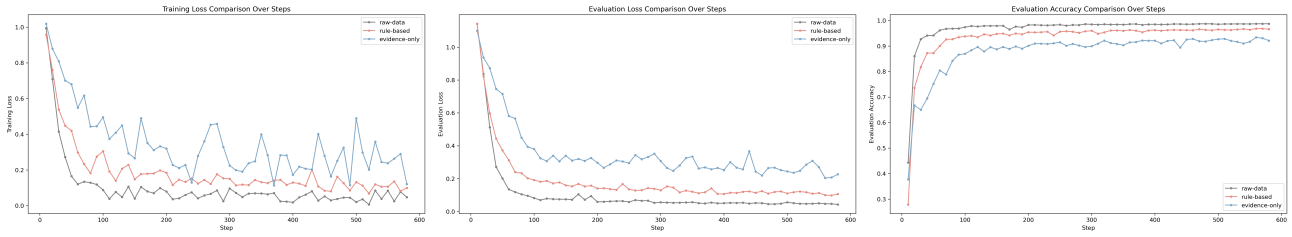
- This missense variant replaces leucine with methionine at codon 9 of the MSH2 protein. Computational prediction is inconclusive regarding the impact of this variant on protein structure and function. Internally defined REVEL score threshold: 0.5 < inconclusive < 0.7 (PMID: 27666373)...
- This missense variant replaces methionine with valine at codon 779 of the MSH2 protein. Computational prediction tool is inconclusive regarding the impact of this variant on protein structure and function. Internally defined REVEL score threshold: 0.5 < inconclusive < 0.7 (PMID: 27666373) ...
- This missense variant replaces lysine with glutamic acid at codon 579 of the MSH2 protein. Computational prediction is inconclusive regarding the impact of this variant on protein structure and function. Internally defined REVEL score threshold: 0.5 < inconclusive < 0.7 (PMID: 27666373)...

As we can see from the example above, with a threshold of 0.95 using MinHash, these submission summaries are identified as highly similar, and we can notice that these summaries are highly template-based and the only difference between each of them is the amino acid and codon information mentioned, and the rest of the evidence being reference to is exactly the same for all of them. These are the summaries that we want to filter out to ensure training text data diversity before sampling training and testing data.

Fine-tuning language models with ClinVar dataset

Model training details We fine-tuned all models using a maximum sequence length (`max_length`) of 512 tokens, which is the maximum token length for BERT models. This configuration was chosen based on the distribution of token lengths in our training data, where the average token count is 155.59, the median is 143.0, and the 90th percentile is 254. We set the learning rate at 2×10^{-5} and applied a weight decay of 0.01 to optimize training.

Training loss, evaluation loss, and evaluation accuracy comparison during training using ClinVar datasets with different text processing methods:



(a) Training Loss (b) Evaluation Loss (c) Evaluation Accuracy

Fig. S1: Comparison of training Loss, evaluation Loss, and accuracy during training among three text processing methods

Model validation with experimental and computational data

Alternative priors for validation with experimental functional data

1. We adopted the observed class distribution estimated from Figure 5 (B) in the AlphaMissense paper 12, where approximately 37% of MAVE prioritized variants are classified as P/LP and 49% as B/LB. Using these proportions, we labeled the top 37% of submissions by normalized prediction scores as P/LP and the bottom 49% as B/LB, with the remaining variants designated as VUS.
2. We sampled 10,000 VUS submissions from ClinVar across the five genes with available DMS functional scores (*BRCA1*, *TP53*, *LDLR*, *HRAS*, and *PTEN*) and applied the AlphaMissense-defined functional thresholds. This yielded a distribution of 61.26% B/LB, 30.57% P/LP, and 8.17% VUS. We then used these proportions as the expected proportions for B/LB, P/LP, and VUS for calibrated model predictions.

Neuron activation pattern analysis

To analyze whether our ClinVar-BERT model focuses on words and phrases indicative of pathogenicity or benignity, we examined its neuron activation pattern. With the model tokenized ClinVar summaries, we used Ecco 27 to visualize and interpret neuron activation patterns. Ecco applies non-negative matrix factorization (NMF), a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional, more interpretable matrix. By specifying the number of components, we visualized the pattern of neuron activation through color-coded tokenizations. These components represent 'concepts' within the ClinVar summaries,

highlighting relationships between words and phrases. This analysis revealed where the model focuses during the prediction task, providing insights into its interpretability.

The Ecco visualization model is configured by specifying the `model_id`, `activations`, and `model_config`. For `model_id`, the path to the trained model was provided, and `activations` were enabled (`True`). The `model_config` was defined with several parameters: `embedding` set to `'embeddings.word_embeddings'`, `type` as `'mlm'`, `activations` as `'intermediate textbackslash.dense'`, `token_prefix` as `', '`, and `partial_token_prefix` as `'##'`.

Submission summary embedding UMAP visualization

To visualize and compare the embeddings from the pre-trained BioBERT 20 model and our fine-tuned ClinVar-BERT model, we analyzed ClinVar submission summary text data, focusing on their submitters and associated submission classification on ClinVar.

We first performed stratified sampling based on the 'submitter' to ensure a balanced representation while maintaining the natural distribution patterns in the data. Specifically, we selected the top 5 submitters by submission volume and sampled a total of 25,000 submissions, with a minimum of 1,000 samples per submitter.

We computed embeddings for each submission summary in our sampled dataset using both BioBERT and ClinVar-BERT models to conduct a comparative analysis of the embedding spaces before and after ClinVar-specific training. We then employed Uniform Manifold Approximation and Projection (UMAP) 28 for dimensionality reduction. The UMAP algorithm was configured with the following parameters: `n_neighbors=50` to balance local and global structure preservation, `min_dist=0.2`, and cosine similarity as the distance metric to effectively capture semantic relationships between texts. The resulting two-dimensional embeddings were visualized as scatter plots, with points colored by three key categorical variables: ClinVar submitter, submission classification made by the submitter, and clinical significance (the interpretation of a variant).

Supplementary Results

Distribution of sentence type across ClinVar dataset

Year	Description	Evidence	Conclusion	Total Sentences	Total Submissions
2018 and earlier	512,802,315 (9.3%)	4,717,817,303 (85.6%)	278,582,558 (5.1%)	5,509,202,176	74,224
2019	42,464,369 (12.1%)	301,547,457 (85.8%)	7,288,223 (2.1%)	351,300,049	18,743
2020	251,239,303 (12.8%)	1,667,787,849 (85.1%)	40,008,969 (2.0%)	1,959,036,121	44,261
2021	146,043,334 (11.8%)	1,085,633,267 (87.4%)	11,167,915 (0.9%)	1,242,844,516	35,254
2022	925,980,086 (13.7%)	5,769,783,363 (85.2%)	78,843,415 (1.2%)	6,774,606,864	82,308
2023	340,706,104 (13.6%)	2,130,711,256 (85.0%)	35,988,116 (1.4%)	2,507,405,476	50,074

Table S1: Distribution of types of sentences across the ClinVar dataset grouped by the year of ClinVar submission creation

Model training and evaluation

We evaluated three text preprocessing strategies by fine-tuning BioBERT-base 20 on: (1) **raw-data** (unprocessed), (2) **rule-based** (template filtering), and (3) **evidence-only** (sentence classification). All models were tested on the same held-out set (20% of raw-data).

All ClinVar-trained models substantially outperformed the pre-trained BioBERT-base (Table S2). However, raw-data and rule-based approaches showed rapid overfitting, converging in < 1000 steps with $> 90\%$ accuracy. Ablation analysis revealed these models learned template-based patterns rather than evidence content. The evidence-only approach converged more slowly (Figure S1c), suggesting more robust learning of clinically relevant patterns. We therefore selected this approach for ClinVar-BERT.

Three text processing methods considered for ClinVar texts:

- **Rule-Based Filtering ('rule-based' dataset)**: This approach uses a rule-based text processing pipeline to remove parts of sentences in the dataset that have pre-defined keywords and phrases that are suggestive of class labels, as specified in Supplementary Methods.
- **Sentence Classification Filtering ('evidence-only' dataset)**: This approach uses a model to classify sentences (SentenceClassifier) as *conclusion*, *description*, or *evidence*, and retains only evidence sentences. We classified and filtered predicted examples of these sentence types, as described in Figure 1b. Figure 1c describes the proportion of each type of sentence in the training data; Figure 1d shows the distribution of three sentence types by year.
- **Original Unfiltered Data ('raw-data' dataset)**: We contrast results with the original raw ClinVar data, which has not been processed.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
BioBERT-base + ClinVar (evidence-only)	0.9720	0.9721	0.9702	0.9711	0.9974
BioBERT-base + ClinVar (rule-based)	0.9805	0.9795	0.9802	0.9798	0.9987
BioBERT-base + ClinVar (raw-data)	0.7207	0.8157	0.7193	0.7264	0.9500
BioBERT-base 20	0.3760	0.1601	0.3133	0.1919	0.5122

Table S2: Performance of fine-tuned BioBERT-base models with ClinVar data using different text processing methods (evidence-only, rule-based filtered, and raw ClinVar data) compared to the pre-trained BioBERT-base model on the test data from the **raw-data** dataset.

Results on the test data for the fine-tuned model using *evidence-only* training data and a pre-trained model.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
BioBERT-large 20 + ClinVar	0.9754	0.9756	0.9730	0.9743	0.9982
ScholarBERT 21 + ClinVar	0.9720	0.9721	0.9702	0.9711	0.9974
ClinicalBERT 22 + ClinVar	0.9666	0.9685	0.9621	0.9651	0.9976
RoBERTa-base 19 + ClinVar	0.9717	0.9734	0.9669	0.9700	0.9979
RoBERTa-large 19+ ClinVar	0.9729	0.9728	0.9705	0.9716	0.9978
BERT-Base 16 + ClinVar	0.9729	0.9724	0.9715	0.9720	0.9979
BERT-Large 16 + ClinVar	0.9754	0.9756	0.9730	0.9743	0.9982
BioBERT-large	0.2000	0.3046	0.3318	0.1136	0.5599
ScholarBERT	0.2309	0.3277	0.3477	0.1789	0.5344
ClinicalBERT	0.4152	0.2801	0.3505	0.2346	0.4686
RoBERTa-base	0.1360	0.0900	0.1537	0.1110	0.4988
RoBERTa-large	0.3030	0.2325	0.3774	0.2552	0.5984
BERT-base	0.4025	0.4503	0.3355	0.1953	0.3768
BERT-large	0.1998	0.2139	0.3132	0.1372	0.4705

Table S3: Performances of fine-tuned BERT models using **evidence-only** training data compared to pre-trained models on ClinVar **raw-data** test data.

DMS validation results

Model	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS					
BioBERT-base + ClinVar (evidence-only)	**0.9272**	**0.8043**	0.5470	**0.7595**	0.4753	0.4930	0.4753	0.4219
BioBERT-base + ClinVar (rule-based)	0.9096	0.7938	0.5377	0.7470	0.4891	0.5098	0.4891	0.4399
BioBERT-base + ClinVar (raw-data)	0.9037	0.7882	0.5826	0.7582	0.4840	0.5306	0.4840	0.4192
BioBERT-base 20	0.3953	0.5428	0.3953	0.4503	0.2713	0.0736	0.2713	0.1158

Table S4: Evaluation results of BioBERT-base models trained with three different text processing and sampling methods (evidence-only, rule-based filtered, and raw ClinVar data), compared to the pre-trained BioBERT-base baseline, evaluated on orthogonally generated DMS data. Ground truth labels are created using functional scores, with the top 27.1% as P/LP, the bottom 27.5% as B/LB, and the rest being VUS.

Functional Score Threshold Strategy	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS					
Functional scores from all genes	0.8816	0.8175	0.5494	0.7495	0.5464	0.5807	0.5464	0.50013
BRCA1 VUS distribution	0.9272	0.8043	0.5470	0.7595	0.4753	0.4930	0.4753	0.4219

Table S5: Evaluation results of **evidence-only** ClinVar-BERT model, using different thresholds for generating ground truth labels for functional scores. We label each variant in the validation set using threshold values based on their functional scores: First row: Variants in any gene that have a functional score of $\geq \mu + \sigma$ are labeled as P/LP, functional scores that are ≤ 0 are labeled as B/LB, and the remaining variants are labeled VUS; Second row: We measured the proportions of ClinVar VUS at time of publication for a large saturation genome editing screen in *BRCA1*, identifying how many are functional, indeterminate, or loss-of-function using published functional screening threshold values. This results in a variant labeling strategy that assigns functional scores that are within the top 27.1% as P/LP, the bottom 27.5% as B/LB, and the remaining are labeled as VUS.

Functional Score Threshold Strategy	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS					
Functional scores from all genes	0.8816	0.8175	0.5494	0.7495	0.5464	0.5807	0.5464	0.50013
BRCA1 VUS distribution	0.9272	0.8043	0.5470	0.7595	0.4753	0.4930	0.4753	0.4219

Table S6: Evaluation results of **evidence-only** ClinVar-BERT model, using different thresholds for generating ground truth labels for functional scores. We label each variant in the validation set using threshold values based on their functional scores: First row: Variants in any gene that have a functional score of $\geq \mu + \sigma$ are labeled as P/LP, functional scores that are ≤ 0 are labeled as B/LB, and the remaining variants are labeled VUS; Second row: We measured the proportions of ClinVar VUS at time of publication for a large saturation genome editing screen in *BRCA1*, identifying how many are functional, indeterminate, or loss-of-function using published functional screening threshold values. This results in a variant labeling strategy that assigns functional scores that are within the top 27.1% as P/LP, the bottom 27.5% as B/LB, and the remaining are labeled as VUS.

Submitting Laboratory	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score	Sample Size
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS						
Labcorp Genetics (formerly Invitae), Labcorp	0.9411	0.8508	0.4623	0.7513	0.5406	0.3919	0.5406	0.4531	2514
Ambry Genetics	0.9330	0.7936	0.5935	0.7734	0.4810	0.5365	0.4810	0.4133	1661
GeneDx	0.9284	0.7965	0.6869	0.8039	0.4115	0.4780	0.4115	0.3671	746
Color Diagnostics, LLC DBA Color Health	0.9414	0.8165	0.6274	0.7951	0.4235	0.6449	0.4235	0.3209	451
Women's Health and Genetics / Laboratory Corporation of America, LabCorp	0.9408	0.8295	0.6831	0.9408	0.4925	0.4886	0.4925	0.4589	335

Table S7: Performance of **evidence-only** ClinVar-BERT evaluated using functional score–based ground truth labels derived from *BRCA1*. Variants are labeled as P/LP if their functional scores fall within the top 27.1%, B/LB if within the bottom 27.5%, and VUS otherwise. The table reports the model's classification performance across submissions from different laboratories.

Year Range	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score	Sample Size
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS						
2022 - 2024	0.9293	0.8182	0.5558	0.7678	0.4552	0.5278	0.4552	0.3877	2478
2019 - 2021	0.9254	0.8121	0.5949	0.7775	0.5260	0.5216	0.5260	0.4769	2152
Before 2019	0.8964	0.7807	0.5778	0.7516	0.5037	0.4869	0.5037	0.4585	1336

Table S8: Performance of **evidence-only** ClinVar-BERT evaluated using functional score–based ground truth labels derived from *BRCA1*. Variants are labeled as P/LP if their functional scores fall within the top 27.1%, B/LB if within the bottom 27.5%, and VUS otherwise. The table reports the model's classification performance across submissions from a range of years.

Variant Origin	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score	Sample Size
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS						
Germline	0.9282	0.8295	0.5592	0.7723	0.5122	0.5112	0.5122	0.4519	5107
Germline/Somatic	0.7954	0.6935	0.5786	0.6891	0.3690	0.6176	0.3690	0.3436	767
Not Provided	0.7437	0.5517	0.5053	0.6002	0.3415	0.2454	0.3415	0.2242	82

Table S9: Performance of **evidence-only** ClinVar-BERT evaluated using functional score–based ground truth labels derived from *BRCA1*. Variants are labeled as P/LP if their functional scores fall within the top 27.1%, B/LB if within the bottom 27.5%, and VUS otherwise. The table reports the model's classification performance across submissions from different variant origins using submissions last updated by 2023.

Variant Origin for PTEN Gene	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score	Sample Size
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS						
Germline	0.9230	0.7247	0.5331	0.7269	0.4561	0.5927	0.4561	0.3720	1151
Germline/Somatic	0.8031	0.6212	0.4800	0.6348	0.3214	0.2139	0.3214	0.2152	84

Table S10: Performance of **evidence-only** ClinVar-BERT evaluated using functional score–based ground truth labels derived from *BRCA1*. Variants are labeled as P/LP if their functional scores fall within the top 27.1%, B/LB if within the bottom 27.5%, and VUS otherwise. The table reports the model's classification performance across submissions from the different variant origins for the PTEN gene using submissions last updated by 2023.

Model	Pair-wise AUC			Avg AUC-ROC	Accuracy	Precision	Recall	F1 Score
	P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS					
BioBERT-large + ClinVar	0.9272	0.8043	0.5470	0.7595	0.4753	0.4930	0.4753	0.4219
ScholarBERT + ClinVar	0.9084	0.8014	0.5641	0.7579	0.4702	0.4840	0.4702	0.4171
ClinicalBERT + ClinVar	0.9057	0.7933	0.4927	0.7306	0.4792	0.4900	0.4782	0.4266
BERT-base + ClinVar	0.9267	0.8135	0.5650	0.7684	0.4829	0.4955	0.4829	0.4330
BERT-large + ClinVar	0.9150	0.8109	0.5085	0.7448	0.4805	0.4967	0.4805	0.4267
RoBERTa-base + ClinVar	0.9217	0.8091	0.5874	0.7728	0.4648	0.4863	0.4648	0.4103
RoBERTa-large + ClinVar	0.8941	0.7810	0.4458	0.7070	0.4754	0.4879	0.4754	0.4219

Table S11: Evaluation results of fine-tuned models trained with the **evidence-only** dataset on orthogonally generated DMS data. Variants are labeled as P/LP if their functional scores fall within the top 27.1%, B/LB if within the bottom 27.5%, and VUS otherwise.

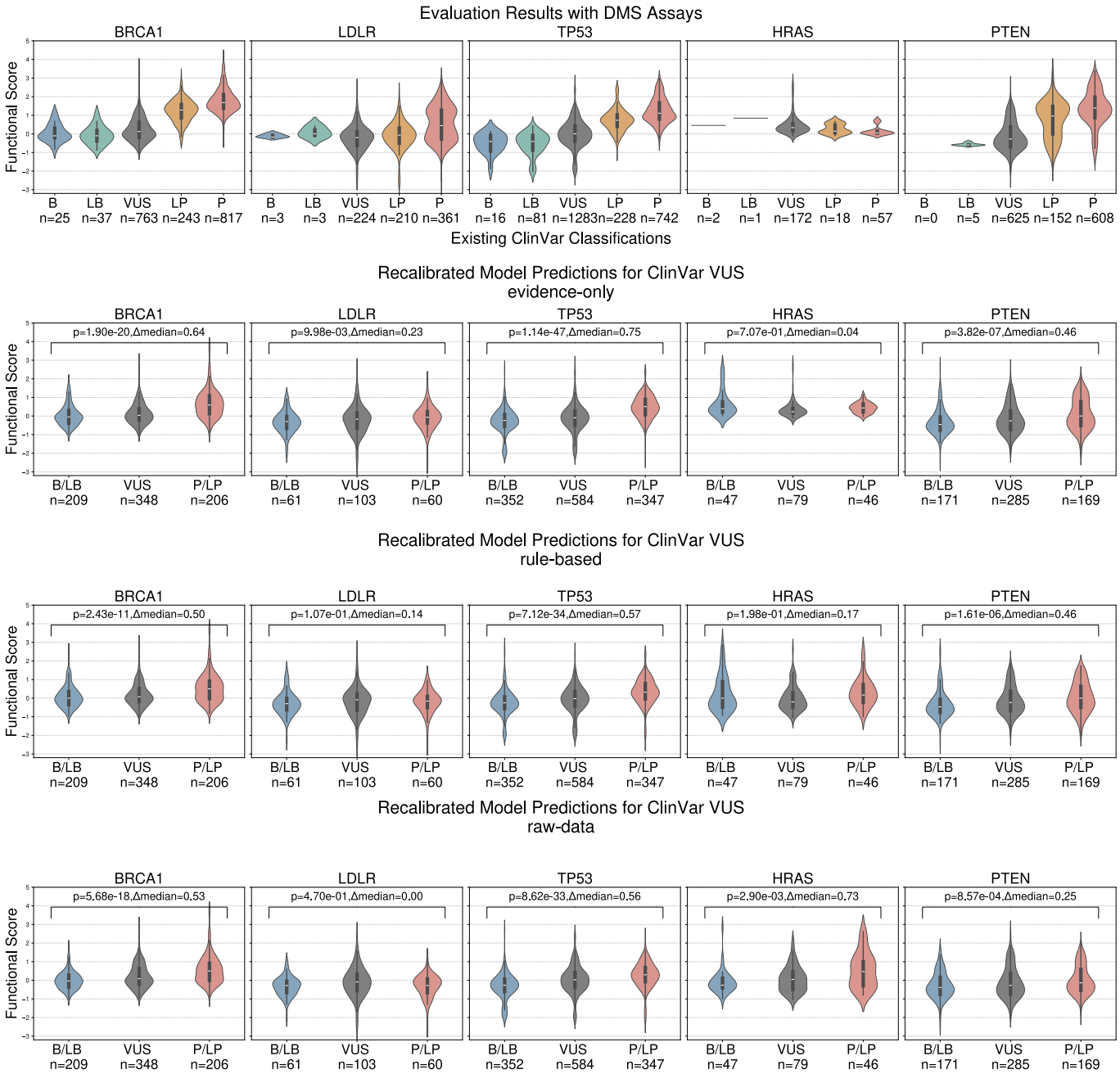


Fig. S1: Top row: Existing classification on ClinVar (B, LB, VUS, LP, and P) on the x-axis and functional scores on the y-axis. The following rows are recalibrated model prediction on **VUS**-labeled submission on the x-axis (B/LB, VUS, and P/LP), and functional scores on the y-axis.

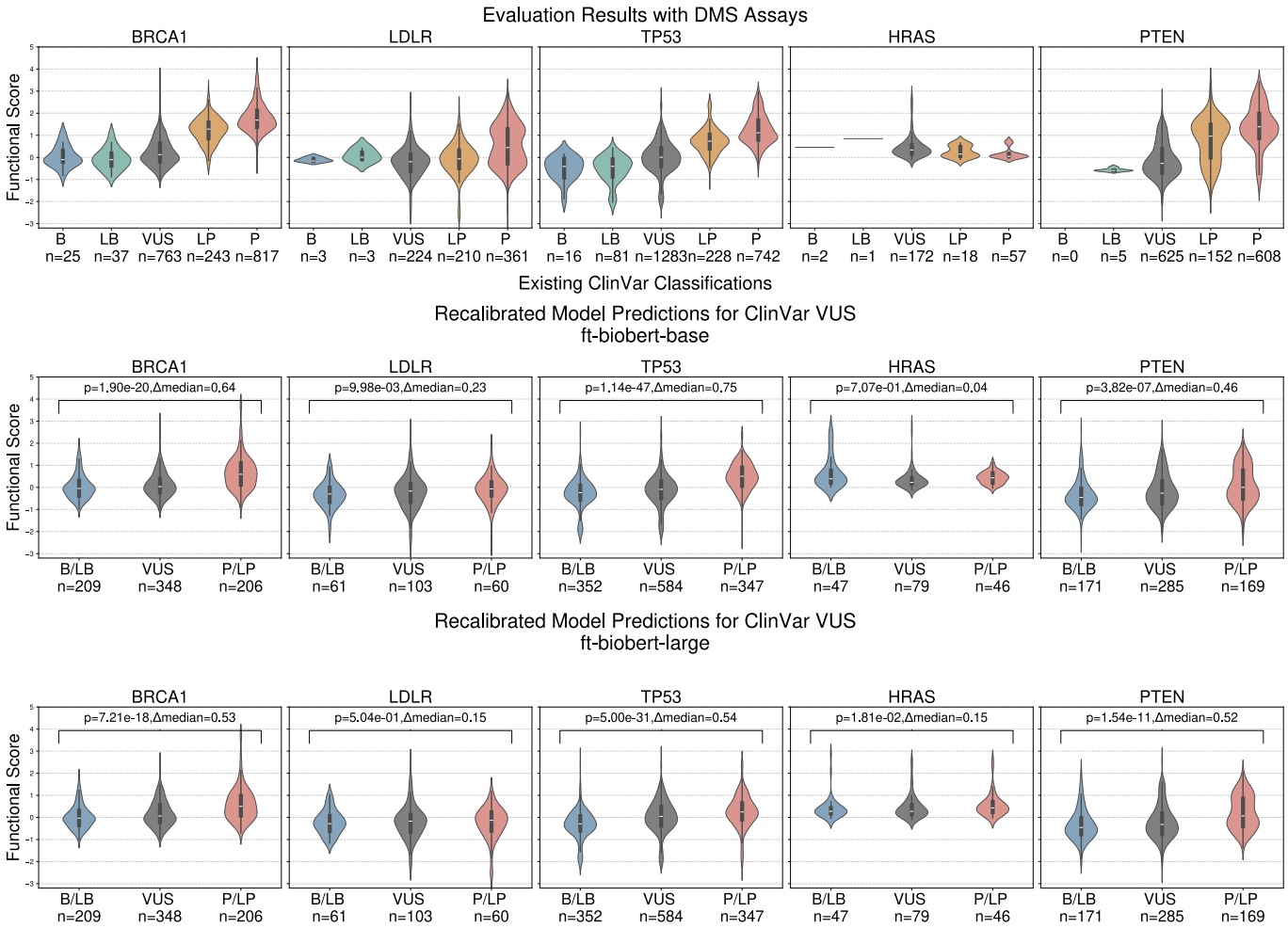


Fig. S2: Comparison of DMS validation results of **evidence-only** trained BioBERT-base and BioBERT-large models on orthogonally generated DMS data

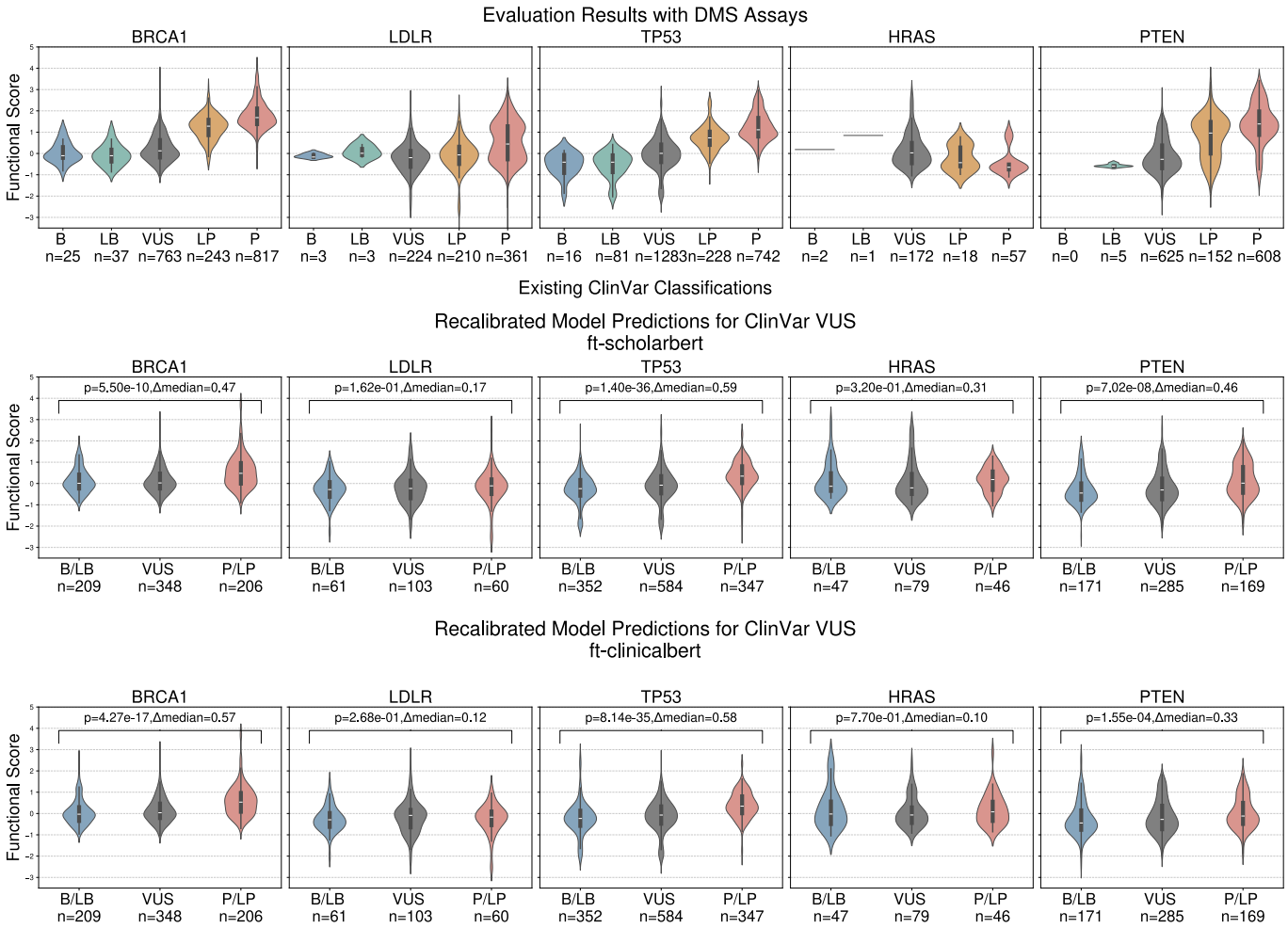


Fig. S3: A comparison of two domain-specific BERT-architecture models (ScholarBERT and ClinicalBERT) fine-tuned with **evidence-only** training data performances on DMS validation.

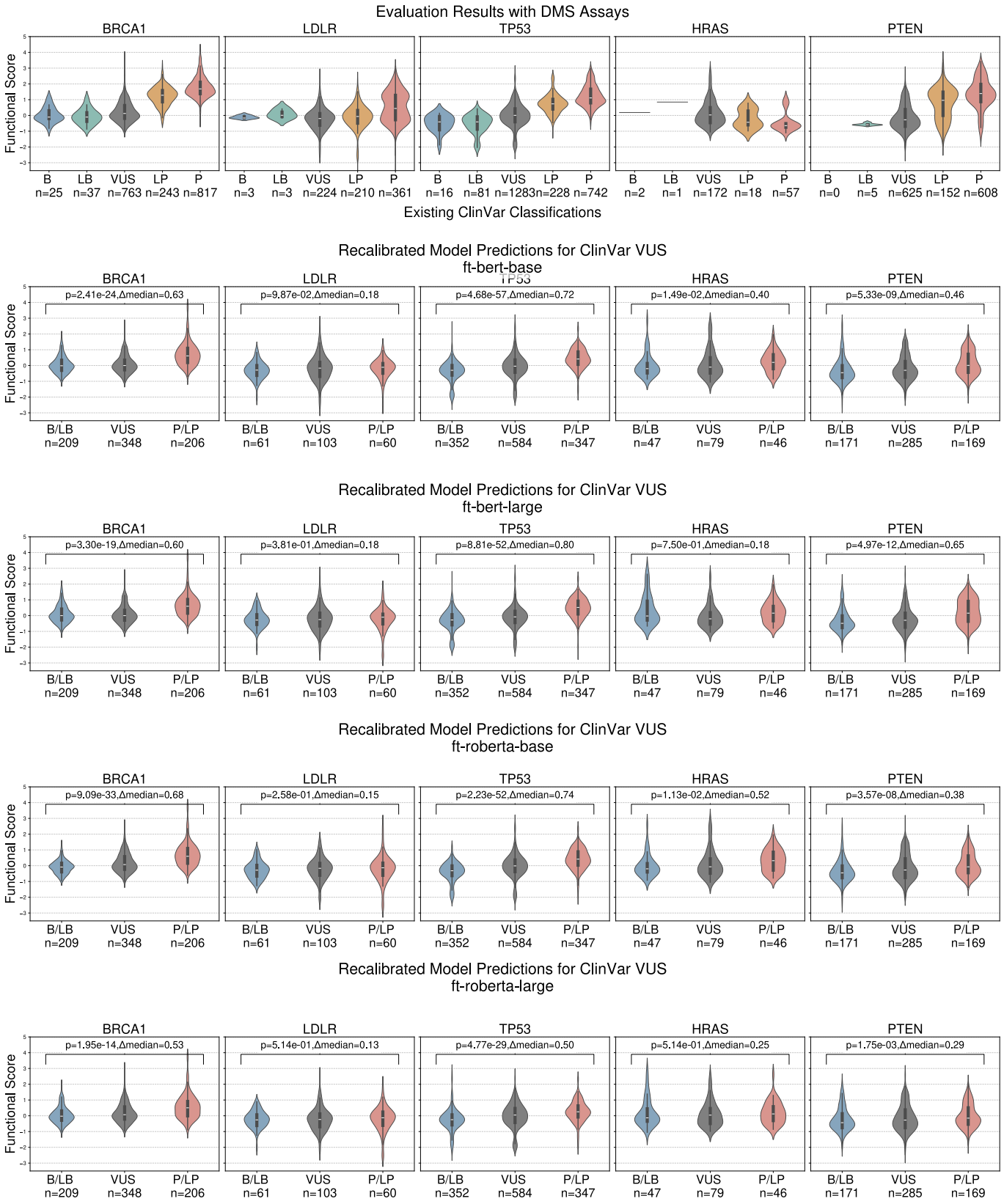


Fig. S4: A comparison of multiple general-domain BERT-architecture models with different model sizes (BERT-base, BERT-large, RoBERTa-based, and RoBERTa-large) fine-tuned with **evidence-only** training data performances on DMS validation.

DMS validation results with different thresholds

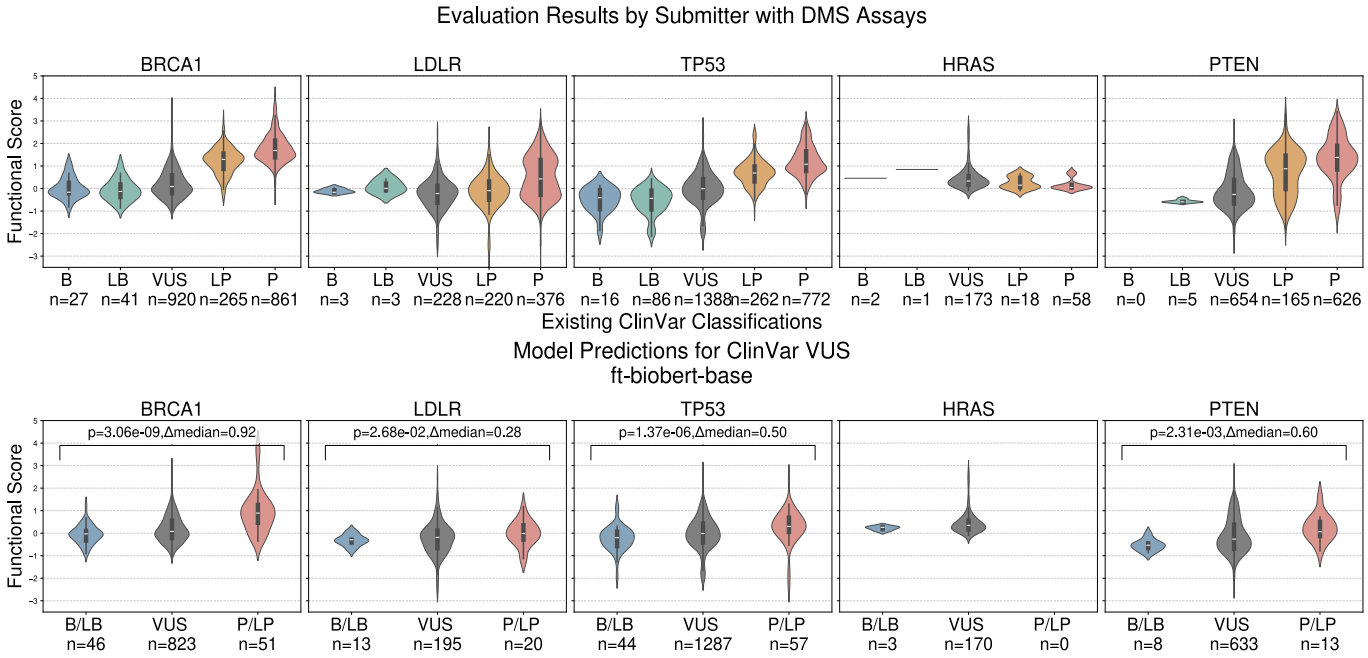


Fig. S5: Top row: Existing classification on ClinVar (B, LB, VUS, LP, and P) on the x-axis and functional scores on the y-axis. Bottom row: directly plot model prediction of **VUS**-labeled submissions on the x-axis (B/LB, VUS, and P/LP) **without recalibration**, and functional scores on the y-axis

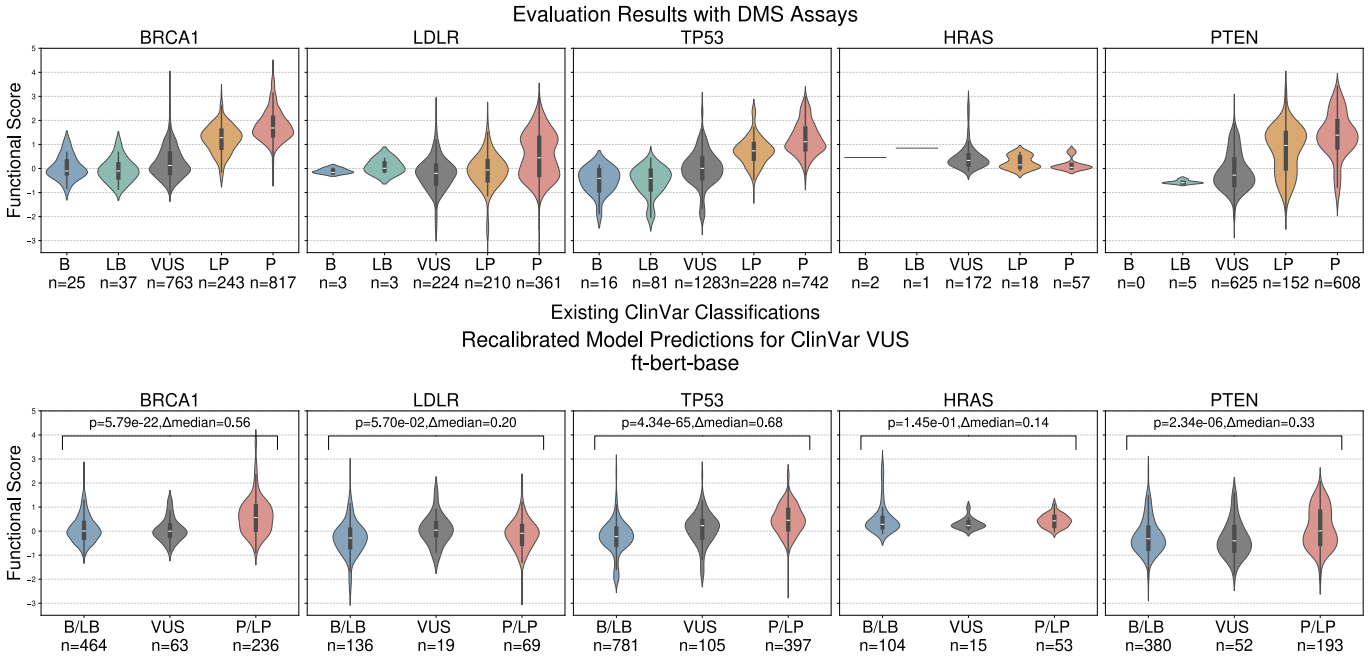


Fig. S6: Top row: Existing classification on ClinVar (B, LB, VUS, LP, and P) on the x-axis and functional scores on the y-axis. Bottom row: recalibrated model prediction using 30.95% percentile for P/LP prediction and 60.91% percentile for B/LB prediction as the threshold (based on the distribution of VUS observed in 10,000 samples using the variant interpretation thresholds defined in AlphaMissense 12) on **VUS**-labeled submission on the x-axis (B/LB, VUS, and P/LP), and functional scores on the y-axis.

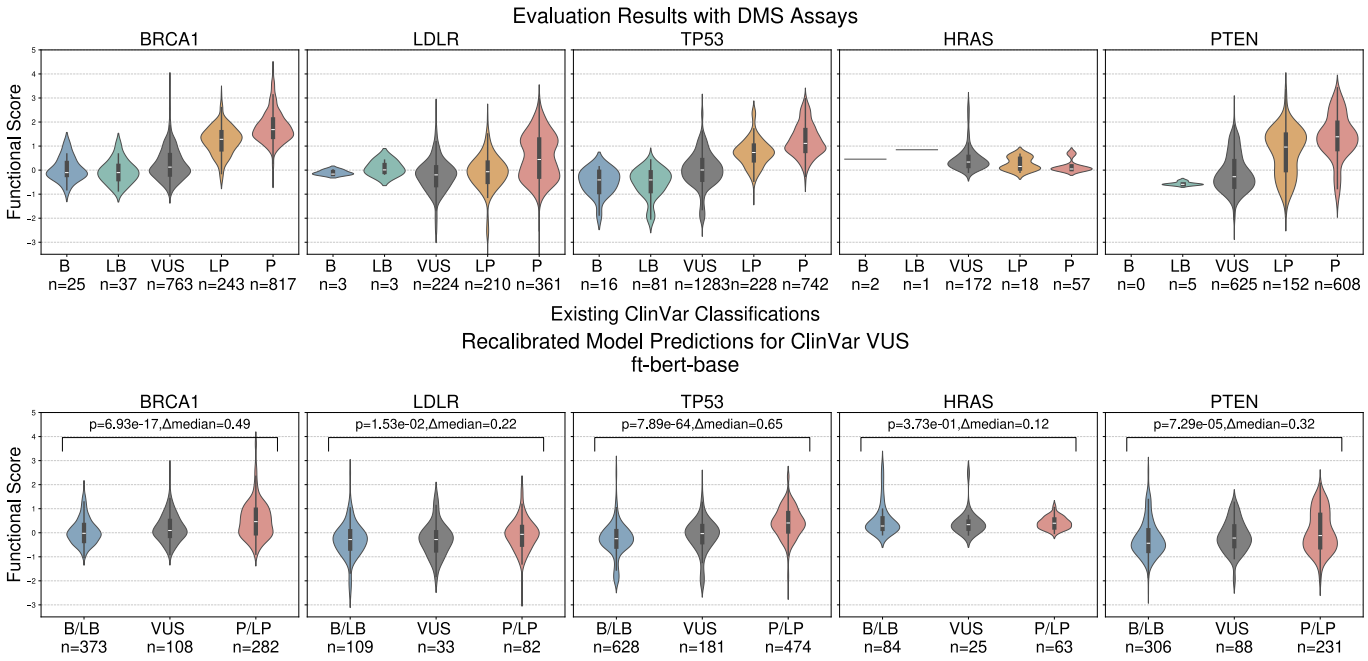


Fig. S7: Top row: Existing classification on ClinVar (B, LB, VUS, LP, and P) on the x-axis and functional scores on the y-axis. Bottom row: recalibrated model prediction using 37% for P/LP prediction and 49% for B/LB prediction as the threshold (based on the distribution estimated from Figure 5 (B) in AlphaMissense 12) on **VUS**-labeled submission on the x-axis (B/LB, VUS, and P/LP), and functional scores on the y-axis.

Table S12: VUS Reclassification Analysis by Expert Panel

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
InSiGHT Hereditary Colorectal Cancer/Polyposis Variant Curation Expert Panel	Strictly	17270	73	42	17155
	2+	17270	308	405	16557
	3+	17270	86	70	17114
Hereditary Breast, Ovarian and Pancreatic Cancer Variant Curation Expert Panel	Strictly	10173	40	58	10075
	2+	10173	168	347	9658
	3+	10173	52	91	10030
Hearing Loss Variant Curation Expert Panel	Strictly	7072	214	163	6695
	2+	7072	348	553	6171
	3+	7072	227	202	6643
Congenital Myopathies Variant Curation Expert Panel	Strictly	6018	162	55	5801
	2+	6018	353	185	5480
	3+	6018	191	79	5748
RASopathy Variant Curation Expert Panel	Strictly	5660	167	41	5452
	2+	5660	288	189	5183
	3+	5660	176	68	5416
Cardiomyopathy Variant Curation Expert Panel	Strictly	5153	131	61	4961
	2+	5153	378	274	4501
	3+	5153	178	95	4880
ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel	Strictly	5098	16	79	5003
	2+	5098	53	483	4562
	3+	5098	19	103	4976
Neurofibromatosis and Schwannomatosis Variant Curation Expert Panel	Strictly	5072	74	10	4988
	2+	5072	139	47	4886
	3+	5072	75	11	4986
Epilepsy Sodium Channel Variant Curation Expert Panel	Strictly	3976	346	32	3598
	2+	3976	561	95	3320
	3+	3976	358	33	3585
Malignant Hyperthermia Susceptibility Variant Curation Expert Panel	Strictly	3358	106	26	3226
	2+	3358	266	116	2976
	3+	3358	135	46	3177
DICER1 and miRNA-Processing Gene Variant Curation Expert Panel	Strictly	3171	5	18	3148
	2+	3171	13	114	3044
	3+	3171	6	26	3139
Parkinson's Disease Variant Curation Expert Panel	Strictly	2289	34	20	2235
	2+	2289	41	45	2203
	3+	2289	34	20	2235
Congenital Heart Disease Variant Curation Expert Panel	Strictly	2072	42	9	2021
	2+	2072	87	73	1912
	3+	2072	48	19	2005

Table S12 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
Limb Girdle Muscular Dystrophy Variant Curation Expert Panel	Strictly	2036	62	16	1958
	2+	2036	104	58	1874
	3+	2036	66	20	1950
GRIN Disorders Variant Curation Expert Panel	Strictly	1948	74	43	1831
	2+	1948	97	74	1777
	3+	1948	74	43	1831
Affiliation	Criteria	Total VUS	Pred. Path.	Pred. Ben.	Pred. VUS
Kidney Cystic and Ciliopathy Disorders Variant Curation Panel	Strictly	1914	211	168	1535
	2+	1914	233	273	1408
	3+	1914	214	182	1518
FBN1 Variant Curation Expert Panel	Strictly	1876	68	8	1800
	2+	1876	118	46	1712
	3+	1876	70	15	1791
Gastric Cancer Variant Curation Expert Panel	Strictly	1828	20	3	1805
	2+	1828	75	64	1689
	3+	1828	26	15	1787
Severe Combined Immunodeficiency Disease Variant Curation Expert Panel	Strictly	1679	30	17	1632
	2+	1679	61	71	1547
	3+	1679	31	26	1622
X-linked Inherited Retinal Disease Variant Curation Expert Panel	Strictly	1550	42	21	1487
	2+	1550	59	34	1457
	3+	1550	42	21	1487
Mitochondrial Diseases Variant Curation Expert Panel	Strictly	1483	52	19	1412
	2+	1483	109	110	1264
	3+	1483	55	26	1402
Brain Malformations Variant Curation Expert Panel	Strictly	1455	63	24	1368
	2+	1455	80	38	1337
	3+	1455	63	25	1367
Rett and Angelman-like Disorders Variant Curation Expert Panel	Strictly	1453	64	44	1345
	2+	1453	99	98	1256
	3+	1453	67	53	1333
Lysosomal Diseases Variant Curation Expert Panel	Strictly	984	29	11	944
	2+	984	65	37	882
	3+	984	38	14	932
Myeloid Malignancy Variant Curation Expert Panel	Strictly	948	11	16	921
	2+	948	31	351	566
	3+	948	12	25	911
Monogenic Diabetes Variant Curation Expert Panel	Strictly	915	81	47	787
	2+	915	116	126	673
	3+	915	84	65	766
Motile Ciliopathy Variant Curation Expert Panel	Strictly	908	13	10	885
	2+	908	20	18	870
	3+	908	15	11	882

Table S12 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
Leber Congenital Amaurosis/early onset Retinal Dystrophy Variant Curation Expert Panel	Strictly	871	30	7	834
	2+	871	44	34	793
	3+	871	31	16	824
ABCA4 Variant Curation Expert Panel	Strictly	757	29	9	719
	2+	757	43	20	694
	3+	757	29	9	719
Optic Nerve Atrophy Variant Curation Expert Panel	Strictly	703	17	10	676
	2+	703	39	36	628
	3+	703	18	11	674
Charcot-Marie-Tooth Disease Variant Curation Expert Panel	Strictly	679	36	7	636
	2+	679	53	21	605
	3+	679	36	9	634
Platelet Disorders Variant Curation Expert Panel	Strictly	660	46	33	581
	2+	660	65	84	511
	3+	660	49	39	572
Familial Hypercholesterolemia Variant Curation Expert Panel	Strictly	660	67	19	574
	2+	660	95	72	493
	3+	660	76	37	547
Cerebral Creatine Deficiency Syndromes Variant Curation Expert Panel	Strictly	657	24	12	621
	2+	657	46	50	561
	3+	657	28	20	609
PTEN Variant Curation Expert Panel	Strictly	647	17	0	630
	2+	647	41	16	590
	3+	647	21	7	619
Antibody Deficiencies Variant Curation Expert Panel	Strictly	622	10	8	604
	2+	622	19	48	555
	3+	622	10	11	601
TP53 Variant Curation Expert Panel	Strictly	570	19	4	547
	2+	570	51	17	502
	3+	570	20	7	543
Endocrine Tumor Predisposition Variant Curation Expert Panel	Strictly	568	15	0	553
	2+	568	27	12	529
	3+	568	16	6	546
Potassium Channel Arrhythmia Variant Curation Expert Panel	Strictly	538	14	6	518
	2+	538	20	61	457
	3+	538	15	15	508
von Willebrand Disease Variant Curation Expert Panel	Strictly	517	47	52	418
	2+	517	58	90	369
	3+	517	49	58	410
Urea Cycle Disorders Variant Curation Expert Panel	Strictly	516	26	3	487
	2+	516	43	9	464
	3+	516	28	3	485
ACADVL Variant Curation Expert Panel	Strictly	433	21	48	364

Table S12 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
	2+	433	34	97	302
	3+	433	24	50	359
VHL Variant Curation Expert Panel	Strictly	432	14	0	418
	2+	432	38	6	388
	3+	432	20	1	411
Hereditary Hemorrhagic Telangiectasia Variant Curation Expert Panel	Strictly	412	23	6	383
	2+	412	36	13	363
	3+	412	26	6	380
Coagulation Factor Deficiency Variant Curation Expert Panel	Strictly	390	93	23	274
	2+	390	107	30	253
	3+	390	93	23	274
Thrombosis Variant Curation Expert Panel	Strictly	354	48	12	294
	2+	354	62	21	271
	3+	354	48	12	294
Pulmonary Hypertension Variant Curation Expert Panel	Strictly	340	4	6	330
	2+	340	9	12	319
	3+	340	4	7	329
Peroxisomal Disorders Variant Curation Expert Panel	Strictly	302	21	6	275
	2+	302	28	18	256
	3+	302	21	7	274
Glaucoma Variant Curation Expert Panel	Strictly	275	7	60	208
	2+	275	14	85	176
	3+	275	7	62	206
Phenylketonuria Variant Curation Expert Panel	Strictly	186	24	7	155
	2+	186	46	14	126
	3+	186	33	9	144
Dopa Decarboxylase (Aromatic L-Amino Acid Decarboxylase) Variant Curation Expert Panel	Strictly	177	3	0	174
	2+	177	5	2	170
	3+	177	4	1	172
Leukodystrophy and Leukoencephalopathy Variant Curation Expert Panel	Strictly	171	14	5	152
	2+	171	20	10	141
	3+	171	14	5	152
Rhodopsin Variant Curation Expert Panel	Strictly	170	20	2	148
	2+	170	21	6	143
	3+	170	20	4	146
Hemoglobinopathy Variant Curation Expert Panel	Strictly	168	31	23	114
	2+	168	61	32	75
	3+	168	38	23	107
Hereditary Angioedema Variant Curation Expert Panel	Strictly	156	5	4	147
	2+	156	8	5	143
	3+	156	5	4	147
Glucose-6-phosphate dehydrogenase Variant Curation Expert Panel	Strictly	113	11	6	96

Table S12 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
	2+	113	12	13	88
	3+	113	11	6	96
Ocular Anterior Segment Disorder Variant Curation Expert Panel	Strictly	109	9	4	96
	2+	109	11	5	93
	3+	109	9	4	96
Galactosemia Variant Curation Expert Panel	Strictly	92	15	5	72
	2+	92	22	8	62
	3+	92	16	5	71
Amyotrophic Lateral Sclerosis Spectrum Disorders Variant Curation Expert Panel	Strictly	44	6	0	38
	2+	44	7	1	36
	3+	44	6	0	38

Criteria	Total VUS Variants	Path. Recl.	Ben. Recl.	Total Recl.	Uncertain
Strict consistency	153,872	4,580	1,941	6,521	147,351
1+ agreement	153,872	7,992	5,974	13,966	139,906
2+ agreement	153,872	7,992	5,974	13,966	139,906
3+ agreement	153,872	5,051	2,364	7,415	146,457

Table S13: VUS reclassification results using calibrated thresholds under different agreement criteria. Results include labeled pathogenic (regardless of evidence strength), benign (regardless of evidence strength), and uncertain, using the calibrated thresholds

Table S14: VUS Reclassification Analysis by Expert Panel

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Total Recl.	Uncertain
InSiGHT Hereditary Colorectal Cancer/Polyposis Variant Curation Expert Panel	Strictly	17270	78	36	114	17156
	2+	17270	330	259	589	16681
	3+	17270	93	47	140	17130
Hereditary Breast, Ovarian and Pancreatic Cancer Variant Curation Expert Panel	Strictly	10173	41	56	97	10076
	2+	10173	180	292	472	9701
	3+	10173	53	79	132	10041
Hearing Loss Variant Curation Expert Panel	Strictly	7072	221	139	360	6712
	2+	7072	361	475	836	6236
	3+	7072	234	162	396	6676
Congenital Myopathies Variant Curation Expert Panel	Strictly	6018	166	42	208	5810
	2+	6018	374	141	515	5503
	3+	6018	199	53	252	5766
RASopathy Variant Curation Expert Panel	Strictly	5660	196	36	232	5428
	2+	5660	348	161	509	5151
	3+	5660	207	58	265	5395
Cardiomyopathy Variant Curation Expert Panel	Strictly	5153	134	54	188	4965
	2+	5153	392	235	627	4526

Table S14 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Total Recl.	Uncertain
	3+	5153	186	78	264	4889
ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel	Strictly	5098	20	74	94	5004
	2+	5098	65	456	521	4577
	3+	5098	23	94	117	4981
Neurofibromatosis and Schwannomatosis Variant Curation Expert Panel	Strictly	5072	75	10	85	4987
	2+	5072	143	44	187	4885
	3+	5072	76	11	87	4985
Epilepsy Sodium Channel Variant Curation Expert Panel	Strictly	3976	356	26	382	3594
	2+	3976	580	80	660	3316
	3+	3976	369	27	396	3580
Malignant Hyperthermia Susceptibility Variant Curation Expert Panel	Strictly	3442	149	7	156	3286
	2+	3442	201	25	226	3216
	3+	3442	155	7	162	3280
TP53 Variant Curation Expert Panel	Strictly	2568	49	15	64	2504
	2+	2568	156	71	227	2341
	3+	2568	68	19	87	2481
ClinGen RASopathy Expert Panel	Strictly	2378	81	12	93	2285
	2+	2378	180	46	226	2152
	3+	2378	94	13	107	2271
Platelet Disorder Variant Curation Expert Panel	Strictly	2203	17	15	32	2171
	2+	2203	70	58	128	2075
	3+	2203	22	16	38	2165
VHL Variant Curation Expert Panel	Strictly	2179	6	3	9	2170
	2+	2179	23	16	39	2140
	3+	2179	6	3	9	2170
Mitochondrial Disease Variant Curation Expert Panel	Strictly	2148	28	9	37	2111
	2+	2148	74	31	105	2043
	3+	2148	32	9	41	2107
ClinGen Epilepsy Expert Panel	Strictly	2076	70	12	82	1994
	2+	2076	155	44	199	1877
	3+	2076	81	14	95	1981
ClinGen Hearing Loss Expert Panel	Strictly	2008	63	25	88	1920
	2+	2008	145	58	203	1805
	3+	2008	71	28	99	1909
ClinGen Clinically Relevant Variant Curation Expert Panel	Strictly	1987	22	11	33	1954
	2+	1987	64	35	99	1888
	3+	1987	25	12	37	1950
ClinGen Malignant Hyperthermia Expert Panel	Strictly	1874	57	4	61	1813
	2+	1874	111	11	122	1752
	3+	1874	61	5	66	1808
ClinGen Familial Hypercholesterolemia Variant Curation Expert Panel	Strictly	1651	34	17	51	1600
	2+	1651	102	44	146	1505
	3+	1651	43	19	62	1589
ClinGen Inherited Cardiomyopathy Expert Panel	Strictly	1538	40	18	58	1480
	2+	1538	115	53	168	1370
	3+	1538	46	20	66	1472

Table S14 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Total Recl.	Uncertain
ClinGen Long QT Syndrome Expert Panel	Strictly	1507	30	12	42	1465
	2+	1507	98	38	136	1371
	3+	1507	37	15	52	1455
ClinGen Inherited Arrhythmia Expert Panel	Strictly	1430	41	10	51	1379
	2+	1430	95	27	122	1308
	3+	1430	44	11	55	1375
ClinGen CFTR Expert Panel	Strictly	1205	32	20	52	1153
	2+	1205	88	45	133	1072
	3+	1205	36	23	59	1146
ClinGen Inherited Cancer Expert Panel	Strictly	1180	22	15	37	1143
	2+	1180	65	34	99	1081
	3+	1180	24	16	40	1140
ClinGen ENIGMA Expert Panel	Strictly	1047	20	12	32	1015
	2+	1047	59	28	87	960
	3+	1047	22	13	35	1012
ClinGen TP53 Expert Panel	Strictly	1012	25	11	36	976
	2+	1012	74	31	105	907
	3+	1012	28	13	41	971
ClinGen PTEN Expert Panel	Strictly	987	18	9	27	960
	2+	987	52	25	77	910
	3+	987	20	10	30	957
ClinGen RYR1 Expert Panel	Strictly	954	27	7	34	920
	2+	954	81	19	100	854
	3+	954	31	8	39	915
ClinGen CDH1 Expert Panel	Strictly	913	16	6	22	891
	2+	913	47	17	64	849
	3+	913	18	7	25	888
ClinGen APC Expert Panel	Strictly	864	20	8	28	836
	2+	864	58	23	81	783
	3+	864	22	9	31	833
ClinGen MYH7 Expert Panel	Strictly	830	24	10	34	796
	2+	830	63	26	89	741
	3+	830	27	11	38	792
ClinGen BRCA1/2 Expert Panel	Strictly	802	19	12	31	771
	2+	802	55	27	82	720
	3+	802	21	13	34	768
ClinGen GAA Expert Panel	Strictly	780	22	8	30	750
	2+	780	60	21	81	699
	3+	780	25	9	34	746
ClinGen LDLR Expert Panel	Strictly	755	15	7	22	733
	2+	755	49	20	69	686
	3+	755	17	8	25	730
ClinGen MLH1 Expert Panel	Strictly	730	13	6	19	711
	2+	730	44	18	62	668
	3+	730	15	7	22	708
ClinGen MSH2 Expert Panel	Strictly	702	14	6	20	682
	2+	702	46	19	65	637
	3+	702	16	7	23	679
ClinGen MSH6 Expert Panel	Strictly	678	12	5	17	661
	2+	678	42	17	59	619
	3+	678	13	6	19	659
ClinGen PMS2 Expert Panel	Strictly	650	11	4	15	635
	2+	650	38	15	53	597
	3+	650	12	5	17	633
ClinGen RET Expert Panel	Strictly	624	10	3	13	611

Table S14 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Total Recl.	Uncertain
	2+	624	36	14	50	574
	3+	624	11	4	15	609
ClinGen RB1 Expert Panel	Strictly	598	9	2	11	587
	2+	598	33	11	44	554
	3+	598	10	3	13	585
ClinGen TSC1 Expert Panel	Strictly	570	8	2	10	560
	2+	570	30	9	39	531
	3+	570	9	2	11	559
ClinGen TSC2 Expert Panel	Strictly	542	7	2	9	533
	2+	542	27	8	35	507
	3+	542	8	3	11	531
ClinGen NF1 Expert Panel	Strictly	518	6	2	8	510
	2+	518	25	7	32	486
	3+	518	7	2	9	509
ClinGen NF2 Expert Panel	Strictly	495	5	1	6	489
	2+	495	23	6	29	466
	3+	495	6	2	8	487
ClinGen SMARCB1 Expert Panel	Strictly	470	5	1	6	464
	2+	470	21	6	27	443
	3+	470	5	1	6	464
ClinGen VHL Expert Panel	Strictly	448	4	1	5	443
	2+	448	20	5	25	423
	3+	448	5	1	6	442
ClinGen WT1 Expert Panel	Strictly	420	4	1	5	415
	2+	420	18	4	22	398
	3+	420	4	1	5	415
ClinGen MEN1 Expert Panel	Strictly	398	3	1	4	394
	2+	398	16	4	20	378
	3+	398	3	1	4	394
ClinGen SDH Expert Panel	Strictly	372	3	1	4	368
	2+	372	15	4	19	353
	3+	372	3	1	4	368
ClinGen FH Expert Panel	Strictly	350	2	1	3	347
	2+	350	13	3	16	334
	3+	350	2	1	3	347

Table S15: VUS Reclassification Analysis by Expert Panel

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
InSiGHT Hereditary Colorectal Cancer/Polyposis Variant Curation Expert Panel	Strictly	17270	73	42	17155
	2+	17270	308	405	16557
	3+	17270	86	70	17114
Hereditary Breast, Ovarian and Pancreatic Cancer Variant Curation Expert Panel	Strictly	10173	40	58	10075
	2+	10173	168	347	9658
	3+	10173	52	91	10030
Hearing Loss Variant Curation Expert Panel	Strictly	7072	214	163	6695
	2+	7072	348	553	6171
	3+	7072	227	202	6643
Congenital Myopathies Variant Curation Expert Panel	Strictly	6018	162	55	5801
	2+	6018	353	185	5480

Table S15 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
	3+	6018	191	79	5748
RASopathy Variant Curation Expert Panel	Strictly	5660	167	41	5452
	2+	5660	288	189	5183
	3+	5660	176	68	5416
Cardiomyopathy Variant Curation Expert Panel	Strictly	5153	131	61	4961
	2+	5153	378	274	4501
	3+	5153	178	95	4880
ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel	Strictly	5098	16	79	5003
	2+	5098	53	483	4562
	3+	5098	19	103	4976
Neurofibromatosis and Schwannomatosis Variant Curation Expert Panel	Strictly	5072	74	10	4988
	2+	5072	139	47	4886
	3+	5072	75	11	4986
Epilepsy Sodium Channel Variant Curation Expert Panel	Strictly	3976	346	32	3598
	2+	3976	561	95	3320
	3+	3976	358	33	3585
Malignant Hyperthermia Susceptibility Variant Curation Expert Panel	Strictly	3358	106	26	3226
	2+	3358	266	116	2976
	3+	3358	135	46	3177
DICER1 and miRNA-Processing Gene Variant Curation Expert Panel	Strictly	3171	5	18	3148
	2+	3171	13	114	3044
	3+	3171	6	26	3139
Parkinson's Disease Variant Curation Expert Panel	Strictly	2289	34	20	2235
	2+	2289	41	45	2203
	3+	2289	34	20	2235
Congenital Heart Disease Variant Curation Expert Panel	Strictly	2072	42	9	2021
	2+	2072	87	73	1912
	3+	2072	48	19	2005
Limb Girdle Muscular Dystrophy Variant Curation Expert Panel	Strictly	2036	62	16	1958
	2+	2036	104	58	1874
	3+	2036	66	20	1950
GRIN Disorders Variant Curation Expert Panel	Strictly	1948	74	43	1831
	2+	1948	97	74	1777
	3+	1948	74	43	1831
Affiliation	Criteria	Total VUS	Pred. Path.	Pred. Ben.	Pred. VUS
Kidney Cystic and Ciliopathy Disorders Variant Curation Panel	Strictly	1914	211	168	1535
	2+	1914	233	273	1408
	3+	1914	214	182	1518
FBN1 Variant Curation Expert Panel	Strictly	1876	68	8	1800
	2+	1876	118	46	1712
	3+	1876	70	15	1791
Gastric Cancer Variant Curation Expert Panel	Strictly	1828	20	3	1805
	2+	1828	75	64	1689
	3+	1828	26	15	1787

Table S15 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
Severe Combined Immunodeficiency Disease Variant Curation Expert Panel	Strictly	1679	30	17	1632
	2+	1679	61	71	1547
	3+	1679	31	26	1622
X-linked Inherited Retinal Disease Variant Curation Expert Panel	Strictly	1550	42	21	1487
	2+	1550	59	34	1457
	3+	1550	42	21	1487
Mitochondrial Diseases Variant Curation Expert Panel	Strictly	1483	52	19	1412
	2+	1483	109	110	1264
	3+	1483	55	26	1402
Brain Malformations Variant Curation Expert Panel	Strictly	1455	63	24	1368
	2+	1455	80	38	1337
	3+	1455	63	25	1367
Rett and Angelman-like Disorders Variant Curation Expert Panel	Strictly	1453	64	44	1345
	2+	1453	99	98	1256
	3+	1453	67	53	1333
Lysosomal Diseases Variant Curation Expert Panel	Strictly	984	29	11	944
	2+	984	65	37	882
	3+	984	38	14	932
Myeloid Malignancy Variant Curation Expert Panel	Strictly	948	11	16	921
	2+	948	31	351	566
	3+	948	12	25	911
Monogenic Diabetes Variant Curation Expert Panel	Strictly	915	81	47	787
	2+	915	116	126	673
	3+	915	84	65	766
Motile Ciliopathy Variant Curation Expert Panel	Strictly	908	13	10	885
	2+	908	20	18	870
	3+	908	15	11	882
Leber Congenital Amaurosis/early onset Retinal Dystrophy Variant Curation Expert Panel	Strictly	871	30	7	834
	2+	871	44	34	793
	3+	871	31	16	824
ABCA4 Variant Curation Expert Panel	Strictly	757	29	9	719
	2+	757	43	20	694
	3+	757	29	9	719
Optic Nerve Atrophy Variant Curation Expert Panel	Strictly	703	17	10	676
	2+	703	39	36	628
	3+	703	18	11	674
Charcot-Marie-Tooth Disease Variant Curation Expert Panel	Strictly	679	36	7	636
	2+	679	53	21	605
	3+	679	36	9	634
Platelet Disorders Variant Curation Expert Panel	Strictly	660	46	33	581
	2+	660	65	84	511
	3+	660	49	39	572

Table S15 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
Familial Hypercholesterolemia Variant Curation Expert Panel	Strictly	660	67	19	574
	2+	660	95	72	493
	3+	660	76	37	547
Cerebral Creatine Deficiency Syndromes Variant Curation Expert Panel	Strictly	657	24	12	621
	2+	657	46	50	561
	3+	657	28	20	609
PTEN Variant Curation Expert Panel	Strictly	647	17	0	630
	2+	647	41	16	590
	3+	647	21	7	619
Antibody Deficiencies Variant Curation Expert Panel	Strictly	622	10	8	604
	2+	622	19	48	555
	3+	622	10	11	601
TP53 Variant Curation Expert Panel	Strictly	570	19	4	547
	2+	570	51	17	502
	3+	570	20	7	543
Endocrine Tumor Predisposition Variant Curation Expert Panel	Strictly	568	15	0	553
	2+	568	27	12	529
	3+	568	16	6	546
Potassium Channel Arrhythmia Variant Curation Expert Panel	Strictly	538	14	6	518
	2+	538	20	61	457
	3+	538	15	15	508
von Willebrand Disease Variant Curation Expert Panel	Strictly	517	47	52	418
	2+	517	58	90	369
	3+	517	49	58	410
Urea Cycle Disorders Variant Curation Expert Panel	Strictly	516	26	3	487
	2+	516	43	9	464
	3+	516	28	3	485
ACADVL Variant Curation Expert Panel	Strictly	433	21	48	364
	2+	433	34	97	302
	3+	433	24	50	359
VHL Variant Curation Expert Panel	Strictly	432	14	0	418
	2+	432	38	6	388
	3+	432	20	1	411
Hereditary Hemorrhagic Telangiectasia Variant Curation Expert Panel	Strictly	412	23	6	383
	2+	412	36	13	363
	3+	412	26	6	380
Coagulation Factor Deficiency Variant Curation Expert Panel	Strictly	390	93	23	274
	2+	390	107	30	253
	3+	390	93	23	274
Thrombosis Variant Curation Expert Panel	Strictly	354	48	12	294
	2+	354	62	21	271
	3+	354	48	12	294
Pulmonary Hypertension Variant Curation Expert Panel	Strictly	340	4	6	330
	2+	340	9	12	319

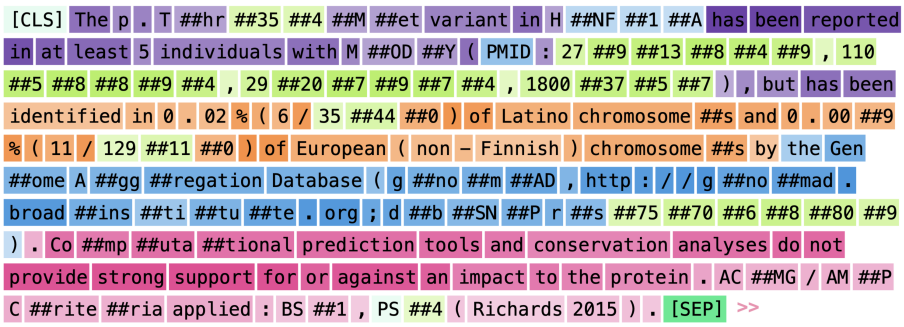
Table S15 (continued)

Expert Panel	Criteria	Total VUS	Path. Recl.	Ben. Recl.	Uncertain
	3+	340	4	7	329
Peroxisomal Disorders Variant Curation Expert Panel	Strictly	302	21	6	275
	2+	302	28	18	256
	3+	302	21	7	274
Glaucoma Variant Curation Expert Panel	Strictly	275	7	60	208
	2+	275	14	85	176
	3+	275	7	62	206
Phenylketonuria Variant Curation Expert Panel	Strictly	186	24	7	155
	2+	186	46	14	126
	3+	186	33	9	144
Dopa Decarboxylase (Aromatic L-Amino Acid Decarboxylase) Variant Curation Expert Panel	Strictly	177	3	0	174
	2+	177	5	2	170
	3+	177	4	1	172
Leukodystrophy and Leukoencephalopathy Variant Curation Expert Panel	Strictly	171	14	5	152
	2+	171	20	10	141
	3+	171	14	5	152
Rhodopsin Variant Curation Expert Panel	Strictly	170	20	2	148
	2+	170	21	6	143
	3+	170	20	4	146
Hemoglobinopathy Variant Curation Expert Panel	Strictly	168	31	23	114
	2+	168	61	32	75
	3+	168	38	23	107
Hereditary Angioedema Variant Curation Expert Panel	Strictly	156	5	4	147
	2+	156	8	5	143
	3+	156	5	4	147
Glucose-6-phosphate dehydrogenase Variant Curation Expert Panel	Strictly	113	11	6	96
	2+	113	12	13	88
	3+	113	11	6	96
Ocular Anterior Segment Disorder Variant Curation Expert Panel	Strictly	109	9	4	96
	2+	109	11	5	93
	3+	109	9	4	96
Galactosemia Variant Curation Expert Panel	Strictly	92	15	5	72
	2+	92	22	8	62
	3+	92	16	5	71
Amyotrophic Lateral Sclerosis Spectrum Disorders Variant Curation Expert Panel	Strictly	44	6	0	38
	2+	44	7	1	36
	3+	44	6	0	38

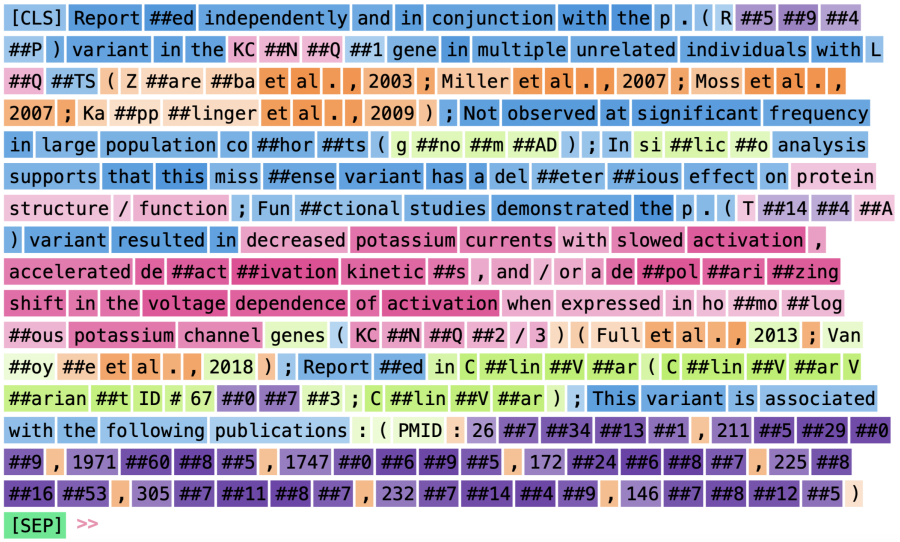
Neuron activation pattern visualization

Through a case review approach, we selected two ClinVar-labeled VUS summaries classified by the model as B/LB and P/LP, respectively. To ensure strong case evidence, these summaries were sorted based on two criteria: high model prediction confidence and sufficient text length. We filtered summaries with a final class

probability greater than 0.8 to ensure high model prediction confidence. Longer summaries were prioritized, as they are more likely to contain comprehensive textual evidence, with selection based on the total string length of the ClinVar summary. After sorting, we manually reviewed the summaries, focusing on those with evidence consistent with the American College of Medical Genetics and Genomics/Association of Molecular Pathology Sequence Variant Interpretation (ACMG/AMP SVI) framework 4.



(a) Example of VUS predicted as B/LB. Purple and orange colors represent case studies, green is PubMed ID references, blue represents population evidence, and pink represents computational scores.



(b) Example of VUS predicted as P/LP. Blue is population case reports, orange is publication references, pink is functional studies, green is ClinVar variant references, and purple is PubMed ID references.

Fig. S8: Neuron activation pattern visualization using Ecco for two variant classification case studies. Both examples are VUS submissions on ClinVar, but are predicted by our ClinVar-BERT model with high confidence (> 0.8) as B/LB and P/LP, respectively. (a) Example of a VUS predicted as B/LB, with strong neuron activations highlighting population frequency data (variant found in Latino and European chromosomes) and computational prediction results suggesting minimal impact on protein function. (b) Example of a VUS predicted as P/LP, with high neuron activations on sections describing functional evidence (decreased potassium currents, activation kinetics), case reports in unrelated individuals, and supporting literature citations, consistent with Long QT Syndrome pathogenicity.

Submission summary model embedding UMAP visualization

To investigate how fine-tuning with ClinVar evidence enhanced the ability of our language models to encode information potentially useful for variant interpretation, we performed a comparative analysis of model text embeddings for a set of submission summaries. We randomly sampled 25,000 ClinVar submission summaries and analyzed both the high-dimensional embedding space using clustering metrics and UMAP visualizations, for *BioBERT* 20 and *ClinVar-BERT*.

Our analysis finds substantial differences between BioBERT and ClinVar-BERT in how they represent ClinVar submission text summaries. As shown in Table S16, ClinVar-BERT produces significantly higher cluster separation for clinically relevant attributes compared to BioBERT. The Silhouette score, which measures how well samples are clustered by comparing the mean intra-cluster distance with the mean nearest-cluster distance (ranging from -1 to 1, with higher values indicating better-defined clusters), reveals important differences between ClinVar-BERT and BioBERT. In Table S16, we computed the average scores from 100 independent and random samples that we used the d-dimensional embeddings directly output from each language model. For clinical significance and submission classification, ClinVar-BERT's raw embeddings achieve substantially higher Silhouette scores (0.4376 and 0.6446, respectively) compared to BioBERT (0.0991 and 0.1170). This difference is also reflected in the Davies-Bouldin (DB) index, a metric that quantifies the average similarity between each cluster and its most similar cluster, where lower values indicate better separation between clusters. ClinVar-BERT shows significantly lower DB index values (1.4352 and 0.5278) than BioBERT (4.4854 and 3.2261), indicating superior cluster separation, as also observed in ClinVar-BERT's visualization in Figure S9e and Figure S9c, compared to BioBERT's visualization as shown in Figure S9f and Figure S9d. However, this pattern is reversed for the Submitter attribute, where BioBERT shows better clustering (Silhouette score 0.1381) than ClinVar-BERT (-0.2696), as we could also observe in ClinVar-BERT's UMAP visualization in Figure S9a compared to BioBERT's Figure S9b. This suggests that fine-tuning may have shifted the model's focus away from capturing submitter-specific patterns toward more variant interpretation-relevant features.

Attribute	ClinVar-BERT		BioBERT	
	Silhouette	Davies-Bouldin	Silhouette	Davies-Bouldin
Clinical Significance	**0.4376 ± 0.0027**	**1.4352 ± 0.0091**	0.0991 ± 0.0026	4.4854 ± 0.0555
Submission Classification	**0.6446 ± 0.0020**	**0.5278 ± 0.0026**	0.1170 ± 0.0029	3.2261 ± 0.0237
Submitter	-0.2696 ± 0.0510	2.8505 ± 0.0883	**0.1381 ± 0.0200**	**1.5892 ± 0.0453**

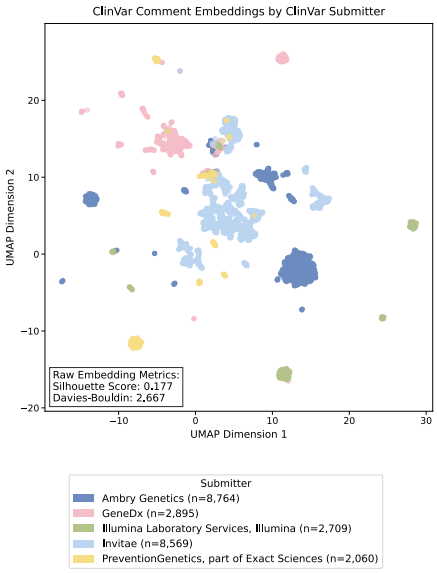
Table S16: A comparison between ClinVar-BERT and BioBERT on embedding clustering quality of language model embeddings in the d-dimensional space. For each attribute, we report Silhouette scores (higher is better) and Davies-Bouldin indices (lower is better) as mean ± standard deviation computed with 100 independent and randomly sampled runs.

We also used UMAP to visualize the reduced embedding spaces of each variant grouping in Figure S9 (a) - (f). We note that dimensionality reduction through UMAP is primarily useful for ease of visualization of these high-dimensional embedding spaces, but the clustering metrics computed on the model output's high-dimensional embeddings offer more reliable evidence of the models' representational properties.

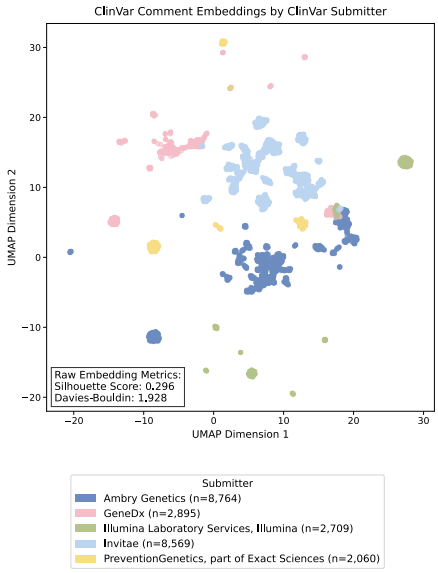
The distribution of VUS records of clinical significance (Figure S9e) from ClinVar-BERT embeddings reveals that the language patterns in VUS summaries systematically differ according to their proximity to P/LP versus B/LB clusters. VUS summaries clustering near P/LP regions tend to include some evidence of pathogenicity, as illustrated by this example: *"The G57R variant has not been published as pathogenic or been reported as benign to our knowledge. The G57R variant is not observed in large population cohorts (Lek et al., 2016; 1000 Genomes Consortium et al., 2015; Exome Variant Server)...This substitution occurs at a position that is conserved across species, and in silico analysis predicts this variant is probably damaging to the protein structure/function."* This example contains population evidence (PM2) and computational evidence (PP3) of pathogenicity. Additionally, this variant has a REVEL score of 0.797 and an AlphaMissense score of 0.9931. Whereas VUS summaries that cluster closer to B/LB regions typically contain language suggesting an absence of pathogenic evidence or evidence of benignity, as seen in this example: *"In summary, the available evidence is currently insufficient to determine the role of this variant in disease... The threonine amino acid residue is found in multiple mammalian species, which suggests that this missense change does not adversely affect protein function..."*.

Moreover, we observed that VUS variants that are closest to a P/LP cluster have multiple forms of evidence of pathogenicity. A series of detailed examples of ClinVar VUS submissions that are closest to a P/LP variant cluster are included in Table 1 in Section Supplementary Results (UMAP visualization clustering examples). We included computational evidence (REVEL and AlphaMissense score) where available, population frequency data from gnomAD, and variant functional consequences (e.g., missense, frameshift).

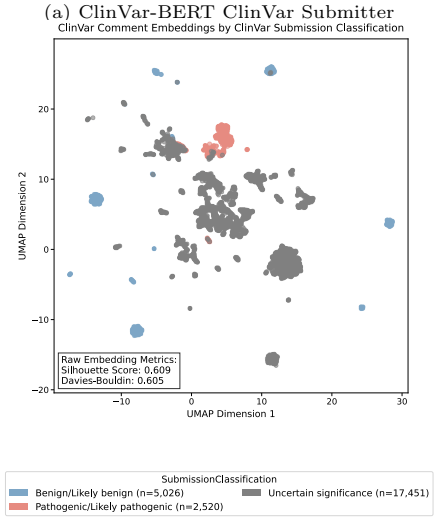
Overall, the quantitative study via cluster quality metrics and qualitative analysis via UMAP visualization both suggest that fine-tuning with ClinVar data helped the model to adapt the embedding space to better capture text related to clinical evidence, and the underlying strength of that evidence. The model's ability to capture the semantic features of clinical evidence can provide valuable insights for prioritizing variants for re-classification.



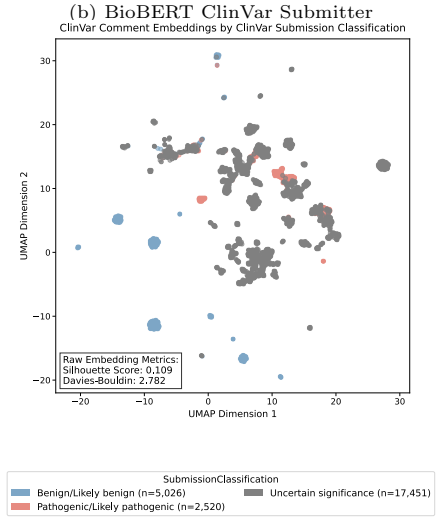
(a) ClinVar-BERT ClinVar Submitter



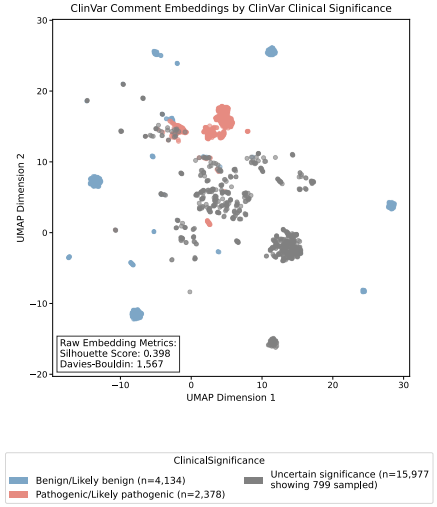
(b) BioBERT ClinVar Submitter



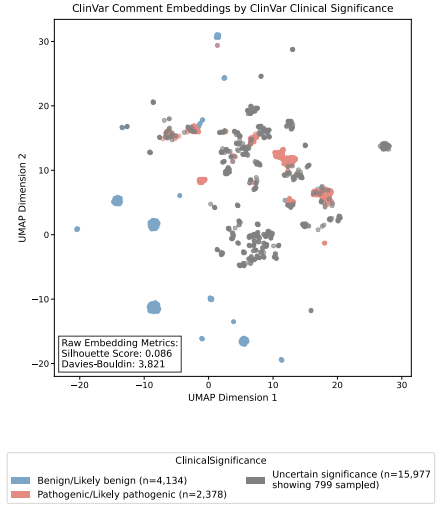
(c) ClinVar-BERT Submission Classification P/LP vs B/LB



(d) BioBERT Submission Classification P/LP vs B/LB



(e) ClinVar-BERT Clinical Significance P/LP vs B/LB



(f) BioBERT Clinical Significance P/LP vs B/LB

Fig. S9: UMAP visualization comparing embeddings from ClinVar-BERT and BioBERT models for 25,000 randomly sampled ClinVar submission summaries. In each figure, we annotate clustering quality metrics computed with d-dimensional embeddings directly from the model output. The left column visualizes ClinVar-BERT embeddings, and the right column visualizes BioBERT embeddings. (a) and (b) are visualizations with respect to the ClinVar submitter. (c) and (d) are UMAP visualizations colored with the submission classification, which is the variant pathogenicity assertion of the submission, showing P/LP and B/LB. (e) and (f) visualize the ClinVar clinical significance at the variant level within ClinVar, which considers all submissions related to the same variant and phenotype.

UMAP visualization clustering examples

Examples of VUS submissions closest to a P/LP cluster in UMAP visualization

REVEL and AlphaMissense scores are not available for loss-of-function variants

- **VCV:** VCV001768964
  SCV: SCV002691514
  Comment: The c.12860_12863delTTCA variant, located in coding exon 29 of the APOB gene, results from a deletion of 4 nucleotides at nucleotide positions 12860 to 12863, causing a translational frameshift with a predicted alternate stop codon (p.I4287Sfs*22). Frameshifts are typically deleterious in nature, however, this frameshift occurs at the 3' terminus of APOB and is not expected to trigger nonsense-mediated mRNA decay. This alteration impacts the last 277 amino acids in the protein C-terminus, which has been indicated to play a role in both lipoprotein assembly and LDLR binding (McCormick SP et al. J. Biol. Chem., 1997 Sep;272:23616-22; Boren J et al. J. Clin. Invest., 1998 Mar;101:1084-93; Borén J et al. J. Biol. Chem., 2001 Mar;276:9214-8). Nevertheless, the exact functional impact of these altered amino acids is unknown at this time. Since supporting evidence is limited at this time, the clinical significance of this alteration remains unclear.
  gnomAD: absent from gnomAD
  Variant type: frameshift variant

- **VCV:** VCV000483568
  SCV: SCV000669586
  Comment: The p.Y97D variant (also known as c.289T¿G), located in coding exon 3 of the MLH1 gene, results from a T to G substitution at nucleotide position 289. The tyrosine at codon 97 is replaced by aspartic acid, an amino acid with highly dissimilar properties. This alteration is observed in an individual whose colorectal tumor demonstrated normal mismatch repair protein expression on immunohistochemistry (Ambry internal data). This alteration was reported in an individual whose colorectal tumor demonstrated low microsatellite instability and family history meeting Amsterdam criteria (Dominguez-Valentin M et al. Hered Cancer Clin Pract, 2013 Dec;11:18; Köger N et al. Genes Chromosomes Cancer, 2018 Jul;57:350-358). This alteration is also reported in an individual whose colorectal tumor demonstrated high microsatellite instability but weak MLH1 protein expression on immunohistochemistry, and family history meeting Bethesda criteria (Hardt K et al. Fam Cancer, 2011 Jun;10:273-84). Functional assays demonstrated reduced MMR activity for p.Y97D compared to wild-type MLH1; however, expression for p.Y97D was similar to wild-type MLH1 (Köger N et al. Genes Chromosomes Cancer, 2018 Jul;57:350-358). Based on internal structural analysis using published crystal structures, p.Y97D is moderately disruptive to the structure (Ambry internal data; (Ban C et al. Cell, 1999 Apr;97:85-97; Hu X et al. FEBS Lett, 2003 Jun;544:268-73; Wu H et al. Acta Crystallogr F Struct Biol Commun, 2015 Aug;71:981-5). This amino acid position is highly conserved in available vertebrate species. In addition, this alteration is predicted to be deleterious by in silico analysis. Since supporting evidence is conflicting at this time, the clinical significance of this alteration remains unclear.
  REVEL score: 0.964
  AlphaMissense score: 0.988
  gnomAD: absent from gnomAD
  Variant type: missense variant

Examples of VUS submissions closest to a P/LP cluster in UMAP visualization (Cont.)

REVEL and AlphaMissense scores are not available for loss-of-function variants

- **VCV:** VCV000066877
  SCV: SCV002720118
  Comment: The p.R654* variant (also known as c.1960C¿T), located in coding exon 11 of the LMNA gene, results from a C to T substitution at nucleotide position 1960. This changes the amino acid from an arginine to a stop codon within coding exon 11. This alteration occurs at the 3' terminus of the LMNA gene, is not expected to trigger nonsense-mediated mRNA decay, and only impacts the last 11 amino acids of the protein. The exact functional effect of this alteration is unknown. This variant was reported in a family with dilated cardiomyopathy (DCM) with segregation in three affected family members, and absence in three other affected family members; an additional cardiac variant in PLN was also detected in the proband and the three affected family members without the LMNA variant (Parks SB et al. Am. Heart J., 2008 Jul;156:161-9; Cowan JR et al. Circ Genom Precis Med, 2018 Jul;11:e002038). This alteration was also detected in an individual with Hutchinson-Gilford Progeria syndrome, who was homozygous for a ZMPSTE24 loss of function variant, as well as in his unaffected mother and brother (Denecke J et al. Hum. Mutat., 2006 Jun;27:524-31). In addition, the variant was reported in an individual with Emery Dreifuss muscular dystrophy type 2 with symptoms of axonal neuropathy (Denecke J et al. Hum. Mutat., 2006 Jun;27:524-31; Bernasconi P et al. Nucleus, 2018 Jan;9:292-304). Functional studies showed formation of protein aggregates, but the functional impact of this finding has not been determined (Cowan J et al. Circ Cardiovasc Genet, 2010 Feb;3:6-14). Since supporting evidence is limited at this time, the clinical significance of this alteration remains unclear.
  gnomAD: 2.464e-05
  Variant type: stop gained

- **VCV:** VCV000546000
  SCV: SCV001185259
  Comment: The p.E17* variant (also known as c.49G¿T), located in coding exon 1 of the VHL gene, results from a G to T substitution at nucleotide position 49. This changes the amino acid from a glutamic acid to a stop codon within coding exon 1. Premature stop codons are typically deleterious in nature; however, an alternate initiation codon exists 37 amino acids downstream from this alteration, and is reported to result in a biologically active isoform known as VHL 19 (Iliopoulos O et al, Proc. Natl. Acad. Sci. U.S.A. 1998 Sep; 95(20):11661-6. Schoenfeld A et al, Proc. Natl. Acad. Sci. U.S.A. 1998 Jul; 95(15):8817-22). Since supporting evidence is limited at this time, the clinical significance of this alteration remains unclear.
  gnomAD: 0
  Variant type: stop gained

- **VCV:** VCV000426716
  SCV: SCV000577240
  Comment: The E227G variant has not been published as a pathogenic variant, nor has it been reported as a benign variant to our knowledge. The E227G variant is not observed in large population cohorts (Lek et al., 2016; 1000 Genomes Consortium et al., 2015; Exome Variant Server). This variant is a non-conservative amino acid substitution, which is likely to impact secondary protein structure as these residues differ in polarity, charge, size and/or other properties. This substitution occurs at a position that is conserved across species, and in silico analysis predicts this variant is probably damaging to the protein structure/function. However, missense variants in nearby residues have not been reported in the Human Gene Mutation Database in association with SEPN1-related disorders (Stenson et al., 2014).
  REVEL score: 0.621
  gnomAD: 6.841e-07
  Variant type: missense variant