From Text to Translation: Using Language Models to Prioritize Variants for Clinical Review

Weijiang Li[1], Xiaomin Li[2], Ethan Lavallee[6], Alice Saparov[3], Marinka Zitnik[4], Christopher Cassa[1]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, 02115, MA, United States.
[2]School of Engineering and Applied Sciences, Harvard University, Boston, 02138, MA, United States.
[3]Institute of Human Genetics, Technical University of Munich, Munich, 80333, Germany.
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, 02115, MA, United States.
[6]Middlebury College.

Contributing authors: weli1@bwh.harvard.edu; xiaominli@g.harvard.edu; elavallee@middlebury.edu; Alice.Saparov@mri.tum.de; marinka@hms.harvard.edu; ccassa@bwh.harvard.edu;

Abstract

Despite rapid advances in genomic sequencing, most rare genetic variants remain insufficiently characterized for clinical use, limiting the potential of personalized medicine. When classifying whether a variant is pathogenic, clinical labs adhere to diagnostic guidelines that comprehensively evaluate many forms of evidence including case data, computational predictions, and functional screening. While a substantial amount of clinical evidence has been developed for these variants, the majority cannot be definitively classified as 'pathogenic' or 'benign', and thus persist as 'Variants of Uncertain Significance' (VUS). We processed over 2.4 million plaintext variant summaries from ClinVar, employing sentence-level classification to remove content that does not contain evidence and removing uninformative summaries. We developed ClinVar-BERT to discern clinical evidence within these summaries by fine-tuning a BioBERT-based model with labeled records. When validated classifications from this model against orthogonal functional screening data, ClinVar-BERT significantly separated estimates of functional impact in clinically actionable genes, including *BRCA1* ($p = \mathbf{1.90 \times 10^{-20}}$), *TP53* ($p = \mathbf{1.14 \times 10^{-47}}$), and *PTEN* ($p = \mathbf{3.82 \times 10^{-7}}$). Additionally, ClinVar-BERT achieved an AUROC of 0.927 in classifying ClinVar VUS against this functional screening data. This suggests that ClinVar-BERT is capable of discerning evidence from diagnostic reports and can be used to prioritize variants for re-assessment by diagnostic labs and expert curation panels.

Keywords: large language models, variant classification, ClinVar, genetic diagnostics

*Please address correspondence to: ccassa@bwh.harvard.edu

Declaration of Interests: The authors declare no competing interests.

Acknowledgments: We gratefully acknowledge funding from NIH R01HG010372 (W.L., E.L., C.C.) and R21HG014015 (W.L, M.Z., C.C.).

1 Introduction

As genomic sequencing becomes increasingly integrated into clinical practice, the pace of variant interpretation and biomedical data production has accelerated. From 2019 to 2024, clinical laboratories have submitted 3.68 million variant classifications to ClinVar, a public archive of human genetic variation linked to clinical disorders [1]. However, even in well-studied disease genes like *BRCA1* and *LDLR*, the majority of variants have only

been observed in a few cases or a single individual [2, 3]. Consequently, these variants often lack definitive human genetic evidence to be classified as pathogenic or benign and are instead classified as 'Variants of Uncertain Significance' (VUS) [4]. Due to this clinical uncertainty, current practice guidelines do not recommend communicating information about VUS to providers or patients outside of a clinical indication for testing [5].

This translational gap prevents many patients who carry such variants in actionable genes from benefiting from genomic medicine at the population level [6]. In nine genes that are responsible for hereditary breast and ovarian cancer (HBOC), Lynch syndrome (LS), and familial hypercholesterolemia (FH), there is a substantial burden of such rare variation in the population. Over 16% or 1 in 6 individuals carry a rare, non-synonymous variant, roughly 18-fold more than those already classified as pathogenic [7]. Although many of these variants may have limited phenotypic effects, some carry a substantial risk of disease, but they lack sufficient evidence to be classified as pathogenic.

The sequence variant interpretation (SVI) process involves expert curation of evidence supporting pathogenicity or benignity, following guidelines developed by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) [4]. Information curated during assessment includes clinical case evidence, computational predictions of variant effect, experimental screening measuring protein function, among others. The available evidence is weighed collectively to reach a clinical classification for a variant and is often compiled into a text summary. Since 2019, 2.1 million of these variant submission text summaries have been deposited to ClinVar [1]. Although they contain a great deal of curated evidence, these reports contain heterogeneous information, lack a consistent structure, and often do not use controlled vocabularies for evidence types or clinical information.

Understanding the evidence contained in these diagnostic reports can be useful for improving variant classification. Recent work has highlighted that as evidence of pathogenicity is developed for a variant, it can be used to sub-classify variants that may ultimately be classified as pathogenic [8]. Here, we trained language models (LMs) to discern the evidence patterns that are indicative of pathogenicity, benignity, or uncertainty contained within variant summaries. Ultimately, this information can be used to identify VUS that have a substantial amount of evidence of pathogenicity to prioritize them for review by expert panels.

2 Results

Our objective is to train a model that learns text representations of evidence of variant pathogenicity, benignity, and uncertainty. With these learned representations of evidence, we then classify clinical text summaries for variants of uncertain significance (VUS), to assess how likely each variant is to contain evidence of being pathogenic, benign, or uncertain.

2.1 Training data: Variant text summaries

We first developed model training data using variants that had been previously classified by a clinical lab and deposited into ClinVar. A central challenge in learning representations of evidence is that ClinVar text summaries are heterogeneous with complex structures. We first deduplicated and filtered highly similar summaries, short, or uninformative summaries, and standardized punctuation, characters, and language (See Methods in Section 4). We also sample summaries across clinical labs, as many submissions come from just a few labs which could potentially contribute to bias and lack of text diversity in model training.

Text summaries from some clinical labs follow a template for how evidence is described. Consequently, summaries from those labs may exhibit high structural similarity. Furthermore, some sentences serve as a *conclusion* of the variant classification *(e.g. 'Based on the supporting evidence, this alteration is interpreted as a disease-causing mutation')* which is a clear proxy of a class label. Other sentences provide a *description* the variant *(e.g. 'The p.L95P pathogenic mutation (also known as c.284T>C), located in coding exon 3 of the SDHD gene, results from a T to C substitution at nucleotide position 284.')*. Both of these examples do not provide evidence of variant pathogenicity and could contribute to bias or overfitting in model training from the presence of specific structural elements.

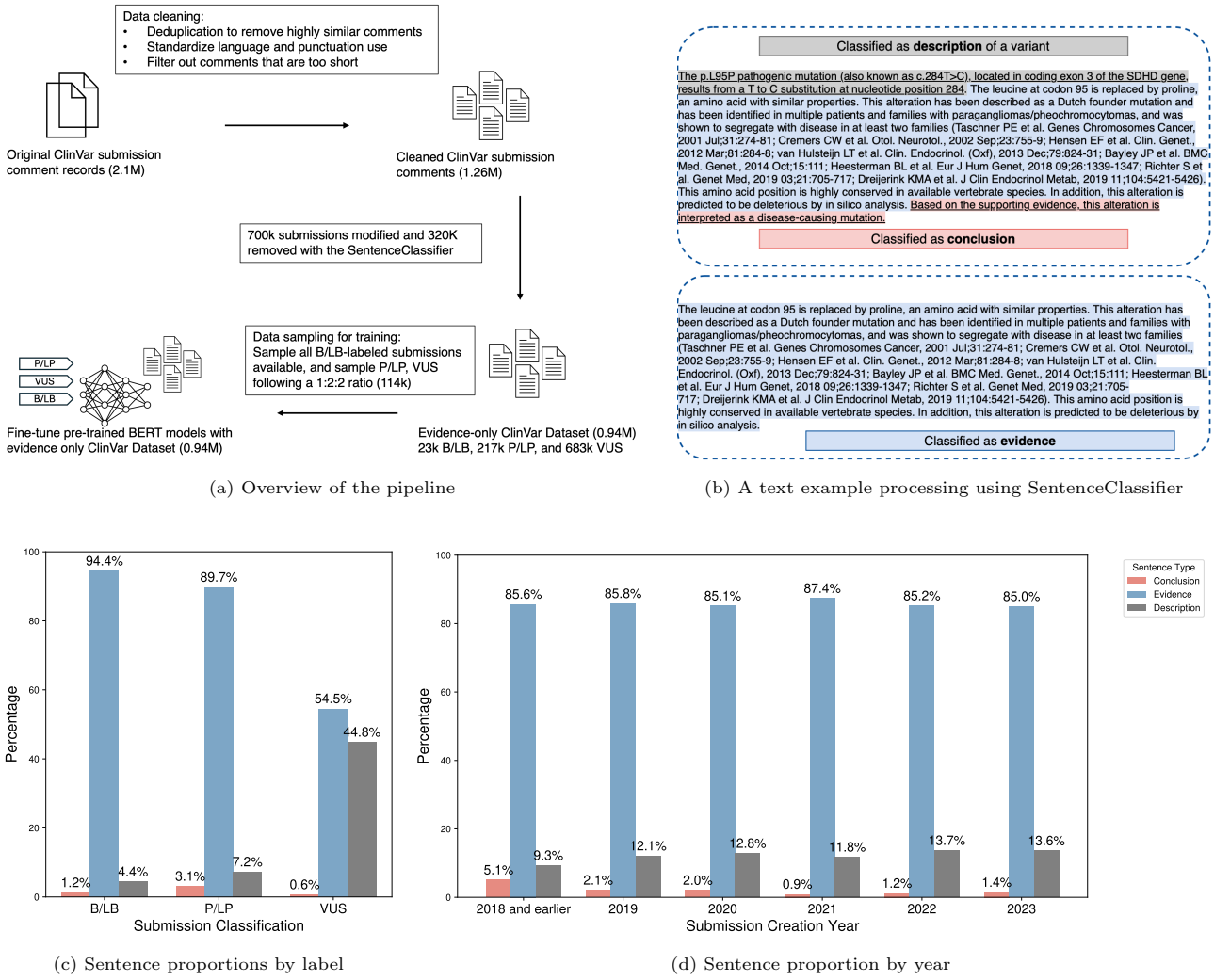
We classified and filtered predicted examples of these sentence types, as shown in Figure 1b. Figure 1c describes the proportion of each type of sentence in the training data; Figure 1d shows the distribution of three sentence types by year.

We evaluated two approaches to mitigating the challenge of embedded class labels and structural similarity: 1) using a rule-based method to filter and 2) using a sentence classifier to filter as illustrated in Figure 1a, and contrasted these with using raw original data.

- **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 Section A (Section A.1.1).

- **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.
- **Original Unfiltered Data ('raw-data' dataset)**: We contrast results with the original raw ClinVar data which has not been processed.

After filtering ClinVar text summaries using each of these three approaches, we created three distinct training sets. We sampled pathogenic or likely pathogenic (P/LP), benign or likely benign (B/LB), or uncertain (VUS) variants in proportions of 2:2:1, given the limited number of B/LB variants. All three sets contained the same number of variants, sampling was done by class, submitting lab, and gene.



(a) Overview of the pipeline

(b) A text example processing using SentenceClassifier

(c) Sentence proportions by label

(d) Sentence proportion by year

Fig. 1: (a) An overview of text processing and record sampling used ahead of fine-tuning BERT models with ClinVar submission text summaries. (b) An example submission summary (SCV002749858): In this submission, the lab describes this variant (gray highlighting) and also classifies the variant as pathogenic (pink highlighting). We trained a sentence classifier to identify and filter these description and conclusion sentences so that only sentences containing evidence (blue highlighting) are used in model training. (c) Sentence type proportion distribution for three submission classification labels (B/LB, VUS, and P/LP) in the training data. Text summaries from the B/LB and P/LP classes have much larger fractions of evidence-labeled sentences, in contrast with VUS-labeled samples, which have a much larger share of description-labeled sentences. (d) Sentence type proportion distribution by ClinVar submission creation year, with pre-2019 years grouped together and individual years shown from 2019 through 2024.

2.2 Evaluation of model performance

Next, we sought to evaluate the performance of each possible training set on *BioBERT-base*, a BERT-based model trained on a broad set of biomedical knowledge, including PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) [9]. We fine-tuned *BioBERT-base* using each of our three training sets and

measured classification performance using testing data, a random held-aside sample of 20% of the **raw-data** dataset without any text processing. We also compared to native *BioBERT-base* model performance without fine-tuning using ClinVar data. We observed a significant performance improvement for models fine-tuned with ClinVar data compared to the pre-trained *BioBERT-base* model. Results are shown in Table 1. This indicates that additional training with ClinVar data enhances the model's ability to learn useful text representations and improves prediction accuracy on test data.

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 [9]	0.3760	0.1601	0.3133	0.1919	0.5122

Table 1: Performance of fine-tuned BioBERT-base models with ClinVar data compared to a pre-trained BioBERT-base model that has not been fine-tuned using ClinVar data on test data from the **raw-data** dataset.

We observed significant evidence of overfitting for the three fine-tuned BioBERT base models we developed. Overfitting was most apparent with the model developed using the *raw-data* training set, both in terms of classification performance and training loss characteristics. We found that training loss converged quickly for the *raw-data* approach, often within 100 steps. While the *rule-based* approach was slower to converge, both the *raw-data* and *rule-based* text processing methods converged quickly during training, achieving over 90% accuracy with fewer than 1,000 examples. In contrast, the model trained using the *evidence-only* approach was the slowest to converge. Training loss, evaluation loss, and accuracy during training for each text processing method can be found in Section A (Figure A1c).

Next, we aimed to understand and mitigate potential causes of overfitting using an ablation study. For each variant summary, we identified the most influential sentence that would change model predictions (e.g., from P/LP to VUS, or *vice versa*), using randomly sampled ClinVar records. Detailed results of this study can be found in Section A (A.2.2). We found that the model often learned to make predictions based on template-based structural characteristics, rather than learning unique or informative evidence of pathogenicity. We used this information to further refine the SentenceClassifier to help remove such template-based sentence structures to mitigate overfitting during training.

We additionally evaluated multiple pre-trained general-domain BERT language models, including BERT [10] and RoBERTa [11] on the **raw-data** test set. We also used BERT models trained with different biomedical text corpora including BioBERT [9], ClinicalBERT [12], and ScholarBERT [13] (See Methods in Section 4), including base and large models where available, on the same test data from the **raw-data** dataset. Overall, pre-trained language models did not perform well in this ClinVar text classification task, there was still a noticeable performance difference between domain-specific language models and general-domain models, as shown in Table B1 in Section B.

Interestingly, we observed that the performance of domain-specific models, such as BioBERT and ScholarBERT, was not as strong as that of general-domain models, without further training on the ClinVar dataset. However, once these models were fine-tuned with ClinVar data, domain-specific models had slightly better performance than general-domain models, and the performance gap became smaller, with all fine-tuned models achieving comparable results.

2.3 Validation using orthogonal functional screening data

We next proceeded to validate model predictions using orthogonal estimates of variant-level functional impact from deep mutational scanning (DMS) screens as 'ground truth' validation data. These experimental screens enable the functional assessment of thousands of coding variants that were installed in a cell line, typically replacing the native gene sequence. We evaluated model performance using DMS assays for five commonly screened genes: *BRCA1*, *LDLR*, *TP53*, *HRAS*, and *PTEN*, using functional score data from MaveDB [14, 15], which was normalized and processed using the FUSE pipeline [16].

We measured our model performance by comparing model predictions of pathogenicity from submission summaries with experimental measurements of variant functional impact from these DMS screening assays. We first developed class labels for 'ground truth' functional score data, using thresholds based on the prior expectation of proportions of single nucleotide ClinVar VUS [17]. These thresholds were predicted to have damaging effects, intermediate effects, or have preserved function, from the *BRCA1* screening dataset. The resulting functional score thresholds set the top 27.1% of variant functional scores to be damaging (equivalent

to our predicted P/LP class) and the bottom 27.5% of variant functional scores to have preserved function (equivalent to our B/LB class), with the rest labeled as having intermediate function (similar to some variants in our VUS class). With each functional score having a ground truth class label, we then compared prediction results from our trained models to the ground truth labels. Results for BioBERT-base models are shown in Table 2; results for all models are included in Section B (Table B2).

Model	Accuracy	Precision	Recall	F1 Score	Pair-wise AUC			Avg AUC-ROC
					P/LP vs B/LB	P/LP vs VUS	B/LB vs VUS	
BioBERT-base + ClinVar (evidence-only)	0.4753	0.4930	0.4753	0.4219	**0.9272**	**0.8043**	0.5470	**0.7595**
BioBERT-base + ClinVar (rule-based)	0.4891	0.5098	0.4891	0.4399	0.9096	0.7938	0.5377	0.7470
BioBERT-base + ClinVar (raw-data)	0.4840	0.5306	0.4840	0.4192	0.9037	0.7882	0.5826	0.7582
BioBERT-base [9]	0.2713	0.0736	0.2713	0.1158	0.3953	0.5428	0.3953	0.4503

Table 2: Evaluation results of trained BioBERT-base models trained with three different text processing and sampling methods (evidence-only, rule-based filtered, and raw ClinVar data) and pre-trained BioBERT-base on orthogonal generated DMS Data

We next analyzed the distributions of functional scores for each existing and predicted class label. We first analyzed DMS functional scores using existing ClinVar variant classifications (not derived from our models). Higher DMS functional scores generally indicate larger impacts on protein function, whereas lower scores indicate smaller impacts on protein function. As expected, in most genes, variants originally classified as B or LB had lower functional scores than VUS variants, which had lower functional scores than variants classified as LP or P (Figure 2, top). Notably, this was not the case for *HRAS* which had limited numbers of variants with classifications in ClinVar with functional scores, and *PTEN* which had no B variants and very few LB variants.

We then used our fine-tuned models trained using the three text processing methods to make predictions on each VUS-labeled ClinVar submission with a DMS functional score. Variant text summaries were assigned B/LB, VUS, and P/LP classifications following specific recalibration proportions derived from published proportions in *BRCA1* [17] so that each model would have an equal number of predictions in each class (See Section 4.3.2). The DMS functional scores of variants that were classified as P/LP were significantly different from the functional scores for variants that were predicted as B/LB, for nearly all model types and genes. For many models, we found that predicted P/LP and B/LB groups showed highly significant differences, with large separations in their median functional scores (Figure 2).

Notably, DMS data from *LDLR* was only recently published in [18], so it was highly unlikely that this DMS information was present in any ClinVar text summaries that we evaluated for this gene. These results suggest that these fine-tuned models had been sufficiently trained to understand information useful for variant classification for variants within ClinVar that did not have sufficient evidence required to be classified as benign or pathogenic. Full visualization results are included in Section B (Figure B2 and Figure B3).

2.3.1 Performance Discussion

We found that the fine-tuned *BioBERT-base* and *BioBERT-large* models had the best performance across the language models we evaluated. We found consistent and highly significant differences between the functional scores of variants classified as B/LB and P/LP in all five genes, with large shifts in median functional scores between these two groups. Results for other fine-tuned models were included in Section A (B.0.2). For the remainder of the paper, we used the *evidence-only* trained version of *BioBERT-base*, which we call *ClinVar-BERT*.

2.4 Model Attention Weight Visualization

Given that the types of evidence and descriptions in these submission summaries were heterogeneous, we aimed to characterize the forms of evidence that our models were identifying. We analyzed attention weights to understand the components of each text summary to identify the specific words or phrases that were influential in driving model classifications. We used Ecco [19], a Python library for interpreting and visualizing language model attention weights, to inspect attention patterns that emerged during classification. Ecco processed input text through our fine-tuned BERT models, extracting the attention weights from each layer and the head of the transformer, and aggregated these weights to identify components within the input that were influential for classification.

Using a case review, we analyzed examples of variants submitted to ClinVar as uncertain (VUS) that were classified by ClinVar-BERT as B/LB or P/LP with high model prediction confidence (probability > 0.8). For the case example where a ClinVar VUS was classified as B/LB by the model, we found high attention weights on parts of the summary that describe evidence of variant benignity. Specifically, the model focused on text describing

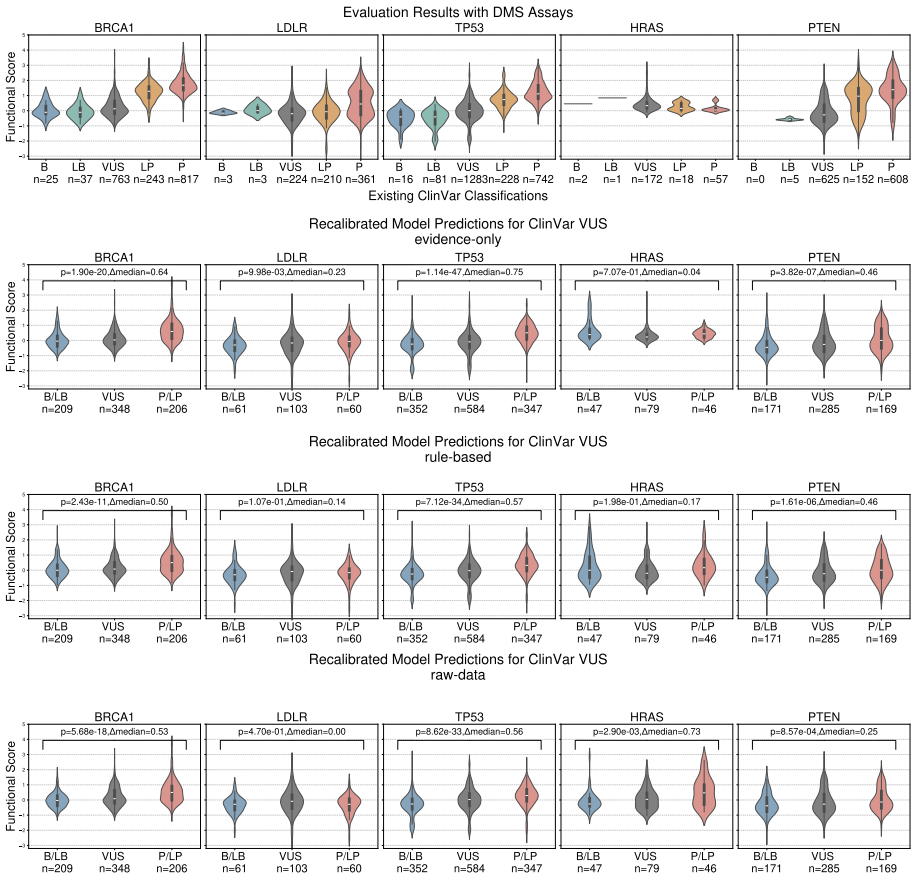
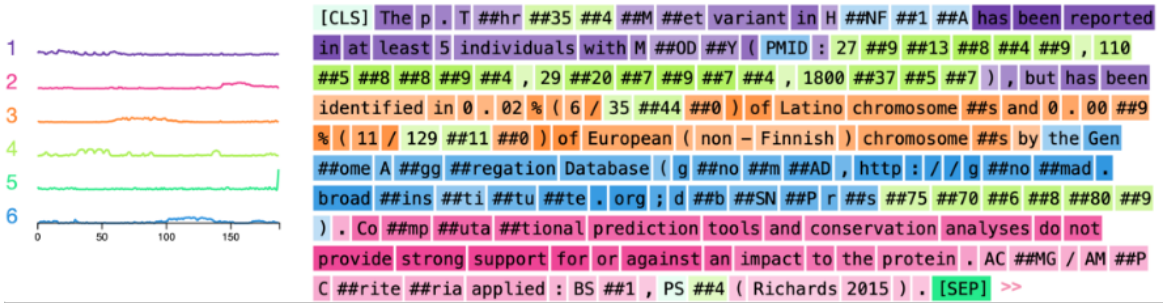
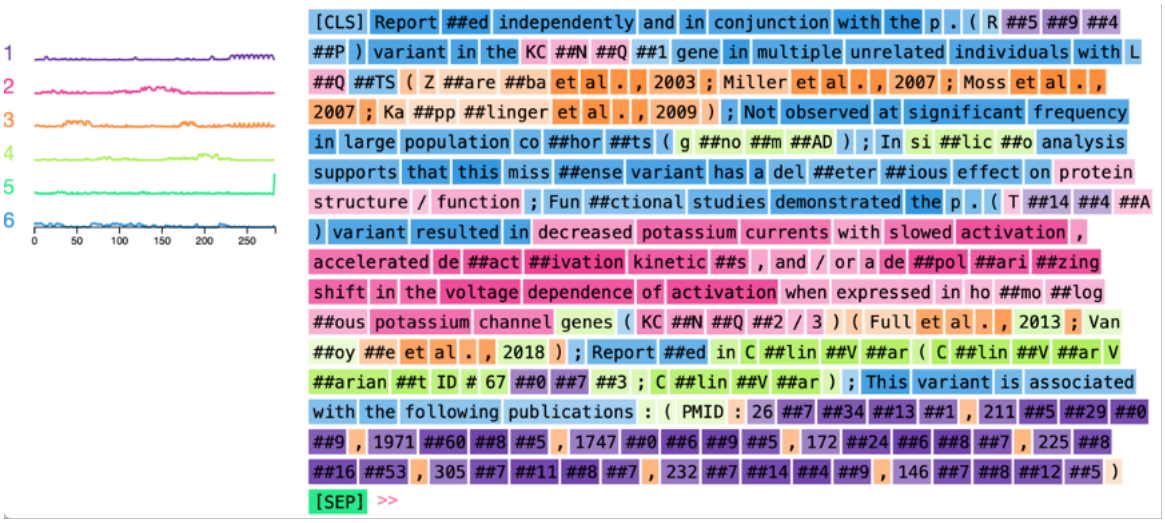


Fig. 2: Top row: Existing classification on ClinVar (B, LB, VUS, LP, and P) on the x-axis and functional scores on the y-axis. The following rows are recalibrated model prediction on **VUS**-labeled submission on the x-axis (B/LB, VUS, and P/LP), and functional scores on the y-axis.

high population frequency and computational evidence predicting no likely impact on protein function. For the case example of a ClinVar-labeled VUS submission summary that was predicted to be P/LP, the model showed high attention weights on different parts of the summary. The model particularly focused on references to prior publications describing the variant, functional screening evidence highlighting the impact on potassium channel function, and case reports for individuals with Long QT Syndrome. Altogether, this supported the notion that the model focused on clinically relevant information, such as specific ACMG evidence types including case reports, functional screening evidence, and computational predictions.



(a) Example of VUS Predicted to be B/LB



(b) Example of VUS Predicted to be P/LP

Fig. 3: Attention visualization using Ecco for two case examples. Text summaries are from variants classified as uncertain (VUS) within ClinVar, which are predicted by our model to be B/LB and P/LP with probabilities over 0.8. (a) Example of a ClinVar-labeled VUS submission summary that is model predicted as B/LB. The model has high attention weights on parts of the summary that describe the variant, highlighting that it is not highly rare in human population datasets, and that computational evidence predicts that it is not likely to impact protein function. (b) Example of a ClinVar-labeled VUS submission summary that is model predicted to be P/LP. The model has high attention weights on parts of the summary that highlight references to prior publications about the variant, functional screening evidence that highlights an impact on potassium channel function, and case reports for individuals with Long QT Syndrome.

2.5 Submission Summary Embedding UMAP Visualization

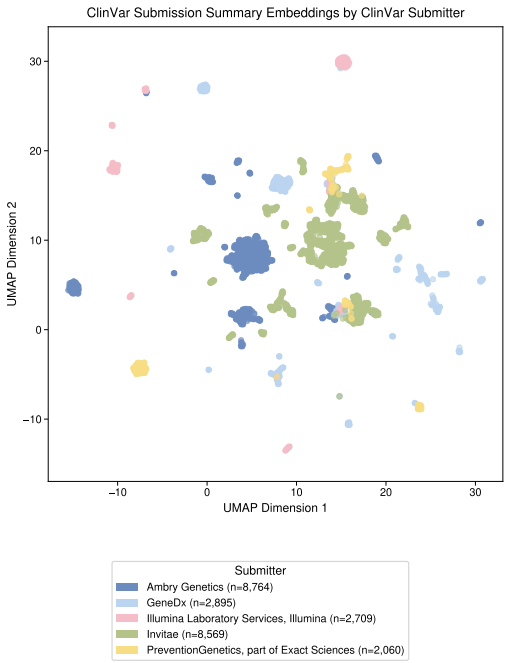
To investigate how fine-tuning with ClinVar data enhances the ability of models to encode clinically meaningful patterns, we performed a comparative analysis of submission summaries. We randomly sampled 25,000 ClinVar submission summaries and then visualized and compared the embedding spaces of *BioBERT* [9] and *ClinVar-BERT* using UMAP [20]. We analyzed these embedding spaces with regard to submitting laboratory, submission classification, and overall clinical significance, as shown in Figure 4.

BioBERT and ClinVar-BERT produce different embedding spaces when representing ClinVar submission texts. In the BioBERT visualization (Figure 4b, 4d, and 4f), submissions show a scattered distribution with loosely defined boundaries. Compared to ClinVar-BERT embedding spaces, each BioBERT space has a lower Silhouette score of 0.061 for clinical significance and 0.094 for submission classification. In contrast, ClinVar-BERT embeddings (Figures 4a, 4c, and 4e) demonstrate significantly tighter clustering, with higher Silhouette Scores of 0.116 for clinical significance and 0.157 for submission classification. The well-defined clusters in the UMAP visualization demonstrate that ClinVar fine-tuning strengthened the model's ability to encode evidence-associated patterns from submission text records.

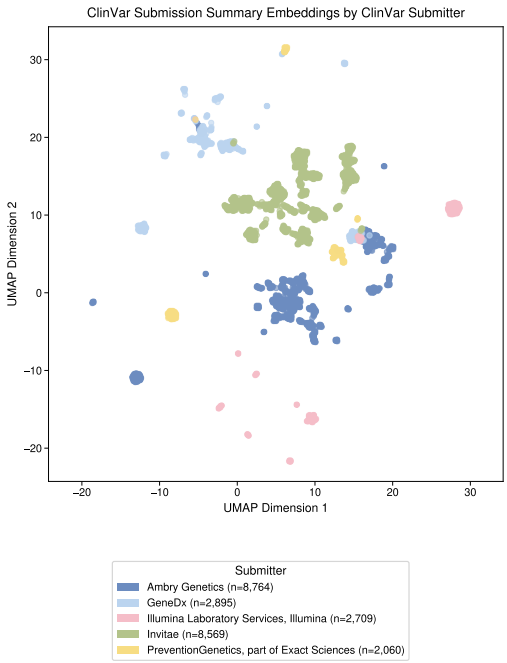
The distribution of VUS records of clinical significance (Figure 4e) from ClinVar-BERT embeddings reveals that the language patterns in VUS summaries systematically differ according to their proximity to P/LP versus

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. 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..."*.

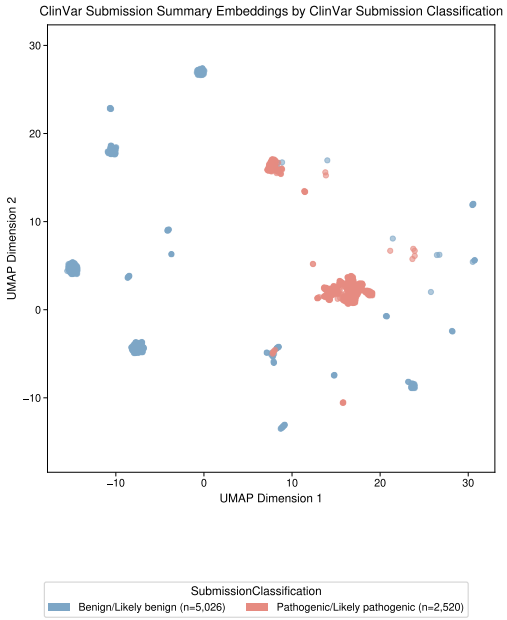
These UMAP visualizations suggest that fine-tuning with ClinVar data has helped to adapt the embedding space to better capture text related to pathogenicity and the underlying strength of evidence. This information appears to be reflected in submitter-specific documentation patterns. The model's ability to capture the semantic features of clinical evidence can provide valuable insights for prioritizing variant re-classification.



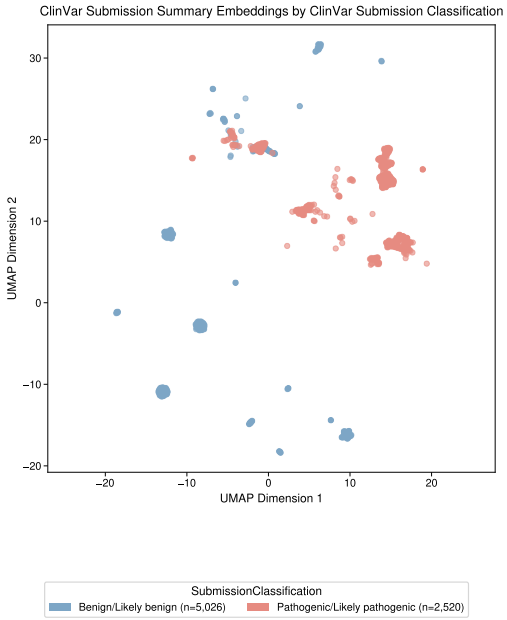
(a) ClinVar-BERT ClinVar Submitter



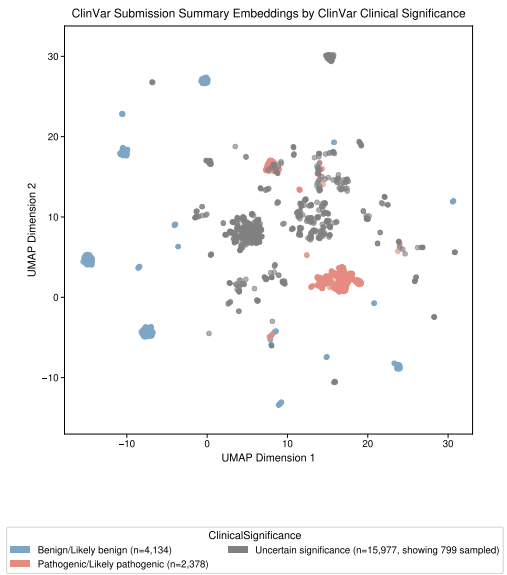
(b) BioBERT ClinVar Submitter



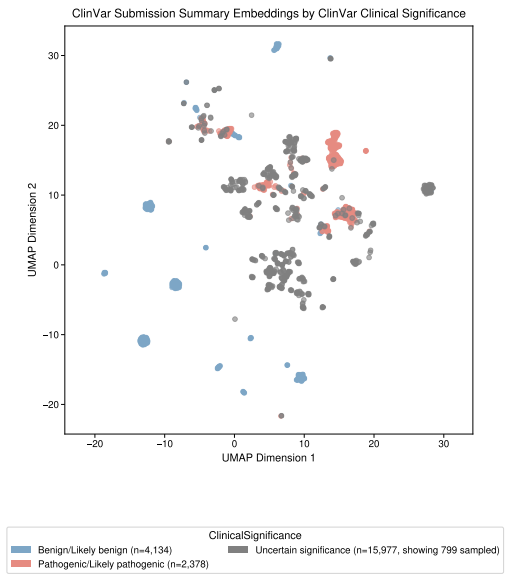
(c) ClinVar-BERT Submission Classification P/LP vs B/LB



(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. 4: UMAP visualization comparing embeddings from ClinVar-BERT and BioBERT models for 25,000

3 Discussion

In this study, we evaluated the potential of training language models to learn generalizable clinical evidence from unstructured variant summary text records in ClinVar. We fine-tuned both general-purpose and domain-specific models using labeled clinical summary text records to discern evidence of variant pathogenicity or benignity. We mitigated model bias and overfitting using a quality control pipeline that identified problematic records and a sentence classifier that filtered sentences that were unlikely to contain evidence of pathogenicity or benignity. We also validated our models using orthogonal data from functional screening assays in genes which were entirely held aside from model training. We found that variants classified as pathogenic and benign using *ClinVarBERT* had significantly different functional assay scores, supporting the hypothesis that this model identified relevant evidence.

We found large improvements in variant classification tasks using models fine-tuned with ClinVar training data over general-domain models. Among the models we evaluated, classification performance is also consistently better when fine-tuning models trained on broad biomedical text corpora, such as *BioBERT-base*, rather than using general-domain models. Given that *BioBERT-base* was trained on a large set of PubMed abstracts and PubMedCentral full-text articles, it appears to be a solid foundation for learning additional generalizable biomedical and clinical evidence within ClinVar text summaries. Training and classification metrics also demonstrate that our sentence classification approach (filtering sentences from text summaries that were unlikely to contain clinical evidence) performed well across validation datasets. This approach reduces overfitting, likely by removing conclusion sentences that contained proxy class labels, as well as by removing sentences with high structural similarity which were common in description sentences.

Our findings underscore the utility of language models in processing and interpreting intricate clinical narratives, offering potential applications in variant prioritization. Our model *ClinVar-BERT* has the potential to identify variants whose text summaries contain meaningful clinical evidence, but which were not yet sufficient for a pathogenic or benign classification. Future research may include integrating information across a set of clinical summaries about the same variant from different diagnostic labs. By analyzing information developed by different labs, such an approach could prioritize variants that collectively had sufficient evidence to be reclassified, but where the evidence provided by any single lab was insufficient for classification. This information could be used to inform ClinGen Variant Curation Expert Panels (VCEPs) about which variants are most likely to change classifications from expert review. Future work could also extract specific forms of evidence which were present within a clinical summary text, and use that to identify evidence gaps.

3.1 Limitations of this work

This study has several limitations. First, some variants have multiple text summaries from different clinical labs, which may need to be harmonized. Because evidence of pathogenicity is often developed over time, more information is likely to be contained in the latest submission summary, but it is not guaranteed to contain all available evidence. In contrast, while a very large number of variants have submission text summaries, many do not.

Second, text summaries are increasingly being generated using lab-standardized templates, which leads to high structural similarity among text summaries. This can lead to bias or overfitting if models learn characteristics about these templates which are correlated with a specific classification, rather than learning about relevant evidence. Pre-processing approaches must meet the challenge of filtering these highly predictive sentence structures. These templates are not uniform across labs, but increasingly VCEPs are recording structured evidence types with reports from expert reviews, which should help mitigate this issue more generally.

Finally, while the accuracy of our classifications is strong, these models are more suited to variant prioritization for expert review. These models have not been calibrated to measure the strength of evidence provided for variant classification. Given that these models were trained on many forms of clinical evidence, it complicates their calibration and use following the existing classification guidelines. Model performance is contingent on the quality and representativeness of the clinical reports available within ClinVar. Class imbalance, particularly in the B/LB category, poses a challenge that must continue to be addressed. Continual improvement of model architectures, training strategies, and dataset diversity is needed to enhance model robustness and generalizability.

4 Methods

We parsed and extracted 2.1 million plaintext submission summaries from ClinVar[1]. These submission summaries were used by diagnostic labs to describe the evidence used during variant interpretation when submitting a classification to ClinVar [1]. We processed these submission summaries by removing potential class labels from

[1] https://www.ncbi.nlm.nih.gov/clinvar/

each submission summary record and filtering short or duplicated records, as described in the following section. This processing step reduced the number of submission summaries linked to variant assertions to about 1.2 million. We then developed training and testing datasets to fine-tune language models to understand evidence of pathogenicity. Next, we used these models to assign probabilities for each text summary for whether the variant is P/LP (pathogenic or likely pathogenic), B/LB (benign or likely benign), or VUS (variant of uncertain significance). Finally, we validated the classification accuracy of these models using text summary records as well as orthogonal functional screening data.

4.1 ClinVar Dataset Processing

4.1.1 Removing short and duplicated text summaries

We filtered out short submission summaries by setting a threshold of 100 characters to ensure each comment contained sufficient evidence and removed duplicate summaries from the dataset.

4.1.2 Removing highly similar submission summaries and standardizing text

Deduplication of text data was shown to significantly improve model performance, especially for models with a large number of parameters, as this step removed redundancy and thereby increased the diversity of the data [21–24]. Deduplication was frequently performed in an embedding space, where hashed numerical encodings were compared using methods such as MinHashLSH [25], which combined MinHash encoding [26] and the Locality Sensitive Hashing algorithm [27]. Based on n-grams (a contiguous sequence of n characters/tokens from a given sample of text), MinHash provided a technique for quickly estimating the Jaccard similarity between two texts. In our study, we considered the sets of n-grams derived from the raw ClinVar text reports and studied the similarity between the reports and between sentences within each report.

We performed deduplication both at the report level and the sentence level. Moreover, rather than relying on exact repetition searches, researchers had increasingly adopted "fuzzy" deduplication methods [23, 28, 29]. This approach identified "nearly" repeating data by measuring similarity and applying a threshold. Fuzzy deduplication became standard practice in the area of LLMs, as exemplified by its use in models such as GPT-3 [30], Llama [31], Falcon [32], Pangu [33], and the Pile dataset [34]. For our ClinVar model training, we incorporated fuzzy deduplication into our text preprocessing pipeline.

To standardize the text in the extracted ClinVar corpus, we took the following steps: First, we removed non-English words and characters by identifying and eliminating such instances from the corpus. Next, we standardized punctuation and symbols to conform to conventional English usage. Finally, we applied MinHash [35] with a Jaccard similarity threshold of 95% to remove highly similar summaries based on groups of submission labs and genes, where template-based summaries stemmed from. Examples of deduplication results from ClinVar text are provided in Section A (A.1.4).

4.1.3 SentenceClassifier

A ClinVar submission summary is a plaintext clinical report generated by a diagnostic lab that describes how a single variant contributes to a specific phenotype or disorder. These submission text summaries often include one or more purely descriptive sentences, for example, describing the gene, location, and type of variant in the report, rather than the evidence that might be used to classify whether it is pathogenic or benign. Additionally, many text summaries include a conclusion sentence that summarizes the evidence described and an assertion about the variant's pathogenicity, which maps to one of our three class labels (B/LB, VUS, and P/LP). Given our goal is to train the model to understand text representations of *evidence* that indicate a variant is more likely to be pathogenic, benign, or uncertain, these conclusions could lead to bias or overfitting. To address this potential bias or overfitting, we fine-tuned a BERT [10] model with our labeled data to train a sentence classifier for labeling sentences in submission summaries as **description**, **evidence**, and **conclusion**. Examples of our labeled data are provided in Section A (A.1.2).

We then used the NLTK sentence tokenizer [36] to split each submission summary into individual sentences. We employed the sentence classifier to identify sentences labeled as `description` and `conclusion` by the model and removed them from each submission summary (see examples in Section A, A.1.4) to reduce the likelihood that our dataset contains assertions of variant pathogenicity.

4.1.4 Model training data processing

Finally, we sampled a subset of the ClinVar dataset to create training and testing datasets. Given the imbalanced distribution of B/LB submission summaries relative to P/LP and VUS summaries (22k B/LB, 216k P/LP, and 682k VUS), we sampled all B/LB summaries and maintained a 1:2:2 ratio for B/LB, P/LP, and VUS summaries. This corpus was split with 80% used for training and 20% for testing the model.

4.2 Fine-Tuning BERT Models for Sequence Classification

Upon obtaining training and testing data, we defined our approach as a sequence classification task. Specifically, we input a submission summary from ClinVar into a language model and task the model with predicting whether the variant is P/LP (pathogenic or likely pathogenic), B/LB (benign or likely benign), or VUS (variants of uncertain significance). We evaluated multiple BERT-based transformer models [37], including BERT [10], RoBERTa [11], BioBERT [9], ScholarBERT [13], and ClinicalBERT [12]. BERT and RoBERTa are general-domain models, whereas BioBERT, ScholarBERT, and ClinicalBERT are domain-specific models pre-trained on large biomedical or clinical text corpora. Configurations and detailed training setups are discussed in Section A (A.2).

4.3 Evaluation with Experimental Functional Screening Data

4.3.1 Deep Mutational Scanning (DMS) dataset construction

To validate the accuracy and generalizability of our fine-tuned models, we evaluated their performance on separately generated experimental screening data. This dataset includes estimates of functional impact for genetic variants derived from high-quality experimental assays in genes with established clinical significance: *BRCA1*, *HRAS*, *LDLR*, *PTEN*, and *TP53*. Submission summaries for variants in these genes are excluded from the training and testing data. The DMS dataset was constructed by matching ClinVar submission summaries to variants with corresponding functional assay scores. These scores, downloaded from MaveDB [14], were processed using the FUSE optimization pipeline [16] and paired with ClinVar submission summaries at the amino acid substitution level.

4.3.2 Validation of model classifications with DMS functional scores

We then applied our fine-tuned models to make predictions on variants with submission summaries in the DMS dataset, where the model would have three probability scores $P(B/LB)$, $P(P/LP)$, and $P(VUS)$ for each prediction. We then normalized prediction scores for B/LB and P/LP by $P(\text{B/LB}) = \frac{P(\text{P/LP})}{P(\text{B/LB})+P(\text{P/LP})}$, $P(\text{P/LP}) = \frac{P(\text{P/LP})}{P(\text{B/LB})+P(\text{P/LP})}$ so that there is no overlap between the prediction scores for each group.

Then we recalibrated our predicted class label frequencies using the expected proportions of impacts from single nucleotide VUS from a well-established functional assay in *BRCA1* [17]. There are three predicted impact types from this functional assay: 'LOF' or loss-of-function which is equivalent to our P/LP class label, 'INT' or intermediate, which is equivalent to our VUS class label, and 'FUNC' or functional, which is equivalent to our B/LB class label. We matched the frequency of variants exactly in proportion to our recalibrated class labels: variant scores that are in the top 21.1 percentile are assigned P/LP, variants below the 72.5 percentile are assigned B/LB, and all remaining variants are assigned VUS. Finally, with our recalibrated prediction labels, we performed Mann-Whitney U tests [38] to assess the statistical significance and median shifts between groups predicted as P/LP or B/LB based on their functional assay scores.

4.4 Model Attention Weights Visualization

Using the chosen preprocessing criteria and model, we generated predictions on the held-aside dataset. To analyze whether our ClinVar-BERT model focuses on words and phrases indicative of pathogenicity or benignity, we examined its attention weights. With the model's attention weights and tokenized ClinVar summaries, we utilized Ecco [19] to analyze attention 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 attention weights 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.

4.4.1 Selection of summaries for attention weights visualization case review

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].

4.5 Submission Summary Embedding UMAP Visualization

To visualize and compare the embeddings from the pre-trained BioBERT [9] 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 '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, so that we could conduct a comparative analysis of the embedding spaces before and after ClinVar-specific training. We then employed Uniform Manifold Approximation and Projection (UMAP) [39] 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).

5 Conclusion

By training language models to discern evidence of pathogenicity from unstructured clinical text, we have introduced a novel approach to prioritize variants for expert review. This research promises to allow clinicians to more readily make use of expert-curated information that is currently prohibitively complicated to use at scale. This information should enable expert panels to classify a larger proportion of variants as pathogenic or benign, allowing more patients to learn about clinically actionable findings. This could advance genomic medicine to the large number of patients who collectively harbor a VUS, potentially improving their clinical management.

Appendix A Supplementary Methods

A.0.1 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 such that it 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.

A.1 Training a Sentence Classifier

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.

A.1.1 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 [40–43]. 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"

A.1.2 Examples of labeled conclusion vs evidence data

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 `evidence` and `conclusion`.

Example Labeled Data

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 signi ficance 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.

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.

A.1.3 Examples of submission summaries before vs. after removing the conclusion-labeled sentence

Example summaries

Before
The c.1329delT pathogenic mutation, located in coding exon 5 of the BARD1 gene, results from a deletion of one nucleotide at nucleotide position 1329, causing a translational frameshift with a predicted alternate stop codon (p.V444Lfs*31). This alteration is expected to result in loss of function by premature protein truncation or nonsense-mediated mRNA decay. **As such, this alteration is interpreted as a disease-causing mutation.**

After
The c.1329delT pathogenic mutation, located in coding exon 5 of the BARD1 gene, results from a deletion of one nucleotide at nucleotide position 1329, causing a translational frameshift with a predicted alternate stop codon (p.V444Lfs*31). This alteration is expected to result in loss of function by premature protein truncation or nonsense-mediated mRNA decay.

With the sentence classifier, the conclusion sentence containing a proxy of the classification label "disease-causing mutation" is removed.

A.1.4 Examples of highly similar submission summaries

With a threshold of **0.95**, the following submission summaries are identified as highly similar by MinHash [35]

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)...

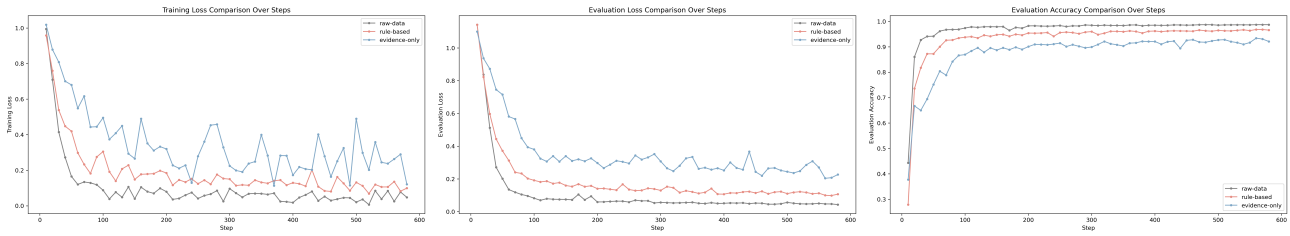
As we can see from the example above, with a threshold of 0.95 using MinHash, these submission summaries are identified as highly similar, and we can notice that these summaries are highly template-based and the only difference between each of them is the amino acid and codon information mentioned, and 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.

A.2 Fine-Tuning Language Models with ClinVar Dataset

A.2.1 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. A1: Comparison of training Loss, evaluation Loss, and accuracy during training among three text processing methods

A.2.2 ClinVar Summary Ablation Study

Removing the sentence changes the prediction label from P/LP to VUS


```
{
    "Comment": "This sequence change replaces arginine, which is basic
    and polar, with glutamine, which is neutral and polar, at codon 495
    of the MYBPC3 protein (p.Arg495Gln). This variant is present in
    population databases (rs200411226, gnomAD 0.006\%). This missense
    change has been observed in individuals with hypertrophic
    cardiomyopathy (PMID: 11499718, 20019025, 22857948, 23396983,
    24093860). ClinVar contains an entry for this variant (Variation ID:
    164113). Algorithms developed to predict the effect of missense
    changes on protein structure and function (SIFT, PolyPhen-2, Align
    GVGD) all suggest that this variant is likely to be disruptive. This
    variant disrupts the p.Arg495 amino acid residue in MYBPC3. Other
    variant(s) that disrupt this residue have been determined to be
    pathogenic (PMID: 18403758, 19659763, 20624503). This suggests that
    this residue is clinically significant, and that variants that
    disrupt this residue are likely to be disease-causing.",
    "removed_sentence": "This missense change has been observed in
    individuals with hypertrophic cardiomyopathy (PMID: 11499718,
    20019025, 22857948, 23396983, 24093860).",
    "Original_Prediction": "P/LP",
    "Ablated_Prediction": "VUS",
    "Prediction_Difference": "Changed",
    "SCV": "SCV000218744",
    "Submitter": "Invitae",
    "Gene": "MYBPC3",
    "ground_truth_classification": "P/LP",
    "prediction_labels_ft": "P/LP",
    "prediction_scores_ft": [
        0.99008584,
        0.009653065,
        0.00026108805
    ],
    "Ablated_Prediction_Scores": [
        8.087950845947489e-05,
        0.9998732805252075,
        4.587376315612346e-05
    ],
    "Score_Difference": 0.9900049604915405
}
```


VUS to P/LP Influential Sentences Examples

- This sequence change creates a premature translational stop signal (p.Arg494*) in the EGF gene.
- This sequence change disrupts the translational stop signal of the FANCM mRNA.
- This sequence change affects an acceptor splice site in intron 5 of the TBX20 gene.
- The V462I variant in the PCCB gene has not been reported previously as a pathogenic variant, nor as a benign variant, to our knowledge.
- This variant results in a copy number gain of the genomic region encompassing exon(s) 50-57 of the NF1 gene.

VUS to B/LB Influential Sentences Examples

- This sequence change replaces aspartic acid, which is acidic and polar, with glycine, which is neutral and non-polar, at codon 650 of the BBS9 protein (p.Asp650Gly).
- This sequence change replaces arginine, which is basic and polar, with threonine, which is neutral and polar, at codon 429 of the PIGV protein (p.Arg429Thr).
- This sequence change replaces aspartic acid, which is acidic and polar, with glycine, which is neutral and non-polar, at codon 435 of the SMCHD1 protein (p.Asp435Gly).
- This sequence change replaces arginine, which is basic and polar, with cysteine, which is neutral and slightly polar, at codon 261 of the DHX32 protein (p.Arg261Cys).
- This sequence change replaces threonine, which is neutral and polar, with proline, which is neutral and non-polar, at codon 249 of the DSC2 protein (p.Thr249Pro).

A.3 Analysis

A.3.1 Attention Weights Visualization Configurations

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 `'##'`.

Appendix B Supplementary Results

B.0.1 Test Data Evaluation Results

Fine-tuned and pre-trained model performances on test data from *evidence-only* processed ClinVar dataset.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
BioBERT-large [9] + ClinVar	0.9754	0.9756	0.9730	0.9743	0.9982
ScholarBERT [13] + ClinVar	0.9720	0.9721	0.9702	0.9711	0.9974
ClinicalBERT [12] + ClinVar	0.9666	0.9685	0.9621	0.9651	0.9976
RoBERTa-base [11] + ClinVar	0.9717	0.9734	0.9669	0.9700	0.9979
RoBERTa-large [11]+ ClinVar	0.9729	0.9728	0.9705	0.9716	0.9978
BERT-Base [10] + ClinVar	0.9729	0.9724	0.9715	0.9720	0.9979
BERT-Large [10] + 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 B1: Performances of fine-tuned BERT models compared to pre-trained models on ClinVar **raw-data** test data.

B.0.2 DMS Evaluation Results

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

Table B2: Evaluation results of fine-tuned models trained with **evidence-only** dataset on orthogonally generated DMS Data

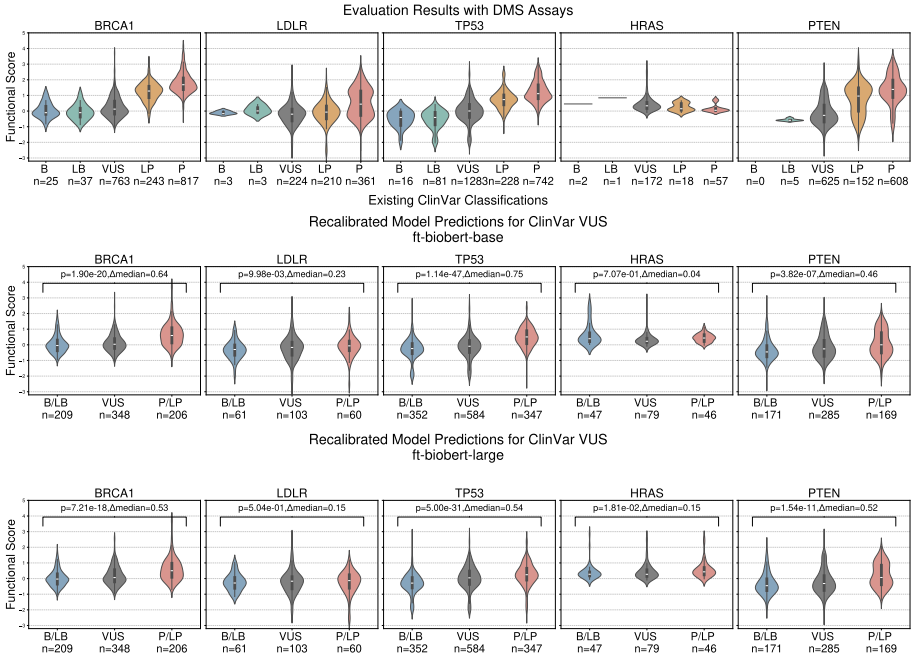


Fig. B2: Comparison of DMS validation results of **evidence-only** trained BioBERT-base and BioBERT-large models on orthogonally generated DMS data

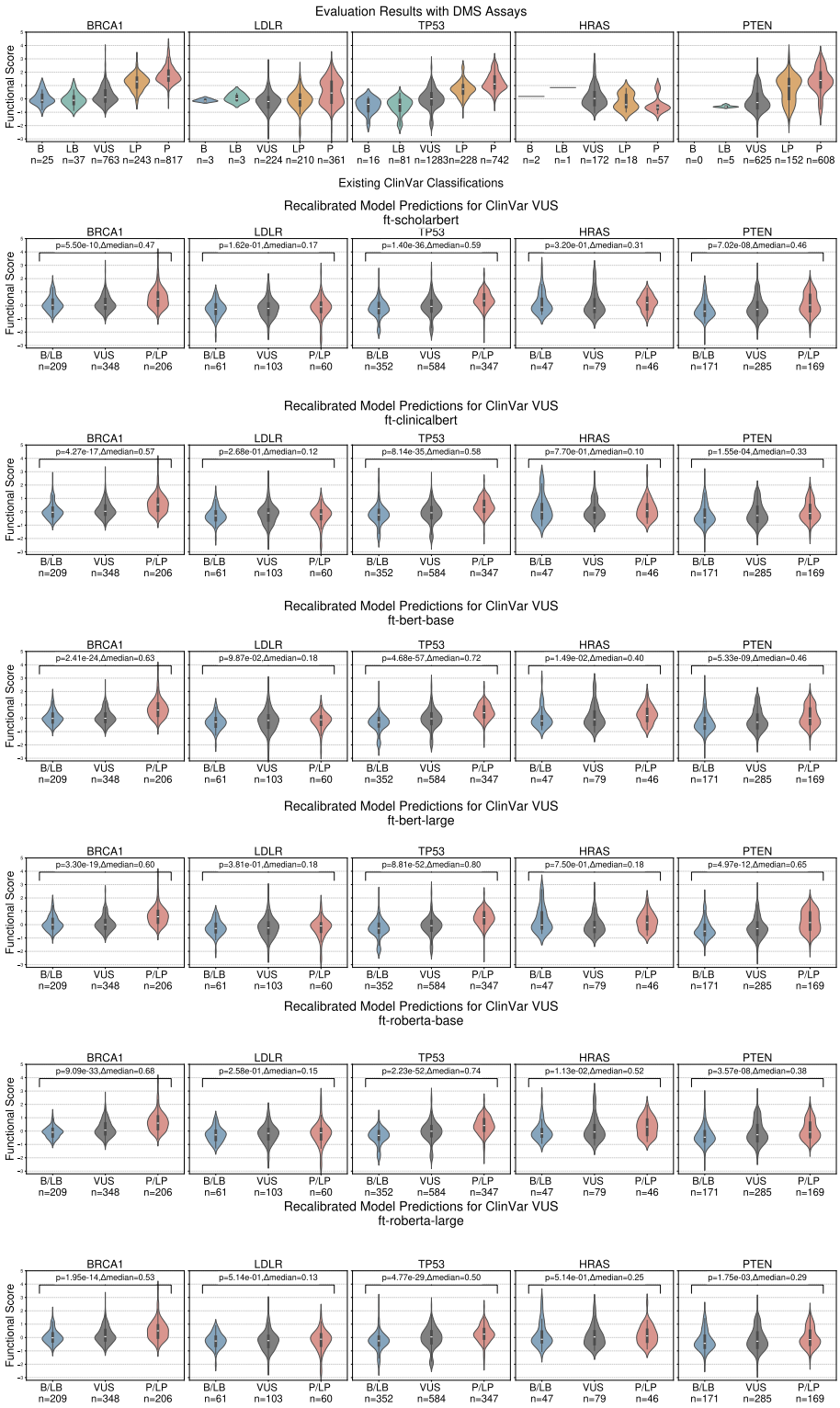


Fig. B3: Comparison of **evidence-only** trained DMS validation results of multiple BERT-architecture language models

References

[1] Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., *et al.*: Clinvar: public archive of interpretations of clinically relevant variants. Nucleic acids research **44**(D1), 862–868 (2016)

[2] Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., *et al.*: Clingen—the clinical genome resource. New England Journal of Medicine **372**(23), 2235–2242 (2015)

[3] Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.*: Analysis of protein-coding genetic variation in 60,706 humans. Nature **536**(7616), 285–291 (2016)

[4] Richards, S., Aziz, N., Bale, S., *et al.*: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. Genetics in Medicine **17**, 405–423 (2015) https://doi.org/10.1038/gim.2015.30

[5] Murray, M.F., Giovanni, M.A., Doyle, D.L., Harrison, S.M., Lyon, E., Manickam, K., Monaghan, K.G., Rasmussen, S.A., Scheuner, M.T., Palomaki, G.E., *et al.*: Dna-based screening and population health: a points to consider statement for programs and sponsoring organizations from the american college of medical genetics and genomics (acmg). Genetics in Medicine **23**(6), 989–995 (2021)

[6] Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., *et al.*: Strategic vision for improving human health at the forefront of genomics. Nature **586**(7831), 683–692 (2020)

[7] Fife, J.D., Tran, T., Bernatchez, J.R., Shepard, K.E., Koch, C., Patel, A.P., Fahed, A.C., Krishnamurthy, S., Center, R.G., Collaboration, D., et al.: A framework for integrated clinical risk assessment using population sequencing data. medRxiv, 2021–08 (2021)

[8] Bennett, G., Karbassi, I., Chen, W., Harrison, S.M., Lebo, M.S., Meng, L., Nagan, N., Rigobello, R., Rehm, H.L.: Distinct rates of vus reclassification are observed when subclassifying vus by evidence level. medRxiv, 2024–11 (2024)

[9] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[12] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)

[13] Hong, Z., Ajith, A., Pauloski, G., Duede, E., Chard, K., Foster, I.: The Diminishing Returns of Masked Language Models to Science (2023)

[14] Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T., Roth, F.P., Fowler, D.M., Rubin, A.F.: Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome biology **20**, 1–11 (2019)

[15] Rubin, A.F., Min, J.K., Rollins, N.J., Da, E.Y., Esposito, D., Harrington, M., Stone, J., Bianchi, A.H., Dias, M., Frazer, J., et al.: Mavedb v2: a curated community database with over three million variant effects from multiplexed functional assays. Biorxiv, 2021–11 (2021)

[16] Yu, T., Fife, J.D., Bhat, V., Adzhubey, I., Sherwood, R., Cassa, C.A.: Fuse: Improving the estimation and imputation of variant impacts in functional screening. Cell Genomics **4**(10) (2024)

[17] Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., Shendure, J.: Accurate classification of brca1 variants with saturation genome editing. Nature **562**(7726), 217–222 (2018)

[18] Ryu, J., Barkal, S., Yu, T., Jankowiak, M., Zhou, Y., Francoeur, M., Phan, Q.V., Li, Z., Tognon, M., Brown, L., et al.: Joint genotypic and phenotypic outcome modeling improves base editing variant effect quantification. Nature Genetics, 1–13 (2024)

[19] Alammar, J.: Ecco: An open source library for the explainability of transformer language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Association for Computational Linguistics, ??? (2021)

[20] McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints (2018) arXiv:1802.03426 [stat.ML]

[21] Allamanis, M.: The adverse effects of code duplication in machine learning models of code. In: Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software, pp. 143–153 (2019)

[22] Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., Carlini, N.: Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499 (2021)

[23] Abbas, A., Tirumala, K., Simig, D., Ganguli, S., Morcos, A.S.: Semdedup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540 (2023)

[24] Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., Zhang, C.: Quantifying memorization across neural language models. arXiv preprint arXiv:2202.07646 (2022)

[25] Ekzhu: DataSketch: MinHash LSH. Accessed: 2024-09-15 (2021). https://ekzhu.com/datasketch/lsh.html

[26] Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), pp. 21–29 (1997). IEEE

[27] Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 253–262 (2004)

[28] Jiang, T., Yuan, X., Chen, Y., Cheng, K., Wang, L., Chen, X., Ma, J.: Fuzzydedup: Secure fuzzy deduplication for cloud storage. IEEE Transactions on Dependable and Secure Computing (2022)

[29] Tirumala, K., Simig, D., Aghajanyan, A., Morcos, A.: D4: Improving llm pretraining via document de-duplication and diversification. Advances in Neural Information Processing Systems **36** (2024)

[30] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[31] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[32] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023)

[33] Zeng, W., Ren, X., Su, T., Wang, H., Liao, Y., Wang, Z., Jiang, X., Yang, Z., Wang, K., Zhang, X., et al.: Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. arXiv preprint arXiv:2104.12369 (2021)

[34] Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027

(2020)

[35] Broder, A.Z.: Identifying and filtering near-duplicate documents. In: Annual Symposium on Combinatorial Pattern Matching, pp. 1–10 (2000). Springer

[36] Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)

[37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[38] MacFarland, T.W., Yates, J.M., MacFarland, T.W., Yates, J.M.: Mann–whitney u test. Introduction to nonparametric statistics for the biological sciences using R, 103–132 (2016)

[39] McInnes, L., Healy, J., Saul, N., Grossberger, L.: Umap: Uniform manifold approximation and projection. The Journal of Open Source Software **3**(29), 861 (2018)

[40] Wettig, A., Gupta, A., Malik, S., Chen, D.: Qurating: Selecting high-quality data for training language models. arXiv preprint arXiv:2402.09739 (2024)

[41] Mu, T., Helyar, A., Heidecke, J., Achiam, J., Vallone, A., Kivlichan, I., Lin, M., Beutel, A., Schulman, J., Weng, L.: Rule based rewards for language model safety

[42] Yuan, W., Pang, R.Y., Cho, K., Sukhbaatar, S., Xu, J., Weston, J.: Self-rewarding language models. arXiv preprint arXiv:2401.10020 (2024)

[43] Li, X., Gao, M., Zhang, Z., Yue, C., Hu, H.: Rule-based Data Selection for Large Language Models (2024). https://arxiv.org/abs/2410.04715