Multi-omic Variant Effect Prediction with Evo 2 and ProGen2

Siddharth Sabata[1] [1]Carnegie Mellon University, Department of Biological Sciences
ssabata@cs.cmu.edu
siddsabata/multi-modal-var

Abstract

Accurate prediction of pathogenic genetic variants remains central to human genetics. Existing tools often treat genomic and protein information separately or rely on hand-crafted features. We explore whether multi-omic embeddings from large DNA and protein language models can improve variant effect prediction when fused carefully. We build a pipeline that extracts DNA sequence context and corresponding protein sequences for each ClinVar single-nucleotide variant (SNV), then encodes them using Evo2-7B and ProGen2. We represent each variant using delta embeddings (variant minus reference) at both DNA and protein levels and train classifiers on DNA-only, protein-only, and multimodal feature sets.

1. Introduction

This project originally aimed to explore multi-modal disease prediction using genomic and clinical foundation models to predict gout in UK Biobank patients, in collaboration with Sedona Health. Due to prolonged approval processes and unexpected data maintenance, we pivoted to variant effect prediction using publicly available datasets while maintaining our focus on multi-modal foundation model approaches in biomedicine.

Interpreting the pathogenicity of genetic variants remains a fundamental challenge in human genetics and precision medicine. High-throughput sequencing has cataloged millions of variants in databases like ClinVar. We focus on single-nucleotide variants (SNVs) from ClinVar as our dataset [5].

Recent biological foundation models offer promising solutions. DNA language models like Evo 2 learn contextual representations of genomic regions, while protein language models like ProGen2 capture rich amino acid sequence representations that correlate with structure and function [1][6].

We investigate: Can we improve pathogenicity prediction by jointly leveraging DNA and protein foundation model embeddings?

Using ClinVar missense SNVs, we evaluate this question with two pretrained models: Evo2-7B for DNA and ProGen2-small for protein. For each variant, we construct delta embeddings capturing the representational change between wild-type and mutant sequences, then evaluate multiple multimodal fusion strategies for binary pathogenicity prediction.

1.1. Contributions

1. **Multimodal embedding dataset for ClinVar missense variants.** We compiled 23,077 initial SNVs from ClinVar, yielding 15,727 variants after filtering. Each variant includes DNA wild-type and mutant context windows embedded by Evo2-7B, protein wild-type and mutant sequences embedded by ProGen2-small, and binary pathogenicity labels.
2. **Systematic comparison of unimodal and multimodal models.** We compare DNA-only (Evo2), protein-only (ProGen2), and multiple fusion strategies including naive concatenation, scaled concatenation, canonical correlation analysis (CCA), feature-selected fusion, and weighted fusion. We test two classifier families: logistic regression and XGBoost [3][2].
3. **Practical insights for multimodal variant effect modeling.** We highlight the importance of per-modality scaling, feature selection, and explicit down-weighting when modalities have very different signal-to-noise profiles.

2. Related Work

2.1. Variant effect prediction

The deep learning revolution has produced numerous biological foundation models showing promise for DNA, RNA, and protein modeling. However, their utility for variant effect prediction remains unclear, particularly given high computational costs, strong results often require massive models. Learning to effectively leverage these models' generalizability for variant effect prediction could significantly advance personalized medicine [4].

2.2. ProGen2

ProGen2 is a family of protein language models trained on over a billion sequences from genomic, metagenomic, and immune-repertoire databases. These models achieve state-of-the-art performance in modeling evolutionary distributions, generating functional proteins, and zero-shot fitness prediction. Like natural language models, ProGen2's success comes from data diversity and scale. Its rich contextual embeddings for wild-type and mutant sequences make it ideal for variant effect prediction and complement DNA-level representations in multimodal pipelines. Specifically we utilize ProGen2-small, the 151-million parameter version [6].

2.3. Evo 2

Evo 2 is a biological foundation model trained on 9.3 trillion nucleotides with a one-million-token context window. It learns genomic patterns directly from sequence data (splice sites, transcription factor motifs, regulatory signals) and achieves strong zero-shot performance predicting variant effects, from noncoding mutations to BRCA1 missense variants. Its single-nucleotide resolution makes it powerful for variant effect prediction. We utilize the 7-billion parameter version of Evo2 in the project [1].

2.4. Multi-modal approaches

Some approaches have been taken to tackle this problem. ModVAR integrates DNA sequences (DNABERT2), predicted protein structures (ESMFold), and cancer omics data to identify somatic driver mutations, achieving strong accuracy on validated drivers and enabling therapeutic prioritization. While both ModVAR and our work use multimodal representations for variant prediction, they differ fundamentally: ModVAR targets somatic drivers using structure and cancer-specific data, whereas we predict germline pathogenicity using only sequence-based embeddings from Evo2 and ProGen2 [7].

3. Methods

3.1. Problem Definition

Given a set of ClinVar missense SNVs, each with a binary label $y \in \{0, 1\}$ (where 0 denotes benign or likely benign and 1 denotes pathogenic or likely pathogenic), we aim to learn a classifier $f : (\Delta z_{\text{DNA}}, \Delta z_{\text{Prot}}) \mapsto \hat{y}$. Here, Δz_{DNA} is a DNA delta embedding derived from Evo2-7B, and Δz_{Prot} is a protein delta embedding derived from ProGen2-small.

We compare three categories of models: a DNA-only model $f(\Delta z_{\text{DNA}})$, a protein-only model $f(\Delta z_{\text{Prot}})$, and multimodal models $f([\Delta z_{\text{DNA}}; \Delta z_{\text{Prot}}])$ under different fusion strategies. Our primary evaluation metric is ROC-AUC, with PR-AUC and standard classification metrics (accuracy, precision, recall, F1-score) serving as secondary measures.

3.2. Data and preprocessing

3.2.1. Data sources and filtering

We obtained single nucleotide variants from ClinVar, using the GRCh38 human genome assembly as our reference genome and UniProt protein sequences for protein-level analysis. Starting from an initial set of 23,077 ClinVar variants, we applied a series of filtering steps to ensure data quality. First, we filtered to missense SNVs with clear pathogenicity labels (pathogenic or likely pathogenic versus benign or likely benign). We then removed variants with uncertain or conflicting labels. Finally, we removed variants where protein embedding extraction failed, indicated by zero vectors, leaving us with a final dataset of 15,727 variants. These failed protein embeddings were due to amino-acid sequences longer than ProGen's context window.

ClinVar provides multiple fields for clinical significance, each encoding pathogenicity at different levels of granularity and curation. In this study, we use `ClinSigSimple`, an integer-valued field defined as follows: a value of 1 indicates that at least one current submission interprets the variant as pathogenic, likely pathogenic, risk allele, or low-penetrance pathogenic; a value of 0 indicates that no submissions classify the variant as pathogenic or risk; and a value of -1 indicates that no meaningful clinical significance records exist (these were removed during preprocessing).

The final dataset exhibits a notable class imbalance, with 4,868 benign variants (31%) and 10,859 pathogenic variants (69%), corresponding to an approximate 2.23:1 ratio of pathogenic to benign cases. This imbalance is handled through appropriate weighting in our classification models.

3.2.2. *DNA delta embeddings (Evo2-7B)*

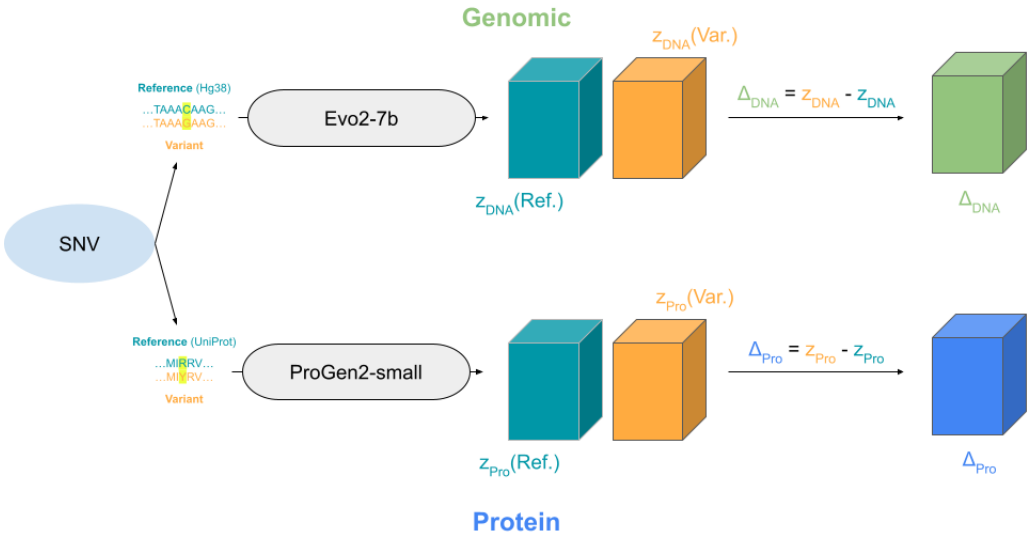

Figure 1 | Embedding construction pipeline for ClinVar missense variants. Starting with 23,077 ClinVar SNVs, we filtered to 15,727 variants with valid embeddings and binary labels (31% benign, 69% pathogenic). For each variant, we extracted 512 bp of flanking genomic context (GRCh38) and obtained wild-type (reference) and mutant (variant) protein sequences from UniProt. DNA embeddings were generated using Evo2-7B: we processed both reference and alternate allele sequences through the model, extracted hidden states from an intermediate MLP layer, and mean-pooled to create 4096-dimensional embeddings. Protein embeddings were generated using ProGen2-small: we tokenized wild-type and mutant sequences, extracted representations from the final transformer layer, and mean-pooled to create 1024-dimensional embeddings. For both modalities, we computed delta embeddings ($\Delta_z = z_{var} - z_{ref}$) to capture the functional impact of each variant. Variants with uncertain clinical labels were excluded from the final dataset.

For each variant, we construct two input sequences: a wild-type DNA sequence containing the reference allele at the center position, and a mutant DNA sequence with the alternate allele at the same position. Each sequence spans a 512 bp flanking region on either side of the SNV, resulting in a total window length of 1025 bp.

To extract embeddings, we perform a forward pass of each sequence through Evo2 and extract hidden states from an intermediate MLP layer (specifically, `blocks.28.mlp.l3`). We then mean-pool these hidden states across all sequence positions to obtain a single 4096-dimensional vector per sequence. The delta embedding is computed as $\Delta z_{\text{DNA}} = z^{\text{wt}}_{\text{DNA}} - z^{\text{alt}}_{\text{DNA}} \in \mathbb{R}^{4096}$, intended to capture how Evo2's representation changes when the nucleotide is mutated.

3.2.3. Protein delta embeddings (ProGen2-small)

For each variant, we similarly construct wild-type and mutant protein sequences. The wild-type protein is the full amino-acid sequence from UniProt, matched via gene or protein ID. The mutant protein is the same sequence with the single amino-acid substitution applied at the appropriate position corresponding to the SNV.

We tokenize the amino acid sequences using ProGen2's vocabulary and perform forward passes of both wild-type and mutant sequences through the transformer. We extract final layer hidden states and mean-pool across all sequence tokens to obtain fixed-dimensional representations. The delta embedding is computed as $\Delta z_{\text{Prot}} = z_{\text{Prot}}^{\text{wt}} - z_{\text{Prot}}^{\text{mut}} \in \mathbb{R}^{\leq 1024}$, capturing how ProGen2's representation of the protein changes under the mutation.

3.3. Prediction models

We evaluate two classifier families to assess the predictive power of our embeddings. First, we use L2-regularized logistic regression with balanced class weights (to compensate for the 2.23:1 class imbalance), the LBFGS solver, and a maximum of 1000 iterations. This model serves as a simple, fast, and interpretable linear baseline.

Second, we use XGBoost, a gradient boosted decision tree classifier, with 300 estimators, maximum depth of 6, and learning rate of 0.05. We set the scale positive weight parameter to 2.23 to reflect the class imbalance and use subsampling rates of 0.8 for both samples and features. This model is motivated by its ability to capture non-linear interactions and its strong performance on tabular feature sets.

3.4. Fusion strategies and research questions

We explore multiple strategies for combining DNA and protein embeddings. The DNA-only baseline applies a classifier directly to Δz_{DNA} after dimensionality reduction. The protein-only baseline applies a classifier to Δz_{Prot}. For multimodal fusion, we evaluate several approaches: naive concatenation, where raw DNA and protein embeddings are concatenated and then reduced via PCA; scaled concatenation, where per-modality StandardScaler is applied before concatenation and PCA; CCA fusion, where each modality is first reduced via PCA, then Canonical Correlation Analysis is applied to find correlated components; feature-selection fusion, where SelectKBest with mutual information is used to select informative DNA features before fusion with full protein features; and weighted fusion, where DNA features are scaled by a weight $w \in [0, 1]$ before concatenation with protein features.

These experiments are designed to address four key research questions. First, how predictive are Evo2 DNA delta embeddings by themselves for ClinVar missense pathogenicity? Second, does adding DNA embeddings to ProGen2 protein embeddings improve performance, or does it simply add noise? Third, can careful fusion strategies such as feature selection, weighting, or CCA recover or surpass the protein-only baseline? Finally, what do these results tell us about the representations learned by Evo2 and ProGen2, and about multimodal modeling in this domain?

4. Experiments

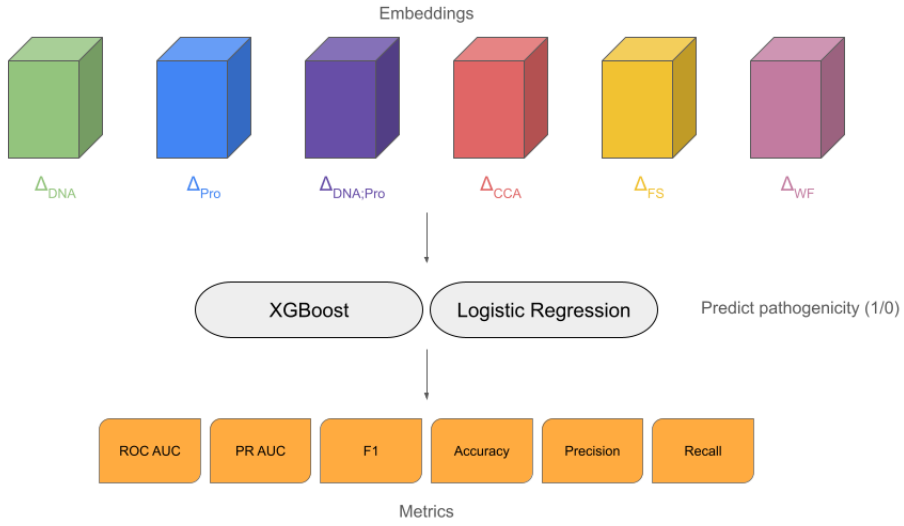

Figure 2 | Pathogenicity prediction task and multimodal fusion strategies. We framed variant classification as a binary task (pathogenic vs benign) using an 80/20 train-test split with 5-fold cross-validation on the training set. Two base classifiers were evaluated: L2-regularized logistic regression (baseline) and XGBoost (n_estimators=300, max_depth=6), both configured to handle the 2.23:1 class imbalance. We compared seven fusion approaches for combining DNA and protein embeddings: (1) DNA-only with PCA dimensionality reduction to 95% variance (dim=15), (2) Protein-only with PCA (dim=28), (3) Concat-Unscaled naïve concatenation that suffers from magnitude imbalance (DNA features are 170× larger than protein, causing PCA to ignore protein signal; dim=15), (4) Concat-Scaled with per-modality standardization before concatenation and PCA (dim=65), (5) CCA Fusion using canonical correlation analysis to project each modality to 20 components before concatenating (dim=40), (6) Feature Selection (FS) applying mutual information to retain only the top K most informative DNA features (K = 500) before combining with full protein features (dim=152), and (7) Weighted Fusion (WF) down-weighting standardized DNA features by a factor w = 0.3 before concatenation (dim=170). All models were evaluated using ROC-AUC (primary), PR-AUC, accuracy, precision, recall, and F1 score.

4.1. Experimental setup

We apply dimensionality reduction to both modalities before classification. Evo2 DNA delta embeddings are 4096-dimensional, and we apply PCA to retain 95% of the variance, which yields 15 components in the DNA-only setup. ProGen2 protein delta embeddings are 1024-dimensional, and PCA to 95% variance yields 28 components in the protein-only setup. In multimodal configurations, PCA is applied after any scaling, selection, or weighting operations and after concatenation.

Our evaluation protocol uses an 80/20 holdout split stratified by label, resulting in 12,581 training variants and 3,146 test variants. We perform 5-fold stratified cross-validation within the training set to tune and evaluate models. To prevent data leakage, we fit all feature selection,

scalers, and PCA transformations only on each CV training fold and apply the learned transforms to the corresponding validation fold. We report ROC-AUC as the primary metric, with PR-AUC, accuracy, precision, recall, and F1-score as secondary metrics. Mean ROC-AUC across CV folds is used for model comparison, and final test results broadly match CV trends.

4.2. Baseline results

We begin by evaluating DNA-only and protein-only baselines to establish the individual predictive power of each modality. The DNA-only configuration uses Evo 2 DNA delta embeddings reduced to 15 components via PCA. Logistic regression achieves a ROC-AUC of approximately 0.69, and XGBoost achieves 0.69. These results indicate that Evo 2 DNA embeddings alone provide some signal for missense pathogenicity prediction but are substantially weaker than protein embeddings.

The protein-only configuration uses ProGen2 protein delta embeddings reduced to 28 components via PCA. Logistic regression achieves a ROC-AUC of approximately 0.56, while XGBoost reaches 0.77, the best overall baseline performance. These results demonstrate that protein mutations captured by ProGen2 are strong predictors of ClinVar pathogenicity with non-linear modeling approaches.

We next evaluate naive multimodal fusion by directly concatenating raw DNA and protein embeddings. The concatenated representation is 5120-dimensional and is reduced to 15 components via PCA. Surprisingly, this configuration achieves a ROC-AUC of only 0.69 with logistic regression and 0.69 with XGBoost, essentially identical to the DNA-only baseline. Analysis reveals that Evo2 features have much larger magnitude than ProGen2 features (mean approximately 1.236 versus 0.0073). As a result, PCA collapses onto the high-variance DNA subspace and effectively ignores the protein signal. Naive concatenation therefore behaves like DNA-only prediction.

To address this issue, we apply per-modality scaling in the scaled concatenation configuration. We apply StandardScaler to each modality separately to achieve zero mean and unit variance per feature, then concatenate and apply PCA to retain 95% variance, yielding 65 components. This configuration achieves a ROC-AUC of 0.70 with logistic regression and 0.73 with XGBoost. Scaling fixes the PCA collapse and allows protein features to contribute, but performance remains approximately 5.5% below the protein-only XGBoost baseline.

We also evaluate CCA fusion as an alternative multimodal integration strategy. We first reduce DNA embeddings from 4096 to 30 dimensions and protein embeddings from 1024 to 30 dimensions using PCA. We then apply Canonical Correlation Analysis to extract 20 canonical components per modality and concatenate them to form a 40-dimensional representation. Logistic regression achieves a ROC-AUC of 0.69, and XGBoost achieves 0.76. CCA identifies correlated patterns between DNA and protein spaces but still slightly underperforms the protein-only baseline by approximately 1% ROC-AUC.

4.3. Advanced fusion strategies

Given that naive and scaled fusion strategies fail to match protein-only performance, we evaluate two strategies to address this issue: feature selection on DNA embeddings and weighted fusion. We only implemented these strategies on XGBoost due to its superior performance from baseline results.

For feature selection, we use mutual information between each DNA feature and the pathogenicity label to select the top K DNA features, where $K \in \{100, 250, 500, 1000, 2000\}$. We find that $K = 500$ yields the best results. In this configuration, we select 500 features from the 4096-dimensional DNA embeddings, apply StandardScaler from SK-Learn, and concatenate with the full 1024-dimensional protein features. PCA to 95% variance yields 152 components. With $K = 500$, XGBoost achieves 0.76. This is very similar to CCA performance, but just slightly lower.

For weighted fusion, we downweight DNA features relative to protein features before concatenation. We scale DNA features by a weight $w \in \{0.1, 0.2, 0.3, 0.5, 0.7\}$ while keeping protein features at full scale. We find that $w = 0.3$ yields the best results. After scaling both modalities with StandardScaler, we apply the weight and concatenate, then apply PCA to 95% variance, yielding 170 components. With $w = 0.3$, XGBoost achieves 0.76. This approach reduces the effective variance of DNA features by scaling them to approximately 30% of the protein magnitude, preventing them from dominating PCA and the classifier. Performance again reaches approximately 99% of the protein-only baseline.

4.4. Results summary

Table 1 | Cross-validation performance (Logistic Regression).

Metric	DNA	Protein	Concat-Unscaled	Concat-Scaled	CCA
Accuracy	0.5980	0.5438	0.5980	**0.6038**	0.5961
Precision	0.8328	0.7325	**0.8328**	0.8276	0.8249
Recall	0.5228	0.5347	0.5228	**0.5385**	0.5269
F1	0.6422	0.6181	0.6422	**0.6524**	0.6430
ROC_AUC	0.6888	0.5574	0.6888	**0.6951**	0.6861
PR_AUC	0.8279	0.7280	0.8279	**0.8308**	0.8212

Table 2 | Cross-validation performance (XGBoost).

Metric	DNA	Protein	Unscaled	Scaled	CCA	FS-k500	WF-w0.3
Accuracy	0.7006	**0.7295**	0.7006	0.7144	0.7348	0.7315	0.7348
Precision	0.7017	0.7215	0.7022	0.7138	0.7290	0.7253	**0.7288**
Recall	0.9853	**0.9905**	0.9832	0.9790	0.9805	0.9837	0.9810
F1	0.8197	0.8349	0.8193	0.8256	0.8362	0.8350	**0.8361**
ROC_AUC	0.6884	**0.7679**	0.6888	0.7260	0.7620	0.7607	0.7607
PR_AUC	0.8276	0.8609	0.8270	0.8442	**0.8611**	0.8621	0.8610

5. Analysis

5.1. Linear model multimodal gains

Linear models underutilize protein embeddings, revealing multimodal gains. In logistic regression, protein-only embeddings performed poorly (ROC-AUC of 0.56) compared to DNA-only models (ROC-AUC of 0.69), suggesting Evo2 embeddings provide more linearly separable information than ProGen2. However, multimodal fusion (Scaled or CCA) slightly outperformed the DNA baseline, indicating protein embeddings contribute complementary signal in a highly nonlinear form that linear models cannot effectively exploit.

5.2. Small DNA windows

Limited DNA performance likely stems from small sequence windows. The DNA-only ROC-AUC of 0.69 is moderate but below expectations for evolution-aware models. The likely cause is Evo2's small input window (1025 bp with 512 bp flanks), which is tiny relative to its training scale. Evo2 excels at modeling long-range context, largely absent in narrow 1 kb regions. Expanding to 5-10 kb windows could capture regulatory features, likely improving both DNA-only and multimodal performance.

5.3. Small foundation models

Small embedding models limit downstream performance. Smaller models have reduced capacity for encoding structural constraints, evolutionary signals, and functional context. The performance gap and difficulty extracting multimodal gains may partly reflect limited representational power. Larger model checkpoints would likely improve results.

5.4. ProGen2 limitations

ProGen2 may not be optimal for embedding generation due to its decoder-only architecture. Encoder models like ESM could produce richer representations better suited for variant effect prediction.

5.5. Delta embedding distortion

Delta embeddings may distort useful information. All experiments used delta embeddings (mutant minus wild-type), which assumes variant effects are captured purely by embedding shifts. This loses absolute context, linearizes nonlinear effects, and exaggerates magnitude differences between modalities. Providing raw embeddings could improve performance.

5.6. Test-train splitting

In this study, we used random train-test splitting, which could bias classification models toward certain protein clusters or families depending on their distribution across splits. ModVar does not address splitting strategy, making this an important consideration when interpreting performance results.

6. Conclusion

This work explored whether combining DNA and protein foundation model embeddings improves missense pathogenicity prediction. Using Evo2-7B for genomic context and ProGen2-small for protein sequence context, we constructed delta embeddings for 15,727 ClinVar variants and evaluated multiple fusion strategies under linear and nonlinear classifiers. Across experiments, protein embeddings provided the strongest standalone signal, DNA embeddings offered moderate but meaningful predictive value, and multimodal fusion produced marginal improvements in the linear setting but did not surpass the protein-only baseline under nonlinear models. Careful pre-

processing, including per-modality scaling, feature selection, and down-weighting, was necessary to prevent DNA noise from overwhelming the protein signal.

Several directions may strengthen future multimodal performance. First, both Evo2 and ProGen2 were used in relatively small configurations; larger DNA and protein foundation models may produce higher-quality embeddings. Second, Evo2 was evaluated only on short 1 kb windows, likely underutilizing its ability to model long-range genomic context; expanding to larger flanking regions could yield more informative DNA representations. Third, delta embeddings may be overly restrictive; exploring alternatives such as concatenating wild-type and mutant embeddings or using more expressive comparison mechanisms may better capture variant effects. Finally, alternative protein embedding models may yield richer, more informative representations for variant effect prediction. Together, these extensions offer promising avenues for building more effective multimodal variant effect predictors.

7. Limitations

This work has several limitations. First, we rely directly on ClinVar labels for pathogenicity, which may contain noise, inconsistencies, or conflicting submissions; additional curation could improve label quality. Second, both Evo2-7B and ProGen2-small are relatively small biological foundation models compared to current state-of-the-art, which likely limits the representational strength of the embeddings. Third, our downstream classifiers are lightweight models (logistic regression and XGBoost) rather than deeper neural architectures, and may not fully capture the complex nonlinear relationships present in biological embeddings. Lastly, it is important to address test-train data splitting to ensure protein families are distributed in an unbiased manner.

References

[1] Garyk Brixi et al. “Genome modeling and design across all domains of life with Evo 2”. In: *bioRxiv* (2025).

[2] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[3] David Roi Hardoon, Sándor Szedmák, and John Shawe-Taylor. “Canonical Correlation Analysis: An Overview with Application to Learning Methods”. In: *Neural Computation* 16 (2004), pp. 2639–2664.

[4] Megha Hegde, Jean-Christophe Nebel, and Farzana Rahman. “Language Modelling Techniques for Analysing the Impact of Human Genetic Variation”. In: *Bioinformatics and Biology Insights* 19 (2025).

[5] Melissa J. Landrum et al. “ClinVar: improving access to variant interpretations and supporting evidence”. In: *Nucleic Acids Research* 46 (2017), pp. D1062 –D1067.

[6] Erik Nijkamp et al. “ProGen2: Exploring the Boundaries of Protein Language Models”. In: *Cell systems* (2022).

[7] Hai Yang et al. “A multimodal framework for comprehensive driver variant prediction in cancer”. In: *Communications Medicine* 5 (2025).

Note: AI tools used for code assistance and writing refinement