Leveraging Vector Embeddings for Rapid and Accurate Pathogenicity Prediction of Genetic Variants

Jiawei Wu[1,2] , Michael Muriello[1], Donald G. Basel[1,2], Xiaowu Gai[1,2]

1. Division of Bioinformatics and Quantitative Child Health, Department of Pediatrics, Medical College of Wisconsin
2. Mellowes Center for Genomic Sciences and Precision Medicine, Medical College of Wisconsin

MEDICAL COLLEGE OF WISCONSIN

Linda T. and John A. Mellowes Center for Genomic Sciences and Precision Medicine

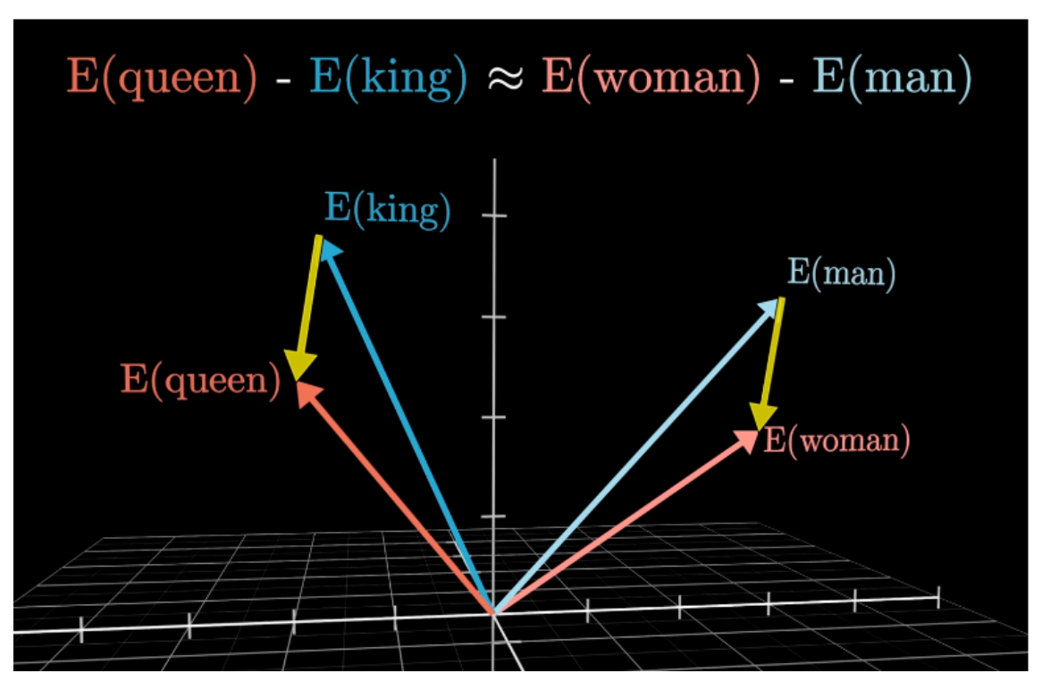
Introduction

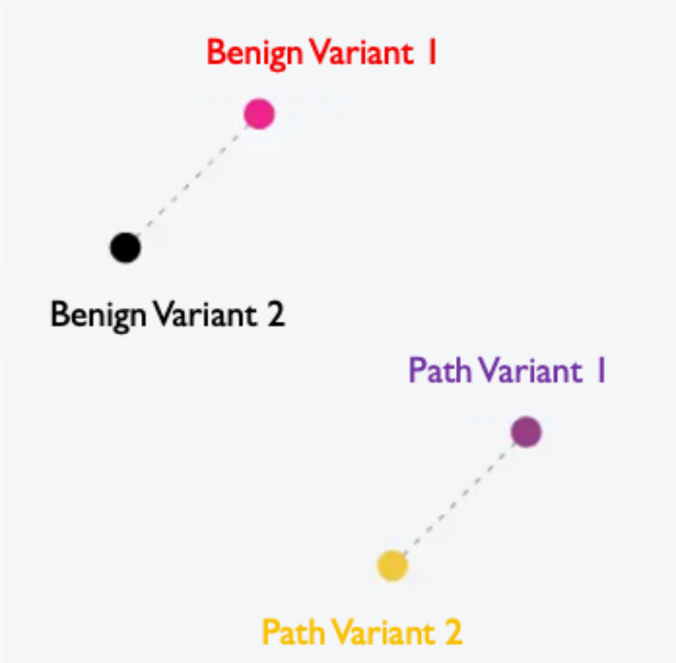
Interpreting the pathogenicity of genetic variants is a critical bottleneck in genomic medicine. Millions of variants of uncertain significance (VUS) hinder the clinical application of genetic findings, creating profound implications for patient care in areas like hereditary cancer screening. While traditional computational tools exist, they often rely on hand-engineered features and struggle to capture the full complexity of genomic data.

Here, we introduce **VUS.Life**, a novel semantic embedding framework that transforms complex variant annotations into natural language. By leveraging pre-trained language models, VUS.Life captures nuanced relationships between different types of genomic evidence, enabling direct and accurate prediction of pathogenicity through representation learning.

Techniques

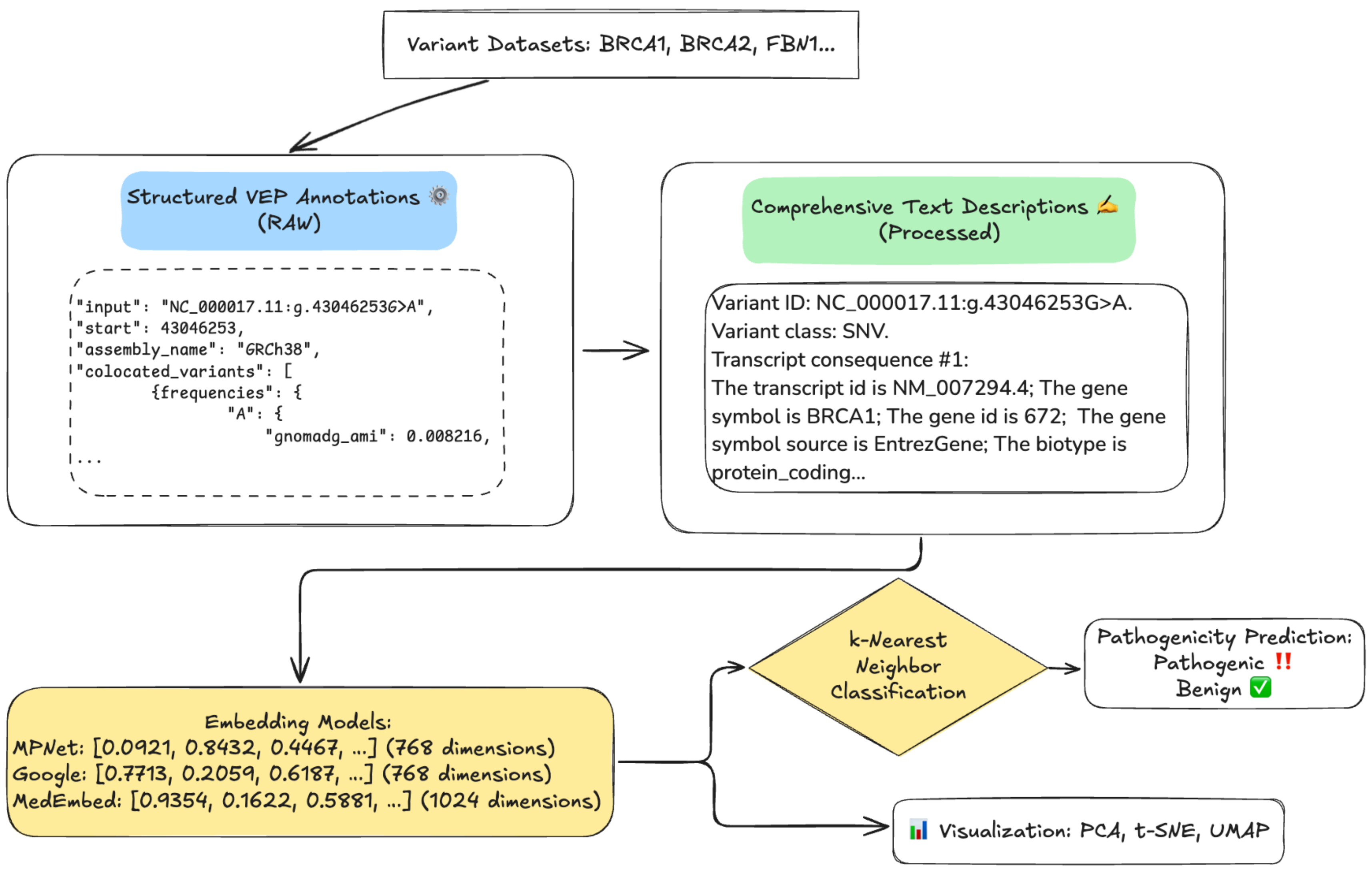
Word Embedding: This AI technique converts text (like variant descriptions) into numerical vectors. It allows the model to understand the context and meaning of genomic terms. Variants with similar biological effects end up "closer" to each other in this numerical space.





k-Nearest Neighbor Classification: A straightforward algorithm that classifies a new variant by looking at its 'k' closest neighbors. If the majority of a variant's neighbors are pathogenic, it is predicted to be pathogenic too.

VUS.Life


Variant Datasets: BRCA1, BRCA2, FBN1...
Structured VEP Annotations ⚙ (RAW)
Comprehensive Text Descriptions ✍ (Processed)
k-Nearest Neighbor Classification
Pathogenicity Prediction: Pathogenic ‼ Benign ✅
📊 Visualization: PCA, t-SNE, UMAP

```
"input": "NC_000017.11:g.43046253G>A",
"start": 43046253,
"assembly_name": "GRCh38",
"colocated_variants": [
        {frequencies": {
                "A": {
                        "gnomadg_ami": 0.008216,
...
```

Variant ID: NC_000017.11:g.43046253G>A.
Variant class: SNV.
Transcript consequence #1:
The transcript id is NM_007294.4; The gene symbol is BRCA1; The gene id is 672; The gene symbol source is EntrezGene; The biotype is protein_coding...

Embedding Models:
MPNet: [0.0921, 0.8432, 0.4467, ...] (768 dimensions)
Google: [0.7713, 0.2059, 0.6187, ...] (768 dimensions)
MedEmbed: [0.9354, 0.1622, 0.5881, ...] (1024 dimensions)

Variant Annotation: The framework systematically converts structured variant data from the Variant Effect Predictor (VEP) into a consistent, human-readable textual format. This process organizes all features, including numerical scores and identifiers, into a unified textual string that serves as the input for language models.

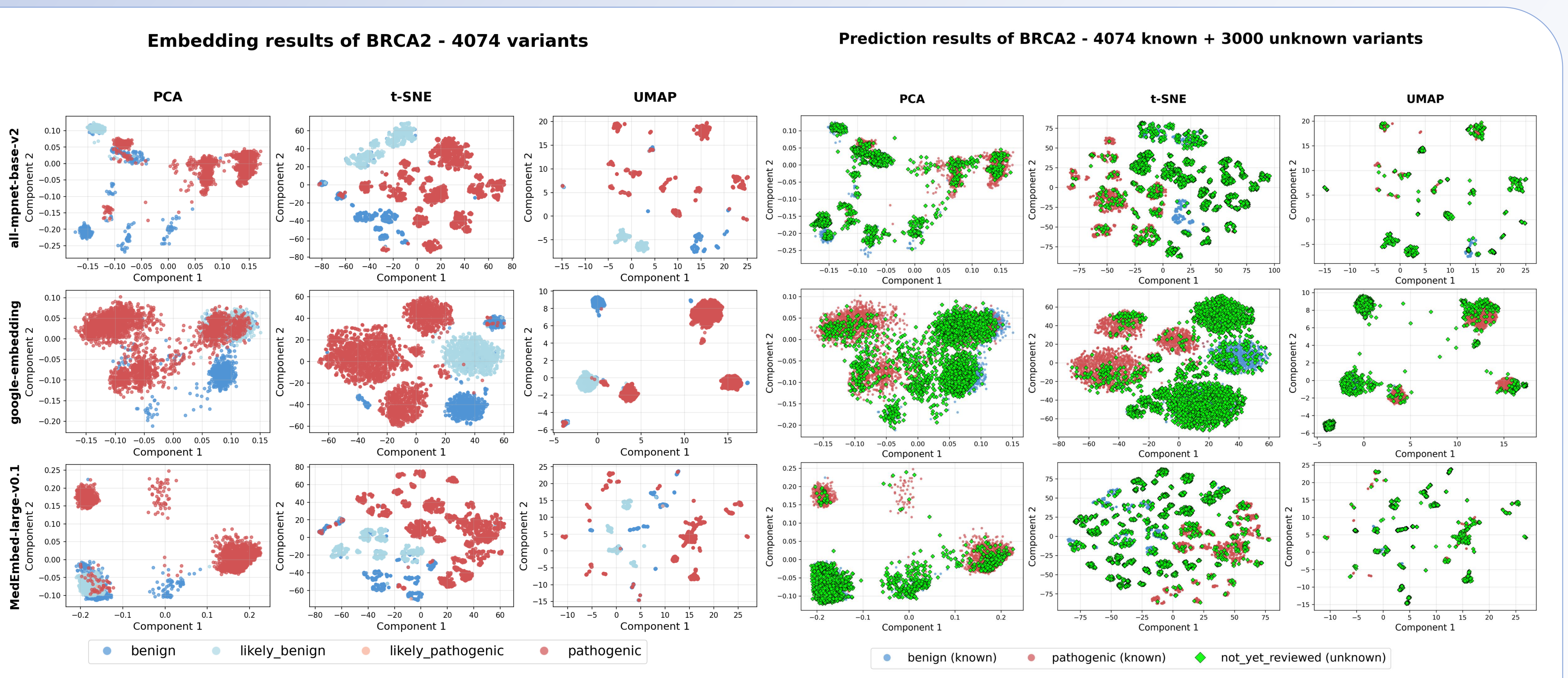
Language Models: VUS.Life utilizes three distinct embedding models (MPNet, Google's text-embedding-004, and MedEmbed) to transform the natural language descriptions of variants into high-dimensional vector representations. These models capture the semantic relationships between different types of genomic evidence.

Pathogenicity Prediction: Pathogenicity for new variants is predicted using a k-nearest neighbor (k-NN) approach. The framework identifies the most similar variants with known classifications in the vector space and assigns a prediction based on a majority vote, enabling a scalable and interpretable classification of VUS.

Results & Implementation

Gene	Model	k=5	k=10	k=15	k=20
BRCA1	MedEmbed	0.981	0.980	0.978	0.977
	google	0.980	0.977	0.974	0.973
	mpnet	0.984	0.983	0.981	0.979
BRCA2	MedEmbed	0.993	0.993	0.992	0.991
	google	0.985	0.983	0.981	0.979
	mpnet	0.992	0.991	0.990	0.990
FBN1	MedEmbed	0.966	0.965	0.964	0.963
	google	0.965	0.964	0.961	0.959
	mpnet	0.967	0.962	0.961	0.959

High-Accuracy Classification: Achieved **>96% accuracy** across all genes evaluated (*BRCA1*, *BRCA2*, *FBN1*). For the gene with large variant set, *BRCA2*, accuracy reached an exceptional **99.1%**.



Effective Variant Separation (Left): Dimensionality reduction techniques (PCA, t-SNE, UMAP) reveal a clear and distinct separation between benign (blue) and pathogenic (red) variant clusters in the high-dimensional vector space. Non-linear methods like t-SNE and UMAP show highly compact and well-separated clusters, confirming that variants in the semantic space are structured by, likely by shared characteristics, including clinical significance.

Rapid Prediction of Unknown Variants (Right): The consistent spatial distribution of unknown variants relative to known benign and pathogenic variants across models and genes suggests that embeddings effectively capture pathogenicity-related semantics.

References

1. Wu, J., Muriello, M., Basel, D. G., & Gai, X. (2025). Predicting Genetic Variant Pathogenicity Using Vector Embeddings.
2. Cheng, J., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science, 381(6664).
3. ClinVar database - https://www.ncbi.nlm.nih.gov/clinvar/
4. BRCA Exchange - https://brcaexchange.org/

Future Work

- ❑ Explore newer transformer-based and domain-specific biomedical models.
- ❑ Enhance variant annotations and expand to somatic variants
- ❑ Incorporating multimodal data, such as protein structure information and functional assay results, for more comprehensive predictions.
- ❑ Validate the approach on a broader range of genes (associated with rare diseases), to establish its universal applicability in clinical genetics workflows.