

Enhancing Missense Variant Classification with AlphaFold2-Generated Mutant Structures

Ph.D. Dissertation

Erda Qorri

Supervisor: Lajos Haracska Ph.D., D.Sc.

Doctoral School of Biology

University of Szeged, Faculty of Science and Informatics



Institute of Genetics

Carcinogenesis and Tumorigenesis Research Group

HUN-REN Biological Research Centre

Szeged

2025

1. Introduction

Recent advances in next-generation sequencing (NGS) technologies have enabled their swift integration into diagnostics, facilitating the identification of thousands of genetic variants, including amino acid substitutions. While many variants have known clinical significance, others remain challenging to classify and are labeled as variants of uncertain significance (VUS), creating uncertainty for both clinicians and patients. In the past decade, various in-silico methods have emerged as an alternative to functional assays, with ongoing efforts focused on further improving their performance.

Structural information has long been recognized as a valuable resource for improving the performance of in-silico tools, but its integration was limited by the scarcity of available protein structures. The development of AlphaFold2 has enabled the accurate prediction of thousands of structures in a relatively short time. Naturally, it also raised the question about whether its predictions can capture the structural impact of variants, and if the generated structures contain information that can be used to develop better-performing models. This thesis evaluates the predictive performance of structural information from thousands of ParaFold-generated native and variant protein structures. It also introduces five novel structural features and HPC-optimized submission strategy for large-scale protein structure prediction.

2. Goals and objectives

The primary aim of this thesis was to evaluate the performance of widely used variant effect predictors (VEPs), identify their limitations, and explore novel directions to enhance their performance. To achieve this, the study focused on three key objectives:

1. Systematic and independent benchmarking of widely used VEPs on a set of clinically relevant missense variants retrieved from ClinVar.
2. Develop a ParaFold-based HPC pipeline for predicting native and variant protein structures to analyze the structural impact of missense variants.
3. Develop and extract machine learning features from the generated structures and evaluate their predictive performance, both independently and combined with well-established sequence-based features.

3. Materials and Methods

Part 1: Systematic and updated benchmarking of existing variant effect predictors

Study design and variant effect predictor selection: Ten variant effect predictors (VEPs) were selected based on three strict criteria: popularity, availability of training datasets, and input format requirements. This resulted in the selection of eight individual VEPs: PANTHER-PSEP, PROVEAN, SIFT, PolyPhen-2, PMut, PhD-SNP, SNPs&GO, and two meta-predictors, META-SNP and PredictSNP.

Benchmarking dataset generation and composition: To evaluate VEP performance, four benchmarking datasets were generated from Clinvar: Expert Panel (EP), CircD, BRCA1, and BRCA2. The EP dataset comprised 404 missense variants from 21 genes with three-star ClinVar status. The BRCA1 and BRCA2 datasets contained 151 and 134 variants, with two- and three-stars. The CircD dataset included 1,053 variants overlapping with the training sets of seven VEPs and was used to evaluate the impact of type 2 circularity on their performance.

Prediction Score Obtention: Prediction scores for each VEP except for SIFT and PhD-SNP were obtained from their respective user interfaces using default parameters. Predictions for SIFT and PhD-SNP were retrieved through PROVEAN. Classification thresholds for each VEP were set as defined by their developers.

Performance Evaluation Analysis: The performance of the evaluated VEPs was assessed using seven machine learning metrics: Matthews Correlation Coefficient (MCC), Area Under the Curve (AUC), Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Accuracy (ACC). These metrics were computed using the *cvms* package in R.

Concordance Analysis: The concordance rates among the evaluated VEPs were assessed within the framework of missense variant interpretation guidelines. This analysis was carried out in the BRCA1 and BRCA2 datasets, where five distinct VEP combinations were examined: Best-performing, Best-performing + PROVEAN, All nine VEPs, and Worst-performing VEPs. These combinations were analyzed across three scenarios: (i) True concordance (agreement on correct predictions), Discordance (disagreement in their predictions), (iii) False concordance (agreement on incorrect predictions). Additionally, the similarity rate of the predictions was evaluated using Cohen's Kappa. The statistical significance threshold was set to $p\text{-value} < 0.05$.

Part 2: Large-Scale Protein Structure Prediction Pipeline

Structure prediction dataset: To investigate whether point mutant structures harbor useful information for improving pathogenicity prediction, a curated dataset of 70,345 missense variants identified across 12,580 proteins (pathogenic: 30,926 and benign: 39,419) was compiled from UniProtKB/Swiss-Prot.

Structure prediction with ParaFold: Native and variant protein structures were predicted using ParaFold, the high-performance computing (HPC) cluster implementation of AlphaFold2. These predictions were performed on the CPU and GPU partitions of the Komondor HPC. Structural comparisons between ParaFold and AlphaFold2 predicted protein structures were assessed using two metrics, TM-score and RMSD.

Submission strategy: Protein prediction followed a two-step job submission strategy. Amino acid sequences were clustered into 14 length-based groups (20-99 to 2500-3500 residues) and first processed in the CPU partition for database queries, multiple sequence alignment, and pair representation. The clusters were then transferred to the GPU partition for model prediction. ParaFold was run with default parameters.

pLDDT distribution statistical analysis: To assess changes in the predicted Local Distance Test (pLDDT), delta pLDDT was calculated as the difference in summed alpha carbon pLDDT values between native and variant structures, normalized by protein length. The distribution of delta pLDDT values was assessed for normality and analyzed for statistical significance using the Kruskal-Wallis test, followed by Wilcoxon rank-sum tests with Bonferroni correction. All statistical analyses were performed in R, with a significant threshold of 0.05.

Part 3: Feature development and testing

Structural feature development: The predicted native and variant protein structures were used to derive five structural features: Alpha Carbon Δ pLDDT, Alpha Carbon Distance, Local Distance RMS (dRMS-Local), Δ SASA, and Miyazawa-Jernigan Potential of the Mutant (MJ Mutant). These features capture distinct aspects of variant effects on protein structure, from geometric alterations to residue-residue energetic changes. All features rely on native vs variant comparison and are implemented in R.

Sequence-based feature extraction: To complement the structural features, four well-established sequence-based features were included: hydrophobicity, Shannon's entropy, Blosom62 substitution matrix, and position-specific scoring matrix (PSSM) at the native

residue. These features were extracted from the amino acid sequence using the PatMut algorithm.

Feature extraction and machine learning model building: The predictive power of the structural features was tested by using three XGBoost models: *SIESTA*, *SIESTA-Str*, and *SIESTA-Seq*. *SIESTA* combined both structure and sequence-derived features, while *SIESTA-Str* and *SIESTA-Seq* focused exclusively on structure and sequence features, respectively. A total of nine features were extracted per variant, and models were trained using an 80:20 training-testing protein split strategy to prevent data leakage and minimize bias.

Complementary analysis: The true positive and true negative predictions of *SIESTA*, *SIESTA-Str*, and *SIESTA-Seq* were compared to determine whether the structure-derived features capture meaningful information not available from sequence data alone.

Feature importance analysis: The contribution of each developed feature to the performance of the *SIESTA* models was quantified using SHAP (Shapley Additive exPlanations) analysis through the SHAP library in Python.

4. Results

A systematic benchmarking of ten widely used variant effect predictors was conducted on a dataset of manually curated missense variants, followed by the development of a large-scale protein structure prediction pipeline along with an optimized job submission strategy on an HPC. From the predicted structures, five structural features were derived and used to train three XGBoost-based machine learning models: *SIESTA*, *SIESTA-Str*, and *SIESTA-Seq* to assess their predictive power.

Evaluation on the Expert Panel Dataset: Based on MCC and AUC values, SNPs&GO, PMut, and PolyPhen2-HumVar were determined to be the top-performing VEPs in the EP dataset, while PhD-SNP showed the poorest performance.

Evaluation in the BRCA1 dataset: SNPs&GO, PANTHER-PSEP, and PMut were determined to be the top-performing VEPs in the BRCA1 dataset based on high MCC and AUC values. In contrast, PROVEAN, HumVar, and HumDiv showed the poorest performance, with PROVEAN failing to predict pathogenic missense variants in BRCA1.

Evaluation in the BRCA2 dataset: PMut outperformed all other VEPs, with an MCC of 0.79, followed by SNPs&GO and HumVar. As with the BRCA1 dataset, PROVEAN performed worst, largely due to a high number of false negatives.

Concordance analysis: Combining the best-performing VEPs from the BRCA1 and BRCA2 datasets minimizes false concordance and discordance rates. Adding PROVEAN to this combination notably reduced true concordance rates, particularly for BRCA1. Additionally, increasing the number of VEPs, results in the highest discordance rates across both genes.

Effect of Type 1 Circularity on VEP performance: Type 1 circularity results in an overestimation of VEP performance when benchmarked on variants included in their training data. This is evident from the overall performance increase of VEPs on the CircD dataset, which consists of variants present in their training set, as well as changes in their ranking. Notably, PMut showed the most significant improvement, moving from third to first place.

Structure prediction pipeline and structure generation: A large-scale protein structure prediction pipeline optimized for HPC systems using ParaFold was developed, along with an HPC-efficient job submission strategy. In total, 77,713 protein structures were generated: 12,101 native and 65,612 variants, covering proteins ranging from 24 to 3,487 amino acids, including clinically relevant proteins such as P53, PTEN, EGFR, BRCA1, and BRCA2.

Quality of the predicted structures: ParaFold predicted 70.6% (54,865) of protein structures with high confidence (pLDDT > 70). In contrast, 29.3% were predicted with low confidence (pLDDT < 70). Analysis of mean pLDDT values revealed that pathogenic structures had significantly lower delta pLDDT values than benign ones.

Feature extraction and correlation: Five structural features, *Alpha carbon distance* (C α -Dist), *Delta alpha carbon pLDDT* (C α - Δ pLDDT), *Delta SASA Normalized* (Δ SASA Normalized), *Miyazawa-Jernigan Potential of the Mutant* (MJ-Mutant), and *dRMS Local* were developed and extracted. Kendall's Tau rank correlation analysis revealed no significant correlations between the features, except for a moderate correlation between C α -Dist and C α - Δ pLDDT.

Feature testing and complementary analysis: The supervised machine learning models SIESTA, SIESTA-Str, and SIESTA-Seq were trained on 39,420 high-confidence variant structures from 6,648 proteins. Structural features alone showed limited discriminative power but improved variant classification when combined with sequence-based features, also enabling the detection of missense variants missed by sequence features alone.

Feature importance analysis: Feature importance analysis of the integrated SIESTA model revealed that PSSM Native, Entropy, and Substitution Matrix were the top contributors, followed by the structural features C α -Dist, C α - Δ pLDDT, and Δ SASA.

5. Conclusions

In this thesis, the performance of ten widely used variant effect predictors was independently benchmarked across four datasets composed of missense variants identified in clinically relevant genes. The best-performing algorithms for each dataset were identified and recommended for use. Additionally, the predictors were also evaluated in the context of variant effect classification guidelines, highlighting their pitfalls and providing recommendations for improvement.

Through the development of an HPC-based large-scale protein structure prediction pipeline, 77,713 protein structures were generated, including 12,101 native and 65,612 variant structures. From these five structural features: Alpha Carbon Distance ($C\alpha$ -Dist), Delta Alpha Carbon pLDDT ($C\alpha$ - Δ pLDDT), Delta SASA Normalized (Δ SASA Normalized), Miyazawa-Jernigan Potential of the Mutant (MJ-Mutant), and dRMS Local were derived and used to train three supervised learning models: SIESTA, SIESTA-Str, and SIESTA-Seq. Comparative analysis of these models highlighted the potential of structural information in enhancing variant effect predictions, as demonstrated by the improved classification performance when integrated with sequence-based features.

6. Acknowledgments

This work received funding from the National Research, Development, and Innovation Office (GINOP-2.3.2-15-2016-00020, GINOP-2.3.2-15-2016-00024, GINOP-2.3.2-15-2016-00026, and RRF-2.3.1-21-2022-00015). Project no. RRF-2.3.1-21-2022-00015 has been implemented with the support provided by the European Union. This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 739593.

7. List of publications

MTMT ID: 10082008

A. Publications used in dissertation:

1. **Qorri, E.**, Takács, B., Gráf, A., Enyedi, M. Z., Pintér, L., Kiss, E., & Haracska, L. (2022). A Comprehensive Evaluation of the Performance of Prediction Algorithms on Clinically Relevant Missense Variants. *International Journal of Molecular Sciences*, 23(14), 7946. **IF: 5.6**
2. Takács, B., Jaksa, G., **Qorri, E.**, Gyuris, Z., Pintér, L., & Haracska, L. (2025). Advancing metagenomic classification with NABAS+: a novel alignment-based approach. *NAR Genomics and Bioinformatics*, 7(3). **IF: 5.0**

B. Other publications:

1. Pesti, I., Varga, V., **Qorri, E.**, Frank, R., Kata, D., Vinga, K., Szarvas, P. A., Menyhárt, Á., Gulya, K., Bari, F., & Farkas, E. (2024). Nimodipine reduces microglial activation in vitro as evidenced by morphological phenotype, phagocytic activity, and next generation RNA sequencing. *British Journal of Pharmacology*. **IF: 6.8**
2. Mórocz M, **Qorri E**, Pekker E, Tick G, Haracska L. Exploring RAD18-dependent replication of damaged DNA and discontinuities: A collection of advanced tools. *J Biotechnol*. 2024 Jan 20;380:1-19. **IF: 4.1**
3. Pekker E*, **Qorri E***, Enyedi M. Z., Szukacsov V., Ayaydin F., Szabó-Kriston É., Csányi B., Sükösd F., Kiss-Tóth E., Haracska L. (2025). Comprehensive Bulk and Single-Cell RNA Sequencing Uncovers Senescence-Associated Biomarkers in Therapeutic Mesenchymal Stem Cells. (manuscript under revision)

*These authors have contributed equally to this study and share first authorship.

C. Other Contributions

Featured as a success story in KIFÜ's (Kormányzati Informatikai Fejlesztési Ügynökség) Echo Magazine, 2024 edition.

Selected as a success story of the National Competence Center (NCC) Hungary (2024). Title: Harnessing HPC in Cancer Genomics: High Throughput Mutated Protein Structure Prediction using AlphaFold2.

D. Conference Talks

Straub Days, Szeged, Hungary (May 2024). Title: Giving significance to unclassified missense variants: Optimized large-scale protein structure prediction through High-Performance Computing.

HPC Matching Day (March 2024). Title: Predict the unknown: Machine learning-driven genetic variant annotation through HPC-aided large-scale protein structure prediction.

MSO-Industry Conference (October 2024). Title: Predict the unknown: Machine learning-driven genetic variant annotation through HPC-aided large-scale protein structure prediction.

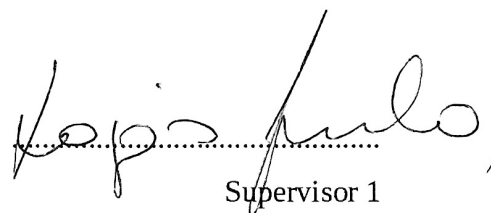
HCEMM 4th PhD & Post-doc Symposium, Bükfürdő, Hungary. (November 2024). Leveraging AlphaFold-Derived 3D Structural Insights to Enhance Machine Learning Algorithms for Missense Variant Classification in Clinical Genetics.

8. Co-Author Waiver

Declaration

I declare that the contribution of Erda Qorri was significant in the listed publications and the doctoral process is based on the publications listed. The results reported in the Ph.D. dissertation and the publications have not been used to acquire any PhD degree previously and will not be used in the future either.

Szeged, 03.02.2025



Supervisor 1