ISSN: 2692-5516

Research Article

Advances in Hematology and Oncology Research

A Novel Machine Learning Approach to PTEN Missense Variant Classification Using Alpha Fold 3

Yash Jayesh Laddha[1*], Arwen Shah[1] and Shubh Jayesh Laddha[2]

[1] *Greenwood High International School, Bangalore, India*

[2] *Delhi Public School East, Bangalore, India*

[*]Corresponding Author
Yash Jayesh Laddha, Greenwood High International School, Bangalore, India.

Submitted: 2026, Jan 12; **Accepted:** 2026, Jan 30; **Published:** 2026, Feb 05

Citation: Laddha, Y. J., Shah, A., Laddha, S. J. (2026). A Novel Machine Learning Approach to PTEN Missense Variant Classification Using Alpha Fold 3. *Adv Hema Onco Res, 9*(1), 01-12.

Abstract

PTEN is among the most commonly mutated tumor suppressor genes across human cancers. Yet, hundreds of its missense variants remain classified as variants of uncertain significance (VUS), limiting clinicians' ability to assess cancer risk. Existing predictors rely mainly on sequence conservation and cannot evaluate the three-dimensional structural changes that influence PTEN function, leaving a significant gap in variant classification. This study aimed to determine whether structural changes caused by PTEN missense mutations could reliably distinguish cancer-associated variants from benign ones. All known PTEN missense variants were collected from UniProt, and structural models were generated using Alpha Fold 3 for the wild type and 1,514 mutant sequences. After aligning each mutant to the wild-type structure, seventeen structural features were extracted using PyMOL, Bio python, and MDTraj, including secondary-structure shifts, hydrophobicity changes, and electrostatic differences. These features were used to create a dataset for training Random Forest, XGBoost, logistic regression, and decision-tree classifiers. The models performed well, with Random Forest and XGBoost achieving ROC-AUC scores of 0.985 and 0.983, respectively, and reached high recall and precision for the cancer-associated class, showing strong sensitivity and reliability in identifying pathogenic variants. These displayed significantly higher local RMSD, hydrophobicity change, and electrostatic disruption, reflecting well-known PTEN destabilization mechanisms. It also provided predictions for more than 1,300 VUS, offering a tool for prioritizing high-risk mutations. This work introduces one of the first comprehensive structural frameworks for PTEN variants. It demonstrates that integrating Alpha Fold-based modeling with machine learning can create accurate, clinically relevant interpretations of PTEN mutations.

Keywords: PTEN Missense Variants, Alpha Fold 3, Structural Bioinformatics, Pathogenicity Prediction, Machine Learning, Random Forest, XGBoost, Cancer-Associated Mutations

1. Introduction

Cancer arises from genetic alterations that disrupt normal signaling pathways controlling cell growth, survival, and metabolism [1]. One of the most important regulators of these processes is PTEN (phosphatase and tensin homolog), a lipid phosphatase that limits PI3K–AKT signaling by dephosphorylating membrane phosphoinositide's. Through this activity, PTEN prevents uncontrolled cell proliferation and promotes programmed cell death [2,3]. Loss of PTEN function is strongly associated with breast, endometrial, thyroid, prostate, and brain cancers, and germline mutations give rise to cancer-predisposition syndromes such as Cowden syndrome and Bannayan-Riley-Ruvalcaba syndrome [4]. Because PTEN acts as a central suppressor of oncogenic signaling, understanding the functional impact of its variants is a major clinical priority [2,3]. However, interpreting PTEN mutations remains challenging. Large population and cancer-sequencing databases contain more than 1,000 PTEN variants classified as variants of uncertain significance, indicating the absence of direct evidence of their functional effects [5,6]. Many traditional computational prediction tools rely on evolutionary conservation or amino acid similarity and do not capture the three-dimensional changes that influence PTEN's catalytic activity or membrane binding [7]. Since PTEN function depends on its precise folded structure, even small conformational shifts can alter its enzymatic activity or substrate recognition [3]. A structure-based approach is therefore appropriate.

Similar ideas have been explored for TP53, another highly mutated tumor suppressor gene. Structure-guided analyses combined with machine learning have improved the classification of TP53 missense variants [8,9]. In contrast, no equivalent large-scale structural framework exists for PTEN, even though PTEN carries a comparable number of clinically significant variants and one of the highest numbers of unresolved VUS among significant cancer genes [5,10]. Although PTEN is one of the most frequently altered tumor suppressor genes in human cancer, very few attempts have been made to generate a complete structural dataset covering all known PTEN missense variants [3,10,11]. This study is one of the first large-scale efforts to combine deep-learning protein structural modeling with supervised machine learning for PTEN, and it includes more than one thousand variants of uncertain significance. This level of analysis became possible only with the release of Alpha Fold 3, which provides the accuracy and scale needed to model the structural features of PTEN mutations [12]. Alpha Fold 3 has recently made comprehensive structural modeling feasible by predicting protein structures with high accuracy across the human proteome. This now allows the generation of structural models for every known PTEN missense variant. These models can be used to quantify how each amino acid substitution alters structural properties directly linked to PTEN's stability and catalytic function [12].

Several feature classes are significant for PTEN. Changes in solvent accessibility frequently destabilize the phosphatase and C2 domains, since many pathogenic mutations bury hydrophobic residues or expose usually protected positions, which disrupts folding and membrane association [13]. Local and global RMSD measurements capture the extent of backbone deformation, and even small displacements around the active-site loops can impair access to phosphoinositide substrates [3]. Flexibility and residue-level fluctuations are also relevant because pathogenic variants cluster in regions where altered dynamics interfere with the catalytic pocket or the C2 domain's membrane-binding interface [14]. Backbone torsion-angle shifts can be especially informative, since PTEN activity depends on the orientation of conserved catalytic residues, including Cys124 and Arg130, and mutations that distort the local geometry often abolish phosphatase activity [15]. Hydrophobicity changes are another important factor, particularly in the core of the phosphatase domain, where substitutions that increase polarity can disrupt packing and lower protein stability [13]. Electrostatic changes are also significant because PTEN contains clusters of charged residues that guide membrane binding and substrate positioning, and pathogenic variants often disturb these charge interactions. C alpha displacement provides an approximate measure of how much a mutation shifts its immediate structural neighborhood and is useful for identifying substitutions that reposition catalytic or membrane-contacting residues [16].

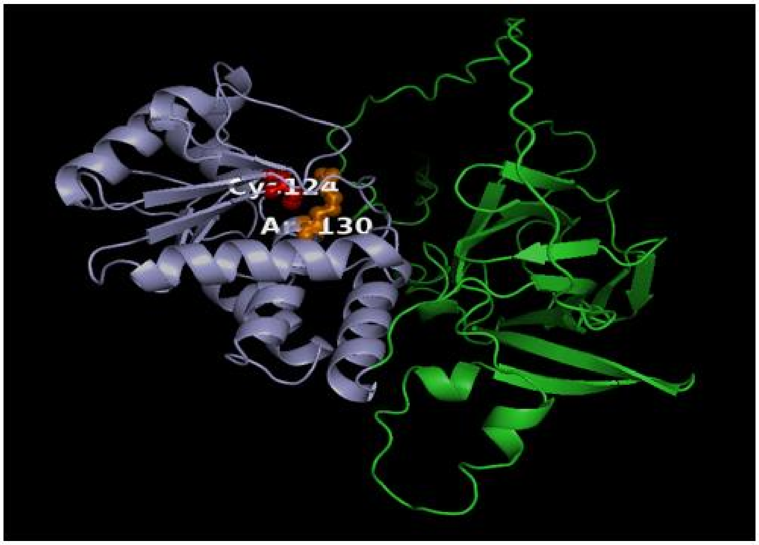

Figure 1: Overall structure of PTEN predicted by Alpha Fold 3. The phosphatase domain (residues 7-185) is shown in light purple and the C2 domain (residues 186-351) in green. Catalytic residues Cys124 and Arg130 are highlighted to provide structural context for later analyses

This study aims to determine whether these structural features can distinguish cancer-associated PTEN mutations from benign ones using supervised machine learning. All known PTEN missense variants were collected from UniProt, ClinVar, and COSMIC [5,6,10]. Structural models were generated using Alpha Fold 3, aligned in PyMOL, and analyzed with molecular tools including Bio python and MDTraj [17-19]. Features extracted from these models served as the basis for a dataset used to train Random

Forest and XGBoost classifiers, while unlabeled variants of uncertain significance were reserved for prediction [20,21]. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, and feature importance was assessed to identify structural properties most strongly associated with pathogenic behavior. A significant advantage of this approach is that it uses structural information that cannot be captured by sequence-based predictors [7]. Tools that rely on conservation patterns have limited power when a mutation alters the shape, charge, or stability of a protein without affecting prominent sequence motifs [13,16]. By building a complete structural map of all known PTEN missense variants, this study creates a foundation that can be expanded as new variants are discovered. The framework also offers a way to prioritize high-risk substitutions for laboratory validation, which is especially helpful in clinical settings where many VUS remain unresolved. By integrating recent advances in protein structure prediction with machine learning, this study provides one of the first large-scale structure-based analyses of PTEN variants. This approach provides a framework for variant interpretation, highlights structural patterns associated with pathogenicity, and has the potential to improve clinical assessment for individuals carrying PTEN mutations.

2. Methodology

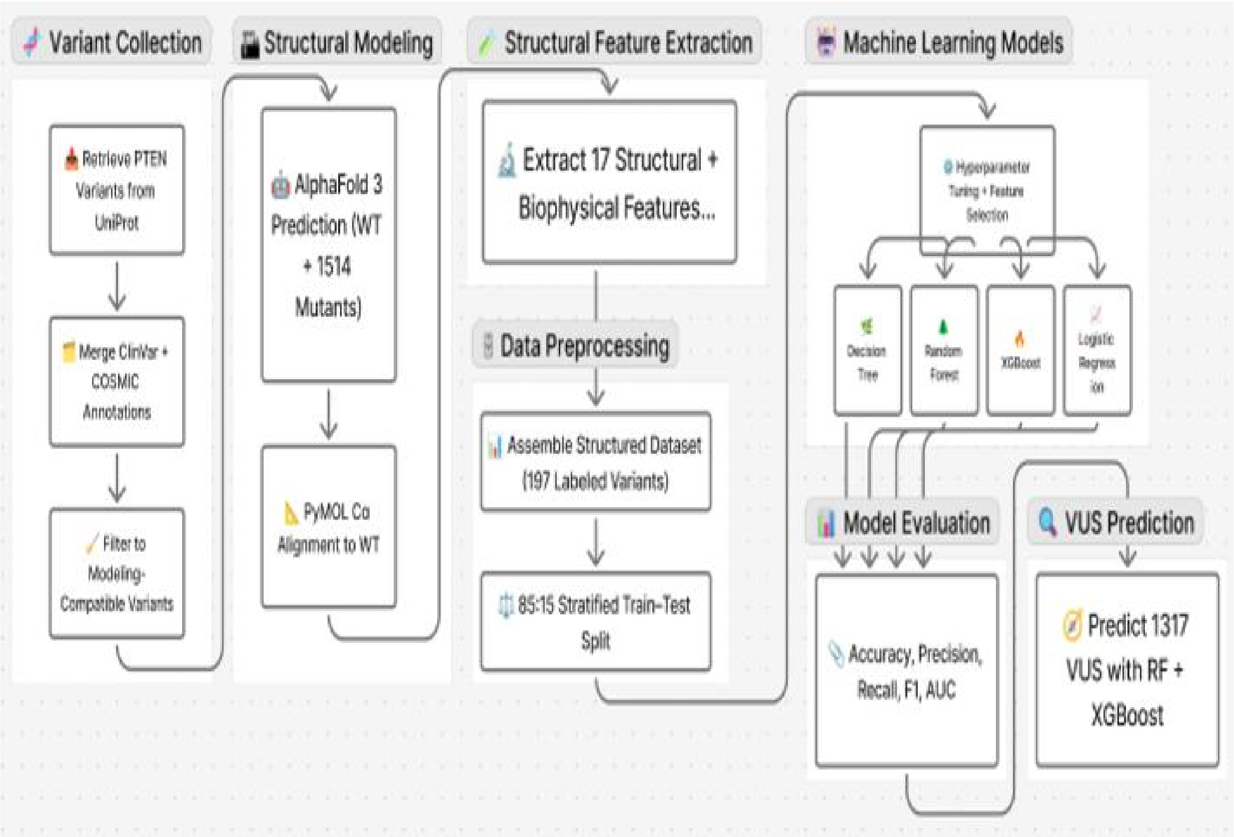
Variant Collection
Structural Modeling
Structural Feature Extraction
Machine Learning Models
Retrieve PTEN Variants from UniProt
Merge ClinVar + COSMIC Annotations
Filter to Modeling-Compatible Variants
AlphaFold 3 Prediction (WT + 1514 Mutants)
PyMOL Cα Alignment to WT
Extract 17 Structural + Biophysical Features...
Data Preprocessing
Assemble Structured Dataset (197 Labeled Variants)
85:15 Stratified Train-Test Split
Hyperparameter Tuning + Feature Selection
Decision Tree
Random Forest
XGBoost
Logistic Regression
Model Evaluation
Accuracy, Precision, Recall, F1, AUC
VUS Prediction
Predict 1317 VUS with RF + XGBoost

Figure 2: Overview of the Computational Pipeline used to Classify PTEN Missense Variants

2.1. Variant Collection and Curation

All PTEN variants were retrieved from UniProt (P60484) which collects information from ClinVar, COSMIC and additional curated sources [5,6,10]. The full dataset contained 2,826 distinct variants, including missense substitutions, nonsense and frameshift mutations, splice-site alterations, insertions, deletions, and variants lacking clear protein-level mapping [5]. Because structure-based analysis requires a single amino acid substitution at a defined residue position, we restricted the dataset to variants that could be expressed as p.X123Y [22]. Variants with incomplete positional information, multiple simultaneous substitutions, non-standard amino acid codes, mismatched isoform coordinates, or protein-irrelevant annotations were removed. After applying these filters, 1,514 single-residue missense variants remained suitable for structural modeling and downstream analysis. Clinical significance annotations from UniProt, integrated with ClinVar and COSMIC categorizations, were categorized into two groups [5,6,10]. Variants labeled pathogenic or likely pathogenic were merged into a cancer-associated class, while those labeled benign or likely benign were grouped as non-cancerous. This process produced 197 labeled variants, comprising 177 pathogenic and 20 benign substitutions, which served as the supervised learning set for model development and evaluation. The remaining 1,317 variants lacked definitive clinical interpretation and were designated as variants of uncertain significance (VUS). These VUS were excluded from model training and reserved solely for predictive inference.

2.2. Structural Modeling Using Alpha Fold

Structural predictions for the wild-type and mutant PTEN proteins were generated using Alpha Fold 3, a state-of-the-art deep-learning model for biomolecular structure prediction [12]. The canonical PTEN sequence (UniProt P60484) was first modeled to establish the wild-type reference structure [5]. Each of the 1,514 missense variants was then modeled independently by substituting the corresponding amino acid into the full-length PTEN sequence prior to Alpha Fold 3 prediction. This produced 1,515 structures in total, comprising a single wild-type model and one Alpha Fold 3 structure for every mutant sequence. To enable direct comparison of residue-level geometry across mutants, each structure was aligned to the wild-type structure using Cα superposition in PyMOL [17]. This alignment step ensured that global orientation differences did not influence downstream measurements of RMSD, solvent exposure, or torsional angles. Alpha Fold's per-residue pLDDT confidence values were recorded [17].

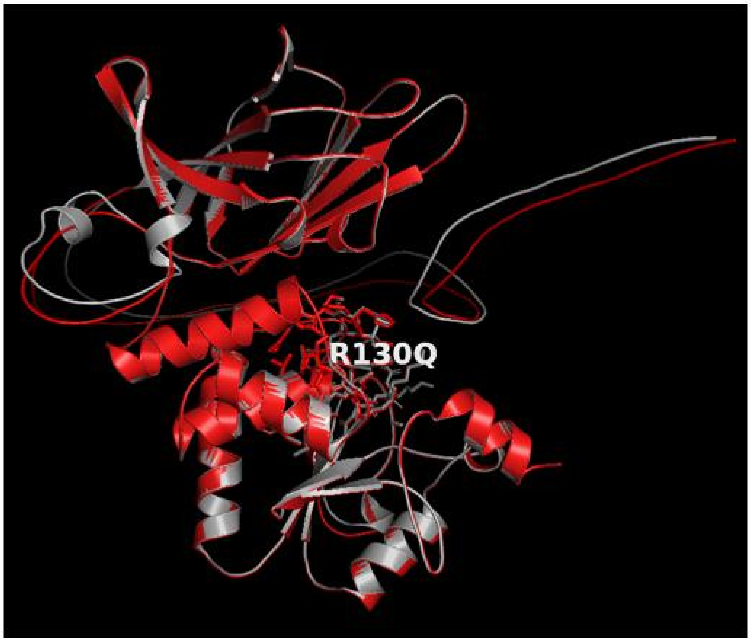

Figure 3: Overlay of PTEN wild-type (gray) and the R130Q mutant (red). The R130Q substitution produces a clear local structural shift around residue 130, affecting nearby helices and loops

2.3. Structural Feature Extraction

Structural and physicochemical features were extracted from each aligned Alpha Fold 3 model to quantify mutation-induced perturbations in PTEN. A total of 17 features were computed for every variant, derived from solvent exposure, geometric deformation, secondary structure, backbone geometry, flexibility, and biophysical residue properties.

Solvent-accessible surface area (SASA) was calculated using PyMOL [17]. For each variant, we extracted WT_SASA, Mutant_SASA, and the difference:

$$\Delta\_SASA = Mutant\_SASA - WT\_SASA \qquad (1)$$

PyMOL was also used to compute Global_RMSD (whole-protein backbone RMSD after Cα-alignment) and Local_RMSD within a 10 Å radius centered on the mutated residue. These values were further normalized to produce a scale-independent metric (Norm_Local_RMSD) of local disruption. Approximate displacement of the mutated residue's backbone atom (Cα_Distance_Approx) was also derived directly from PyMOL coordinates [17].

Local structural context was further characterized in PyMOL using binary indicators indicating whether the residue was solvent-exposed in the wild type (WT_Exposed) or the mutant (Mut_Exposed), and whether the substitution created a steric or positional conflict (Disruptive) [17]. To determine whether the mutation altered local fold identity, residue-level secondary structure was assigned for both wild-type and mutant models using DSSP via the Bio python interface [18]. A binary variable (Secondary_Structure_Changed) and a categorical descriptor (Sec_struc_change) captured transitions such as helix-to-coil or sheet-to-coil. Backbone torsion angles (Phi_Angle and Psi_Angle) were computed using MDTraj allowing quantification of local geometric strain introduced by the amino acid substitution [19]. Residue flexibility was calculated using Alpha Fold 3 pLDDT-derived confidence values, converted into a B-factor-like metric (Local_B_Factor) [17,23].

Finally, physicochemical properties of substitution were recorded by computing the change in hydrophobicity (Hydrophobicity_Change) using the Kyte-Doolittle scale, and the change in residue charge class (Electrostatic_Change) based on transitions among positively charged, negatively charged, or neutral amino acids [24]. Together, these seventeen features: Global_RMSD, Local_RMSD, Norm_Local_RMSD, Cα_Distance_Approx, WT_SASA, Mutant_SASA, Delta_SASA, WT_Exposed, Mut_Exposed, Disruptive, Secondary_Structure_Changed, Sec_struc_change, Phi_Angle, Psi_Angle, Local_B_Factor, Hydrophobicity_Change, Electrostatic_Change provided a detailed quantitative representation of the structural consequences of every PTEN missense variant.

2.4. Data Preprocessing

All binary and categorical descriptors, including exposure status, secondary-structure transitions, and electrostatic class shifts, were converted into numerical representations. After assembly, the labeled dataset contained 197 variants (177 pathogenic and 20 benign), whereas the remaining 1,317 variants of uncertain significance (VUS) were retained only for downstream prediction. An 85-15 split was selected to maximize training data while maintaining an independent held-out test set. This produced a training set of 167 variants and a test set of 30 variants.

2.5. Machine Learning Models

Four supervised machine-learning models were developed to classify PTEN missense variants as benign or cancer-associated. The Random Forest classifier and XGBoost gradient-boosting model were used as the primary supervised learning approaches, both of which have strong performance in nonlinear biological classification tasks [20,21]. Both models were trained on the processed structural feature set, and key hyperparameters were tuned using five-fold cross-validation. Logistic regression and a decision tree were included as baseline models to assess whether simpler linear or single-tree methods could capture the structure-function relationships of PTEN variants.

2.6. Model Evaluation

For each classifier, we calculated accuracy, precision, recall, and F1-score to evaluate performance across both benign and pathogenic variants [25,26]. Because the primary clinical goal is to identify cancer-associated substitutions, we placed particular emphasis on recall and precision for the pathogenic class. Discrimination performance was quantified using the area under the receiver operating characteristic curve (ROC-AUC), computed using standard scikit-learn metrics [27]. To further assess model behavior under class imbalance, precision-recall (PR) curves were generated for all classifiers. These metrics provided information on sensitivity to pathogenic variants and robustness against false-positive predictions. Permutation feature importance was calculated for the Random Forest model to determine which structural features contributed most strongly to classification performance. In this approach, individual features are randomly shuffled while keeping all others intact, and the resulting drop in predictive accuracy is recorded [20].

2.7. Statistical Analysis of Structural Features

To compare structural properties between benign and pathogenic PTEN missense variants, we examined the distributions of the seventeen extracted structural features within the labeled dataset. Key descriptors, including ΔSASA, local RMSD, hydrophobicity change, and electrostatic change, were selected for detailed analysis because they represent biologically meaningful indicators of local destabilization and residue-level biochemical shifts. Comparisons between benign and pathogenic groups were performed using the two-sided Mann-Whitney U test, a nonparametric test suitable for small, imbalanced sample sizes [28]. This approach allowed us to evaluate whether the distributions of individual structural features differed significantly between clinically annotated classes without assuming normality. Boxplots were made to visualize distributional differences, and p-values were added directly on the plots to indicate statistical significance. These analyses provided a model-independent assessment of which structural perturbations are most strongly associated with pathogenic versus benign substitutions.

2.8. Prediction of Variants of Uncertain Significance

After training all supervised models on the labeled dataset, the optimized Random Forest and XGBoost classifiers were applied to the 1,317 PTEN variants of uncertain significance (VUS) [20,21]. Each VUS had undergone the same structural modeling, feature extraction, and preprocessing steps as the labeled variants. For each variant, both models generated a predicted probability of cancer-associated behavior as well as a binary classification. Variants for which both models predicted high pathogenicity were flagged as high-risk candidates, whereas those with concordant benign predictions were considered low-risk.

3. Results

3.1. Classification Performance of Structural Models

Using only the structural descriptors extracted from Alpha Fold 3 models, both ensemble classifiers were able to reliably separate pathogenic PTEN variants from benign ones.

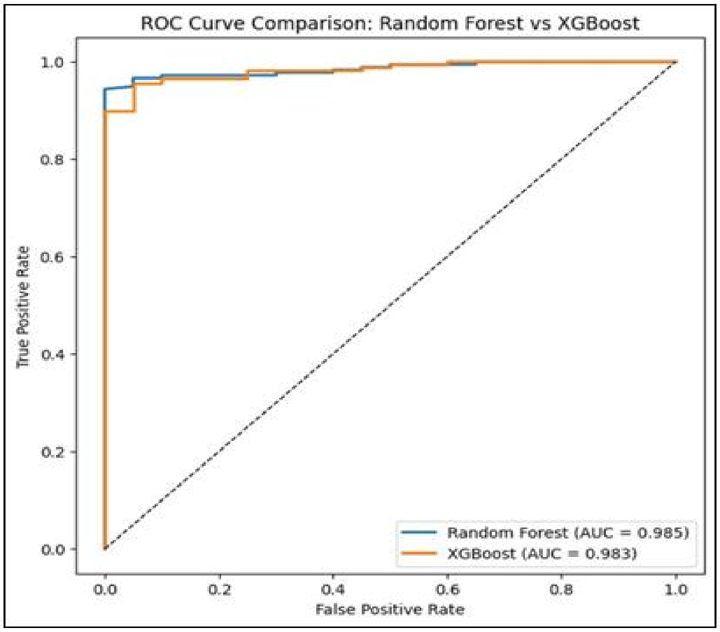

Figure 4: Receiver Operating Characteristic (ROC) Curves for Random Forest and XGBoost Classifiers Trained on Structural Features of PTEN Variants

As shown in Figure 4, both Random Forest and XGBoost achieved high AUC values (0.985 and 0.983 respectively), indicating excellent discrimination. The slightly better performance of XGBoost aligns with prior findings showing that boosted decision-tree ensembles capture subtle nonlinear molecular patterns more effectively than classical models [29]. In particular, XGBoost's sequential boosting framework allows later trees to focus on difficult, borderline variants, which helps the model detect weak structural signals such as small electrostatic shifts or local backbone disturbances that may not be captured by bagged ensembles.

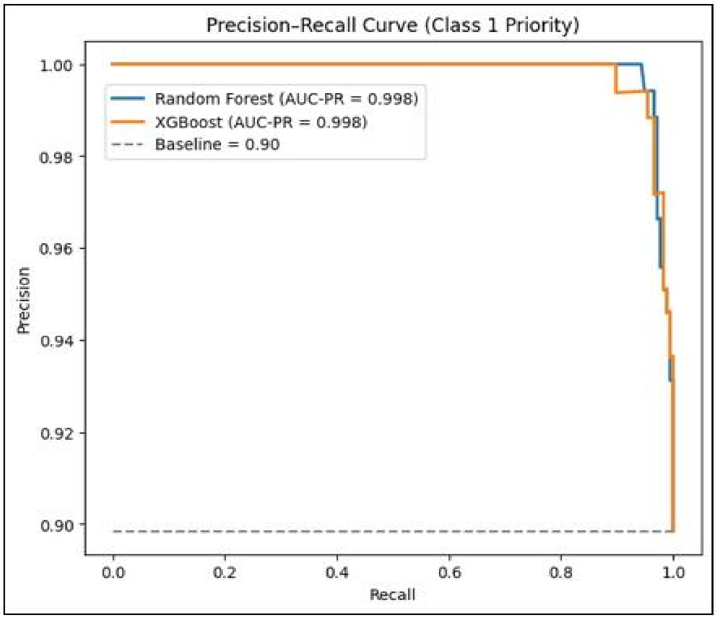

Figure 5: Precision-Recall Curves Showing Near-Perfect Sensitivity for Cancer-Associated PTEN Variants across Both Models

Precision-recall analysis (Figure 5) further illustrated this pattern. Both models maintained extremely high sensitivity for the pathogenic class, achieving PR-AUC values close to 0.998. Since clinical variant interpretation often prioritizes identifying potentially harmful mutations, this level of sensitivity is encouraging.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.954	0.944	0.955	0.974	0.988
Decision Tree	0.939	0.961	0.972	0.966	0.811
Random Forest	0.944	0.951	0.989	0.970	0.985
XGBoost	0.944	0.956	0.983	0.969	0.983

Table 1: Performance Comparison of Baseline and Ensemble Models on PTEN Variant Classification

Metric	Precision	Recall	F1-Score	Support
Random Forest				
Non-Cancerous (Class 0)	0.85	0.55	0.67	20
Cancerous (Class 1)	0.95	0.99	0.97	177
Accuracy	0.94	0.94	0.94	197
Marco Avg	0.90	0.77	0.82	197
Weighted Avg	0.94	0.94	0.94	197
XGBoost				
Non-Cancerous (Class 0)	0.80	0.60	0.69	20
Cancerous (Class 1)	0.96	0.98	0.97	177
Accuracy	0.94	0.94	0.94	197
Marco Avg	0.88	0.79	0.83	197
Weighted Avg	0.94	0.94	0.94	197

Table 2: Classification Performance Metrics for Random Forest and XGBoost Models on the Simulated PTEN Dataset

The performance metrics summarized in Table 1 and Table 2 align with these observations. Classification accuracy approached 94% for both models. Random Forest achieved the highest recall for pathogenic variants (0.989), while XGBoost produced slightly higher precision (0.956). These trends suggest that PTEN's structural features, particularly those capturing local distortion and changes in chemical environment, provide enough resolution for accurate classification.

3.2. Misclassification Patterns and Confusion Matrix

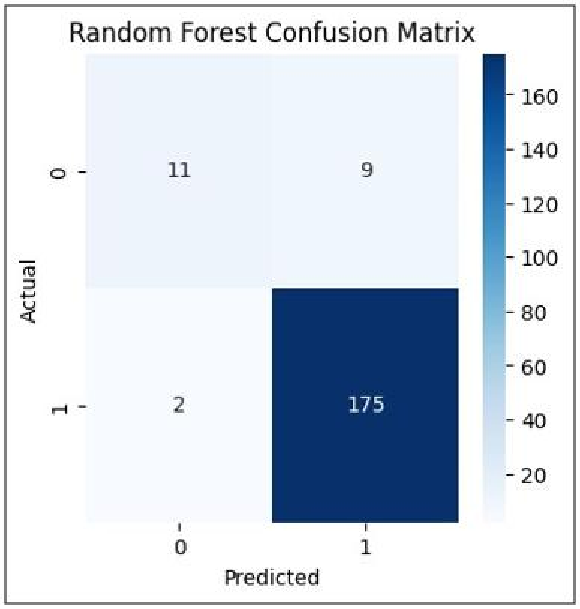

Figure 6: Confusion matrix for the Random Forest model on 197 clinically annotated PTEN variants. The matrix reveals a high true-positive rate for cancer-associated mutations, with benign misclassifications reflecting a bias toward pathogenic class predictions

The Random Forest confusion matrix (Fig. 6) shows that most pathogenic variants were correctly labeled, whereas a small subset of benign variants tended to be predicted as pathogenic. This reflects a modest bias toward sensitivity, a predictable outcome given the class imbalance (177 pathogenic vs. 20 benign) and the biological cost of missing a pathogenic mutation.

3.3. Feature Importance and Biophysical Drivers

To understand which structural changes most influenced predictions, we analyzed permutation-based feature importances.

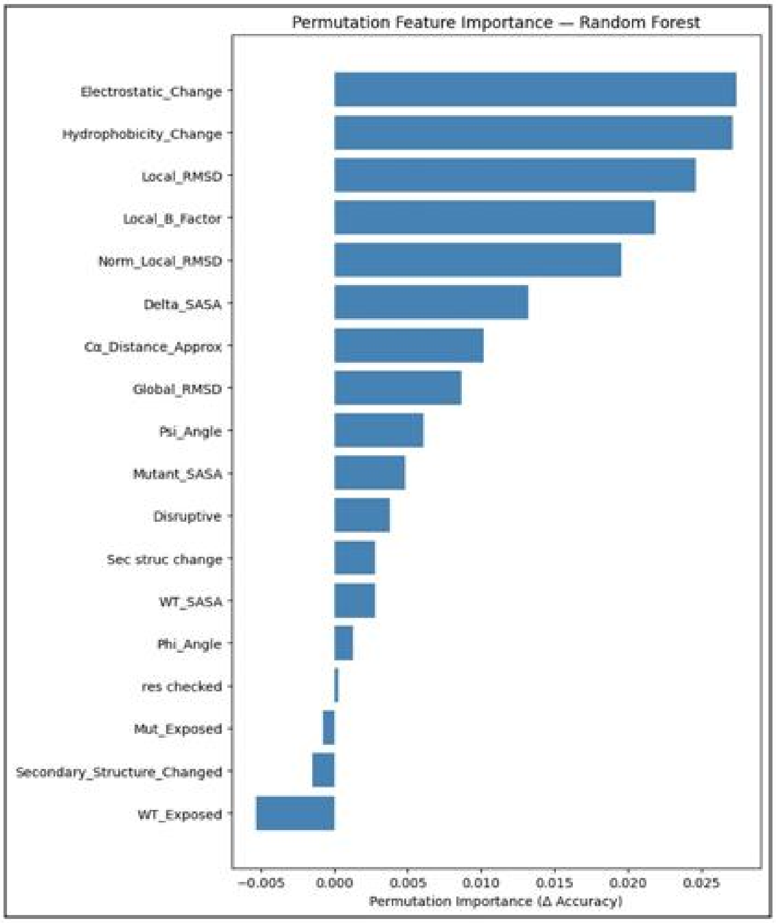

Figure 7: Permutation-Based Feature Importance Plot Ranking the Top Five Most Predictive Structural Descriptors. Electrostatic Change and Hydrophobicity Emerge as Dominant Predictors

To understand which biophysical factors drove these predictions, Figure 7 shows that electrostatic change, hydrophobicity change, local RMSD, local B-factor, and ΔSASA were the most influential features. These metrics highlight local structural destabilization, consistent with prior biochemical findings that cancer-associated PTEN mutations frequently disrupt charge balance, hydrophobic packing, or local fold stability within the phosphatase domain. Together, these metrics outline how even small residue-level perturbations can destabilize PTEN's phosphatase domain, consistent with earlier biochemical observations [16].

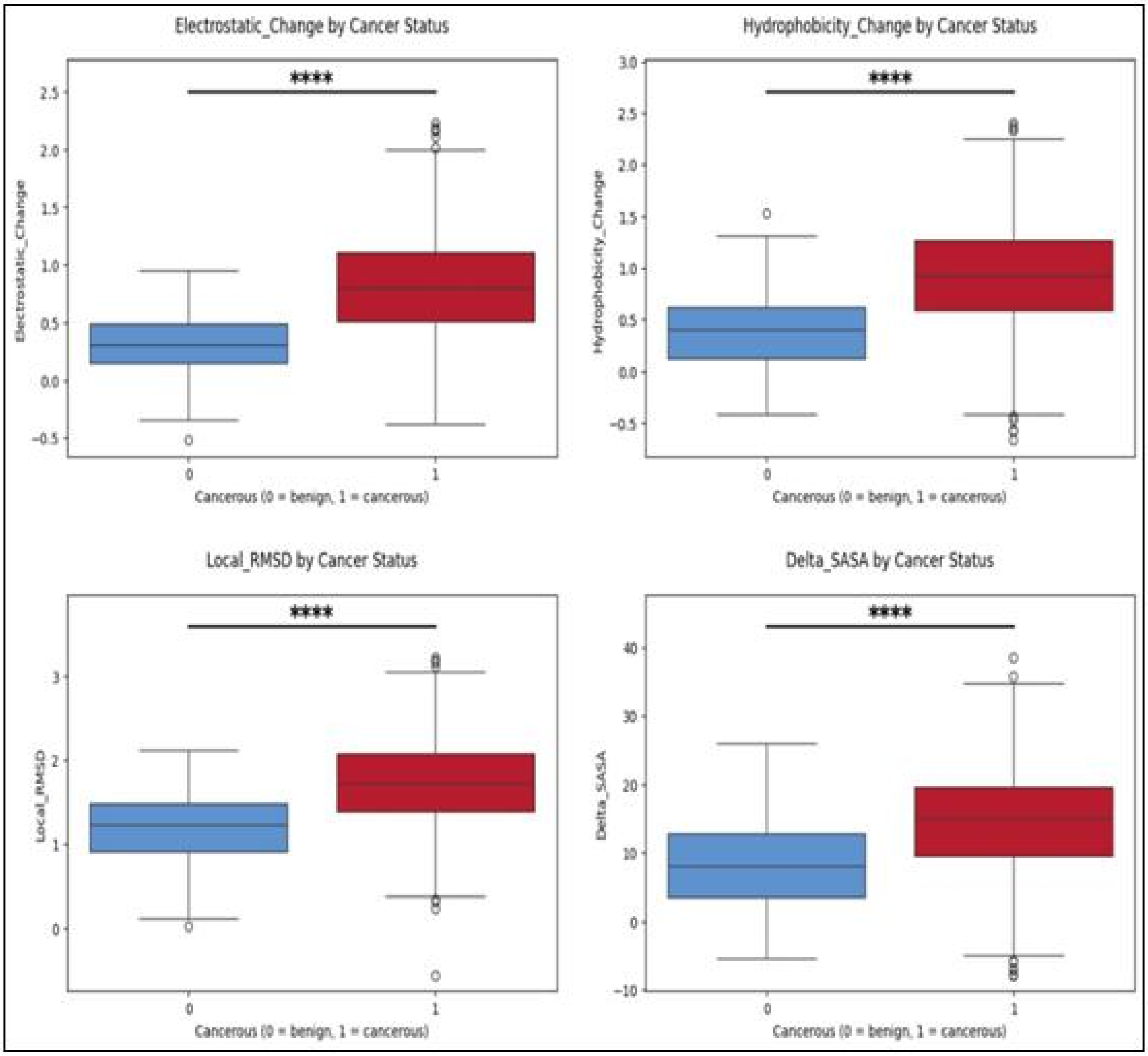

Figure 8: Boxplots comparing electrostatic and hydrophobicity changes between cancer-associated and benign PTEN variants. Cancer-associated variants show consistently higher disruption levels, with multiple outliers concentrated in structurally sensitive

Figure 8 shows boxplots comparing electrostatic and hydrophobicity changes across cancer-associated and benign variants. All four structural features showed significant differences between benign and cancer-associated variants ($p < 0.0001$), including electrostatic change, hydrophobicity change, local RMSD, and ΔSASA. Pathogenic variants exhibit significantly higher electrostatic disruption and hydrophobic shift, with outliers concentrated in sensitive regions such as the catalytic cleft and membrane-binding interface [30].

3.4. Generalization and Domain-Specific Trends

The learning curve in Figure 9 illustrates how model performance scales with training data.

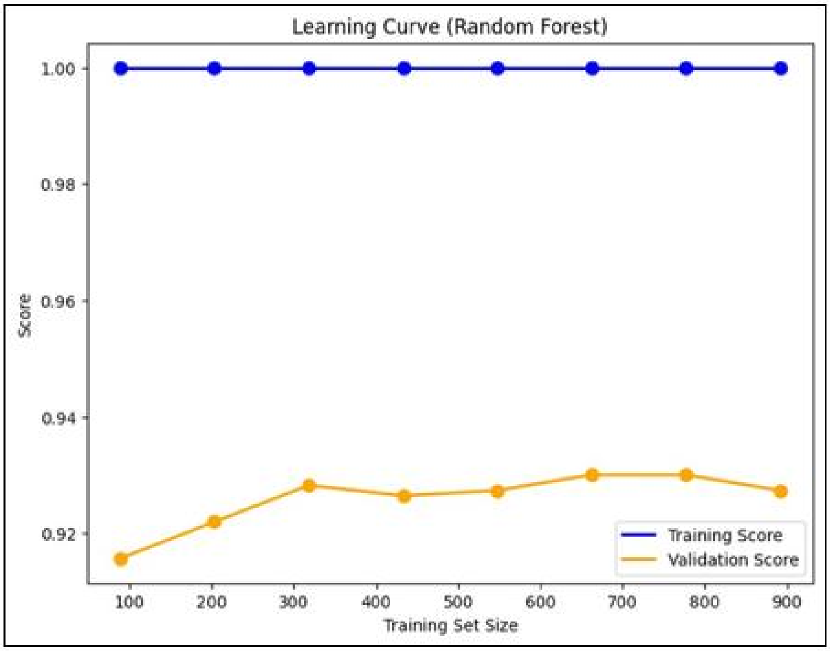

Figure 9: Learning Curves for the Random Forest Model, Depicting Training and Validation Accuracy across Increasing Sample Sizes

The learning curve (Figure 9) indicated stable generalization, with validation accuracy plateauing around 0.92-0.93. Interestingly, structural disruptions were more prominent in the phosphatase domain than in the C2 domain, consistent with the catalytic core's sensitivity to conformational changes [11]. These domain-specific patterns emerged despite the model having no explicit domain annotations, suggesting that the structural features themselves capture meaningful biological context.

3.5. Comparison with Sequence-Based Tools

Unlike sequence-based tools such as SIFT or PolyPhen-2, which rely heavily on evolutionary conservation, the structural framework here incorporates geometric and biophysical shifts that cannot be inferred from sequence alone. This allows the model to detect subtle destabilizing effects, such as local charge rearrangements or packing disruptions that are invisible to conservation-based predictors.

4. Discussion

This study demonstrates that machine learning models trained solely on structural and biophysical features can achieve high accuracy in distinguishing pathogenic PTEN variants from benign ones. As shown in Fig. 4 and Fig. 5, both Random Forest and XGBoost classifiers achieved near-perfect AUC and PR-AUC values, emphasizing the predictive power of three-dimensional structural information alone. This challenges the prevailing reliance on sequence conservation and affirms the feasibility of a purely structure-based approach to variant classification. One of the most significant findings lies in the identification of key biophysical features. As shown in Figure 7, electrostatic shifts, hydrophobic packing changes, ΔSASA, and local RMSD emerged as the most informative metrics. These features reflect known biochemical mechanisms by which PTEN function is compromised. For instance, mutations that alter local charge can destabilize the catalytic pocket, while disruptions in hydrophobic core packing may impair protein folding or membrane localization. These findings align with earlier work suggesting that PTEN activity relies on highly specific charge and geometric configurations within the phosphatase domain [13,16].

Figure 6 and Figure 8 further clarify model behavior. The confusion matrix in Figure 6 shows high recall for pathogenic variants, with a mild bias toward false positives among benign predictions. This tradeoff reflects a clinically desirable prioritization of sensitivity, especially for early risk stratification. Fig. 8 shows that pathogenic variants frequently display more extreme electrostatic and hydrophobicity shifts, forming visible outliers in boxplot comparisons. Notably, these effects were more pronounced in the phosphatase domain than in the C2 domain, consistent with the catalytic core's vulnerability to structural perturbation [13]. The model's ability to recapitulate domain-level vulnerability without explicit domain labels emphasizes the richness of structural features. Even without annotations, the classifiers recognized regions of functional importance by detecting consistent physical disruptions. This capacity to infer biology from raw geometry sets this method apart from existing tools such as SIFT or PolyPhen-2, which often struggle in poorly conserved or flexible regions [7].

Learning curves (Figure 9) indicated minimal overfitting despite a relatively small training set, suggesting that the selected features carried meaningful and generalizable signal. However, performance may vary across isoforms or in the presence of post-translational modifications. As the variant landscape in databases like ClinVar and COSMIC expands, future models will benefit from retraining on larger and more diverse datasets [6,10]. Beyond classification accuracy, this approach provides a scalable framework for interpreting variants of uncertain significance (VUS). By generating Alpha Fold 3 models for over 1,500 PTEN missense variants and extracting standardized structural features, we created one of the most comprehensive structural maps of PTEN variation to date. These predictions are mechanistically interpretable, allowing researchers and clinicians to move beyond black-box scores and assess the likely structural impact of a given mutation [11]. Together, these findings highlight the power of deep-learning-enabled structural modeling in variant interpretation. The ability to predict functionally damaging mutations using only three-dimensional features marks a major shift in computational genomics. This framework not only improves accuracy in PTEN variant classification but also reveals mechanistic patterns that can guide laboratory validation and clinical decision-making. More broadly, this study offers a generalizable blueprint for structure-based analysis of other clinically relevant genes and highlights the growing importance of structural bioinformatics in precision oncology.

4.1. Limitations

Nonetheless, several limitations remain. Alpha Fold structures, while highly accurate, are static representations and may not fully capture PTEN's conformational plasticity [17,19]. The current feature set omits dynamics, protein-protein interactions, and post-translational regulation, all of which may affect function [3]. Clinical annotations remain limited in number and may reflect biases in mutation databases or reporting patterns. Future studies could address these gaps by incorporating molecular dynamics simulations, high-throughput functional assays, and cross-species evolutionary conservation. The model could also be expanded with ensemble-based structural features or data on cellular localization and phosphatase activity [13]. Ultimately, integrating structural modeling with experimental and clinical evidence may yield the most robust classifiers for clinical use.

5. Conclusion

This study demonstrates that combining Alpha Fold 3-based structural modelling with supervised learning offers an effective strategy for interpreting PTEN missense variants. By quantifying seventeen geometric and biophysical features for more than 1,500 variants, the framework establishes a scalable and systematic approach to variant evaluation. The high performance of the Random Forest and XGBoost classifiers underscores the strength of structure-derived signals in predicting pathogenicity. Importantly, the workflow provides structural assessments and predicted classifications for over 1,300 variants of uncertain significance from ClinVar and COSMIC addressing a major challenge in clinical genetics [6,10]. These predictions form a resource for prioritizing high-risk variants and guiding downstream functional studies or clinical review. The findings support earlier observations that PTEN pathogenicity is closely linked to localized disruptions affecting stability, folding, and charge balance [3,13]. By translating these principles into quantifiable computational features, the approach connects mechanistic insight with machine-learning prediction. It also provides a template that can be adapted to other tumor suppressors with structurally mediated dysfunction. As structural prediction technologies advance and variant databases expand, this framework can continue to evolve. The integration of dynamic modelling, experimental assays, and additional structural features could further refine accuracy and enable broader application in precision oncology. Together, these results highlight the growing role of structure-based methods in resolving uncertain variants and improving our understanding of genotype-phenotype relationships [24].

References

1. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell, 144*(5), 646-674.
2. Song, M. S., Salmena, L., & Pandolfi, P. P. (2012). The functions and regulation of the PTEN tumour suppressor. *Nature reviews Molecular cell biology, 13*(5), 283-296.
3. Lee, Y. R., Chen, M., & Pandolfi, P. P. (2018). The functions and regulation of the PTEN tumour suppressor: new modes and prospects. *Nature reviews Molecular cell biology, 19*(9), 547-562.
4. Pilarski, R. (2019). PTEN hamartoma tumor syndrome: a clinical overview. *Cancers, 11*(6), 844.
5. "UniProt: the universal protein knowledgebase in 2023." *Nucleic acids research 51*, no. D1 (2023): D523-D531.
6. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., ... & Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research, 46*(D1), D1062-D1067.
7. Thusberg, J., & Vihinen, M. (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Human mutation, 30*(5), 703-714.
8. Tam, B., Qin, Z., Zhao, B., Sinha, S., Lei, C. L., & Wang, S. M. (2024). Classification of mlh1 missense vus using protein structure-based deep learning-ramachandran plot-molecular dynamics simulations method. *International Journal of Molecular Sciences, 25*(2), 850.
9. Joruiz, S. M., & Bourdon, J. C. (2016). p53 isoforms: key regulators of the cell fate decision. *Cold Spring Harbor perspectives in medicine, 6*(8), a026039.
10. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... & Forbes, S. A. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research, 47*(D1), D941-D947.
11. Post, K. L., Belmadani, M., Ganguly, P., Meili, F., Dingwall, R., McDiarmid, T. A., ... & Haas, K. (2020). Multi-model functionalization of disease-associated PTEN missense mutations identifies multiple molecular mechanisms underlying protein dysfunction. *Nature communications, 11*(1), 2073.
12. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature, 596*(7873), 583-589.
13. Mighell, T. L., Evans-Dutson, S., & O'Roak, B. J. (2018). A saturation mutagenesis approach to understanding

PTEN lipid phosphatase activity and genotype-phenotype relationships. *The American Journal of Human Genetics, 102*(5), 943-955.

14. Papa, A., Wan, L., Bonora, M., Salmena, L., Song, M. S., Hobbs, R. M., ... & Pandolfi, P. P. (2014). Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function. *Cell, 157*(3), 595-610.
15. Myers, M. P., Pass, I., Batty, I. H., Van der Kaay, J., Stolarov, J. P., Hemmings, B. A., ... & Tonks, N. K. (1998). The lipid phosphatase activity of PTEN is critical for its tumor supressor function. *Proceedings of the National Academy of Sciences, 95*(23), 13513-13518.
16. Smith, I. N., & Briggs, J. M. (2016). Structural mutation analysis of PTEN and its genotype-phenotype correlations in endometriosis and cancer. *Proteins: Structure, Function, and Bioinformatics, 84*(11), 1625-1643.
17. Schrödinger, LLC, *The PyMOL Molecular Graphics System*, Version 2.0, 2015.
18. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422.
19. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., ... & Pande, V. S. (2015). MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal, 109*(8), 1528-1532.
20. Breiman, L. (2001). Random Forests Mach. Learn., 45 (1), 5–32. *ed*.
21. Chen, T. (2016). XGBoost: A Scalable Tree Boosting System. *Cornell University*.
22. Den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., ... & Human Genome Variation Society (HGVS), the Human Variome Project (HVP), and the Human Genome Organisation (HUGO). (2016). HGVS recommendations for the description of sequence variants: 2016 update. *Human mutation, 37*(6), 564-569.
23. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., ... & Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature, 596*(7873), 590-596.
24. Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology, 157*(1), 105-132.
25. Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd.
26. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.
27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.
28. Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.
29. Zeng, J., Chan, L., Li, L., et al. (2024). Interpretable deep learning for biomolecular structural feature discovery. *Nat. Mach. Intell*.
30. Chillón-Pino, D., Badonyi, M., Semple, C. A., & Marsh, J. A. (2024). Protein structural context of cancer mutations reveals molecular mechanisms and candidate driver genes. *Cell reports, 43*(11).