AlphaFold-derived local structural confidence in BRCA1 missense variant interpretation: a ClinVar-based computational analysis

Yaroslav Yasinskyi

yasyar@knu.ua

Taras Shevchenko National University of Kyiv

Mariana Chopei

Taras Shevchenko National University of Kyiv

Andrei Sivolob

Taras Shevchenko National University of Kyiv

Research Article

Keywords: BRCA1, variant of uncertain significance, ClinVar, AlphaFold, pLDDT, missense variants, variant interpretation, computational genomics

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10346088/v1

Additional Declarations: No competing interests reported.

Abstract

Background

Breast Cancer 1 (*BRCA1*) gene variations are known to be associated with breast and ovarian cancers. At the same time, interpreting *BRCA1* missense variants remains challenging, particularly for variants of uncertain significance (VUS), because sequence-based predictors do not fully capture local structural context. AlphaFold-derived predicted Local Distance Difference Test (pLDDT) scores provide residue-level confidence estimates that distinguish well-ordered regions from regions of lower structural confidence. We assessed whether local pLDDT is associated with the clinical classification of BRCA1 missense single-nucleotide variants and whether it stratifies the performance of commonly used in silico predictors.

Results

BRCA1 missense variants collected in the ClinVar database were filtered, annotated with the Ensembl Variant Effect Predictor, mapped to residue-level pLDDT values from the AlphaFold BRCA1 model, and integrated with UniProt domain annotations. The final dataset comprised 2,399 unique missense variants, 2,390 of which had mappable pLDDT values. Pathogenic and likely pathogenic variants were strongly enriched in regions with pLDDT > = 70 relative to benign and likely benign variants (odds ratio 30.13, 95% confidence interval 17.95–50.60; Fisher's exact test $p = 1.41 \times 10^{-61}$). This association remained significant within the BRCT region (odds ratio 9.35, 95% confidence interval 2.04–42.93; p = 0.00109) and in ClinVar subsets with stronger review support. Continuous pLDDT distributions also differed markedly between pathogenic/likely pathogenic and benign/likely benign variants (Mann–Whitney U test $p = 1.31 \times 10^{-60}$). In receiver operating characteristic analysis, SIFT outperformed PolyPhen overall (AUC 0.872 vs 0.692), whereas predictor performance differed across pLDDT strata.

Conclusions

In BRCA1, AlphaFold-derived local confidence was strongly associated with clinical missense variant classification and may serve as useful contextual structural evidence rather than as a standalone pathogenicity score. pLDDT may therefore help refine variant prioritization and support the interpretation of BRCA1 VUS in conjunction with existing computational and clinical evidence.

Background

BRCA1 is among the most clinically important cancer susceptibility genes and a central component of the hereditary breast and ovarian cancer syndrome framework. At the protein level, BRCA1 contains functionally critical regions, most notably the N-terminal RING domain and the C-terminal BRCT repeats, which mediate ubiquitin-related signaling, phosphoprotein interactions, and DNA damage response

functions. Although loss-of-function variants in BRCA1 are often clinically interpretable, the classification of missense variants remains substantially more difficult because single amino acid substitutions can have heterogeneous effects on protein stability, intermolecular interactions, and downstream cellular function. This uncertainty directly affects genetic counseling, surveillance strategies, and decisions about clinical risk reduction [1, 2].

Public databases such as ClinVar have greatly improved access to variant interpretations and the evidence supporting them, yet uncertainty remains a major challenge in missense variant interpretation. In BRCA1, many variants encountered in diagnostic practice remain difficult to classify confidently, particularly when functional evidence is incomplete or unavailable. High-throughput functional studies, including saturation genome editing and multiplexed assays, have advanced the field substantially and shown that the effects of BRCA1 variants are strongly shaped by local protein context. Even so, these resources do not remove the need for scalable computational approaches that can help prioritize variants for follow-up and assist interpretation in routine analysis [3–5].

Accordingly, computational prediction remains an important component of missense variant assessment under the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) framework. Widely used tools such as SIFT and PolyPhen-2 are valuable because they provide rapid estimates of the likely functional consequences of amino acid substitutions, but they are driven primarily by sequence conservation and related handcrafted features rather than by explicit gene-specific structural context. Consequently, performance can vary across genes, protein domains, and variant sets. The stronger performance of gene-specific models such as BRCA-ML compared with more general predictors suggests that biological context plays an important role in predicting the effects of BRCA1 variants [2, 5–7].

Recent advances in AI-based protein structure prediction provide a potentially useful way to capture that context. In particular, AlphaFold-derived models provide residue-level confidence estimates through the predicted local distance difference test (pLDDT). Importantly, pLDDT is not a pathogenicity score. Rather, it reflects confidence in local model geometry and has been widely used to distinguish well-ordered structural regions from regions that are low-confidence, flexible, or disordered. This makes pLDDT attractive as a contextual feature for variant interpretation, especially in proteins such as BRCA1 that combine structured functional domains with long regions of lower structural certainty. At the same time, low pLDDT should not be taken to mean that a region is biologically unimportant, as some low-confidence regions may still be functional or become structured under particular conditions [8, 9].

For BRCA1, this raises a specific and clinically relevant question: whether local AlphaFold confidence is associated with the observed clinical pathogenicity of missense variants, and whether that structural context modifies the discriminative value of established in silico predictors. This question is especially important because BRCA1 has a rich body of functional and clinical evidence, a well-defined domain architecture, and a substantial burden of variants that remain uncertain in practice. A focused gene-level analysis of pLDDT as a source of contextual evidence could therefore help determine whether AlphaFold

confidence scores can improve missense variant interpretation and guide the prioritization of BRCA1 VUS for further analysis [3–5].

In this study, we examined ClinVar-annotated BRCA1 missense single-nucleotide variants in the context of AlphaFold-derived residue-level pLDDT. Specifically, we asked whether pathogenic/likely pathogenic and benign/likely benign variants differed in their local pLDDT distributions and whether pLDDT stratification altered the performance of standard missense effect predictors. By treating pLDDT as contextual structural evidence rather than as a standalone classifier, this work aimed to assess its potential value for prioritizing and interpreting BRCA1 variants of uncertain significance.

Methods

Study design

This study was designed as a reproducible computational analysis of publicly available data to evaluate whether local structural confidence from AlphaFold, measured as per-residue predicted Local Distance Difference Test (pLDDT), is associated with the clinical interpretation of BRCA1 missense single-nucleotide variants (SNVs) and whether pLDDT stratifies the discriminative performance of commonly used in silico missense predictors. The primary comparison contrasted pathogenic/likely pathogenic (P/LP) versus benign/likely benign (B/LB) BRCA1 missense variants, whereas variants of uncertain significance (VUS) were retained for descriptive analyses and exploratory prioritization [3, 10, 11].

Data sources

Clinical variant data were obtained from the ClinVar GRCh38 variant call format (VCF) release dated 1 February 2026, which provides variant-level assertions of clinical significance together with review-status metadata [3]. Structural confidence values were derived from the AlphaFold Protein Structure Database model for human BRCA1 (UniProt accession P38398; model AF-P38398-F1), accessed on 4 February 2026, using residue-level pLDDT values encoded in the PDB structure file [11]. Protein domain annotations for BRCA1 were retrieved from the UniProt REST API for accession P38398 on 4 February 2026 and were used to define RING and BRCT domain intervals [12].

Variant extraction and filtering

Variant extraction and filtering
ClinVar records were filtered to retain only BRCA1 missense SNVs. Variants were required to meet all of the following criteria: the GENEINFO field contained BRCA1; the molecular consequence (MC) field contained the Sequence Ontology term missense_variant (SO:0001583); and both the reference and alternate alleles were single nucleotides. Clinical significance labels were normalized and collapsed into five categories: pathogenic, likely pathogenic, benign, likely benign, and uncertain significance. Records

labeled as conflicting interpretations of pathogenicity were excluded. Duplicate entries were removed at the allele level using the genomic key (chromosome, position, reference allele, alternate allele). Review-status annotations from CLNREVSTAT were retained for subsequent robustness analyses [3].

Variant annotation with Ensembl VEP

Filtered BRCA1 missense alleles were exported to a minimal VCF file and then annotated using the Ensembl Variant Effect Predictor (VEP), version 115.2. This was done in a Docker-based environment using an offline GRCh38 cache. The tabular output from VEP was used to obtain transcript-specific consequence annotations, protein positions, HGVS nomenclature, and predictor outputs, including SIFT and PolyPhen scores [6, 7, 10]. When the CANONICAL field was available, only canonical transcript annotations were retained before merging. The ClinVar and VEP tables were then merged on chromosome, genomic position, and alternate allele.

Protein coordinates were parsed preferentially from the Protein_position field; when this field was unavailable or ambiguous, the amino acid position was recovered from the protein-level HGVS annotation (HGVSp). Variants without a mappable protein coordinate could not be assigned a residue-level pLDDT value and were therefore excluded from analyses involving pLDDT.

Extraction of residue-level pLDDT and structural context assignment

The AlphaFold BRCA1 structure was downloaded through the AlphaFold API, and the longest standard-protein chain in the structure file was selected for residue parsing. Per-residue pLDDT values were extracted from the PDB B-factor field, which AlphaFold repurposes to store residue-level confidence values. If a Cα atom was present for a residue, its B-factor was used directly; otherwise, the mean B-factor across atoms in the residue was used. A residue-indexed pLDDT array was then created and mapped to each variant according to its protein position [11].

In descriptive analyses, pLDDT values were categorized into four groups: 90 or above, 70–89.999, 50–69.999, and below 50. For the main association analysis, pLDDT was dichotomized using a cutoff of 70, a standard threshold distinguishing well-ordered from less reliable AlphaFold regions [8, 11]. Additionally, UniProt domain intervals were used to assign each protein position to a specific domain, such as RING, BRCT, or no annotated domain [12].

Construction of the master analysis table

A master table was generated by combining the filtered ClinVar dataset, the VEP annotation output, the residue-level pLDDT mapping, and the UniProt domain assignments. Because transcript-aware variant annotation may still yield more than one row per genomic allele after merging, two complementary representations were retained: a merged analysis table for downstream statistical processing and a deduplicated allele-level table based on (chromosome, position, reference allele, alternate allele) for

unique-variant summaries. This approach allowed descriptive reporting at the unique-allele level while preserving the canonical merged output of the annotation pipeline.

Outcome definitions and analytical subsets

Outcome definitions and analytical subsets
For binary pathogenicity analyses, variants were coded as positive (1) if classified as pathogenic or likely pathogenic and as negative (0) if classified as benign or likely benign. Variants classified as uncertain significance were excluded from the primary binary comparisons. Variants with missing pLDDT values were excluded from pLDDT-dependent tests. Review-status sensitivity analyses were performed on two higher-confidence ClinVar subsets: reviewed by expert panel and multiple submitters with no conflicts.

For ROC analyses, only entries with non-missing numeric SIFT and PolyPhen values, non-missing pLDDT, and binary class labels were retained. No a priori power calculation was performed because the study included all eligible publicly available BRCA1 missense SNVs from the selected dataset release.

Statistical analysis

The primary association framework was applied overall and within higher-confidence ClinVar review-status subsets. A domain-restricted inferential analysis was retained for the BRCT region. A separate threshold-based comparison for the RING region was not reported because the domain annotation step did not yield RING as an independent analytical subset, and inspection of RING-region variants showed that all evaluable variants fell within the high-pLDDT stratum.

To assess whether the association extended beyond a single threshold-based representation, continuous pLDDT distributions were compared between P/LP and B/LB variants using a two-sided Mann–Whitney U test.

Predictive performance of SIFT and PolyPhen was evaluated using receiver operating characteristic (ROC) analysis and the area under the ROC curve (AUC). To align score directionality across predictors, SIFT values were sign-inverted before ROC/AUC computation because lower SIFT scores indicate greater predicted deleteriousness, whereas higher PolyPhen scores indicate greater predicted damagingness [6, 7]. AUC estimates were obtained overall and after stratification by pLDDT (< 70 vs. >=70; additional descriptive pLDDT bins were evaluated when sample size permitted). Ninety-five percent confidence intervals for AUC were estimated by nonparametric bootstrap with 2,000 resamples using percentile intervals and a fixed random seed for reproducibility.

Exploratory prioritization of VUS

As an exploratory downstream analysis, VUS were screened to identify candidates warranting external validation. Variants were prioritized when they met all of the following criteria: SIFT < 0.05, PolyPhen >

0.90, and pLDDT > = 70. Candidate variants were then ranked by descending PolyPhen score and ascending SIFT score to generate a shortlist for potential follow-up.

Software

All processing steps were implemented in a custom script-based workflow written in Python. Data manipulation was performed with pandas and NumPy; statistical testing used SciPy and statsmodels; ROC/AUC analyses used scikit-learn; plotting was performed with matplotlib; ClinVar VCF parsing used vcfpy; structure parsing used Biopython; and data retrieval from external resources used requests. The workflow covered ClinVar download, BRCA1 missense extraction, VCF generation for VEP, AlphaFold pLDDT extraction, UniProt domain retrieval, master-table construction, statistical analyses, and figure generation. The archived analysis scripts and supporting data are available in Zenodo [13].

Results

Cohort composition and variant coverage

After ClinVar filtering, VEP annotation, and assembly of the merged analysis table, the dataset comprised 2,851 annotated records corresponding to 2,399 unique BRCA1 missense SNVs (Table 1). Residue-level pLDDT values could be assigned to 2,390 unique variants; nine variants lacked an unambiguous protein position and were therefore excluded from pLDDT-dependent analyses. At the unique-allele level, the cohort included 167 benign, 239 likely benign, 132 pathogenic, 100 likely pathogenic, and 1,761 variants of uncertain significance (VUS).

Table 1
Composition of the BRCA1 missense variant cohort at the unique-allele level and residue-level pLDDT coverage.

Class	Unique variants	With pLDDT	Coverage, %
Benign	167	167	100.0
Likely benign	239	235	98.3
Likely pathogenic	100	100	100.0
Pathogenic	132	132	100.0
Uncertain significance	1,761	1,756	99.7
Total	2,399	2,390	99.6

Pathogenic variants were enriched in high-confidence structural regions

Across ClinVar classes, pathogenic and likely pathogenic variants were concentrated in higher-confidence AlphaFold regions, whereas benign, likely benign, and especially VUS were more frequently

observed in lower-confidence regions. This pattern was evident in the class-stratified pLDDT bin distributions (Fig. 1) and supported the treatment of pLDDT as a contextual structural variable in subsequent analyses.

Using a dichotomous threshold of pLDDT > = 70, pathogenic/likely pathogenic (P/LP) variants were strongly enriched in high-confidence regions relative to benign/likely benign (B/LB) variants (Fig. 2). Among unique variants with available pLDDT, 213 of 232 P/LP variants (91.8%) were located in regions with pLDDT > = 70, compared with 109 of 402 B/LB variants (26.8%). This yielded an odds ratio (OR) of 30.13 (95% CI [17.95, 50.60]; Fisher's exact test $p = 1.41 \times 10^{-61}$, Table 2).

This association was not confined to a threshold-based representation. When pLDDT was analyzed as a continuous variable, the distributions differed markedly between P/LP and B/LB variants, with a clear shift of P/LP variants toward higher values (Mann–Whitney U test $p = 1.31 \times 10^{-60}$).

The association remained detectable in BRCT and in higher-confidence ClinVar subsets

To test whether the observed enrichment simply reflected broad differences between highly structured and low-confidence regions, a domain-restricted analysis was performed for the BRCT region. Within BRCT, the enrichment of P/LP variants at positions with pLDDT > = 70 remained significant (OR = 9.35, 95% CI [2.04, 42.93]; p = 0.00109), indicating that the signal persisted even within a functionally important structured domain (Table 2).

The main effect also remained robust when analyses were restricted to ClinVar entries with stronger review support. In the reviewed by expert panel subset, the OR was 58.64 (95% CI [22.76, 151.10]; $p = 8.95 \times 10^{-29}$). In the multiple submitters with no conflicts subset, the OR was 30.46 (95% CI [12.07, 76.88]; $p = 4.80 \times 10^{-17}$). These findings argue against the overall association being driven primarily by lower-confidence ClinVar assertions.

Table 2
Enrichment of pathogenic/likely pathogenic BRCA1 missense variants in high-confidence structural regions (pLDDT ≥ 70).

Analysis subset	P/LP, ≥ 70	P/LP, < 70	B/LB, ≥ 70	B/LB, < 70	Odds ratio (95% CI)	p value
Overall	213	19	109	293	30.13 (17.95–50.60)	1.41×10^{-61}
BRCT domain	120	2	77	12	9.35 (2.04–42.93)	0.00109
Expert panel	68	6	23	119	58.64 (22.76–151.10)	8.95×10^{-29}
Multiple submitters, no conflicts	81	9	13	44	30.46 (12.07–76.88)	4.80×10^{-17}

Predictor performance varied across structural-confidence strata

Receiver operating characteristic (ROC) analysis was performed on the subset of variants with available binary class labels, pLDDT values, and numeric SIFT and PolyPhen scores (n = 603). In the full evaluable subset, SIFT showed stronger discrimination between P/LP and B/LB variants than PolyPhen, with an AUC of 0.872 (95% CI [0.847, 0.897]) compared with 0.692 (95% CI [0.651, 0.732]) for PolyPhen.

When the same analysis was stratified by pLDDT, predictor performance changed non-uniformly (Table 3). In the pLDDT < 70 subset (n = 308), SIFT achieved an AUC of 0.828 (95% CI [0.752, 0.900]), whereas PolyPhen achieved an AUC of 0.605 (95% CI [0.453, 0.746]). In the pLDDT > = 70 subset (n = 295), SIFT achieved an AUC of 0.775 (95% CI [0.717, 0.830]), whereas PolyPhen achieved an AUC of 0.713 (95% CI [0.649, 0.777]). Thus, pLDDT did not act as a universal indicator of predictor quality; instead, it functioned as a stratifying contextual variable associated with different predictor behavior in different structural environments.

The intermediate pLDDT range of 50-69.999 contained only 13 variants in the ROC-ready subset and was therefore not interpreted separately.

Table 3
Discriminative performance of SIFT and PolyPhen-2 overall and after stratification by local pLDDT.

Subset	n	SIFT AUC (95% CI)	PolyPhen-2 AUC (95% CI)
Overall	603	0.872 (0.847–0.897)	0.692 (0.651–0.732)
pLDDT < 70	308	0.828 (0.752–0.900)	0.605 (0.453–0.746)
pLDDT ≥ 70	295	0.775 (0.717–0.830)	0.713 (0.649–0.777)

Exploratory prioritization of VUS

As an exploratory downstream step, the analysis pipeline generated a ranked shortlist of VUS candidates for external follow-up. Variants were prioritized when they met all three criteria simultaneously: SIFT < 0.05, PolyPhen > 0.90, and pLDDT > = 70. The top 15 candidates are provided in Table 4.

Rank	cDNA	Protein	pLDDT	SIFT	PolyPhen	ClinVar review status
1	c.89T > G	p.Leu30Trp	90.56	0.00	0.997	single submitter
2	c.5346G > C	p.Trp1782Cys	95.56	0.04	0.991	single submitter
3	c.185C > T	p.Pro62Leu	86.19	0.00	0.990	multiple submitters, no conflicts
4	c.152T > C	p.Leu51Pro	84.69	0.00	0.989	multiple submitters, no conflicts
5	c.5549T > C	p.Leu1850Pro	93.38	0.00	0.988	multiple submitters, no conflicts
6	c.5102T > C	p.Leu1701Pro	95.81	0.00	0.988	multiple submitters, no conflicts
7	c.5098A > C	p.Thr1700Pro	95.69	0.00	0.988	multiple submitters, no conflicts
8	c.5525T > C	p.Val1842Ala	96.94	0.00	0.987	single submitter
9	c.5445G > C	p.Trp1815Cys	86.94	0.00	0.987	single submitter
10	c.5146T > C	p.Tyr1716His	97.25	0.00	0.987	multiple submitters, no conflicts
11	c.107C > G	p.Ser36Cys	91.69	0.00	0.987	single submitter
12	c.185C > G	p.Pro62Arg	86.19	0.00	0.986	multiple submitters, no conflicts
13	c.89T > C	p.Leu30Ser	90.56	0.00	0.986	single submitter
14	c.155T > C	p.Leu52Pro	83.94	0.00	0.985	multiple submitters, no conflicts
15	c.5384T > A	p.Leu1795His	88.81	0.02	0.984	single submitter

Discussion

The present study showed that BRCA1 pathogenic and likely pathogenic missense variants were strongly enriched in regions with higher AlphaFold local confidence, as measured by residue-level pLDDT, and that this association remained detectable in a functionally important structured region of the protein and in ClinVar subsets with stronger review support. At the same time, pLDDT stratified the observed performance of commonly used in silico predictors, indicating that local structural confidence is not merely descriptive but may serve as a practically useful contextual variable for interpretation. Taken

together, these findings suggest that pLDDT may help refine the interpretation of computational evidence during BRCA1 missense variant assessment, particularly in the context of VUS prioritization.

Biologically, this association is plausible. BRCA1 contains functionally critical structured regions, including the RING and BRCT domains, where amino acid substitutions can disrupt protein stability, phosphopeptide recognition, and DNA damage response functions. In that context, it is plausible that clinically pathogenic missense variants would accumulate preferentially in better-ordered, more structurally constrained parts of the protein. The persistence of the association within BRCT is especially important because it argues against an overly simplistic explanation in which the overall effect is driven only by broad differences between highly structured domains and flexible low-confidence regions elsewhere in the protein. Instead, the results suggest that local structural confidence captures information relevant to functional constraint even within a known disease-relevant domain.

These findings should not, however, be interpreted as evidence that pLDDT is itself a pathogenicity predictor. pLDDT is a model-confidence metric, and low values often correspond to disordered, flexible, or conditionally structured regions rather than to biological insignificance. Accordingly, the principal value of pLDDT in this study is contextual: it helps distinguish positions at which structure-based reasoning is more or less likely to be informative. This interpretation is consistent with broader discussions in the AlphaFold literature, where pLDDT is often useful for identifying regions of lower structural certainty but must be interpreted cautiously, particularly in proteins with intrinsic disorder or context-dependent conformations.

The predictor-stratified results reinforce that point. SIFT outperformed PolyPhen overall in the present dataset, whereas PolyPhen showed greater discrimination in the higher-pLDDT subset than in the lower-pLDDT subset. Although these differences were not formally tested as between-stratum AUC contrasts, they support the broader idea that predictor performance depends on the structural environment in which a substitution occurs. This interpretation is also consistent with prior work showing that gene-specific models can outperform general-purpose missense predictors in BRCA1 and BRCA2, implying that local biological context is a major determinant of predictive performance [5]. In practical terms, the current results suggest that confidence in a predictor score may reasonably depend on whether the affected residue lies in a well-modeled or low-confidence region.

From a translational perspective, the most realistic use case is not direct reclassification but prioritization. ClinVar contains many BRCA1 missense VUS, and functional assays, segregation studies, and expert reinterpretation resources remain limited relative to the volume of uncertain variants. In that setting, pLDDT may serve as a useful auxiliary feature for triaging variants for downstream follow-up, especially when combined with conventional predictors and domain context. The shortlisted VUS candidates generated by the present workflow illustrate this applied value: rather than claiming clinical resolution, the pipeline identifies variants for which external validation may be particularly informative or urgent. Such a role fits well within the ACMG/AMP framework, in which computational evidence is considered supportive but is not sufficient on its own to establish pathogenicity.

Several limitations should be acknowledged. First, the study is based on ClinVar assertions, which are highly valuable but heterogeneous in evidentiary depth and submission history; although the main effect remained evident in the expert-panel and multiple-submitters subsets, residual classification noise cannot be excluded. Second, the present analysis is restricted to a single gene (BRCA1), so the generalizability of the observed pattern to other clinically important genes remains uncertain. Third, the structural feature used here is limited to residue-level pLDDT from a single AlphaFold model and does not incorporate other potentially informative descriptors such as solvent accessibility, interface location, estimates of changes in the protein structure free energy, AlphaMissense-like scores, or experimentally measured functional readouts. Fourth, the merged analysis pipeline is transcript-aware, and although the inferential statistics were grounded in allele-level deduplication where appropriate, future analyses should include an explicitly exported final analysis table containing one row per unique variant. Finally, the study did not include an external validation cohort or formal functional confirmation of prioritized VUS candidates.

These limitations also point to the most useful next steps. The clearest extension would be to benchmark pLDDT-informed interpretation against stronger contemporary comparators, including gene-specific or ensemble predictors, and to test whether the same contextual effect is observed in additional hereditary cancer genes. A second important step would be to evaluate whether pLDDT strata correspond to differences in externally measured functional outcomes, for example from multiplexed assays or curated experimental datasets, to determine whether the observed ClinVar-based association reflects a deeper relationship between structural confidence and mutational intolerance. Interpreted in this context, the present study does not position pLDDT as a replacement for established classification frameworks; rather, it supports its use as an interpretable contextual feature that may help organize uncertainty in BRCA1 missense variant assessment.

Conclusions

In this study, BRCA1 pathogenic and likely pathogenic missense variants were markedly enriched in regions with higher AlphaFold local confidence, as measured by residue-level pLDDT. This association was strong in the overall dataset, remained detectable within the BRCT region, and persisted in ClinVar subsets with stronger review support, indicating that the observed pattern was not driven solely by lower-confidence clinical assertions.

In this analysis, pLDDT did not serve as a standalone pathogenicity score. Instead, it functions as a contextual structural feature that helps distinguish regions with different apparent functional constraints and in which conventional in silico predictors may perform differently. In particular, the discriminative performance of SIFT and PolyPhen varied across pLDDT strata, supporting the interpretation of pLDDT as a stratifying variable rather than a universal measure of predictor reliability.

Taken together, these findings suggest that AlphaFold-derived local confidence scores may serve as a useful auxiliary feature for BRCA1 missense variant interpretation and for prioritizing VUS candidates for

downstream validation. However, pLDDT should be applied as supportive contextual evidence rather than as an independent clinical criterion of pathogenicity.

Abbreviations

AUC
Area under the curve
B/LB
Benign/Likely benign
BRCA1
Breast cancer type 1 susceptibility gene
BRCT
BRCA1 C-terminal domain
CI
Confidence interval
ClinVar
Clinical Variation database
HGVS
Human Genome Variation Society
OR
Odds ratio
P/LP
Pathogenic/Likely pathogenic
pLDDT
Predicted Local Distance Difference Test
ROC
Receiver operating characteristic
SNV
Single-nucleotide variant
VEP
Variant Effect Predictor
VUS
Variant of uncertain significance

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified, aggregate data from ClinVar together with publicly accessible protein structure and protein annotation resources. It did not involve direct interaction with

human participants, access to identifiable personal data, or the use of human tissue. Therefore, ethics approval and informed consent to participate were not required.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

Author Contribution

Y.Y. conceived the study, developed the analysis pipeline, performed data processing and statistical analyses, prepared the figures and tables, and drafted the manuscript. M.C. contributed a molecular biology perspective, provided consultation on BRCA1 functional context and variant interpretation, and critically reviewed the manuscript. A.S. contributed to study design, interpretation of the results, and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The datasets and scripts supporting the conclusions of this article are available in the Zenodo repository, https://doi.org/10.5281/zenodo.19789707.

References

1. Clark SL, Rodriguez AM, Snyder RR, Hankins GDV, Boehning D. Structure-function of the tumor suppressor BRCA1. Comput Struct Biotechnol J. 2012;1(1):e201204005. 10.5936/csbj.201204005.
2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of

Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24. 10.1038/gim.2015.30.

3. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7. 10.1093/nar/gkx1153.
4. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562(7726):217–22. 10.1038/s41586-018-0461-z.
5. Hart SN, Polley EC, Shimelis H, Yadav S, Couch FJ. Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. NPJ Breast Cancer. 2020;6:13. 10.1038/s41523-020-0159-x.
6. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4. 10.1093/nar/gkg509.
7. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9. 10.1038/nmeth0410-248.
8. Wilson CJ, Choy WY, Karttunen M. AlphaFold2: a role for disordered protein/region prediction? Int J Mol Sci. 2022;23(9):4591. 10.3390/ijms23094591.
9. Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. Protein Sci. 2022;31(11):e4466. 10.1002/pro.4466.
10. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122. 10.1186/s13059-016-0974-4.
11. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439–44. 10.1093/nar/gkab1061.
12. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2025;53(D1):D609–17. 10.1093/nar/gkae1010.
13. Yasinskyi Y. Bioinformatic pipeline and research data for AlphaFold-derived local structural confidence in BRCA1 missense variant interpretation: a ClinVar-based computational analysis [Data set]. Zenodo. 2026. https://doi.org/10.5281/zenodo.19789707.

Figures

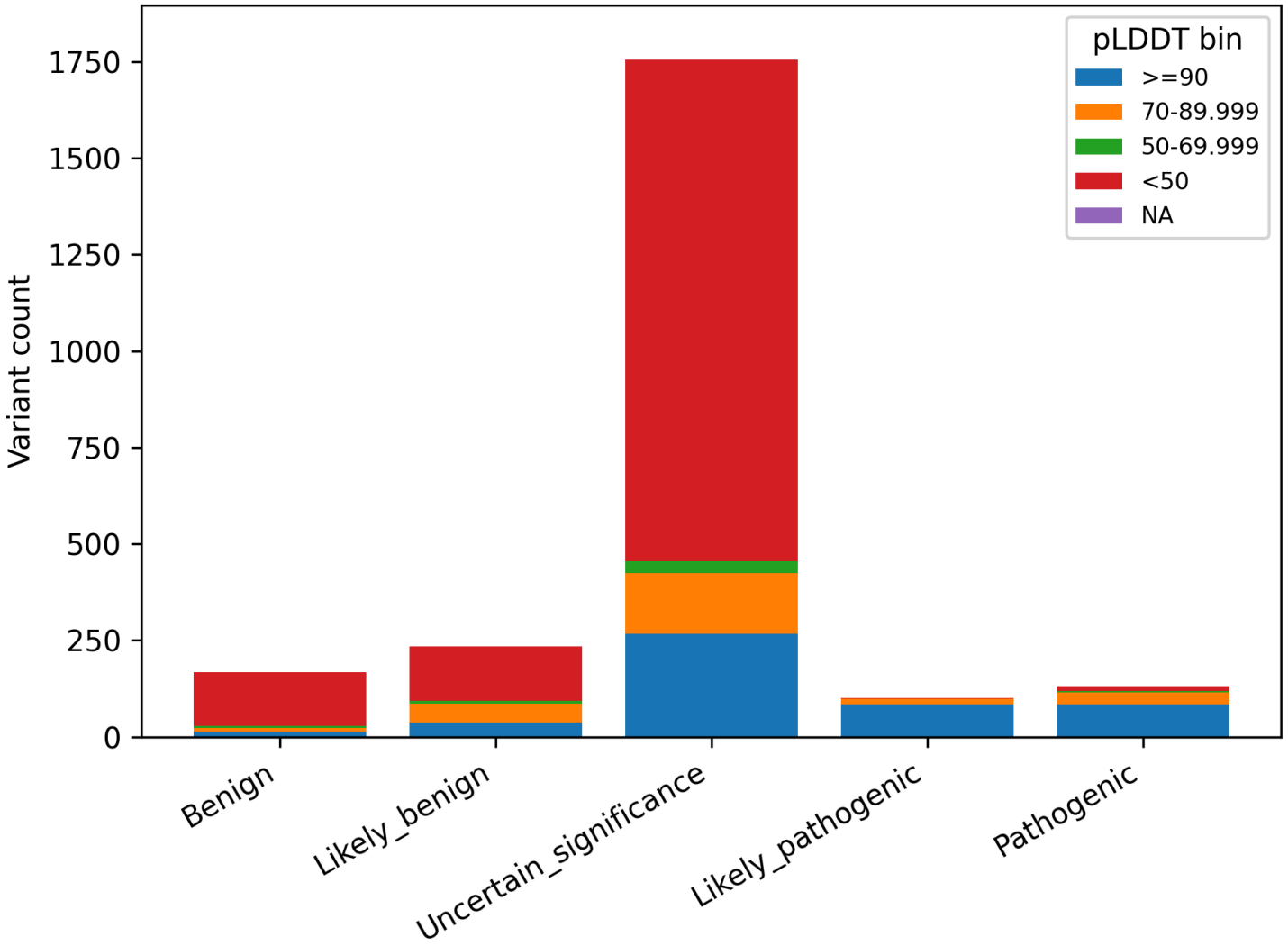

Figure 1

Count distribution of BRCA1 missense variants across clinical classes stratified by residue-level pLDDT bins.

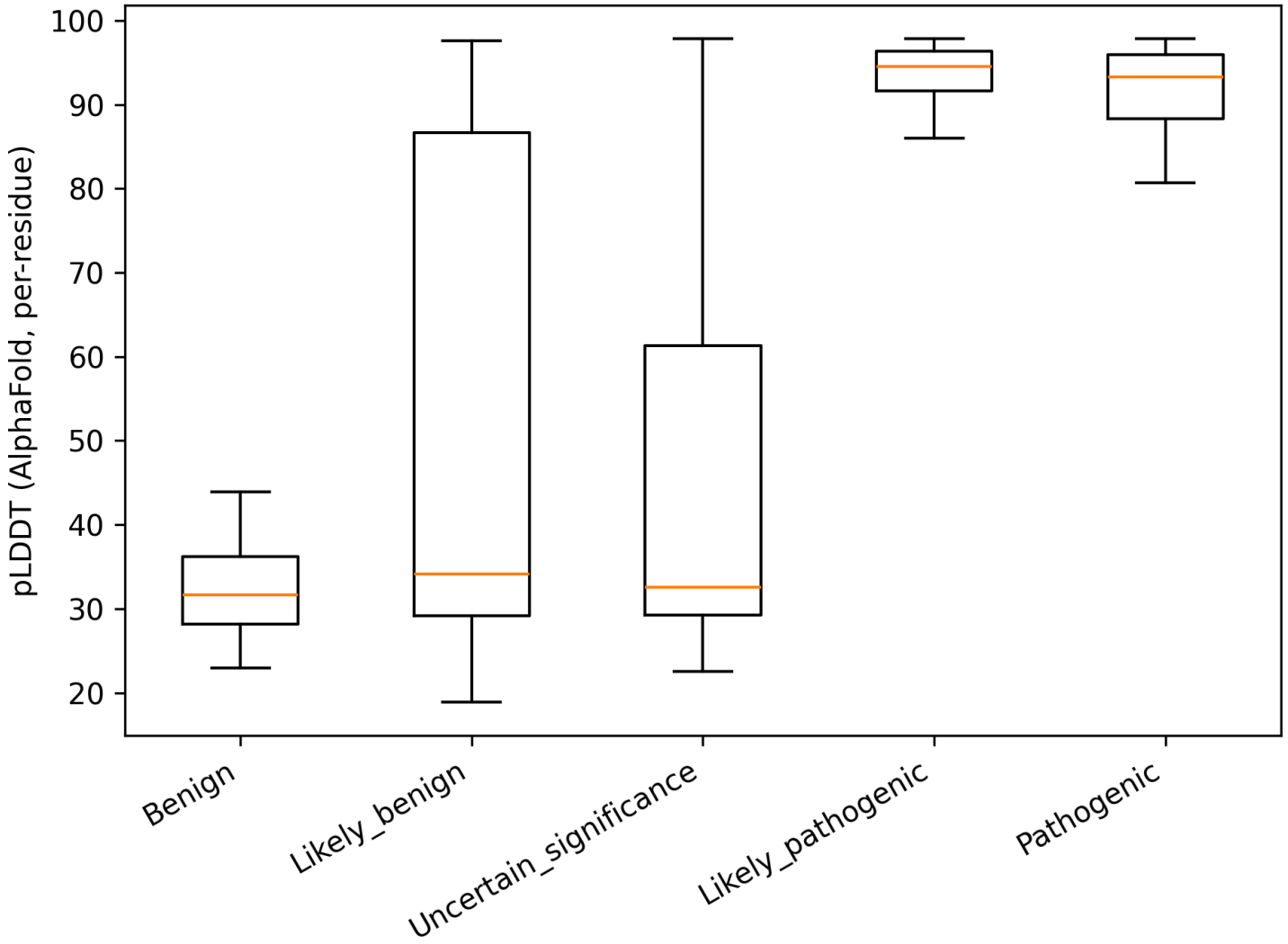

Figure 2

Distribution of BRCA1 residue-level AlphaFold pLDDT values across ClinVar clinical significance classes.