Interpreting the Effects of DNA Polymerase Variants at the Structural Level

Matteo Arnaudi[1,2], Karolina Krzesińska[1,2], Ludovica Beltrame[1,2], Matteo Tiberti[1], Mef Nilbert [3,4], Anna Rohlin[5,6] *, Elena Papaleo[1,2] *

[1] Cancer Structural Biology, Danish Cancer Institute, 2100, Copenhagen, Denmark
[2] Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark,2800, Lyngby, Denmark
[3] Institute of Clinical Medicine, Dept. Oncology and Pathology, Lund University, Sweden
[4] Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre, Copenhagen, Denmark
[5] Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Sweden
[6] Department of Laboratory Medicine, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Corresponding authors:

Elena Papaleo, elenap@cancer.dk, elpap@dtu.dk
Anna Rohlin, anna.rohlin@vgregion.se

Abstract

Genetic variants in the DNA polymerase enzymes POLE and POLD1 can profoundly impact protein function, potentially disrupting stability, catalysis, DNA-binding, and interactions with other molecules. Understanding the structural basis of these functional impacts at the atomic level is essential for comprehensively interpreting variant effects. In this study, we utilize MAVISp, a modular, structure-based framework, to systematically characterize over 60,000 missense variants within the structured domains of POLE and POLD1. By integrating predictions related to protein stability, local misfolding, active-site or DNA and protein binding alterations, we provide a detailed structural interpretation of variants reported across various databases, including ClinVar, COSMIC, and cBioPortal. Moreover, we predict the functional consequences of numerous variants not found in disease-related mutation databases, creating a comprehensive catalogue for future studies. Our approach identifies structural mechanisms underlying established pathogenic mutations and flags over 300 Variants of Uncertain Significance (VUS) as potentially damaging, integrating them with already published experimental data (https://poled-db.org/). Collectively, our findings underscore the effectiveness of integrative structural analyses in elucidating the functional impact of genetic variants. This approach not only advances our understanding of protein variant effects at the atomic scale but also facilitates future studies aimed at variant classification, prioritization of variants for experimental studies, and functional interpretation in diverse biological contexts.

Introduction

DNA polymerases ε (POLE) and δ (POLD1) play critical roles in DNA replication and ensure replication fidelity through their base selectivity and proofreading activities [1]. Genetic variants in these proofreading polymerases can disrupt their structural integrity and function, potentially leading to significant alterations in DNA replication accuracy [2]. Variants affecting POLE and POLD1 have been implicated in conditions characterized by increased genomic instability [3] highlighting the need for detailed structural and functional characterizations.

POLE is involved in the replication of the leading strand of the DNA, while POLD1 synthesizes the Okazaki fragments of the lagging strand of genomic DNA [4,5]. The exonuclease domains (residues 268-472 in POLE, and residues 309-526 in POLD1) are primarily responsible for the correction of rare errors which might occur during the DNA replication process [1]. The proofreading activity ensures high-fidelity replication, with remarkably low error rates in the range of 10^{-7}-10^{-8} errors per base pair [6]. Moreover, the two enzymes interact with several accessory subunits, which ensure their stability and functionality, and operate in concert with other proteins to form the complex replisome machinery.

Our group recently proposed a structure-based framework, MAVISp (Multi-layered Assessment of Variants by Structure for proteins), to characterize and classify protein variants found in ClinVar [7] and other cancer datasets in a high-throughput manner [8]. MAVISp is founded on the principle that mutations in coding regions can have diverse consequences for the protein itself and the cellular function of the resulting protein. These may include changes in structural stability, altered interactions with other proteins or biomolecules (e.g., DNA), disruption of catalytic or cofactor-binding sites, and changes in post-translational regulation.

POLE and POLD1 thus represent compelling candidates for investigation using the MAVISp framework. In this study, we applied MAVISp to systematically characterize all possible missense mutations across the structured regions of POLE and POLD1. By integrating saturation mutagenesis data with structural, functional, and interaction-based analyses, we aim to propose possible molecular mechanisms underlying both known and uncertain variants in POLE and POLD1. The MAVISp framework enables us to provide mechanistic insights that can support the classification of VUS, prioritize variants for experimental validation, and enhance clinical interpretation for both cancer risk surveillance and precision medicine recommendations.

Results and Discussion

Overview of variants characterized in the study and structure selection

To orient the reader, we first provide a summary of the main data included in the study and the protein complexes selected for analysis (**Figure 1**). Using an approach based on *in-silico* saturation mutagenesis, we generated data for 43415 variants in POLE (residues 2-2286, **Figure 1A**) and 21014 variants in POLD1 (residues 2-1107, **Figure 1B**). Among these, 4291 POLE (**Figure 1C**) and 2025 POLD1 (**Figure 1D**) variants were reported in ClinVar. Additionally, 1966 POLE and 1990 POLD1 variants have been annotated in ClinVar as associated with colorectal cancer, polymerase proofreading-associated polyposis (PPAP), or hereditary cancer-predisposing syndrome (see OSF repository). The three-dimensional (3D) models shown in **Figure 1A** and **1B** for POLE and POLD1, respectively, were used as input structures to analyze the effects of variants on structural stability, functional sites, long-range interactions, and phosphorylation.

As previously mentioned, POLE and POLD1 are the catalytic subunits of the DNA replication machinery, responsible for leading and lagging strand elongation, respectively. Their activities are tightly regulated through interactions with a range of protein partners that are essential for proper replication. Beyond replication, POLE and POLD1 also contribute to multiple DNA repair pathways, including base/nucleotide-excision repair, mismatch repair, post-replication repair, and homologous recombination [9] [10] [11]. Given these critical functions, we analyzed the effects of variants on the protein-protein complexes involving POLD1 and four key interactors (PCNA, POLD4, POLD2, and LASP1), as well as POLE and five interactors (PCNA, CDC45, POLE2, MCM2, and MCM5). As detailed in the Methods, these interactors were selected as they were the only ones for which 3D structures of protein-protein complexes that satisfied the quality control step of the MAVISp protocol were available. Furthermore, we also assessed the impact of amino acid substitutions on the interaction with DNA for both POLE and POLD1. To this end, we utilized experimentally determined 3D structures of the POLE and POLD1 polymerase domains in complex with DNA, as well as with five additional POLE structures representing key intermediates during base pair mismatch recognition and excision (see **Materials and Methods**). We applied both the *simple* and *ensemble* modes of MAVISp across all the available modules, except for local interactions analysis in the *ensemble* mode, which was excluded due to computational time constraints. Unless otherwise specified, results discussed in the text refer to the *ensemble* mode, as it captures the dynamic nature of protein structures. Where available, we also provided experimental validation of the variants by querying the PolED database (https://poled-db.org), which aggregates literature-based data on the functional impact of POLE and POLD1 variants on various aspects of the protein, including biochemical properties, cellular activity, and more.

Given the complex roles of *POLE* and *POLD1*, and the fact that most ClinVar-annotated VUS in these genes are linked to hereditary cancer predisposition syndrome, colorectal cancer, and PPAP, we contextualized MAVISp predictions within a cancer-relevant framework. This

disease-oriented perspective helps guiding the interpretation of structural and functional impacts predicted by MAVISp.

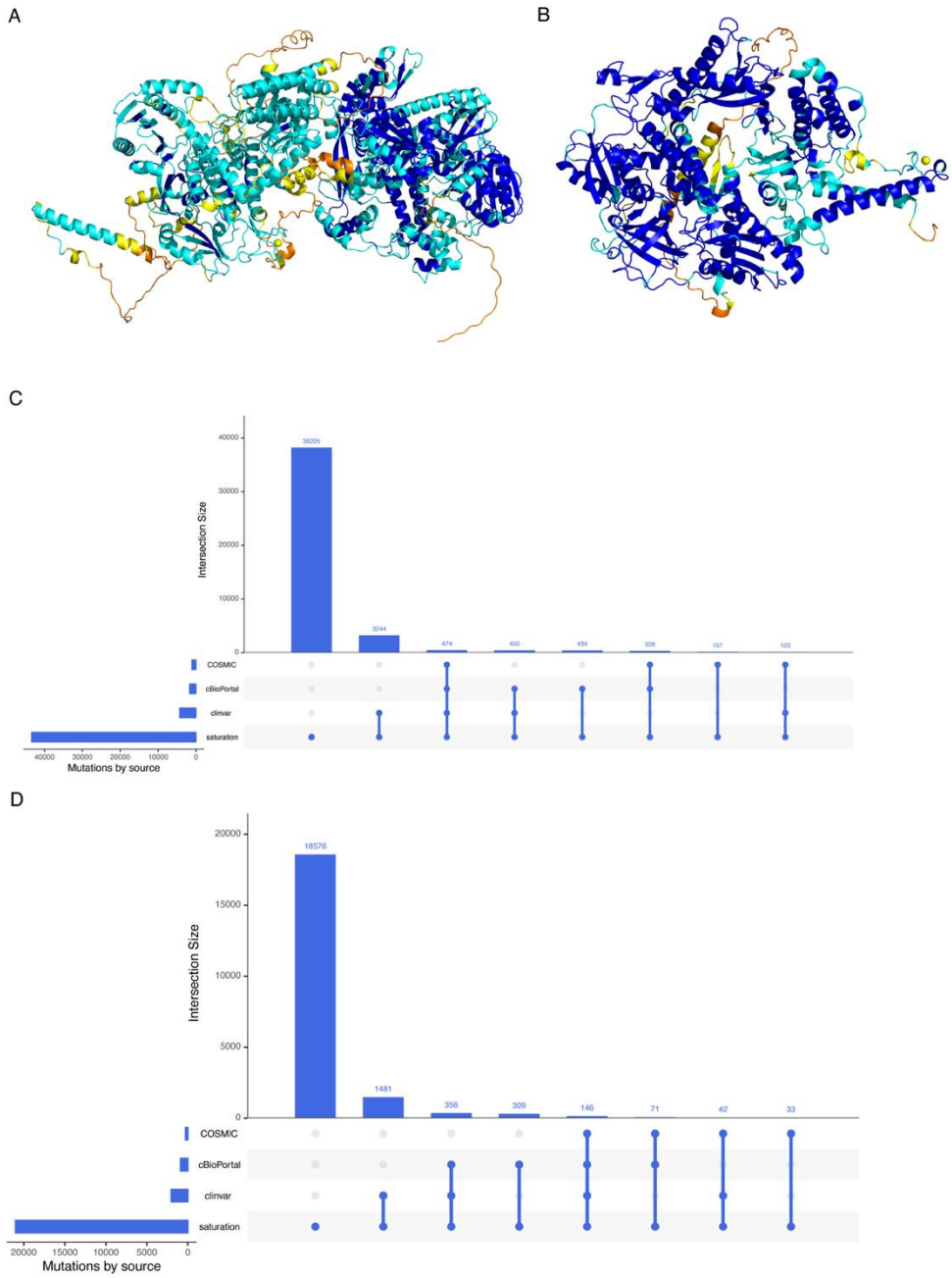


Figure 1 Overview of the initial inputs for the MAVISp framework on POLE and POLD1

***A**) Three-dimensional (3D) structure of POLE (1-2286) as retrieved from the AlphaFold (AF) Database (AFDB), used in the* simple *mode calculation of the MAVISp framework, colored according to the pLDDT score, using the same color scale as the AFDB. Zn^{2+} ions have been modelled through superimposition with the experimental structure. Additional details on the starting structure and structural trimming are reported in the Materials and Methods. **B)** 3D structure of POLD1 (1-1107) structure as retrieved by the AlphaFold (AF) Database (AFDB), used in the* simple *mode calculation of the MAVISp framework, colored according to the pLDDT score, using the same color scale as AFDB. Zn^{2+} ion has been modelled through AlphaFill. Additional details on the starting structure and structural trimming are reported in the Materials and Methods. **C)** UpSet plot showing the number of mutations in POLE analyzed with MAVISp grouped by source type. The plot displays the distribution of variants originating from*

ClinVar, cBioPortal, COSMIC, and saturation mutagenesis data included in MAVISp. ***D)*** *UpSet plot showing the number of mutations in POLD1 analyzed with MAVISp grouped by source type. The plot displays the distribution of variants originating from ClinVar, cBioPortal, COSMIC, and saturation mutagenesis data included in MAVISp.*

Interpretation of MAVISp Results

The MAVISp database and its underlying framework are designed to predict the mechanistic effects of missense mutations on critical protein features. These predictions are complemented by pathogenicity scores from AlphaMissense and EVE [12,13], as well as functional impact assessments from GEMME and DeMaSk [14,15]. However, since the framework is intended for high-throughput application across diverse proteins, pathogenicity and functional predictions should always be interpreted within the specific biological context of the protein under investigation. This is particularly crucial for proteins composed of multiple functional domains, where mutations in distinct domains may result in divergent phenotypic outcomes. Given that both AlphaMissense and EVE make use of ClinVar classification data, either as a validation set (AlphaMissense) or proposing a protocol to integrate their prediction to the ACMG/AMP guidelines (EVE), their predictions are, in principle, aligned with the ACMG/AMP guidelines [16] that underpin ClinVar annotations. However, in practice, these predictions do not always align with phenotype-specific interpretations, thereby necessitating critical evaluation by the user in a disease-relevant context. The same applies to gain- and loss-of-function predictions from GEMME and DeMaSk, which should be interpreted with respect to the functional role of the affected domain in multidomain proteins, or to the overall function of the protein in cases of single-domain architecture. In all cases, functional classifications should be considered in relation to the phenotype in which the variant has been observed.

The DNA polymerases POLE and POLD1 are good examples for which the functional consequences of mutations are highly context-dependent. In the context of cancer, the role of POLE and POLD1 as tumour suppressors is still not completely defined. Accordingly, predictions from AlphaMissense, EVE, DeMaSk, and GEMME concerning variants in these genes should be interpreted with caution when evaluating potential pathogenicity. In hereditary cancer syndromes and hypermutated tumors, mutations in the exonuclease (proofreading) domain, known to impair replication fidelity, are typically classified as loss-of-proofreading-function and are associated with a tumorigenic phenotype. In such cases, loss-of-function predictions from DeMaSk and GEMME are likely to reflect biologically relevant mechanisms and may provide meaningful insight. By contrast, loss-of-function mutations in

the polymerase domain are rarely observed in these cancers, as intact polymerase activity is generally required to sustain malignant proliferation[17,18]. Consequently, loss-of-function predictions in the polymerase domain should be interpreted with caution, as they may be misleading or counterintuitive. Indeed, mutations in this domain may fine-tune fidelity without abolishing enzymatic activity altogether. Therefore, within MAVISp, a loss-of-function prediction for a polymerase domain variant should not be assumed to indicate complete inactivation, especially in cancer contexts where residual activity may be necessary for cell viability. In such cases, a more detailed structural analysis, guided by the mechanistic insights provided by MAVISp, should be undertaken to more accurately evaluate the potential impact of the variant and its relevance to the specific phenotype under investigation. This approach can help contextualize loss- or gain-of-function predictions and determine whether they are consistent with the biological scenario in question. The considerations outlined for POLE and POLD1 can be broadly extended to other proteins in which functional interpretation of variants is highly dependent on phenotypic context. Given that all such predictions are inherently context-dependent, variants that may appear uninformative in oncogenic or clinical settings could still hold substantial relevance in non-clinical contexts. For example, polymerase domain mutations predicted to impair function may prove valuable for basic research into essential catalytic residues, or protein engineering applications aimed at precisely modulating polymerase activity.

A similar consideration applies to mutations predicted to impair protein folding or disrupt critical protein-protein interactions. A clarification regarding how MAVISp predicts destabilizing mutations is essential for accurate interpretation. Mutations are classified as destabilizing when the predicted $\Delta\Delta G$ from FoldX and RaSP [19,20] exceeds 3 kcal/mol, a threshold that has been shown to correlate with reduced protein abundance at the cellular level in several studies [21] [22] [23] [24]. However, this correlation has only been validated in a limited number of proteins and mutation types, and is, therefore, not broadly generalizable across all proteins.

Furthermore, FoldX and RaSP [19,20] rely their predictions on local structure changes but cannot predict effects on the global fold or structure. Mutations predicted as destabilizing might affect specific structural domains causing local misfolding and should not be automatically interpreted as destabilizing in a way that reduces the overall protein abundance. Instead, they should be evaluated in terms of their potential to disrupt functionally critical regions, particularly those involved in specific molecular or cellular phenotypes.

Keeping these considerations in mind is crucial for interpreting such mutations in phenotype-specific contexts. For example, while MAVISp may classify certain cancer-associated variants

in POLE or POLD1 as destabilizing or misfolding-prone, these mutations are unlikely to be positively selected in cancer, where partial preservation of function is typically required for replication competence. Unless there is direct experimental evidence supporting reduced abundance, such variants should not be interpreted as leading to protein degradation. Rather, in the absence of experimental validation, prioritization should be based on the mutation's location within the protein and its potential to impair functional domains relevant to the phenotype under investigation.

Accordingly, MAVISp provides mechanistic predictions alongside, when available, information about the phenotypic context in which each variant has been reported. It remains the user's responsibility to interpret these predictions within the appropriate biological and clinical context, considering domain architecture, mutation position, and phenotype-specific functional relevance.

Analysis of variants reported as pathogenic or benign in ClinVar

We started by focusing on variants in POLE and POLD1 that are classified as pathogenic or benign in ClinVar, to assess how well MAVISp can identify disease-relevant mutations. In MAVISp, the first step in evaluating variants of uncertain significance (VUS) is to prioritize those that may be pathogenic using pathogenicity prediction such as AlphaMissense [12] and EVE [13]. In the context of *POLE* and *POLD1*, which have been linked to colorectal cancer, the key question is whether mutations give rise to an error-prone phenotype by compromising the replication fidelity of the enzymes. This is followed by assessment of potential loss or gain of protein function, particularly relevant in carcinogenesis where pathogenic mutations often result in impaired proofreading activity or increased DNA synthesis. For this, MAVISp integrates functional predictors such as DeMaSk [15] and GEMME [14]. To test the accuracy of this approach in the POLE and POLD1 context, we compared the predictions from DeMaSk and GEMME within MAVISp with 61 ClinVar-classified variants, 39 in POLE and 22 in POLD1, that have available clinical annotations (benign, likely benign, pathogenic, or likely pathogenic; see **Table S1).** Most of these variants are associated with hereditary cancer predisposition syndromes, including colorectal cancer susceptibility and polymerase proofreading-associated polyposis. Additional cancer types, such as familial ovarian cancer, endometrial carcinoma, and breast cancer, have also been linked to some of these variants. Notably, a subset of variants has been associated with non-cancer phenotypes, including facial dysmorphism-immunodeficiency-livedo-short stature syndrome and intellectual disability. Since the number of well-characterized variants is limited for each gene, we combined data from both proteins for this analysis. As shown in **Table 1 and Table S1**, both AlphaMissense

and EVE [12,13] performed well in predicting pathogenicity, particularly concerning the cancer-related and syndromic phenotypes described above. AlphaMissense was more sensitive, capturing more pathogenic variants, although with a slightly lower specificity. Overall, AlphaMissense was available for a broader range of relevant mutations than EVE in both POLE and POLD1, highlighting its utility in variant prioritization (**Table S1**).

Next, we took a closer look at the cases where ClinVar classifications did not match the pathogenicity predictions provided by the MAVISp tools. To help interpret and identify these discrepancies, MAVISp provides a visual summary of all prediction results using dot plots. These plots offer an overview of the predictions and annotations from the various MAVISp modules, making it easier to explore and compare variant effects. Dot plots for the ClinVar classified pathogenic or benign variants are shown in **Figures 2** and **3**, for POLE and POLD1, respectively. Results are shown using both the MAVISp *simple* (**Figures 2A and 3A**) and the *ensemble* modes (**Figures 2B and 3B**).

One example where both EVE and AlphaMissense disagree with the ClinVar classification is the POLE variant E1956V (p.Glu1956Val), associated with facial dysmorphism-immunodeficiency-livedo-short stature syndrome. Both AlphaMissense and EVE predict this variant to be benign, while ClinVar lists it as likely pathogenic (**Figure 2A**). Furthermore, according to DeMaSK and GEMME, there is no evidence that this variant causes loss or gain of function, nor are there any mechanistic indicators suggesting a damaging effect in the context of replication fidelity. We also verified that this variant was not predicted to affect splicing (**Tables S2 and S3**). Importantly, the ClinVar entry for E1956V has a low review status (1 star) and was classified as likely pathogenic by the submitting group based on the American College of Medical Genetics and Genomics (ACMG) guidelines [25]. However, the entry does not provide supporting details such as allele frequency, family history, or even functional data. Without experimental validation or additional context, the classification relies solely on the submitter's interpretation. Two other cases where MAVISp predictions differ from ClinVar classifications are the POLE variant V240L (p.Val240Leu) associated with hereditary cancer and syndromic features and the POLD1 variant T441K (p.Thr441Lys) associated with hereditary cancer, shown in **Figures 2A** and **3A**. These variants are located at the N-terminus and the exonuclease domains, respectively. AlphaMissense predicts both to be pathogenic, while EVE classifies V240L as uncertain and T441K as benign. Notably, neither variant is predicted to affect splicing (**Tables S2 and S3**). In contrast to AlphaMissense, ClinVar classifies both variants as benign or likely benign. V240L has a review status of 2, supported by three independent submissions. These classify the variant as germline and base their assessment on guidelines from ACMG, AMP, or GeneDx ([25], https://www.ncbi.nlm.nih.gov/clinvar/submitters/26957/). Supporting this, MAVISp does not predict any damaging effects for V240L, and both DeMaSk and GEMME also indicate a neutral

impact (**Figure 2A**). T441K, on the other hand, is classified as benign in ClinVar based on a single submission (review status 1) using the Ambry Genetics classification scheme. While AlphaMissense classifies it as pathogenic and GEMME suggests a potential loss-of-function impact for this variant, DeMaSk predicts it having a neutral effect. This is consistent with MAVISp, which finds no mechanistic evidence of damage (**Figure 3A**). Additionally, according to the evidence reported in ClinVar, the population frequency of both variants in healthy individuals does not support a pathogenic role, further reinforcing their likely benign status. Overall, these results highlight the value of a comprehensive approach like the one offered by MAVISp, where predictions of pathogenicity are considered alongside functional and structural evidence.

The pathogenic and likely pathogenic missense variants reported in ClinVar for POLE that occur within its exonuclease domain are predicted by DeMaSK and GEMME to result in loss of function. This observation supports the yet unconfirmed hypothesis that the exonuclease domain of POLE may function as a tumor suppressor, in line with evidence from the COSMIC database [26]. To further explore the potential molecular mechanisms underlying these pathogenic variants, we used MAVISp to analyze a subset of well-characterized *POLE* mutations, specifically P286H (p.Pro286His), S297F (p.Ser297Phe), L424V (p.Leu424Val), S459F (p.Ser459Phe), and N363K (p.Asn363Lys) (**Figures 2** and **4**). To support interpretation, MAVISp succinctly visualizes the predicted broad mechanistic effects in a *lollipop plot* (**Figure 4A**).

Using the MAVISp *simple* mode, we predicted that the POLE variant S297F (p.Ser297Phe) affects structural stability (**Figures 2A** and **4A**). However, in the *ensemble* mode, the variant has been classified as Uncertain and shows conformation-dependent changes in folding free energy, indicating a context-specific destabilizing effect (**Figure 4B**). Predictions of destabilizing effects are unlikely to indicate complete POLE misfolding, particularly given the inherent limitations of FoldX and Rosetta, which model only the local environment surrounding the mutation site. Furthermore, the fact that MAVISp predicts that S297F has only a detrimental effect on stability in the exonuclease domain, together with the fact that the variants are associated with autosomal dominant POLE-related disorders, in which the cancer-prone phenotype is believed to require a functionally active POLE with reduced replication fidelity, supports a model of localized destabilization rather than global misfolding. Such partial destabilization may preserve the overall domain architecture while nonetheless impairing proofreading activity. Supporting this hypothesis, S297F is a recurrent cancer driver mutation and has been associated with hypermutated tumor profiles [27][28]. Its damaging impact is also supported experimentally by a yeast-based functional assay [29], linking it to significantly elevated mutation rates, consistent with a loss-of-proofreading functional phenotype (**Figure 2A,** Experimental Data Classification row). To further explore the structural impact of S297F,

we analyzed the conformations used for the calculations performed in the *ensemble* mode. Visual inspection of these models suggests that the bulky phenylalanine substitution at position 297 may cause a conformational shift in the nearby catalytic residues D275 and E277, potentially altering active site geometry (see OSF repository, **Figure 4C**). To assess the local quality of the structural model around position S297, we performed a control "self-mutation" analysis (see **Materials and Methods**), in which the serine residue was replaced with itself (**Figure 4B**). This type of control is expected to result in minimal changes in folding free energy, typically close to 0 kcal/mol [30]. In our case, the predicted ΔΔG values for the self-mutation at position 297 were never below ~0.4 kcal/mol, suggesting that the model used in this calculation is at an energy minimum. However, it's important to note that the method used for the ΔΔG estimate does not account for larger structural rearrangements in the protein backbone, which may be relevant for this variant. As such, further molecular dynamics simulations would be needed to confirm whether the S297F substitution induces a conformational change in the active site.

We then investigated, with MAVISp, the effects of the variants on POLE-DNA interactions using five distinct 3D structures representative of the structures adopted during the mismatch excision process [31]. The results suggest that L424V (p.Leu424Val) weakens DNA binding in both the frayed and mismatch excision states. In contrast, P286H (p.Pro286His), S459F (p.Ser459Phe) and N363K (p.Asn363Lys) appear to affect DNA interaction only in the mismatch excision state (**Figure 2A, Figure 4A**). In the frayed state, the terminal bases of the newly synthesized DNA strand are loosely held within the exonuclease domain and deviate from the template strand. This separation is stabilized by α-helix 3, which physically blocks the direct path between the polymerase and exonuclease active sites. In contrast, during the mismatch excision state, the mismatched base is correctly positioned within the exonuclease site for removal [31]. Given the proximity of residues L424, N363, and S459 to the path traversed by the 3′ terminal DNA bases en route to the exonuclease catalytic site [31], mutations at these positions could compromise proper strand positioning within the exonuclease domain, potentially impairing proofreading efficiency. Additionally, we identified residues L424 and S459 as being in contact with the catalytic residues D462 and D277 (Figure 4C, left panel). Notably, this interaction is observed only in the free state of the protein (**Figure 4C**, left panel), and is lost upon DNA binding (**Figure 4C**, right panel). Thus, although MAVISp predicts a change in the interactions with catalytic residues upon mutations (**Figure 2A**), this occurs exclusively in the DNA-free structure. It is therefore unlikely that the functional impact of the mutation arises from disrupted coordination. Rather, the molecular mechanism underlying the pathogenic effect is more plausibly explained by the MAVISp-predicted impairment of DNA binding (**Figures 2A and 4A**). Experimental studies support these predictions, showing that L424V is associated with increased mutation rates in both humans

and yeast [32] [33]. Biochemical assays further demonstrate that the variant has reduced exonuclease activity but increased polymerase activity, an imbalance that is not typically seen with standard exonuclease defects [33]. While L424V (p.Leu424Val) retains 89% of wild-type proofreading capacity, it also shows an increased ability to bypass DNA hairpins, compared to both the wild-type protein and exonuclease-deficient mutants [33]. This likely reflects a combination of reduced proofreading activity and elevated polymerase activity [33].

The S459F (p.Ser459Phe) variant is known to cause a complete loss of proofreading activity along with a modest increase in polymerase activity [33]. In addition, it has also been proposed to interfere with the formation of functional exonuclease complexes. These disruptions contribute to an increased ability to extend mismatched DNA strands, ultimately leading to higher mutation rates in vivo [34]. S459F has also been linked to strong tumor susceptibility in mouse models [34] and is associated with elevated mutation rates in yeast and mismatch repair-deficient human cell lines [35][36]. These findings are consistent with MAVISp predictions indicating compromised DNA binding upon mutation (**Figures 2A** and **4A**). Indeed, substitution of this residue located near the catalytic site with an amino acid of differient size and chemical properties could perturb the local structural environment, potentially altering the catalytic geometry and compromising DNA binding.

Although residue N363 is not part of the core set of catalytic residues in the POLE active site, we focused on N363K (p.Asn363Lys) as the yeast model of this variant was found to increase the rate of mutation at *CAN1*, *his7-2*, and *lys2-InsE*$_{A14}$ reporters. 32-, 95-, and 12-fold, respectively (**Figure 2A**, Experimental Data Classification row) [32,37]. Among all experimentally assessed *POLE* variants [29] [36] [38] [39] [40], the mutator effect of N363K (p.Asn363Lys) is second only to that of the most common and strongly pathogenic variant P286R (p.Pro286Arg) [18]. An engineered N363K (p.Asn363Lys) mutation also increases the mutation rate in human cells [32]. The strong mutator effect is consistent with the severe cancer predisposition in N363K (p.Asn363Lys) carriers [41][42].[43][37][44][45]. Yeast protein-based assays demonstrated that N363K (p.Asn363Lys) reduces exonuclease activity modestly on single-stranded DNA and severely on double-stranded DNA, suggesting a defect in DNA partitioning to the exonuclease active site [37]. Molecular dynamics simulations suggested that the impaired functionality of the N363K (p.Asn363Lys) variant is due to steric hindrance caused by the lysine substitution, which interferes with proper DNA insertion into the exonuclease domain [37]. Consistent with these findings, MAVISp predicts that the N363K variant disrupts DNA binding specifically in the mismatch excision state (**Figure 2A**).
Of note, the three variants discussed above are not predicted to alter splicing sites (**Tables S2 and S3**).

Lastly, the well-studied P286R (p.Pro286Arg) variant has been shown to reduce exonuclease activity, elevate mutation rates, and increase tumor susceptibility in experimental models [46][18][38][47][48][49] (see **Text S1**, Experimental Data Classification row). These experimental findings are consistent with MAVISp predictions, for which the variant has an effect on the interaction with DNA in the mismatch excision state (**Text S1**) and further support its utility in identifying functionally relevant variants. While the MAVISp framework provided mechanistic insights for most of the pathogenic *POLE* variants located within the exonuclease domain except for P436S, no clear mechanism could be assigned to the ClinVar-pathogenic variants E830G (p.Glu830Gly) and likely pathogenic E1956V (p.Glu1956Val), for which clinical information is also limited and that have a low associated review status in ClinVar. Similarly, none of the pathogenic *POLD1* variants annotated in ClinVar were annotated with interpretable mechanistic predictions in MAVISp (see Figures 3A and 3B). These observations suggest that further investigation is needed, particularly into alterations affecting DNA and protein interactions, to elucidate the functional consequences of *POLD1* variants as well as *POLE* variants such as P436S (p.Pro436Ser), E830G (p.Glu830Gly), and E1956V (p.Glu1956Val) in their clinical context.

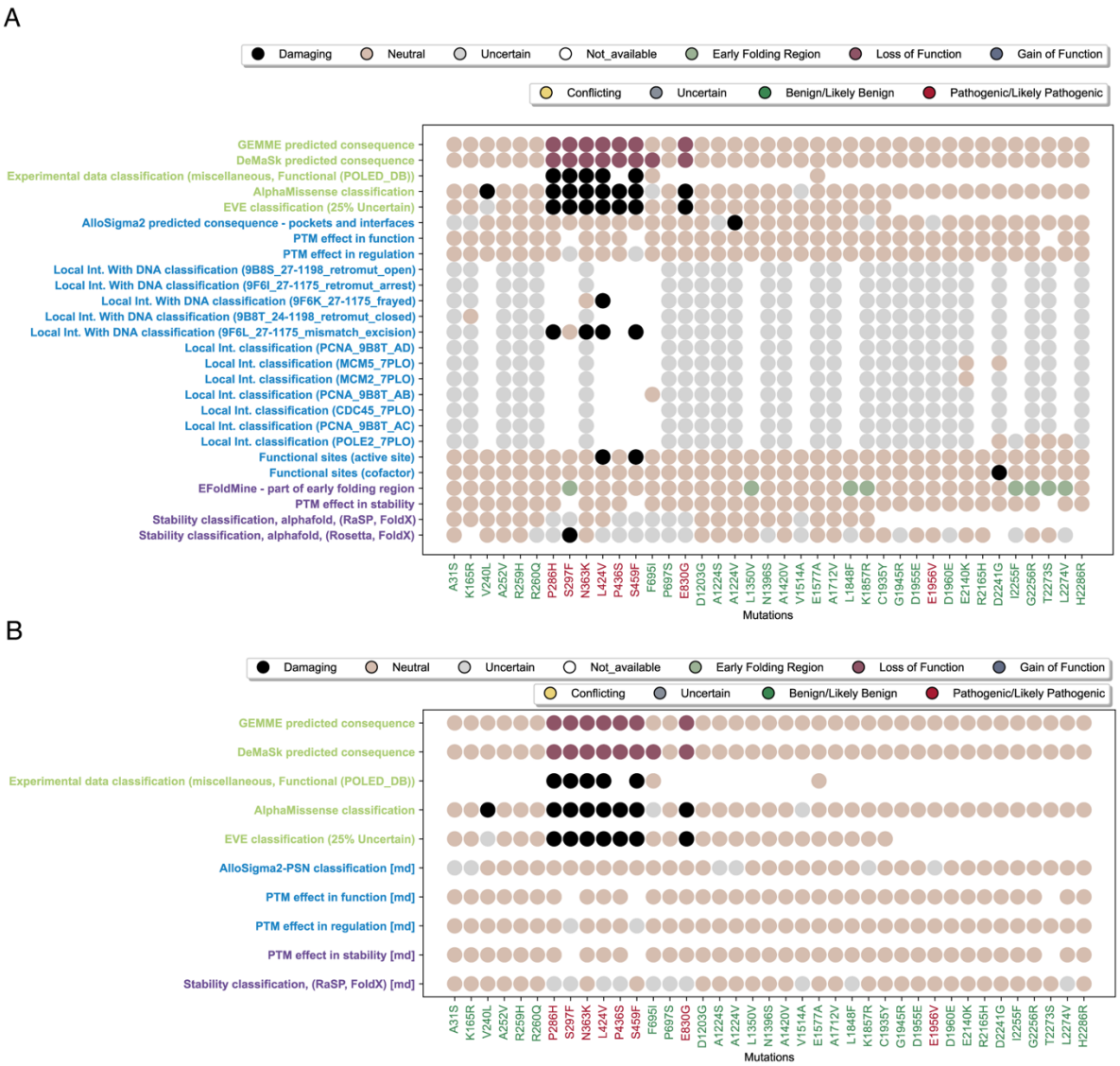


Figure 2 ***MAVISp-predicted effects of ClinVar-reported variants in POLE.*** *Each column in the dot plot represents a specific variant, with dots on the plot corresponding to the predicted effects or annotation by MAVISp and integrated tools for that variant. The MAVISp features related to effects on structural stability are reported in purple, whereas the MAVISp features related to various functional effects are reported in blue. The rows in light green report the results and annotations of external tools and available experimental data.* ***A)*** *Dot plot showing the predicted effects of POLE ClinVar-variants from calculations in the* simple *mode of MAVISp.* ***B)*** *Dot plot showing the predicted effects of POLE ClinVar-variants from calculations in the* ensemble *mode of MAVISp.*

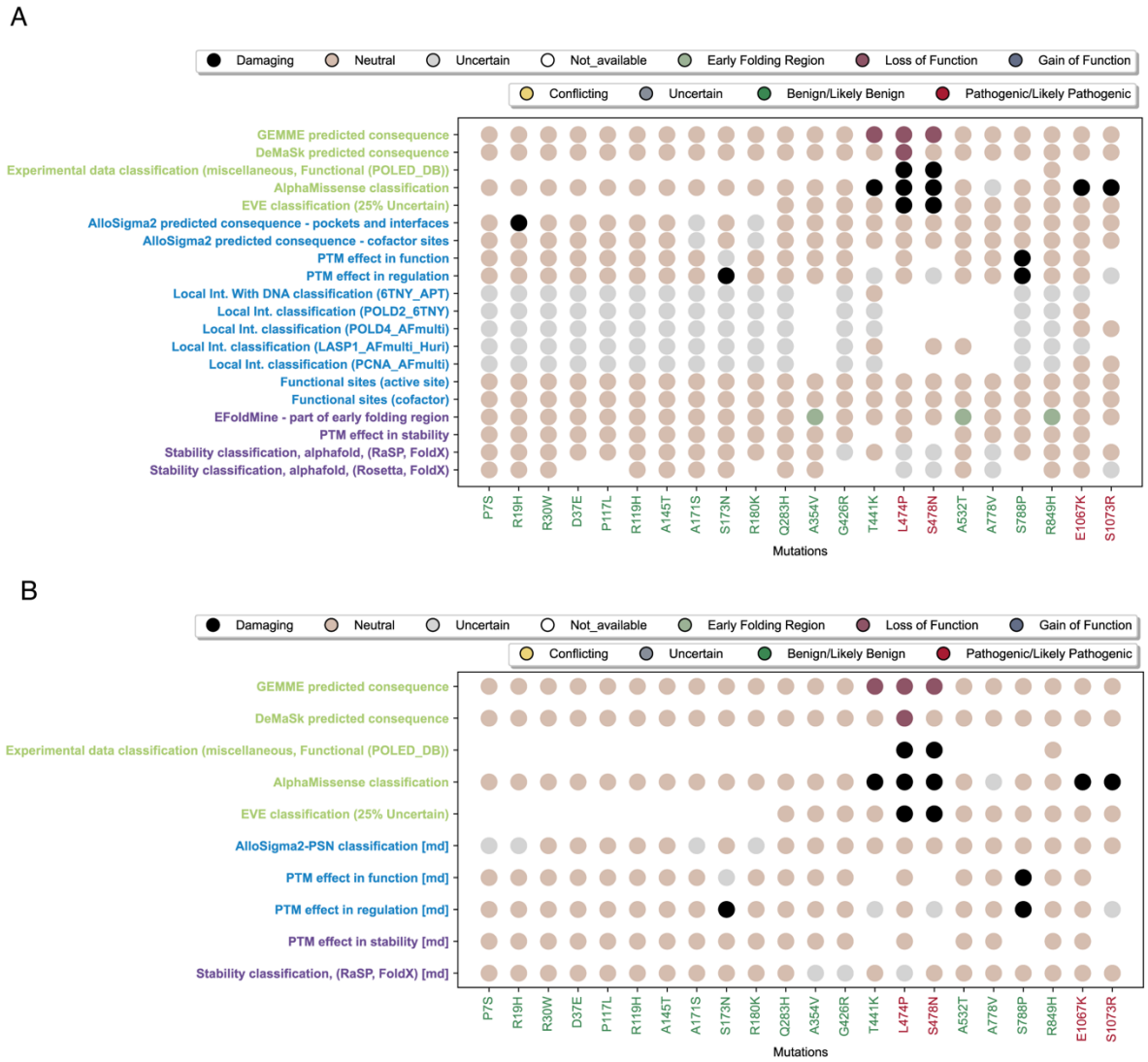
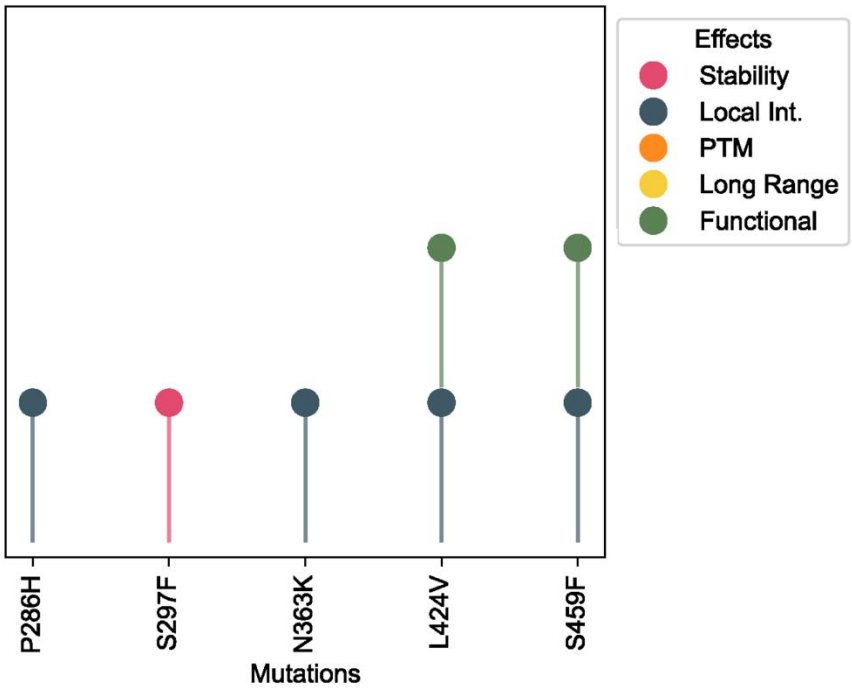
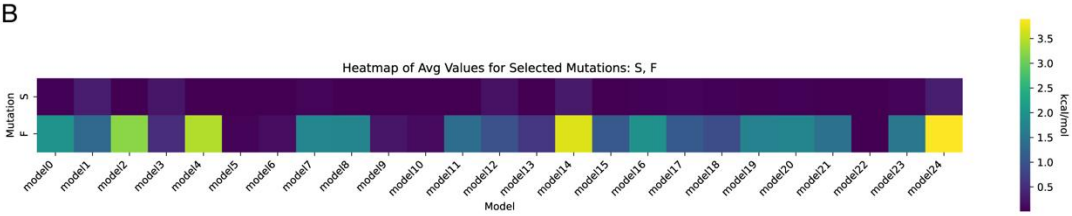


Figure 3 ***MAVISp-predicted effects of ClinVar-reported variants in POLD1.*** *For an explanation on how to read the dot plot see the description in the caption of Figure 2.* ***A)*** *Dot plot showing the predicted effects of POLD1 ClinVar-variants from calculations in the* simple *mode of MAVISp.* ***B)*** *Dot plot showing the predicted effects of POLD1 ClinVar-variants from calculations in the* ensemble *mode of MAVISp.*

A



B



C

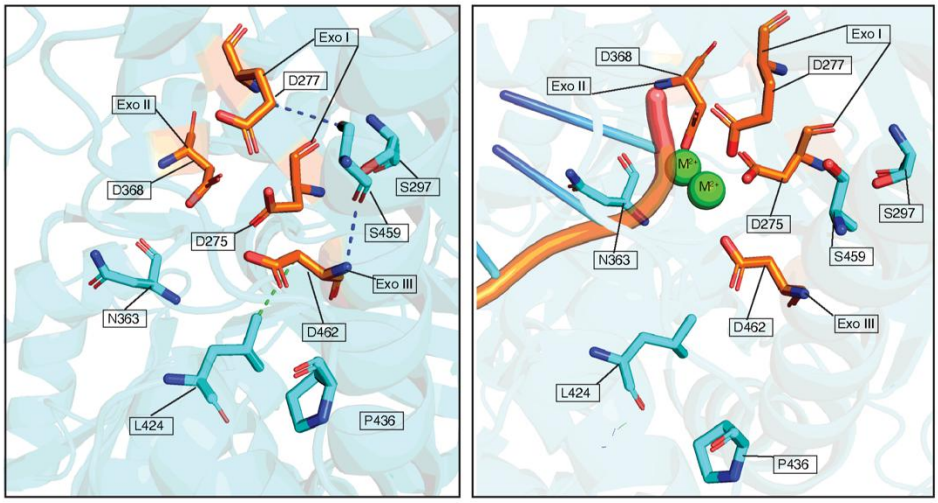


Figure 4 ***Mechanistic indicators for POLE known pathogenic variants. A)*** *The lolliplot represents the identified broad classes of mechanistic indicators identified by MAVISp to ClinVar-pathogenic variants of POLE.* ***B)*** *Heatmap displaying the ΔΔG values for the POLE S297F mutation as well as the self-mutation of residue in position S297, computed with FoldX across 25 models extracted from the simulation of the POLE.* ***C)*** *Structural representation of the POLE exonuclease domain catalytic site in the absence (left) and presence (right) of DNA in its mismatch excision state (PDB ID 9F6L). Catalytic residues are highlighted in orange, second-sphere residues in light blue, with key interactions shown as hydrogen bonds (blue) and van der Waals interactions (green).*

Characterization of variants with uncertain significance or conflicting evidence in ClinVar

We next analyzed all POLE and POLD1 variants reported in ClinVar as either VUS or with conflicting classifications, a total of 4251 for POLE and 2003 variants for POLD1. We aimed to evaluate whether MAVISp could aid in identifying variants predicted to be pathogenic by AlphaMissense and EVE, and classified as loss-of-function by DeMaSK and GEMME, and to link these predictions to specific structural or functional effects. Particular emphasis was placed on variants associated with cancer phenotypes, with the goal of highlighting mechanisms such as loss of proofreading activity or, more broadly, reduced replication fidelity (**Table S4**; dot plots available in **Texts S1–S4**). We identified 190 POLE variants that are predicted to be pathogenic according to AlphaMissense and associated with loss or gain of function by DeMaSK or GEMME, for which the assigned mechanistic indicator is an alteration of the protein structural stability in the *ensemble* mode (**Text S5**). Structural stability, which is closely linked to both the abundance and function of a protein, represents a critical feature often impacted by missense mutations. In the context of DNA polymerases, however, the relationship between pathogenicity and loss of protein stability is nuanced and need to be interpreted in light of the specific functional context of the affected domain. For example, destabilizing mutations within the exonuclease domain are more straightforwardly linked to impaired proofreading activity. Conversely, mutations within the polymerase domain are oncogenic primarily if they preserve robust polymerase activity; local misfolding in this region may promote cancer only if it disrupts the mismatch recognition system. In this regard, further experimental validation is essential to elucidate the relationship between variants predicted to be destabilizing, their impact on polymerase activity, and their contribution to disease phenotypes, particularly when computational predictions alone do not fully explain clinical outcomes. One such example is the F931S (p.Phe931Ser) variant (**Figure 5A**), which has a review status of 1 in ClinVar and is associated with a hereditary cancer-predisposing syndrome. Notably, MAVISp predicts F931S (p.Phe931Ser) to cause a large destabilizing effect, with a folding free energy change greater than 5 kcal/mol. Similar levels of destabilization in DNA repair proteins have been linked to damaging effects at the cellular level [21] [22] [23] [24]. However, this association has not been validated for POLE and POLD1, which are substantially larger than previously studied proteins. According to our analysis of the AlphaFold (AF) structure of POLE, F931 forms contacts with D626, a critical residue for the coordination of Mg2+ which is essential for the correct orientation of the incoming nucleotide (Figure XXX). Notably, this contact is present only in the absence of DNA or when

the exonuclease domain is bound to DNA in a closed conformation but is lost when a mismatched template resides within the active site. Consequently, disruption of this contact by a mutation introducing an amino acid with different size and physicochemical properties is unlikely to significantly impair replication fidelity. Instead, given the mutation's location within the palm domain responsible for positioning the incoming deoxynucleotide triphosphate (dNTP), it is more plausible that it induces a local conformational change, potentially altering dNTP specificity and compromising the enzyme's ability to recognize mismatches. Further functional and in vitro assays are required to validate the MAVISp predictions and to establish a mechanistic link between the structural alteration and the pathogenicity of the variant.

The A456P (p.Ala456Pro) variant is also classified as pathogenic by AlphaMissense and EVE and loss or gain of function by either DeMaSK and GEMME (Figure **5A, Text S1**). Our analysis suggests a potential impact on protein stability (**Figure 5A, Text S1**). The residue is located in the exonuclease domain, where potential local misfolding could impair its functionality, affecting proofreading activity. While in vitro assays are required to validate this hypothesis, the pathogenicity of the variant, predicted by AlphaMissense, is supported by experimental evidence from yeast, where A456P (p.Ala456Pro) was tested using three independent reporter systems (CAN1 forward mutation, *his7-2* reversion, and *lys2-InsE*$_{A14}$ reversion). The variant was shown to cause an increased mutation rate [39] (**Figure 5A**, Experimental Data Classification row), consistently with a damaging effect on replication fidelity.

A characteristic phenotype associated with mutations in *POLE* and *POLD1* is polymerase proofreading-related adenomatous polyposis, which is linked to colorectal cancer and hypermutated tumors [50]. Therefore, we focused our investigation more specifically on these variants. Among the variants predicted to affect POLE protein stability, 21 were initially identified using the *simple* mode in MAVISp. This list was reduced to 11 variants in the *ensemble* mode, all of which are associated with colorectal cancer susceptibility according to ClinVar. Of these, Y85C (p.Tyr85Cys), W243G (p.Trp243Gly), Y298C (p.Tyr298Cys), Q352P (p.Gln352Pro), Y623C (p.Tyr623Cys), Y726D (p.Tyr726Asp), Y1400C (p.Tyr1400Cys), I1585N (p.Ile1585Asn) and G1927R (p.Gly1927Arg) were consistently annotated as destabilizing by both modes, while two were flagged as damaging only in the *ensemble* mode (Figure 5A, Texts S1, S2, S5, and S6). However, based on the phenotypes associated with these mutations, it is unlikely that polymerase and proofreading activities are fully abrogated due to protein misfolding. These variants should be interpreted in the context of their localization in a domain, particularly if they reside within the exonuclease domain, where partial misfolding may compromise proofreading efficiency, or within critical regions of the polymerase domain involved in mismatch recognition. Nonetheless, without *in vitro* validation

linking predicted destabilization to a measurable loss of replication fidelity, such variants are of limited relevance in the context of tumorigenesis.

An important feature assessed by the MAVISp framework and closely linked to protein functionality is whether a given residue is predicted to serve as a nucleation point for the initial folding of the domain, therefore being involved in an early folding (EF) event. Mutations at such sites may disrupt the folding trajectory, potentially leading to domains with altered conformations and compromised functionality. A representative case is the Y85C (p.Tyr85Cys) variant, which is not only predicted by MAVISp to have destabilizing effects, but also involves a residue, Y85, predicted as a key early folding initiator (Figure 5A, EFoldMine row). This suggests that Y85 plays a critical role in guiding proper domain folding and maintaining the structural stability of the N-terminal and exonuclease domains of POLE.

Among the variants predicted to affect the functional properties of POLE, such as protein-protein interactions, DNA binding, or catalytic activity, we identified 255 variants of uncertain significance (VUS) that are predicted to be pathogenic by AlphaMissense and classified as loss or gain of function by either DeMaSK or GEMME (**Texts S1 and S5**). Of these, 17 have been reported in ClinVar as being associated with colorectal cancer susceptibility (**Figure 5B**). Within the MAVISp framework, most of these variants are consistently predicted to disrupt DNA binding, based on structural analyses spanning five distinct conformational states of the enzyme (see below for details). A representative subset of the most compelling variants prioritized for experimental validation is presented in Table 2, along with the predicted molecular mechanism of disruption and the corresponding protein domain affected.

One notable exception is D462N (p.Asp462Asn), where the mutation directly affects the catalytic site by substituting a residue with markedly different physicochemical properties, making it a strong candidate for further functional validation. Another exception is T466A (p.Thr466Ala), which appears to exert its effect primarily by perturbing the coordination of the catalytic residue D462 (**Figure 5B**). Structural analysis reveals that T466 establishes a van der Waals interaction with the backbone oxygen of D462. This interaction is conserved across all mismatch excision conformational states, suggesting a potential role in stabilizing the spatial arrangement of the catalytic residue.

Some variants display combined effects. For example, V460M (p.Val460Met) is predicted to impact both structural stability and active site conformation (**Figure 5A**). Similarly, G977R (p.Gly779Arg), Y726D (p.Tyr726Asp), and D1129G (p.Asp1129Gly) likely disrupt DNA interaction through local destabilization of the polymerase domain. potentially impairing the efficient switching of the DNA strand template between the polymerase and exonuclease sites (**Figure 5A**). Another variant, M295V (p.Met295Val), affects both DNA binding in one of the exonuclease conformations and the structural organization of the catalytic site (**Figure 5A,** [46][18][38][47][48][49]).

In addition to the variants associated with colorectal cancer, we also identified potential mechanistic effects for two POLE variants, L2216P (p.Leu2216Pro) and W2276C (p.Trp2276Cys), for which no clinical condition has been reported**.** The two variants are classified as VUS under the Sherloc classification system [16]. Both variants are predicted to weaken the interaction with POLE2, a key subunit of the POLE ε holoenzyme essential for processive DNA synthesis (see **Texts S1** and **S6**) [51]. In addition, L2216P is also predicted to destabilize the protein 3D structure (**Texts S1, S5,** and **S6**). Another variant, I1183N (p.Ile1183Asn) associated to hereditary cancer predisposing syndrome, which is predicted pathogenic by Alphamissense and EVE and with loss or gain of function by DeMaSK and GEMME, appears to disrupt the binding with PCNA, a factor critical for polymerase processivity (**Text S1 and S6**) [40][41][42][4][43][44]. Given the impact on accessory POLE subunits essential for enzymatic processivity, these mechanisms are unlikely to be cancer-promoting. We also examined the VUS POLE variant L1831P (p.Leu1831Pro), which is of particular interest as it was identified in a patient with suspected Lynch syndrome. The patient was diagnosed at 38 years with colorectal cancer, but without any family history of cancer. In addition, the patient also had a VUS in MSH6 [52]. This variant is predicted as uncertain for stability in the *ensemble* mode (**Text S2**). Moreover, it was predicted with neutral and uncertain consequences across all other MAVISp modules applied, suggesting limited mechanistic support for a damaging effect (Text S1).

A

B

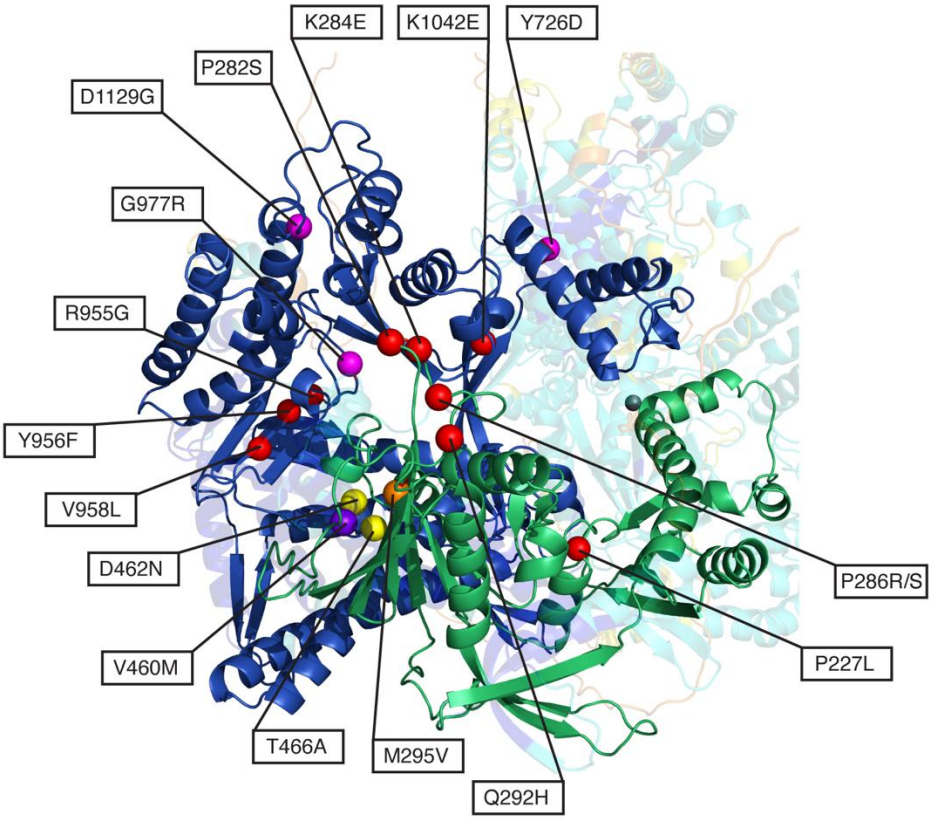


Figure 5 ***Subset of VUS in POLE predicted as pathogenic and loss-of-function and supporting MAVISp mechanistic effects.*** *For an explanation on how to read the dotplot see the description in the caption of Figure 2.* ***A)*** *Dot plot illustrating 23 VUS in POLE associated with colorectal cancer and other variants discussed in the main text, which have been classified as pathogenic by AlphaMissense and loss or gain of function based on at least*

one loss/gain-of-function predictor (i.e., GEMME or DeMaSK) and predicted to be affected by changes in structural stability. ***B)*** *Structural representation highlighting the 17 POLE VUS mutations impacting protein functional properties according to MAVISp, associated with susceptibility to colorectal cancer. These mutations are classified as potentially pathogenic by AlphaMissense and as loss or gain of function by at least one loss/gain-of-function predictor (i.e., GEMME or DeMaSK). The POLE structure is color-coded according to the structural domains as indicated by PFAM annotation, i.e., the exonuclease domain in green and the polymerase domain in blue. The mutation sites of interest are represented as spheres centered around the Cα atoms, colored based on the impact of the corresponding amino acid substitutions. Red and yellow indicate substitutions affecting interactions and functional sites, respectively. On the other hand, orange represents substitutions impacting both interactions and functional sites, whereas fuchsia is related to substitutions influencing both interaction and stability, and purple for substitutions altering both stability and functional sites.*

For POLD1, the predictive power of MAVISp is more limited, primarily due to the absence of suitable structural models, in contrast to the availability of detailed structural data for POLE. We identified 87 POLD1 variants predicted to be potentially pathogenic according to AlphaMissense and classified as loss or gain of function by either DeMaSK or GEMME with a mechanism related to changes in structural stability (**Text S7**). Of these, 72 have been reported in ClinVar as associated with colorectal cancer susceptibility (**Table S4**). One example is Y396S (p.Tyr396Ser), which has a ClinVar review status of 1. MAVISp predicts a substantial destabilizing effect for this variant, with changes in folding free energy ranging from 4.4 to 4.9 kcal/mol (Figure 6A). The residue is located within the exonuclease domain, and the mutation may induce local misfolding, potentially compromising the proofreading activity of POLD1. Supporting this prediction, a yeast-based assay using the *CAN1*, *trp1-289*, and *his7-2* reporters confirmed that Y396S leads to a high mutation rate [53] (**Figure 6A**, Experimental Data Classification row). Notably, among the 72 colorectal cancer-associated POLD1 variants with predicted destabilizing effects, several are also predicted to act as nucleation points for early folding events (**Text S3**). These include L266R, Q336P, W381C/R/S, S382P, S408P, F418C, M459K, L513P, L547P, I598T, T620R, V751E, Y801C/R, L851P, L872P, L967P and F970S (**Texts S7 and S8**). Their involvement in early folding suggests a key role in maintaining the structural integrity of the N-terminal, exonuclease and polymerase domains of POLD1 [54][33][55][56].

We next employed MAVISp to investigate the functional impact of POLD1 variants beyond structural stability. The lack of structural data for human POLD1 in the context of mismatch excision significantly limits MAVISp's ability to identify functionally relevant mutations associated with hypermutated cancers and thus restricts opportunities for experimental validation and potential variant classification. However, MAVISp provided predictions on functional effects on POLD1 variants that can be used for other purposes. These include potential disruptions to protein-protein interactions, DNA binding, catalytic activity, or cofactor binding. Following the same approach used for POLE, we focused on variants predicted to be

pathogenic by AlphaMissense and associated with loss or gain of function according to either DeMaSK or GEMME (**Text S8**). From this group, in the *simple* mode, we identified 65 variants with at least one functional mechanistic indicator unrelated to stability. Of these, 49 are reported in ClinVar as linked to colorectal cancer susceptibility (**Table S4**). Most of these variants are located at or near functional sites, including 20 within the exonuclease domain and three in the polymerase domain (**Table S4**). These are predicted to interfere with catalytic activity by affecting residues coordinating the catalytic site. Two variants, I317N (p.Ile317Asn) and Q336P (p.Gln336Pro), also show destabilizing effects on the 3D protein structure (**Text S8**). In addition, we identified mutations in residues that coordinate the Zn^{2+} cluster, C1015R (p.Cys1015Arg), C1029R (p.Cys1029Arg), C1029Y (p.Cys1029Tyr), and C1012Y (p.Cys1012Tyr), with C1012Y (p.Cys1012Tyr), also predicted to destabilize domain structure (**Table S4, Text S8, Figure 6A and 6B**). Mutations in surrounding residues, such as G1014D (p.Gly1014Asp) and R1032P (p.Arg1032Pro), may further impair Zn^{2+} binding, potentially compromising structural integrity or enzymatic function (**Figure 6B**). These mutations are unlikely to promote a tumorigenic phenotype. In particular, variants disrupting Zn^{2+} coordination, essential for maintaining the structural integrity of the PCNA-interacting domain, are expected to compromise DNA synthesis.

Another potential mechanism affecting protein stability involves changes in phosphorylation. In the wild-type protein, phosphorylation at specific sites can regulate stability [57], either by promoting degradation to turn off function or by enhancing stability to maintain activity [58]. Mutations at these sites can disrupt this regulation, leading to constant destabilization or stabilization, or even reversing the intended effect of phosphorylation. The MAVISp PTM module facilitates the exploration of these regulatory mechanisms. In POLD1, we focused on four variants, S314G (p.Ser314Gly), S314T (p.Ser314Thr), Y511F (p.Ser511Phe), and S783L (p.Ser783Leu), predicted to be pathogenic by Alphamissense and associated with loss or gain of function according to either DeMaSK or GEMME (**Text S3**). These sites have been reported in proteomics datasets, and phosphorylation of S783 has been specifically confirmed in a published study [59]. However, their functional roles and the responsible kinases remain undefined. Predictions from NetPhos [60] suggest PKA and CDC2 as potential kinases for S314 and S783, with confidence scores of 0.6 and 0.7, respectively. In MAVISp, the PTM regulation module classifies their variants as "uncertain" in both the simple and ensemble modes, primarily due to their low solvent accessibility, suggesting they are unlikely to be phosphorylated in the protein major state (**Text S3** and **S4**). An exception is S314T (p.Ser314Thr), which is predicted to have a neutral effect (**Text S3** and **S4**) since threonine is still a phosphorylatable residue type. To explore this further, we analyzed a 500-nanosecond molecular dynamics (MD) simulation to assess how solvent accessibility changes over time

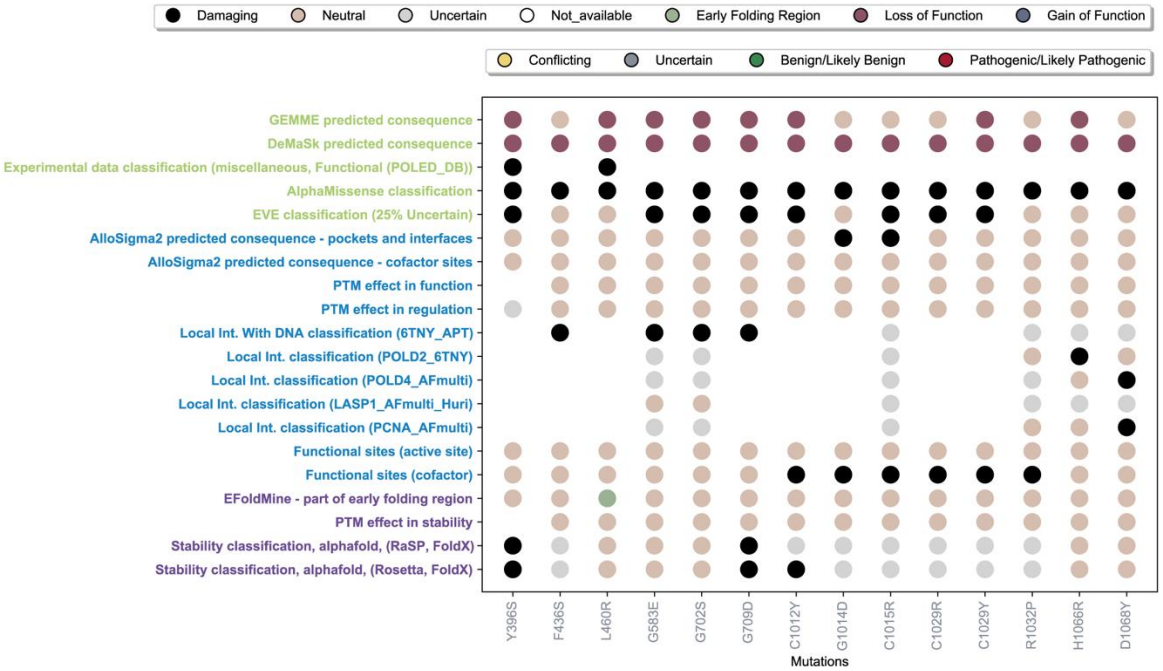
for these phosphosites. We calculated the time series and density plots of side-chain solvent accessibility for Y511, S783, and S314 (**Figure S1**). Results showed that S314 remained buried throughout the simulation, while Y511 and S783 were only briefly solvent-exposed (SASA > 20%). These findings suggest that these phosphorylation sites are unlikely to be modified in POLD1 major unbound state. Instead, phosphorylation may occur during conformational changes triggered by binding to partner proteins or in minor states accessible during longer timescales of protein dynamics.

To better understand how mutations affect the function of POLD1, we used MAVISp to analyze variants that disrupt local interactions with key binding partners. Disruption of the POLD1–POLD2 interaction can impair the stability of the polymerase δ holoenzyme and reduce its processivity, increasing the risk of replication and repair defects. One such variant, H1066R (p.His1066Arg), is predicted to weaken the interaction with POLD2, a subunit essential for holoenzyme stability and for acting as a scaffold for PCNA binding (**Figure 6A**). Residue H1068 in POLD1 is particularly important, as it is known to mediate interactions with both POLD4 and PCNA [61]. POLD4 plays a regulatory role in DNA repair, especially during the recovery of stalled replication forks through homologous recombination. In the context of colorectal cancer susceptibility, we identified D1068Y (p.Asp1068Tyr) as a potentially pathogenic, loss-of-function variant based on predictions from AlphaMissense and either DeMaSK or GEMME. However, the destabilization of the interaction between POLD1 and its accessory subunit POLD4 and PCNA is unlikely to promote a tumorigenic phenotype, given the critical role of this interaction in ensuring replication processivity. This suggests that further investigation is required to determine whether the mutation is associated with a tumor-compatible functional effect.

Since the catalytic activity of POLD1 and POLE also depends on DNA binding, mutations that disrupt this interaction are likely to impair polymerase function and compromise both replication and DNA damage response. Using MAVISp, we assessed the impact of variants on DNA binding (**Figure 6A** and **Text S8**) and identified among the several VUS in POLD1, F436S (p.Phe436Ser), G583E (p.Gly583Glu), G702S (p.Gly702Ser), and G709D (p.Gly709Asp), predicted to be pathogenic according to AlphaMissense, with impaired DNA interaction identified as the primary mechanistic effect (**Figure 6A and Text S8**). Notably, G709D (p.Gly709Asp) also affects protein stability, suggesting that multiple mechanisms may contribute to its functional impact. Importantly, both G702S (p.Gly702Ser) and G709D (p.Gly709Asp) have been reported in association with colorectal cancer (**Table S4**) and are located in the fingers and palm domains of POLD1, regions critical for coordinating DNA binding and mismatch sensing (**Figure 6B**). In the context of the associated phenotype, the

destabilizing effect is unlikely to be directly exploited by cancer cells to promote tumor progression. However, given its distance from the catalytic site, the residue may play a structural role in ensuring proper DNA positioning during replication. A mutation that weakens its interaction with DNA could therefore compromise activity or replication fidelity. Notably, the prediction reflects a local destabilization effect, implying that only the DNA region proximal to the mutation is affected rather than the entire molecule. Experimental validation would be required to establish a correlation between the predicted impact on DNA binding and replication fidelity loss, as a function of mutation position. The pathogenic predictions for these two variants can thus be interpreted in the context of impaired replication accuracy, highlighting them as compelling candidates for experimental validation and potential reclassification (Table 2).

A



B

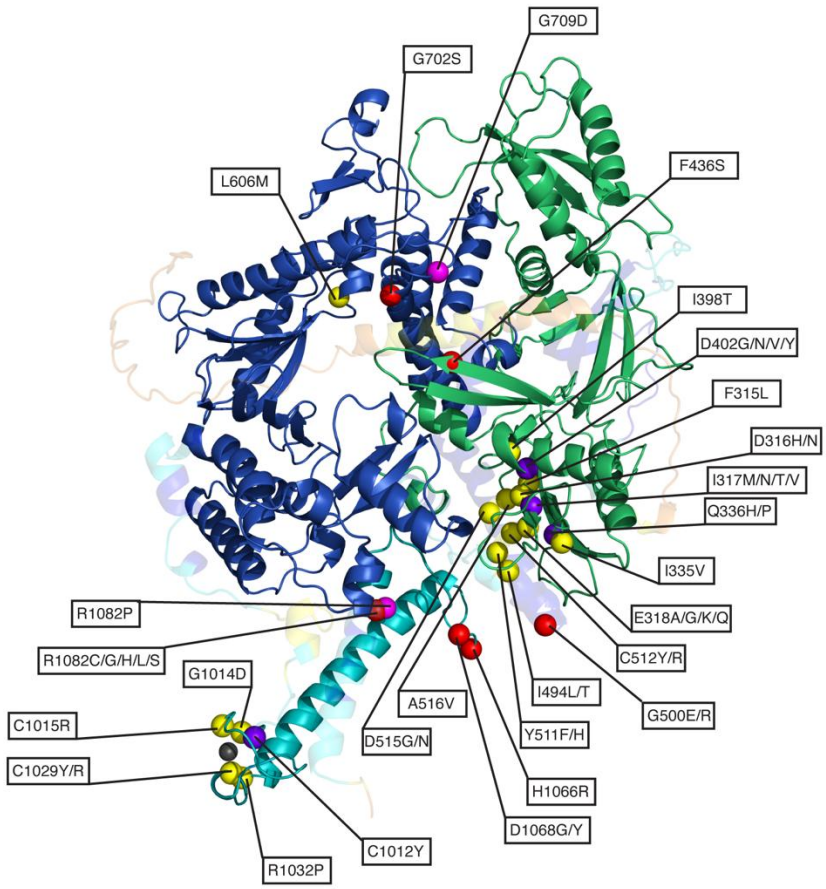


Figure 6 ***Selected VUS in POLD1 predicted as pathogenic and loss or gain of function and with supporting MAVISp mechanistic effects.*** *For an explanation on how to read the dotplot see the description in the caption of*

Figure 2. ***A)*** *Dot plot illustrating a selection (14) of VUS from POLD1 of interest in terms of their mechanistic indicators and/or association with colorectal cancer. All the variants reported have been filtered at first for their pathogenic potential according to AlphaMissense and predicted loss/gain-of-function with GEMME or DeMaSK.* ***B)*** *Diagram illustrating the 49 POLD1 VUS mutations impacting protein functional properties according to MAVISp, associated with susceptibility to colorectal cancer. These mutations are classified as potentially pathogenic by AlphaMissense and as loss or gain of function by GEMME or DeMaSK. The POLD1 structure is shown with critical domains highlighted as indicated by PFAM annotation: the exonuclease domain in green and the polymerase domain in blue. C-alpha positions of the affected residues are represented as spheres, colored based on their impact: red for mutations affecting interactions, yellow for mutations affecting functional sites, orange for mutations impacting both interactions and functional sites, fuchsia for mutations influencing both interaction and stability, and purple for mutations altering both stability and functional sites.*

In summary, our systematic analysis of POLE and POLD1 VUS using MAVISp demonstrates how structure-based mechanistic insights can complement traditional pathogenicity predictions to prioritize variants of uncertain significance or conflicting evidence. By integrating diverse data, including stability and DNA binding, protein-protein interactions, and phosphorylation-dependent regulation, MAVISp provides a detailed understanding of how specific mutations may compromise polymerase function.

Characterization of other variants

MAVISp also facilitates the interpretation of variants reported in external databases such as COSMIC and cBioPortal [26,62], offering insights into the potential molecular mechanisms underlying their functional or pathogenic effects. These analyses, available in the OSF repository linked to this study (see the *downstream analyses* folder therein), use the same dot plot and lolliplot visualizations applied to the ClinVar variants. For POLE, we identified 240 variants from cBioPortal and 171 from COSMIC that are associated with predicted molecular mechanisms, including loss and gain of-function and pathogenic predictions from GEMME and DeMaSK and from Alphamissense and EVE, respectively. Similarly, for POLD1, we identified 104 variants from cBioPortal and 37 from COSMIC with associated mechanistic indicators. Since MAVISp applies a comprehensive saturation mutagenesis approach, it also highlights additional potentially pathogenic variants that have yet to be reported in disease-related databases. In particular, we identified 20,585 and 11,449 previously unreported variants in POLE and POLD1, respectively, occurring at positions already known to be mutated and annotated in COSMIC, cBioPortal, or ClinVar. These variants are classified as pathogenic by AlphaMissense and predicted to result in loss- or gain-of-function by either DeMaSK or GEMME. These variants, along with their predicted effects, are accessible for exploration via the POLE and POLD1 entries in the MAVISp database (https://services.healthtech.dtu.dk/services/MAVISp-1.0/). Notably, several POLE variants such as I276N (p.Ile276Asn) and S459Y (p.Ser459Tyr) from cBioPortal, and F274C

(p.Phe274Cys) from COSMIC along with G364R (p.Gly364Arg), G364W (p.Gly364Trp), and L424F (p.Leu424Phe), which are found in both databases, are located in the exonuclease domain and predicted to impair interaction, structural stability, or catalytic site function. These variants represent strong candidates for further experimental validation. Likewise, the POLD1 variants Y405C (p.Tyr405Cys) and A1024D (p.Ala1024Asp) are predicted to affect structural stability and key functional regions, including the catalytic residues of the exonuclease domain and the coordination of Zn^{2+}, respectively.

Conclusions

Recognition of pathogenic variants in polymerase proofreading functions is critical for understanding genomic stability and DNA replication fidelity.
Here, we provide a comprehensive structural analysis of all possible missense variants in the DNA polymerases POLE (2-2286) and POLD1 (2-1107), using the MAVISp (Multi-layered Assessment of VarIants by Structure for proteins) framework. By integrating predictions of changes in protein stability, functional site perturbation, and disruption of DNA and protein–protein interactions, we provide mechanistic insights into both known pathogenic variants and VUS. While VUS in other proteins may be linked to diverse phenotypes or investigated for protein engineering applications, variants in POLE and POLD1 are primarily associated with colorectal cancer, polymerase proofreading-associated polyposis (PPAP), and hereditary cancer syndromes. In contrast, somatic POLE mutations are more frequently observed, and clinically significant, in endometrial cancer, where they play a critical role in patient management. We, therefore, interpreted the molecular consequences of these variants in light of their potential to promote a tumorigenic phenotype. A recurrent explanatory mechanism for the effect of the studied mutations emerging from our analysis is the alteration of structural stability. However, interpreting this effect in the context of cancer for these genes is particularly complex. Given the essential role of POLE and POLD1 in DNA replication, complete protein misfolding leading to degradation is unlikely to support a tumor-promoting phenotype. Mutations that markedly reduce polymerase abundance within the cell are therefore not expected to confer a selective advantage in cancer, and are more plausibly interpreted as heterozygous passenger variants, functionally compensated by the remaining wild-type allele. Accordingly, homozygous mutations observed in cancer samples for which MAVISp predicts only a strong destabilizing effect on protein stability ($\Delta\Delta G > 3$ kcal/mol), and no other mechanistic consequence, should not be automatically interpreted as causing polymerase depletion. This is particularly relevant given that $\Delta\Delta G$-based predictors, including those implemented in MAVISp, model only local structural perturbations around the mutated residue

and cannot account for global folding dynamics, especially in large, multidomain proteins such as replicative polymerases.

These observations suggest that partial or local misfolding affecting enzymatic function, rather than full destabilization or degradation, is a more plausible mechanism in the context of tumorigenesis. Variant interpretation must also consider the functional domain in which the mutation occurs. In both POLE and POLD1, local destabilization of the exonuclease domain may impair proofreading activity while preserving polymerase function, thereby contributing to a hypermutator phenotype that facilitates tumor progression. Conversely, mutations in the polymerase domain require more nuanced interpretation: local destabilization may not abolish DNA synthesis outright, but could selectively reduce replication fidelity. Altogether, these considerations underscore the importance of experimentally validating variants predicted to affect structural stability, particularly through assays assessing mutation rates and cellular proliferation. Such efforts are essential both to clarify their oncogenic potential and to benchmark the accuracy of MAVISp in prioritizing clinically actionable variants.

Importantly, our framework identified several compelling candidates for further experimental validation. These include variants in the exonuclease domain predicted to impair DNA binding or indirectly affect exonuclease catalysis. In *POLE*, examples include P227L (p.Pro227Leu), Q292H (p.Gln292His), M295V (p.Met295Val), D462N (p.Asp462Asn), T466A (p.Thr466Ala), V460M (p.Val460Met). Additionally, we identified compelling candidate variants within the polymerase domain predicted to affect the DNA binding interface, particularly in regions that may be critical for mismatch recognition. In *POLE*, G977R (p.Gly977Arg), Y726D (p.Tyr726Asp), and D1129G (p.Asp1129Gly), and in POLD1, G702S (p.Gly702Ser) and G709D (p.Gly709Asp), stand out as particularly promising (Table 2). Nonetheless, our integrative approach, combining structure-informed mechanistic predictions with phenotypic context, provides a robust framework for prioritizing variants for functional validation and potential clinical reclassification. By guiding the design of targeted functional assays, this strategy supports both the validation of computational predictions and the generation of evidence that may contribute to variant interpretation under the ACMG/AMP guidelines. In conclusion, our work highlights distinct structural mechanisms through which POLE and POLD1 variants may alter enzymatic function. While MAVISp effectively captured critical functional and structural features of POLE variants, its application to POLD1 is currently limited by the lack of DNA-bound human structures. As structural data becomes available, particularly for excision repair intermediates, the MAVISp database can be complemented using this new information, therefore improving its effectiveness. Taken together, this framework offers a

promising platform to enhance variant interpretation, guide targeted experimental validation and ultimately inform and advance molecular diagnostics for cancer and hereditary syndromes.

Materials and Methods

Identification and aggregation of variants for the study

We used Cancermuts [63] to retrieve variants from COSMIC [26] and cBioPortal [64], as well as to aggregate variants from ClinVar [7] [65]. We filtered variants downloaded from ClinVar, keeping only those annotated as having a protein mutation in specific protein sequences, namely RefSeq IDs NP_006222 (POLE) and NP_002682 (POLD1). These correspond to the sequences of the main UniProt isoform. The data was collected on 24/06/2023. We applied the Cancermuts workflow (https://github.com/ELELAB/cancermuts) and a Python script clinvar.py (available in the mavisp_templates folder of https://github.com/ELELAB/MAVISp_automatization) to retrieve the variants of interest and their annotations.

Retrieval of structure from the Protein Data Bank and the AlphaFold Protein Structure Database

We applied the STRUCTURE_SELECTION module of MAVISp [8] to retrieve the AlphaFold (AF) models from the AlphaFold Protein Structure Database [66], along with information on structures available in the Protein Data Bank (PDB) [67]. We compared the quality of the AF models against the available experimental structures of both POLE and POLD1, and ultimately selected the catalytic subunits of both proteins ($POLE_{2\text{-}2286}$ and $POLD1_{2\text{-}1107}$) in the AF models, for our analyses. Zinc ions were modelled on the structures by superimposition to an X-ray structure (PDB ID 5VBN, chain A [68] and a Cryo-EM structure (PDB ID 6TNY, chain A [69]), to POLE and POLD1, respectively. The sulfur-iron cluster (SF4) in the C-terminal CysB motif of POLD1, which is required for successful DNA synthesis and repair mediated by the enzyme, and whose loss implies a slight destabilization of the protein [70], has not been included in the structure. This exclusion is due to its inability to fit its structure in the AF model, so that it would be correctly bound to the four conserved cysteine residues (Cys1058, Cys1061, Cys1071, Cys1076) that coordinate it, as well as the lack of parameters in the tools used in the MAVISp *simple* mode to represent this cofactor.

Starting structures for analyses of protein complexes

Concerning the selection of structures to be used in the LOCAL INTERACTION module, we first identified interactors for POLE and POLD1 by mining the Mentha database [71], using the Mentha2PDB script from the MAVISp toolkit (https://github.com/ELELAB/PPI2PDB), and focused on interactors with a high Mentha score and available structural data. In the case of POLE, we used a Cryo-EM structure with 2.80 Å resolution (PDB ID 7PLO; [72]), in which POLE (chain A) interacts with POLE2 (chain B; $POLE_{1371-2280}$-$POLE2_{1-527}$), MCM2 (chain D; $POLE_{1371-2280}$-$MCM2_{169-893}$), MCM5 (chain E; $POLE_{1371-2280}$-$MCM5_{21-651}$), and CDC45 (chain C; $POLE_{1371-2280}$-$CDC45_{1-566}$) and a cryo-EM structure with 2.95 Å resolution for the POLE-PCNA complex (PDB ID 9B8T [73]). We could not include the Zn^{2+} ion coordinated by Cys2158, Cys2161, Cys2187, and Cys2190 in the model by superimposition, as it would have created steric clashes even after optimization of the rotamers of the surrounding residues. Regarding POLD1, we selected a Cryo-EM structure of resolution 3 Å (PDB ID 6TNY [69] to study the interaction between POLD1 (chain A) and POLD2 (chain B) ($POLD1_{78-1107}$-$POLD2_{1-462}$). We excluded the iron-sulfur cluster from the C-terminal domain of POLD1, due to the same limitations we encountered for POLE. As the interfaces between POLD1 and POLD4 or PCNA included a few missing residues, we modelled the complex using AlphaFold-multimer and selected the top-ranking model after a quality control step based on the Predicted Aligned Error (PAE) score of AlphaFold and the pDOCKQ2 score (>0.23) [74]. For POLD1, we identified two pre-calculated and experimentally validated models [75] with pDOCKQ2 scores > 0.3, POLD1-C9ORF169 and POLD1-LASP1. Of these, only POLD1-LASP1 was included in the LOCAL_INTERACTION analyses, having discarded the POLD1-C9ORF169 model due to its high level of disorder. Additionally, we generated models for all the remaining interactors reported by Mentha2PDB using AlphaFold3 (https://alphafoldserver.com/) and retained for the LOCAL_INTERACTION only the models that passed the quality control in terms of different scores (as explained in the original publication of MAVISp [75]. No additional structure passed the quality control steps with respect to the ones described above.

Starting structures for analyses of protein-DNA complexes

We retrieved the structure of POLD1 in complex with DNA from a Cryo-EM structure ($POLD1_{78-1107}$-DNA) with a resolution of 3.08 Å (PDB ID 6TNY, chains A, P and T [69]). We retrieved the structures of POLE bound to DNA within the polymerase domain in both the open (PDB ID 9B8S, residues 27-1198, chains A, P and T [73]) and closed (PDB ID 9B8T, residues 24-1198, chains A,P and T states [73]). The corresponding PDB files were processed to remove molecules not supported by FoldX [76]. Specifically, the iron-sulfur cluster was removed from both structures, and the dTTP residue representing the entering dNTP was excluded from the 9B8T structure. Additionally, the A275 and A277 residues were retro-mutated to D275 and

E277, respectively, using the *mutatagenesis wizard* function of PyMOL [77], selecting optimal rotamers based on clash and steric hindrance considerations. In addition, four structures were employed to study the impact of mutations on the binding of POLE's exonuclease domain to DNA in various states: post-insertion (PDB ID 9F6I, residues 27-1175, chains A, P and T, with a resolution 3.30 Å, [31], arrest (PDB ID 9F6J, residues 27-1175, chains A, P and T, with a resolution of 3.90 Å [31]), frayed substrate (PDB ID 9F6K, residues 27-1175, chains A, P and T, with a resolution of 4.20 Å [31]), and mismatch excision (PDB ID 9F6L, residues 27-1175, chains A, P and T, with a of resolution of 3.90 Å [31]). All PDB files were processed to remove the 4Fe4S clusters. In the 9F6I structure, the substrate analog DDS and a Ca^{2+} ion were removed, and, in both 9F6I and 9F6J, the A275 and A277 residues were retro-mutated to D275 and E277.

Molecular dynamics simulations

The AF models of the catalytic subunits of POLE ($POLE_{2-2286}$) and POLD1 ($POLD1_{2-1107}$), as previously detailed, (see previous section) were used as starting structures for molecular dynamics (MD) simulations. MD simulations were performed using GROMACS 2022 or 2024 [78] in the NVT ensemble at 298 K. Histidine residues were protonated following the assignment provided by GROMACS, with the exception of H404 in POLE, which was manually protonated on Nδ (HSD). We carried out all-atom explicit-solvent MD simulations of POLE and POLD1 in their free state (i.e., DNA-unbound) using the CHARMM36m force field [79]. We used a dodecahedral box of TIP3P water molecules with a minimum distance between the protein and the box edges of 15 Å, applying periodic boundary conditions. We used a concentration of NaCl of 150 mM, neutralizing the net charge of the system. Bond lengths involving hydrogen atoms were constrained using the Linear Constraint Solver (LINCS) algorithm [80]. Non-bonded interactions were handled using a 0.9 nm cut-off for both van der Waals and electrostatics, with long-range electrostatics treated using the particle-mesh Ewald (PME) method [81] [82]. The system was equilibrated in multiple steps. Initially, temperature equilibration was performed under NVT conditions with position restraints applied to the heavy atoms of the protein. This was followed by pressure equilibration in the NPT ensemble in two steps: first with restraints on the protein backbone, and then without any restraints. In all simulations, temperature was kept constant at 298 K using the velocity-rescale thermostat, and pressure was coupled to 1 bar using either the Berendsen [83] or Parrinello-Rahman [84] barostat. We collected 500 ns of MD simulation for each system, according to the default MAVISp protocol when systems contain more than 100,000 atoms upon solvation. The input, parameters, and trajectory files are reported in the OSF repository (https://osf.io/z8x4j) for each protein.

Effects of mutations on protein structural stability

To estimate the changes in folding free energy upon mutation (ΔΔGs), we used the STABILITY module in *simple mode* of MAVISp [8] using the AF models of the catalytic subunits of POLE ($POLE_{2-2286}$) and POLD1 ($POLD1_{2-1107}$). We therefore applied both MutateX [30], which employs FoldX 5 [76], as well as RosettaDDGPrediction, which utilizes the Rosetta cartddg2020 protocol [85] with the ref2105 energy function [86]. In addition, we applied RaSP [20], a machine learning model trained on Rosetta. While Rosetta was only applied to a selection of variants due to limitations in available computational power, both MutateX and RaSP were employed for performing saturation mutagenesis scans. We used the workflows of the MAVISp toolkit for each of these methods (https://github.com/ELELAB/mutatex, https://github.com/ELELAB/RosettaDDGPrediction, and https://github.com/ELELAB/RaSP_workflow). According to the MAVISp framework, variants were classified as destabilizing if consensus ΔΔG predictions from MutateX and RaSP or from MutateX and Rosetta exceeded 3 kcal/mol, neutral if ΔΔG predictions ranged between -1 and 1 kcal/mol, and as stabilizing if both predictions agreed on ΔΔG ≤ -3 kcal/mol. Furthermore, we labelled the predicted effect as uncertain if any of either values fell between -3 < ΔΔG < -1 kcal/mol, or 1 < ΔΔG < 3 kcal/mol, or in case of disagreement between the methods.

Additionally, we used the self-scan function of MutateX [30] with and without the repair step of FoldX as a control experiment to evaluate the quality of the initial structures before and after the repair. This approach consisted of mutating each residue to its own type. The ΔΔG values associated with self-mutations are expected to be as close as possible to 0 kcal/mol. Any deviations from this value in the repaired model (i.e., repair function active) or the initial structure (i.e., repair function inactive) indicates an initial unfavorable orientation of the sidechain and/or the surrounding residues. The self-scans, with the repair function on and off, confirmed the quality of the initial structures. No ΔΔG values were found above or below the cutoffs of ±3 kcal/mol in either the X-ray structures or in the AF models.

Effect of mutations on early folding events

We employed the EFOLDMINE module of MAVISp which predicts whether a specific mutation site constitutes a seed of an early folding event based on EfoldMine [87]. We applied a threshold of 0.169 for the EfoldMine score to define sites involved in early folding events and considered regions with a minimum length of three early folding residues to avoid isolated peaks [87].

Effects of mutations on binding free energies

To estimate the changes in binding free energies upon mutation, we applied the LOCAL_INTERACTION module of MAVISp [8] based on FoldX 5 [76] and the RosettaDDGPrediction flexddg protocol [85][88] with the talaris2014 energy function [89] [90] [91]. The calculations were carried out for each complex (see OSF repository), retaining only mutations within 10 Å of the protein-protein interface to accommodate the local sampling carried out by both FoldX and Rosetta. The MutateX [30] and RosettaDDGPrediction workflows used for the data collection are available and maintained on GitHub (https://github.com/ELELAB/mutatex and https://github.com/ELELAB/RosettaDDGPrediction).

Effects of mutations at phosphorylation sites

We applied the MAVISp PTM module to predict the effects of mutations at known phosphorylatable sites on the protein [8]. Briefly, the module uses multiple sources of information and custom decision logic to predict whether each mutation could have consequences on regulation by phosphorylation (PTM_regulation), structural stability (PTM_stability), or function (PTM_function) due to the interplay between the abolished phospho-site and the effect of the amino acid substitution. The data sources used to infer this includes the solvent accessibility of the mutation site calculated using NACCESS [92] on the protein structure, whether the mutation is part of a phosphorylatable linear motif as determined by ELM [93] using Cancermuts [63], as well as, the predicted changes of stability or binding free energies as calculated by FoldX 5 [76]. In the interpretation of the results from the PTM modules, as the PTM_regulation module integrates proteomics data with structural analysis of solvent-accessible surface area (SASA) to estimate the likelihood of phosphorylation at a given site, it should be applied as an initial filter to exclude residues that are sufficiently buried and thus unlikely to be accessible to kinases. In cases where the effect at this layer is predicted neutral or uncertain, the results from the other two modules should be disregarded.

Effects of mutations on long-range communication to functional sites

We used the MAVISp LONG_RANGE module to predict the mutations with allosteric effects in both the *simple* and *ensemble* modes [8]. Briefly, the module is based on the coarse-grained model of AlloSigMA2 [94], which simulates the effects of perturbations across the protein structure. The model represents mutations as either UP or DOWN, relating to the change in side-chain size (increase; UP and decrease; DOWN), upon mutation, respect to the wild type residue. The module applies a systematic filtering workflow to identify candidate mutations with allosteric effects (https://github.com/ELELAB/MAVISp_allosigma2_workflow). Namely,

for each mutation that resulted in a significant change in residue size (at least 5 $Å^3$), we identified all the predicted response residues in the protein that displayed a significant positive or negative change in allosteric free energy ($|\Delta g| > 2$ kcal/mol). We further filtered this list of residues by removing those in close proximity to the mutation site (C_α-C_α distance < 15 Å), to remove possible local-contacts. Furthermore, the response residues were further filtered to retain only those solvent accessible (relative side chain solvent accessible surface as calculated by NACCESS [92] >= 20%) and predicted to be localized within pockets by FPocket [95,96]. Finally, the *simple* mode module classifies each mutation as stabilizing if it solely caused negative changes in allosteric free energy, destabilizing if the changes were positive, mixed effects if both changes were observed in both directions, and neutral if no significant change was predicted. Mutations that did not result in significant changes in residue side-chain volume were classified as uncertain.

In the *ensemble* mode, the LONG_RANGE module employs graph-based methodologies on ensemble structures to refine and validate predictions of the *simple* mode module. The module uses atomic contact-based protein structure networks (acPSN), initially proposed by Vishveshwara's group [96], constructed using the PyInteraph2 framework [97] for analysis of molecular dynamics (MD) simulations. To define the edges of the network, we considered residue pairs with a sequence separation greater than one and a distance between their atoms of less than 4.5 Å. Residue-specific normalization factors were applied to account for their inherent propensity to form contacts. The interaction strength between two residues was calculated by dividing the number of atom pairs within 4.5 Å by the square root of the product of the normalization factors for each residue, then multiplying by 100 [96]. If the interaction strength exceeded a predefined threshold (I_{min} = 2.5 for our simulations), an edge was created, weighted by the calculated strength. Furthermore, only edges present in more than 50% of the frames across the ensemble were retained in the final acPSN. The acPSN is then weighted according to the average interaction strength. Subsequently, we then utilized the *path_analysis* tool of PyInteraph2 [98] to identify whether long-range communication routes exist between each identified mutation and the respective response sites, reported in the *simple* mode predictions. Shortest paths of communication are calculated, utilizing the Dijkstra algorithm [99] as implemented in NetworkX [100]. The workflow then filters the identified paths, retaining only those that span four or more residues in length, to ensure only long-range communication paths are considered. We generated the data using the code available for the LONG_RANGE data collection in the *ensemble* mode at https://github.com/ELELAB/MAVISp_allosigma2_workflow. The module then classifies the given mutation as having a damaging long-range effect if a path is identified for at least one the predicted response sites, serving as a validation of the *simple* mode results.

Effect of mutations on active and cofactor binding sites

We employed the MAVISp FUNCTIONAL_SITES module to predict the impact of mutations occurring at residues within the second coordination sphere in relation to active site residues or their corresponding binding sites. To perform the contact analysis, we utilized the Arpeggio software [101] on the POLE and POLD1 wild-type structures, as well as $POLE_{30-522}$ S297F protein structures. Before analysis, structure optimization was conducted using a Conjugate Gradients minimization method [102] with 50 minimization steps, employing the MMFF94 force field [103] using a van der Waals cutoff of 0.1, an interacting cutoff of 5.0 Å, and a physiological pH of 7.4. Subsequently, the output was preprocessed to exclude clashing contacts, attributed to poorly modelled residues, as well as contacts classified as "proximal," which the Arpeggio [101] framework deems non-critical (https://github.com/ELELAB/mavisp_accessory_tools). We classified as damaging variants of residues identified by this analysis, and as neutral the remaining residues.

Analysis of variants affecting splicing sites

We evaluated the potential impact of each mutation on splice sites using SpliceAI [104] and Pangolin software [105]. Specifically, we employed the SpliceAI_lookup API from the Broad Institute (https://spliceailookup.broadinstitute.org) and used a Python script to automate the collection of results (https://github.com/ELELAB/mavisp_accessory_tools). In detail, we predicted the effects of genomic mutations within a 500-nucleotide window upstream and downstream of the mutation site. Using SpliceAI [104], we obtained predictions on whether a specific nucleotide gains or loses splicing acceptor activity, as well as whether it gains or loses splicing donor activity. With Pangolin [105], we assessed whether a specific nucleotide acquires or loses splicing activity. Each prediction is based on scores representing the difference between the reference score (in the absence of mutation) and the alternate (ALT) scores (in the presence of mutation) obtained from SpliceAI and Pangolin for a given position. These scores reflect the likelihood of the position becoming an acceptor or donor site or losing its acceptor or donor activity (SpliceAI) or acquiring or losing splicing site properties (Pangolin). The results for POLE and POLD1 were subsequently aggregated into an Excel file and uploaded as a supplementary **Tables S2** and **S3**.

Acknowledgements

Our research has been supported by Danmarks Grundforskningsfond (DNRF125), Novo Nordisk Fonden Bioscience and Basic Biomedicine (NNF20OC0065262) to the E.P. group. Part of the calculations have been supported by a EuroHPC Benchmark Access Grant (EHPC-BEN-2023B02-010) and a EuroHPC Regular Grant (EHPC-REG-2023R01-051) on Discoverer. The work is also supported by A PhD Fellowship from the Danish Data Science Academy (DDSA) to MA. AR by The Assar Gabrielsson's foundation and The Healthcare Board, Region Västra Götaland.

Author Contributions

Conceptualization: EP *Data Curation*: MA, LB, KK, AR, EP *Formal Analysis*: MA, LB, EP *Investigation:* MA, AR, EP *Funding Acquisition:* EP, AR, MN, MA *Methodology*: MA, LB, KK, MT, EP *Project administration:* EP. *Resources*: EP, AR. *Software*: MA, LB, KK, MT, EP *Supervision*: EP. *Validation*: All the co-authors. *Visualization*: MA, EP. *Writing – Original Draft:* MA, EP. *Writing – Review and Editing*: All the coauthors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used OpenAI ChatGPT 4.0 to improve the language of the manuscript. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability statement

All data and code, including input files and simulation trajectories, are freely available on OSF (https://osf.io/z8x4j) and the different GitHub repositories provided in the Materials and Methods Section. The aggregated data for POLE and POLD1 are also available as entries in the MAVISp database (https://services.healthtech.dtu.dk/services/MAVISp-1.0/).

Tables

Table 1. Performance metrics of the pathogenicity predictors (EVE, AlphaMissense) and loss/gain-of-function predictors (GEMME, DeMaSk) implemented in MAVISp, in classifying pathogenic and benign variants of POLE and POLD1, as reported in ClinVar. The label 'filtered' indicates the analysis carried out on the subset of variants for which both AlphaMissense and EVE provided a classification different from ambiguous/uncertain.

Predictor	Total Variants	Sensitivity	Specificity	Accuracy	Precision	F1 score
AlphaMissense	58	0.917	0.957	0.948	0.846	0.88
AlphaMissense - filtered	35	1.0	0.958	0.971	0.917	0.957
EVE	38	0.818	1.0	0.947	1.0	0.900
EVE - filtered	35	0.818	1.0	0.943	1.0	0.900
DeMaSK	61	0.667	0.98	0.918	0.889	0.762
GEMME	61	0.75	0.98	0.934	0.9	0.818

Table 2. Selected candidate variants for experimental validation. This table lists high-priority missense mutations in POLE and POLD1 for which MAVISp predicted a mechanistic impact consistent with a tumorigenic phenotype. All listed variants were also independently predicted to be pathogenic by AlphaMissense and to result in loss or gain of-function by at least one evolutionary constraint-based method (GEMME or DeMaSK).

Mutation	Protein	Domain	Predicted MAVISp Effect
P227L	POLE	Exonuclease	DNA binding during mismatch excision
P282S	POLE	Exonuclease	DNA binding during mismatch excision
K284E	POLE	Exonuclease	DNA binding during mismatch excision
P286S	POLE	Exonuclease	DNA binding during mismatch excision
Q292H	POLE	Exonuclease	DNA binding during mismatch excision
M295V	POLE	Exonuclease	DNA binding during mismatch excision + abolishment of contact with E277 across all the mismatch states
V460M	POLE	Exonuclease	abolishment of contact with D462 across all the mismatch states
D462N	POLE	Exonuclease	Catalytic site involved in Mg2+ coordination
T466A	POLE	Exonuclease	abolishment of contact with D462 in all the mismatch states
Y726D	POLE	Polymerase	DNA binding during mismatch excision
R955G	POLE	Polymerase	DNA binding during mismatch excision
Y956F	POLE	Polymerase	DNA binding during mismatch excision
V958L	POLE	Polymerase	DNA binding during mismatch excision
G977R	POLE	Polymerase	DNA binding during mismatch excision
K1042E	POLE	Polymerase	DNA binding during mismatch excision
D1129G	POLE	Polymerase	DNA binding during mismatch excision
G702S	POLD1	Polymerase	DNA binding in the polymerase domain
G709D	POLD1	Polymerase	DNA binding in the polymerase domain

Supplementary material

Supplementary Figure S1. **Solvent Accessibility (SASA) Analysis of S314, Y511, S783 in POLE.** The time series for the relative solvent accessibility of the side chain atoms of S314 (A), Y5111 (B) and S783 (C) in the MD simulation of POLE are reported on the left, whereas the corresponding density plots are reported on the right panels.

Supplementary Text S1. MAVISp dotplots in the *simple* mode for POLE VUS variants. The results from each module of the MAVISp *simple* mode for POLE variants annotated as VUS in ClinVar are depicted in the dotplots. Each column in the plot represents a specific variant, with dots on the plot corresponding to the predicted effect or provided annotation by MAVISp for that variant. The MAVISp features related to effects on structural stability are reported in blue, whereas the MAVISp features related to different functional effects are reported in purple. The row in light green indicates the results of modules not directly referred to the mechanistic indicators but either to prediction of pathogenicity, loss/gain-of-function or available experimental data.

Supplementary Text S2. MAVISp dotplots in the *ensemble* mode for POLE VUS variants. The results from each module of the MAVISp *ensemble* mode for POLE variants annotated as VUS in ClinVar are depicted in the dotplots. Each column in the plot represents a specific variant, with dots on the plot corresponding to the predicted effect or provided annotation by MAVISp for that variant. The MAVISp features related to effects on structural stability are reported in blue, whereas the MAVISp features related to different functional effects are reported in purple. The row in light green indicates the results of modules not directly referred to the mechanistic indicators but either to prediction of pathogenicity, loss/gain-of-function or available experimental data.

Supplementary Text S3. MAVISp dotplots in the *simple* mode for POLD1 VUS variants. The results from each module of the MAVISp *simple* mode for POLD1 variants annotated as VUS in ClinVar are depicted in the dotplots. Each column in the plot represents a specific variant, with dots on the plot corresponding to the predicted effect or provided annotation by MAVISp for that variant. The MAVISp features related to effects on structural stability are reported in blue, whereas the MAVISp features related to different functional effects are reported in purple. The row in light green indicates the results of modules not directly referred to the mechanistic indicators but either to prediction of pathogenicity, loss/gain-of-function or available experimental data.

Supplementary Text S4. MAVISp dotplots in the *ensemble* mode for POLD1 VUS. The results from each module of the MAVISp *ensemble* mode for POLD1 variants annotated as VUS in ClinVar are depicted in the dotplots. Each column in the plot represents a specific variant, with dots on the plot corresponding to the predicted effect or provided annotation by MAVISp for that variant. The MAVISp features related to effects on structural stability are reported in blue, whereas the MAVISp features related to different functional effects are reported in purple. The row in light green indicates the results of modules not directly referred to the mechanistic indicators but either to prediction of pathogenicity, loss/gain of function or available experimental data.

Supplementary Text S5. POLE VUS with loss of function, predicted pathogenic and with available mechanistic indicators as predicted by MAVISp in the *ensemble* mode.

The lolliplot illustrates the variants in POLE with a linked mechanistic indicator which are also potentially pathogenic based on AlphaMissense predictions, as well as loss or gain of function alterations identified by GEMME or DeMaSk.

Supplementary Text S6. POLE VUS with loss of function, predicted pathogenic and with available mechanistic indicators as predicted by MAVISp in the *simple* mode. The lolliplot illustrates the variants in POLE with a linked mechanistic indicator which are also potentially pathogenic based on AlphaMissense predictions, as well as loss or gain of function alterations identified by GEMME or DeMaSk.

Supplementary Text S7. POLD1 VUS with loss or gain of function, predicted pathogenic and with available mechanistic indicators as predicted by MAVISp in the *ensemble* mode. The lolliplot illustrates the variants in POLD1 with a linked mechanistic indicator which are also potentially pathogenic based on AlphaMissense predictions, as well as loss or gain of function alterations identified by GEMME or DeMaSk.

Supplementary Text S8. POLD1 VUS with loss or gain of function, predicted pathogenic and with available mechanistic indicators as predicted by MAVISp in the *simple* mode.The lolliplot illustrates the variants in POLD1 with a linked mechanistic indicator which are also potentially pathogenic based on AlphaMissense predictions, as well as loss or gain of function alterations identified by GEMME or DeMaSk.

Supplementary Table S1. Variants of POLE and POLD1 reported as pathogenic, likely pathogenic, benign, or likely benign in ClinVar at the date of 03/06/2024.

Supplementary Table S2. Predictions with SpliceAI. We included the genomic coordinates of the mutation, which were provided as input to SpliceAI, listed under the "genomic_coordinates" column. The table also provides the position within a 500-nucleotide window (default: ±500 bp) that may be affected by the variant, as specified in the "affected_nucleotide_position" column. The predicted types of splicing alterations caused by the variants are detailed in the "Δ_type" column. These include "acceptor gain", indicating that the nucleotide specified in the "affected_nucleotide_position" gains splicing acceptor activity; "donor gain", where the nucleotide gains splicing donor activity; "acceptor loss", where the nucleotide loses splicing acceptor activity; and "donor loss", where the nucleotide loses splicing donor activity. The "Δ_score" column reflects the difference between the reference (REF) and alternate (ALT) scores obtained from SpliceAI for a given position. These scores represent the likelihood of a position becoming an acceptor or donor site or losing its acceptor or donor activity. The "REF_score" and *ALT_score* columns provide the computed probabilities of a given position being a splice acceptor or donor based on the reference and alternate haplotype sequences, respectively, as predicted by SpliceAI. The table also includes the *gene_id* column, which lists the Ensembl gene code representing the gene containing the mutation, and the *transcript_id* column, which provides the Ensembl transcript code corresponding to the transcript affected by the mutation. Lastly, the *ref_seq_id* column includes the RefSeq identifier of the protein (if applicable) affected by the mutation, assuming the transcript encodes a protein.

Supplementary Table S3. Predictions with Pangolin of effects induced upon mutation on splicing sites. We included the genomic coordinates of the mutation, which were provided as input to Pangolin, listed under the "genomic_coordinates" column. The table also provides

the position within a 500-nucleotide window (default: ±500 bp) that may be affected by the variant, as specified in the “affected_nucleotide_position” column. The predicted types of splicing alterations caused by the variants are detailed in the “Δ_type” column. These include “splice gain”, where the nucleotide acquires splicing activity, and “splice loss”, where the nucleotide loses splicing activity. The “Δ_score” column reflects the difference between the reference (REF) and alternate (ALT) scores obtained from Pangolin for a given position. These scores represent the likelihood of a position acquiring or losing splicing site properties. The “REF_score” and *ALT_score* columns provide the computed probabilities that the position is a splice site based on the reference and alternate haplotypes. The table also includes the *gene_id* column, which lists the Ensembl gene code representing the gene containing the mutation, and the *transcript_id* column, which provides the Ensembl transcript code corresponding to the transcript affected by the mutation. Lastly, the *ref_seq_id* column includes the RefSeq identifier of the protein (if applicable) affected by the mutation, assuming the transcript encodes a protein.

Supplementary Table S4. Annotations of 4251 variants for POLE and 2003 variants for POLD1, classified as Variants of Uncertain Significance (VUS) or with conflicting interpretations of pathogenicity in ClinVar. The table includes information regarding the domains in which the mutations occur, ClinVar variant IDs, ClinVar interpretations, associated conditions, and the ClinVar review status, which reflects the reliability of the interpretation.

References

1. Bębenek A, Ziuzia-Graczyk I. Fidelity of DNA replication—a matter of proofreading. Current Genetics. 2018. doi:10.1007/s00294-018-0820-1
2. Strauss JD, Pursell ZF. Replication DNA polymerases, genome instability and cancer therapies. NAR Cancer. 2023;5. doi:10.1093/narcan/zcad033
3. Mur P, García-Mulero S, del Valle J, Magraner-Pardo L, Vidal A, Pineda M, et al. Role of POLE and POLD1 in familial cancer. Genetics in Medicine. 2020;22. doi:10.1038/s41436-020-0922-2
4. Henninger EE, Pursell ZF. DNA polymerase ε and its roles in genome stability. IUBMB Life. 2014. doi:10.1002/iub.1276
5. Nicolas E, Golemis EA, Arora S. POLD1: Central mediator of DNA replication and repair, and implication in cancer and other pathologies. Gene. 2016. doi:10.1016/j.gene.2016.06.031
6. Prindle MJ, Loeb LA. DNA polymerase delta in dna replication and genome maintenance. Environmental and Molecular Mutagenesis. 2012. doi:10.1002/em.21745
7. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020;48: D835–D844. doi:10.1093/NAR/GKZ972
8. Arnaudi M, Utichi M, Degn K, Beltrame L, Scrima S, Krzesińska K, et al. MAVISp: A Modular Structure-Based Framework for Protein Variant Effects. BioRxiv. 2022. pp. 1–23. doi:10.1101/2022.10.22.513328

9. Li H, Xie B, Zhou Y, Rahmeh A, Trusa S, Zhang S, et al. Functional roles of p12, the fourth subunit of human DNA polymerase δ. Journal of Biological Chemistry. 2006;281. doi:10.1074/jbc.M600322200
10. Meng X, Zhou Y, Lee EYC, Lee MYWT, Frick DN. The p12 subunit of human polymerase Δ modulates the rate and fidelity of DNA synthesis. Biochemistry. 2010;49. doi:10.1021/bi100042b
11. Lin SHS, Wang X, Zhang S, Zhang Z, Lee EYC, Lee MYWT. Dynamics of enzymatic interactions during short flap human Okazaki fragment processing by two forms of human DNA polymerase δ. DNA Repair (Amst). 2013;12. doi:10.1016/j.dnarep.2013.08.008
12. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381: eadg7492. doi:10.1126/science.adg7492
13. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599: 91–95. doi:10.1038/s41586-021-04043-8
14. Laine E, Karami Y, Carbone A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. Mol Biol Evol. 2019;36: 2604–2619. doi:10.1093/molbev/msz179
15. Munro D, Singh M. DeMaSk: A deep mutational scanning substitution matrix and its use for variant impact prediction. Bioinformatics. 2020;36: 5322–5329. doi:10.1093/bioinformatics/btaa1030
16. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho YY, et al. Sherloc: A comprehensive refinement of the ACMG-AMP variant classification criteria. Genetics in Medicine. 2017;19. doi:10.1038/gim.2017.37
17. Park VS, Pursell ZF. POLE proofreading defects: Contributions to mutagenesis and cancer. DNA Repair. 2019. doi:10.1016/j.dnarep.2019.02.007
18. Xing X, Kane DP, Bulock CR, Moore EA, Sharma S, Chabes A, et al. A recurrent cancer-associated substitution in DNA polymerase ε produces a hyperactive enzyme. Nat Commun. 2019;10. doi:10.1038/s41467-018-08145-2
19. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;35: 4168–4169. doi:10.1093/bioinformatics/btz184
20. Blaabjerg LM, Kassem MM, Good LL, Jonsson N, Cagiada M, Johansson KE, et al. Rapid protein stability prediction using deep learning representations. Elife. 2023;12: e82593. doi:10.7554/eLife.82593
21. Abildgaard AB, Stein A, Nielsen S V, Schultz-Knudsen K, Papaleo E, Shrikhande A, et al. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. Elife. 2019;8: e49138. doi:10.7554/eLife.49138
22. Nielsen S V., Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, et al. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. Nichols KE, editor. PLoS Genet. 2017;13: e1006739. doi:10.1371/journal.pgen.1006739
23. Schulze TK, Lindorff-Larsen K. Effects of residue substitutions on the cellular abundance of proteins. 2024. doi:10.7554/eLife.103721.1
24. Høie MH, Cagiada M, Beck Frederiksen AH, Stein A, Lindorff-Larsen K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and

conservation. Cell Rep. 2022;38: 110207. doi:10.1016/J.CELREP.2021.110207/ATTACHMENT/EF288B6D-E8ED-4AB3-94AD-763EFB0AA108/MMC1.PDF
25. 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: 405–424. doi:10.1038/GIM.2015.30
26. COSMIC, Catalogue Of Somatic Mutations In Cancer, https://cancer.sanger.ac.uk.
27. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell. 2017;171. doi:10.1016/j.cell.2017.09.048
28. Zou Y, Liu FY, Liu H, Wang F, Li W, Huang MZ, et al. Frequent POLE1 p.S297F mutation in Chinese patients with ovarian endometrioid carcinoma. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. 2014;761. doi:10.1016/j.mrfmmm.2014.01.003
29. Herzog M, Alonso-Perez E, Salguero I, Warringer J, Adams DJ, Jackson SP, et al. Mutagenic mechanisms of cancer-associated DNA polymerase alleles. Nucleic Acids Res. 2021;49. doi:10.1093/nar/gkab160
30. Tiberti M, Terkelsen T, Degn K, Beltrame L, Canter Cremers T, da Piedade I, et al. MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. Brief Bioinform. 2022; 1–16. doi:10.1093/bib/bbac074
31. Roske JJ, Yeeles JTP. Structural basis for processive daughter-strand synthesis and proofreading by the human leading-strand DNA polymerase Pol ε. Nat Struct Mol Biol. 2024. doi:10.1038/s41594-024-01370-y
32. Labrousse G, Perre P Vande, Parra G, Jaffrelot M, Leroy L, Chibon F, et al. The hereditary N363K POLE exonuclease mutant extends PPAP tumor spectrum to glioblastomas by causing DNA damage and aneuploidy in addition to increased mismatch mutagenicity. NAR Cancer. 2023;5. doi:10.1093/narcan/zcad011
33. Barbari SR, Beach AK, Markgren JG, Parkash V, Moore EA, Johansson E, et al. Enhanced polymerase activity permits efficient synthesis by cancer-Associated DNA polymerase variants at low dNTP levels. Nucleic Acids Res. 2022;50. doi:10.1093/nar/gkac602
34. Galati MA, Hodel KP, Gams MS, Sudhaman S, Bridge T, Zahurancik WJ, et al. Cancers from novel pole-mutant mouse models provide insights into polymerase-mediated hypermutagenesis and immune checkpoint blockade. Cancer Res. 2020;80. doi:10.1158/0008-5472.CAN-20-0624
35. Hodel KP, Sun MJS, Ungerleider N, Park VS, Williams LG, Bauer DL, et al. POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. Mol Cell. 2020;78. doi:10.1016/j.molcel.2020.05.012
36. Barbari SR, Kane DP, Moore EA, Shcherbakova P V. Functional analysis of cancer-associated DNA polymerase ε variants in Saccharomyces cerevisiae. G3: Genes, Genomes, Genetics. 2018;8. doi:10.1534/g3.118.200042
37. Dahl JM, Thomas N, Tracy MA, Hearn BL, Perera L, Kennedy SR, et al. Probing the mechanisms of two exonuclease domain mutators of DNA polymerase ε. Nucleic Acids Res. 2022;50. doi:10.1093/nar/gkab1255

38. Kane DP, Shcherbakova P V. A common cancer-associated DNA polymerase ε mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from loss of proofreading. Cancer Res. 2014;74. doi:10.1158/0008-5472.CAN-13-2892
39. Lee M, Eng G, Barbari SR, Deshpande V, Shcherbakova P V., Gala MK. Homologous Recombination Repair Truncations Predict Hypermutation in Microsatellite Stable Colorectal and Endometrial Tumors. Clin Transl Gastroenterol. 2020;11. doi:10.14309/ctg.0000000000000149
40. Ostroverkhova D, Tyryshkin K, Beach AK, Moore EA, Masoudi-Sobhanzadeh Y, Barbari SR, et al. DNA polymerase ε and δ variants drive mutagenesis in polypurine tracts in human tumors. Cell Rep. 2024;43. doi:10.1016/j.celrep.2023.113655
41. Rohlin A, Zagoras T, Nilsson S, Lundstam U, Wahlström J, Hultén L, et al. A mutation in POLE predisposing to a multi-tumour phenotype. Int J Oncol. 2014;45: 77–81. doi:10.3892/ijo.2014.2410
42. Djursby M, Madsen MB, Frederiksen JH, Berchtold LA, Therkildsen C, Willemoe GL, et al. New Pathogenic Germline Variants in Very Early Onset and Familial Colorectal Cancer Patients. Front Genet. 2020;11. doi:10.3389/fgene.2020.566266
43. Hamzaoui N, Alarcon F, Leulliot N, Guimbaud R, Buecher B, Colas C, et al. Genetic, structural, and functional characterization of POLE polymerase proofreading variants allows cancer risk prediction. Genetics in Medicine. 2020;22. doi:10.1038/s41436-020-0828-z
44. Herr AJ, Ogawa M, Lawrence NA, Williams LN, Eggington JM, Singh M, et al. Mutator suppression and escape from replication error-induced extinction in yeast. PLoS Genet. 2011;7. doi:10.1371/journal.pgen.1002282
45. Williams LN, Herr AJ, Preston BD. Emergence of DNA polymerase ε antimutators that escape error-induced extinction in yeast. Genetics. 2013;193. doi:10.1534/genetics.112.146910
46. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res. 2014;24. doi:10.1101/gr.174789.114
47. Ma X, Riaz N, Samstein RM, Lee M, Makarov V, Valero C, et al. Functional landscapes of POLE and POLD1 mutations in checkpoint blockade-dependent antitumor immunity. Nat Genet. 2022;54. doi:10.1038/s41588-022-01108-w
48. Li HD, Cuevas I, Zhang M, Lu C, Alam MM, Fu YX, et al. Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load. Journal of Clinical Investigation. 2018;128. doi:10.1172/JCI122095
49. Li HD, Lu C, Zhang H, Hu Q, Zhang J, Cuevas IC, et al. A PoleP286R mouse model of endometrial cancer recapitulates high mutational burden and immunotherapy response. JCI Insight. 2020;5. doi:10.1172/JCI.INSIGHT.138829
50. Church JM. Polymerase proofreading-associated polyposis: A new, dominantly inherited syndrome of hereditary colorectal cancer predisposition. Dis Colon Rectum. 2014;57. doi:10.1097/DCR.0000000000000084
51. Vipat S, Gupta D, Jonchhe S, Anderspuk H, Rothenberg E, Moiseeva TN. The non-catalytic role of DNA polymerase epsilon in replication initiation in human cells. Nat Commun. 2022;13. doi:10.1038/s41467-022-34911-4
52. Kayser K, Degenhardt F, Holzapfel S, Horpaopan S, Peters S, Spier I, et al. Copy number variation analysis and targeted NGS in 77 families with suspected Lynch

syndrome reveals novel potential causative genes. Int J Cancer. 2018;143. doi:10.1002/ijc.31725
53. Murphy K, Darmawan H, Schultz A, Da Silva EF, Reha-Krantz LJ. A method to select for mutator DNA polymerase δs in Saccharomyces cerevisiae. Genome. 2006;49. doi:10.1139/G05-106
54. Foley MC, Couto L, Rauf S, Boyke A. Insights into DNA polymerase δ's mechanism for accurate DNA replication. J Mol Model. 2019;25. doi:10.1007/s00894-019-3957-z
55. Shcherbakova P V., Kunkel TA. Mutator Phenotypes Conferred by MLH1 Overexpression and by Heterozygosity for mlh1 Mutations . Mol Cell Biol. 1999;19. doi:10.1128/mcb.19.4.3177
56. Tran HT, Keen JD, Kricker M, Resnick MA, Gordenin DA. Hypermutability of Homonucleotide Runs in Mismatch Repair and DNA Polymerase Proofreading Yeast Mutants. Mol Cell Biol. 1997;17. doi:10.1128/mcb.17.5.2859
57. Swaney DL, Beltrao P, Starita L, Guo A, Rush J, Fields S, et al. Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. Nat Methods. 2013;10: 676–682. doi:10.1038/nmeth.2519
58. VerPlank JJS, Goldberg AL. Regulating protein breakdown through proteasome phosphorylation. Biochemical Journal. 2017. doi:10.1042/BCJ20160809
59. Rigbolt KTG, Prokhorova TA, Akimov V, Henningsen J, Johansen PT, Kratchmarova I, et al. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. Sci Signal. 2011;4. doi:10.1126/scisignal.2001570
60. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294: 1351–1362. doi:10.1006/jmbi.1999.3310
61. Lee MYWT, Zhang S, Lin SHS, Wang X, Darzynkiewicz Z, Zhang Z, et al. The tail that wags the dog: P12, the smallest subunit of DNA polymerase δ, is degraded by ubiquitin ligases in response to DNA damage and during cell cycle progression. Cell Cycle. 2014;13. doi:10.4161/cc.27407
62. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2: 401–404. doi:10.1158/2159-8290.CD-12-0095
63. Tiberti M, Di Leo L, Vistesen MV, Kuhre RS, Cecconi F, De Zio D, et al. The Cancermuts software package for the prioritization of missense cancer variants: a case study of AMBRA1 in melanoma. Cell Death Dis. 2022;13: 872. doi:10.1038/s41419-022-05318-2
64. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6: pl1. doi:10.1126/scisignal.2004088
65. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42. doi:10.1093/NAR/GKT1113
66. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589. doi:10.1038/S41586-021-03819-2
67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. doi:10.1093/NAR/28.1.235

68. Baranovskiy AG, Gu J, Babayeva ND, Kurinov I, Pavlov YI, Tahirov TH. Crystal structure of the human Polε B-subunit in complex with the C-terminal domain of the catalytic subunit. Journal of Biological Chemistry. 2017;292: 15717–15730. doi:10.1074/jbc.M117.792705
69. Lancey C, Tehseen M, Raducanu VS, Rashid F, Merino N, Ragan TJ, et al. Structure of the processive human Pol δ holoenzyme. Nat Commun. 2020;11. doi:10.1038/s41467-020-14898-6
70. Jozwiakowski SK, Kummer S, Gari K. Human DNA polymerase delta requires an iron–sulfur cluster for high-fidelity DNA synthesis. Life Sci Alliance. 2019;2. doi:10.26508/lsa.201900321
71. Calderone A, Castagnoli L, Cesareni G. Mentha: A resource for browsing integrated protein-interaction networks. Nat Methods. 2013;10: 690–691. doi:10.1038/nmeth.2561
72. Jenkyn-Bedford M, Jones ML, Baris Y, Labib KPM, Cannone G, Yeeles JTP, et al. A conserved mechanism for regulating replisome disassembly in eukaryotes. Nature. 2021;600: 743–747. doi:10.1038/s41586-021-04145-3
73. He Q, Wang F, Yao NY, O'Donnell ME, Li H. Structures of the human leading strand Polε–PCNA holoenzyme. Nature Communications . 2024;15. doi:10.1038/s41467-024-52257-x
74. Zhu W, Shenoy A, Kundrotas P, Elofsson A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. Cowen L, editor. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad424
75. Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, et al. Towards a structurally resolved human protein interaction network. Nat Struct Mol Biol. 2023;30. doi:10.1038/s41594-022-00910-8
76. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: Working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019; 1–2. doi:10.1093/bioinformatics/btz184
77. Pasi M, Tiberti M, Arrigoni A, Papaleo E. xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. J Chem Inf Model. 2012;279: 1–6. doi:10.1021/ci300213c
78. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;2: 19–25. doi:10.1016/j.softx.2015.06.001
79. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017;14: 71–73. doi:10.1038/nmeth.4067
80. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 1997;18: 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H
81. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys. 1995;103: 8577–8593. doi:10.1063/1.470117
82. Darden T, York D, Pedersen L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. J Chem Phys. 1993;98: 10089–10092. doi:10.1063/1.464397
83. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984;81: 3684–3690. doi:10.1063/1.448118

84. Nosé S, Klein ML. Constant pressure molecular dynamics for molecular systems. Mol Phys. 1983;50: 1055–1076. doi:10.1080/00268978300102851
85. Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. Journal of Physical Chemistry B. 2018;122: 5389–5399. doi:10.1021/acs.jpcb.7b11367
86. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017;13: 3031–3048. doi:10.1021/acs.jctc.7b00125
87. Raimondi D, Orlando G, Pancsa R, Khan T, Vranken WF. Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. Sci Rep. 2017;7: 8826. doi:10.1038/s41598-017-08366-3
88. Sora V, Otamendi Laspiur A, Degn K, Arnaudi M, Utichi M, Beltrame L, et al. RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. Protein Science. 2023;32: e4527. doi:10.1002/pro.4527
89. Shapovalov M V., Dunbrack RL. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure. 2011;19: 844–858. doi:10.1016/j.str.2011.03.019
90. Song Y, Tyka M, Leaver-Fay A, Thompson J, Baker D. Structure-guided forcefield optimization. Proteins: Structure, Function and Bioinformatics. 2011;79: 1898–1909. doi:10.1002/prot.23013
91. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, Dimaio F, et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput. 2015;11: 609–622. doi:10.1021/ct500864r
92. Hubbard SJ, Thornton JM,. NACCESS. 1993. Available: Department of Biochemistry and Molecular Biology, University College London
93. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020. doi:10.1093/nar/gkz1030
94. Tan ZW, Guarnera E, Tee WV, Berezovsky IN. AlloSigMA 2: Paving the way to designing allosteric effectors and to exploring allosteric effects of mutations. Nucleic Acids Res. 2020;48: W116–W124. doi:10.1093/NAR/GKAA338
95. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 1–11. doi:10.1186/1471-2105-10-168/TABLES/1
96. Bhattacharyya M, Ghosh S, Vishveshwara S. Protein Structure and Function: Looking through the Network of Side-Chain Interactions. Curr Protein Pept Sci. 2016;17: 4–25. doi:10.2174/1389203716666150923105727
97. Tiberti M, Invernizzi G, Lambrughi M, Inbar Y, Schreiber G, Papaleo E. PyInteraph: A framework for the analysis of interaction networks in structural ensembles of proteins. J Chem Inf Model. 2014;54: 1537–1551. doi:10.1021/ci400639r
98. Sora V, Tiberti M, Robbani SM, Rubin J, Papaleo E. PyInteraph2 and PyInKnife2 to analyze networks in protein structural ensembles. bioRxiv. 2020. doi:10.1101/2020.11.22.381616
99. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math (Heidelb). 1959;1: 269–271. doi:10.1007/BF01386390

100. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. 7th Python in Science Conference (SciPy 2008). 2008; 11–15. doi:http://conference.scipy.org/proceedings/SciPy2008/paper_2
101. Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J Mol Biol. 2017;429: 365–371. doi:10.1016/j.jmb.2016.12.004
102. Richard J, August S. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Science (1979). 1994.
103. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem. 1996;17. doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P
104. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176. doi:10.1016/j.cell.2018.12.015
105. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. 2022;23. doi:10.1186/s13059-022-02664-4