Clinical & Molecular Biomedicine

An International Journal of Translational Medicine and Molecular Research

www.editorypress.uz | Volume 1, Issue 1 | June 2026

Clinical & Molecular Biomedicine

An International Journal of Translational Medicine and Molecular Research

Volume 1 • Issue 1 • June 2026

Editory Press

https://www.editorypress.uz/find-a-journal/clinical-molecular-biomedicine

Editorial Board

Editorial Board

Publisher: Editory Press **Address:** Toshkent shahri, Mirzo Ulug'bek tumani, Ulug'bek shaharchasi Oqibat MFY, Mustaqillik shoh ko'chasi, 7-uy 2.
https://www.editorypress.uz/find-a-journal/clinical-molecular-biomedicine

Contents

Continuous article pagination: 1–55.

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 1–8, 2026
Open Access © 2026 Editory Press

Computational and Biophysical Characterization of Limonene as a Potential Natural Inhibitor of CDK6 for Therapeutic Targeting of Cancer

Zulfareen[1,*], **Sumra**[2] **and Shagufta Jahan**[1]

[1]*Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India*
[2]*Department of Biotechnology, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India*
[*]*Correspondence: zulfareengaur@gmail.com*

Abstract

Cyclin-dependent kinase 6 (CDK6) plays a central role in G1–S phase cell cycle progression and is frequently dysregulated in various cancers, making it an established therapeutic target. Although selective CDK4/6 inhibitors are clinically available, exploration of natural compounds targeting CDK6 remains limited. The present study aimed to investigate the binding mechanism and inhibitory potential of limonene against CDK6 using integrated computational and experimental approaches. Recombinant CDK6 was cloned, expressed, and purified, followed by molecular docking, fluorescence spectroscopy, and kinase inhibition assay. Docking analysis revealed a binding free energy of -6.3 kcal mol^{-1} with a calculated pK_i of 4.62 and ligand efficiency of 0.63 kcal mol^{-1} per non-hydrogen atom. Fluorescence quenching studies demonstrated strong binding affinity ($K = 4.1\times10^7$ M^{-1}), while enzymatic assays confirmed dose-dependent suppression of CDK6 activity. Collectively, these findings indicate that limonene directly interacts with and functionally inhibits CDK6, highlighting its potential as a natural scaffold for the development of CDK6-targeted anticancer therapeutics.

Keywords: *Cyclin-dependent kinase 6, Limonene, Molecular docking, Fluorescence spectroscopy, Kinase inhibition assay, Protein–ligand interaction, Cell cycle regulation, Anticancer therapy*

Received: March 10th, 2025 **Accepted:** July 20th, 2025 **Published:** January 5th, 2026

1. INTRODUCTION

Cancer is characterised by uncontrolled cell proliferation, driven largely by dysregulation of the cell cycle. The cell cycle comprises a series of phases, and key regulators include cyclins, cyclin-dependent kinases (CDKs), and CDK inhibitors (CKIs), which work together to control progression through these phases. The disruption of these components of the cell cycle is a core hallmark of cancer, enabling uncontrolled cell proliferation and tumorigenesis (Hanahan & Weinberg, 2011; **?**).

Cyclin-dependent kinases (CDKs) are serine/threonine protein kinases that are involved in the cell cycle, transcription, and other biological processes like translation, neurogenesis, and apoptosis (Malumbres, 2014). Among CDK family members, cyclin-dependent kinase-6 (CDK6) and CDK4 play important roles in mammalian cell proliferation by driving cells into the DNA synthetic (S) phase of the cell division cycle (Sherr et al., 2016). CDK4 and CDK6 associate with D-type cyclins (D1, D2, and D3), which phosphorylate retinoblastoma (RB), thereby releasing E2F transcription factors. E2F-dependent gene activation allows for G1 to S phase progression and DNA synthesis (Classon & Harlow, 2002), helping cells progress through the early G1 phase of the cell cycle (Nebenfuehr et al., 2020). Beyond its function in cell cycle control, CDK6 contributes to cell differentiation, tissue development, and hematopoietic lineage commitment (Tigan et al., 2016). Emerging evidence also suggests that CDK6 modulates anti-tumour immunity by regulating immune cell proliferation and immune checkpoint molecule expression (Petroni et al., 2020). Figure 1 illustrates the transcriptional, post-transcriptional, and post-translational regulation of Cyclin D–CDK4/6.

CDK6 and CDK4 form larger protein complexes with heat shock protein 90 (Hsp90) and its co-chaperone cell division cycle 37 (Cdc37), which secure correct protein folding, assembly, maturation, and stability of a large number of proteins (Hallett et al., 2017). Hyperactivation of CDK6 has been linked to various malignancies, including breast cancer, leukaemia, lymphoma, and glioblastoma. Experimental disruption of CDK6 prolongs the G1 phase and inhibits S-phase entry. G1-phase prolongation causes premature differentiation in neuroepithelial cells, which explains reduced proliferation after CDK6 loss (O'Sullivan, 2026).

Structurally, CDK6 is a 326-amino-acid protein (~36.5 kDa) encoded on chromosome 7 (Russo et al., 1998). It comprises several key domains: the N-terminal lobe (residues 1–98) contains a regulatory cyclin-binding domain, while the C-terminal lobe forms the catalytic core (Sielecki et al., 2000). The activation (T-) loop (residues 156–172) within the catalytic domain regulates kinase activity through phosphorylation, while additional regulatory motifs, including a C-terminal PEST sequence, contribute to protein stability (Li et al., 2015).

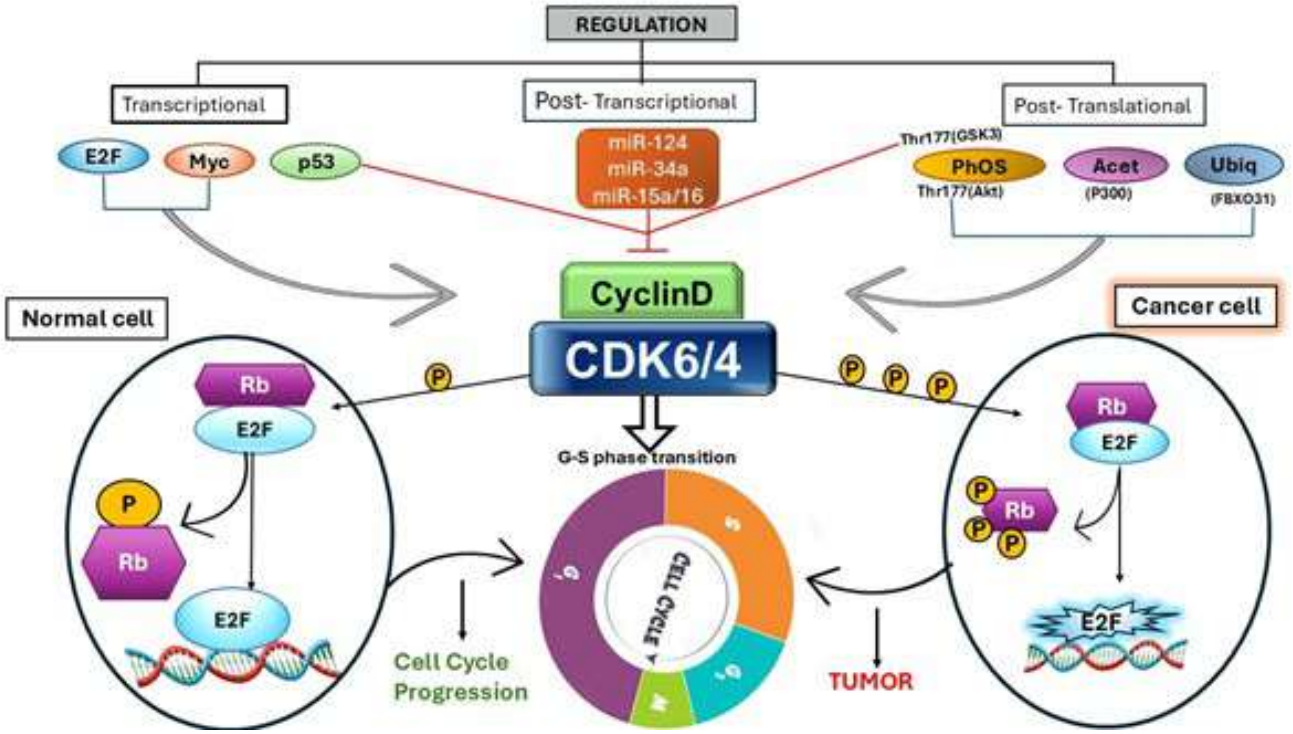

Figure 1. The schematic illustration of transcriptional, post-transcriptional, and post-translational regulation of Cyclin D–CDK4/6. In normal cells, controlled CDK4/6 activation leads to the regulated phosphorylation of the retinoblastoma protein (Rb), releasing E2F to permit orderly G1–S progression. In contrast, hyperactivation of CDK4/6 in cancer cells leads to excessive Rb phosphorylation, sustained E2F hyperactivation, and uncontrolled cell cycle progression, contributing to tumor development.

Activation of CDK6 requires interaction with cyclin D and phosphorylation by CDK-activating kinase (CAK), whereas its activity is negatively controlled by INK4 and Cip/Kip family inhibitors. Beyond phosphorylation, CDK6 is finely regulated at transcriptional, post-transcriptional (e.g., microRNAs such as miR-124 and miR-34a), and post-translational levels, including phosphorylation, acetylation, and ubiquitination (Bachs et al., 2018; Garzon et al., 2010).

Due to its primary function in tumour growth and survival, CDK6 has been recognised as a validated therapeutic target (Yousuf et al., 2022). Selective CDK4/6 inhibitors, Palbociclib, Ribociclib, and Abemaciclib, have shown considerable clinical benefit and have been approved by regulators for hormone receptor-positive, HER2-negative metastatic breast cancer (Rencuzogullari et al., 2020; Shah et al., 2018). These substances competitively bind the ATP-binding pocket of CDK4/6, inhibit Rb phosphorylation, and induce G1 cell cycle arrest. Combination strategies integrating CDK6 inhibitors with endocrine therapy, chemotherapy, or immunotherapy have further enhanced therapeutic efficacy and addressed resistance mechanisms. CDK6 disruption prolonged the G1 phase and inhibited S-phase entry in these cells. G1-phase lengthening has been shown to induce premature differentiation in neuroepithelial cells, explaining the decreased proliferation associated with CDK6 disruption (Pavani et al., 2016; Spring et al., 2020).

Natural products have also gained interest as potential modulators of CDK6 activity (Figure 2). Bioactive phytochemicals that modulate cell cycle and apoptosis-related pathways, such as flavonoids (e.g., quercetin, fisetin), resveratrol, curcumin, sulforaphane, and monoterpenes like limonene, exhibit anti-proliferative properties (Srivastava & Saxena, 2025). These compounds affect cancer-associated signaling networks, including RB-E2F, PI3K/Akt/mTOR, NF-κB, and MAPK pathways, and may contribute to CDK6 inhibition either directly or indirectly (Kim et al., 2013; de Figueiredo et al., 2015; Sa & Das, 2008; Spring et al., 2020).

Limonene, a naturally occurring monocyclic monoterpene abundant in citrus oils, promotes the GST (Glutathione S-transferase) system in the liver and small bowel, which helps in eliminating carcinogens (Sun, 2007). Limonene 1,2-diol has been reported to inhibit tumor growth through inhibition of p21-dependent signaling, induce apoptosis via the induction of the transforming growth factor beta-signaling pathway, inhibit post-translational modification of signal transduction proteins, result in G1 cell cycle arrest, and modulate genes associated with cell cycle progression and apoptosis (Sun, 2007).

Collectively, CDK6 occupies a central position at the intersection of cell cycle control, oncogenic signaling, and immune regulation. Its structural features, multilayered regulatory mechanisms, and proven druggability highlight its significance as both a mechanistic driver of tumorigenesis and a compelling therapeutic target. Continued exploration of CDK6 biology and inhibition strategies, synthetic and natural, holds substantial promise for advancing precision oncology.

2. MATERIALS AND METHODS

2.1. Chemicals and reagents

The expression vector pET-28a(+) and *Escherichia coli* BL21 (DE3) competent cells were procured from Qiagen, while *E. coli* DH5α cells were obtained from Invitrogen. Analytical-grade reagents,

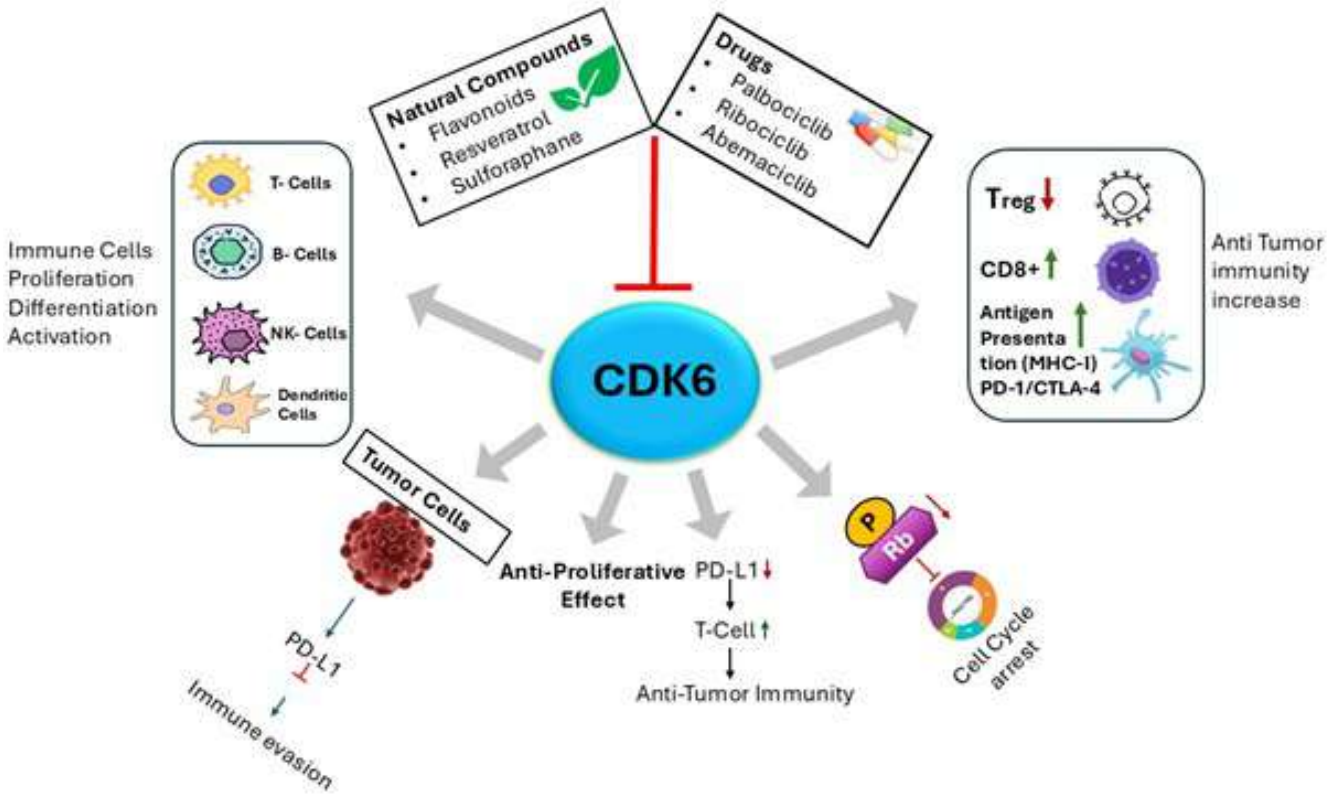

Figure 2. The diagram illustrates the central role of CDK6 in coordinating tumor cell proliferation and immune regulation. CDK6 promotes cell cycle progression by phosphorylating Rb, facilitating tumor growth and immune evasion via PD-L1 regulation. Inhibition of CDK6 by approved inhibitors or bioactive natural compounds induces cell-cycle arrest, reduces PD-L1 expression, and suppresses tumor growth. In parallel, CDK6 modulation influences immune cell proliferation, differentiation, and activation, including those of T cells, B cells, NK cells, and dendritic cells. Therapeutic targeting of CDK6 decreases regulatory T cells (Tregs) while enhancing $CD8^+$ T-cell responses and antigen presentation (MHC-I), thereby strengthening anti-tumor immunity.

including sodium chloride (NaCl), ethylenediaminetetraacetic acid (EDTA), and other routine chemicals, were purchased from Merck (India). Luria–Bertani (LB) broth was obtained from Merck (Darmstadt, Germany). Kanamycin and isopropyl-β-D-thiogalactopyranoside (IPTG) were sourced from Sigma-Aldrich (St. Louis, MO, USA). Nickel–nitrilotriacetic acid (Ni–NTA) affinity chromatography columns and Ni–NTA agarose beads were acquired from Bio-Rad and Qiagen (QIAexpress), respectively. Limonene used in this study was purchased from Sigma-Aldrich (USA).

2.2. Molecular docking

Molecular docking was performed to investigate the binding interactions between Conformer3D_COMPOUND_CID_22311 (limonene) and the crystal structure of CDK6 (PDB ID: 3NUP). Docking simulations were conducted using InstaDock, an automated platform for structure-based virtual screening (Mohammad et al., 2021). Binding affinity calculations were carried out using QuickVina-W (Hassan et al., 2017), a modified version of AutoDock Vina (Trott & Olson, 2010) that integrates empirical and knowledge-based scoring functions. A blind docking approach was employed to allow unbiased exploration of potential ligand-binding sites across the protein surface. The predicted binding free energy (ΔG, kcal mol^{-1}) obtained from docking was used to calculate the inhibitory constant ($K_{i,\mathrm{pred}}$) and pK_i values according to the following thermodynamic relationships:

$$\Delta G = RT \ln K_{i,\mathrm{pred}} \tag{1}$$

$$K_{i,\mathrm{pred}} = e^{(\Delta G/RT)} \tag{2}$$

$$pK_i = -\log K_{i,\mathrm{pred}} \tag{3}$$

where ΔG represents the docking-derived binding free energy, R is the universal gas constant (1.98 cal mol^{-1} K^{-1}), T is the temperature (298.15 K), and $K_{i,\mathrm{pred}}$ is the predicted inhibition constant (Shityakov & Foerster, 2014).

To further evaluate ligand binding efficiency, ligand efficiency (LE) was calculated using:

$$LE = -\frac{\Delta G}{N} \tag{4}$$

where LE denotes ligand efficiency (kcal mol^{-1} per non-hydrogen atom), ΔG is the binding free energy, and N corresponds to the number of non-hydrogen atoms in the ligand. This parameter provides a normalized assessment of binding strength relative to molecular size (Hopkins et al., 2004).

2.3. Plasmid isolation

The overnight bacterial culture was pelleted by centrifugation at 6000 rpm for 5 minutes at room temperature, then resuspended in the resuspension buffer. Lysis was performed to release cellular contents, including plasmid DNA, using the lysis buffer. Proteins

were then digested using proteinase K. Purification of the plasmid was done using a phenol-chloroform extraction kit. Finally, the concentration and purity of plasmid DNA were checked using Nanodrop at 260 nm. Purified bands of isolated plasmid were also confirmed using agarose gel electrophoresis (Ausubel et al., 1992).

2.4. Agarose gel electrophoresis

A 0.8% agarose gel was prepared by adding TAE buffer and heating in a microwave. EtBr was added to the cooled gel solution, and then the cooled gel was poured into the gel tray. Gel was allowed to solidify for about 20 min. Samples were prepared by adding loading dye (bromophenol blue) to the isolated plasmid DNA, and loading the samples into the wells after removing the comb. The tank was filled with TAE buffer, and the leads were connected to the power supply. After applying a constant voltage of 100–150 V for about 45–60 minutes, the gel was removed from the gel box, placed on a UV transilluminator, and visualized using a gel documentation system.

2.5. Competent cell preparation and transformation

A single colony was inoculated into LB broth containing kanamycin and incubated overnight at 37 °C, then transferred to a larger volume of fresh LB broth and incubated until the OD reached 0.4–0.6 at 600 nm. Metabolic activity was reduced by cooling the culture on ice for 15–30 minutes. Culture was centrifuged, and the supernatant was discarded. Pellet was resuspended in ice-cold 10% glycerol solution. Competent cells formed were aliquoted in microcentrifuge tubes and stored at −80 °C (Russell & Sambrook, 2001).

The recombinant plasmid construct pET-28a(+) harboring the CDK6 gene was transformed into *Escherichia coli* BL21 (DE3) competent cells using the heat-shock method. Briefly, chemically competent BL21 (DE3) cells were thawed on ice, and 50–100 ng of plasmid DNA was gently mixed with the cells. The mixture was incubated on ice for 20 min to facilitate DNA adsorption. Subsequently, cells were subjected to a heat shock at 42 °C for 90 s, followed by immediate ice incubation for 5 min to stabilize the membranes. Thereafter, 900 μL of Luria–Bertani (LB) broth was added, and the cells were allowed to recover at 37 °C with shaking at 220 rpm for 1 h. The transformed cells were then spread onto LB agar plates supplemented with kanamycin and incubated overnight (12–16 h) at 37 °C to allow colony formation (Yousuf et al., 2022).

2.6. Expression and purification

The full-length CDK6 gene (981 nucleotides) was cloned into the pET-28a(+) expression vector and sequence-verified prior to protein expression. The recombinant plasmid was transformed into *Escherichia coli* BL21 (DE3) cells for heterologous protein production. Transformed cells were cultured in LB medium containing kanamycin at 37 °C until the desired optical density was reached, then induced with 0.25 mM isopropyl-β-D-thiogalactopyranoside (IPTG). After induction, cells were harvested by centrifugation and resuspended in lysis buffer (25 mM Tris–HCl, pH 8.0, 200 mM NaCl, 1 mM DTT, and 10 mM PMSF). Cell disruption was performed by sonication on ice, and the lysate was centrifuged at 9,000 rpm for 20 min at 4 °C to separate soluble and insoluble fractions. The pellet containing inclusion bodies was collected and washed three times with Milli-Q water to remove impurities.

The inclusion bodies were subsequently solubilized in solubilization buffer (25 mM Tris–HCl, 200 mM NaCl, and 0.5% sarcosine) and loaded onto a pre-equilibrated Ni–NTA affinity chromatography column for purification. After binding, the column was washed to remove non-specifically bound proteins, and recombinant CDK6 was eluted using elution buffer containing 150 mM imidazole (25 mM Tris–HCl, pH 8.0, 200 mM NaCl). Protein purity and molecular weight were analyzed using 12% SDS–PAGE (Yousuf et al., 2020).

2.7. Kinase inhibition assay

The inhibitory effect of limonene on recombinant CDK6 was evaluated using a malachite green–based ATPase assay. The reaction mixture (final volume: 50 μL) contained purified CDK6 (1 μM) and freshly prepared ATP (200 μM). Increasing concentrations of limonene were added to assess dose-dependent inhibition. The reaction mixtures were incubated at 37 °C for 45 min to allow ATP hydrolysis. The reaction was terminated by adding 100 μL of malachite green reagent, followed by incubation at room temperature for 15–20 min to enable color development resulting from inorganic phosphate release. Absorbance was measured at 600 nm using a microplate ELISA reader in a 96-well format. Enzyme activity was calculated based on phosphate release, and percentage inhibition was determined relative to control reactions lacking inhibitor (Voura et al., 2019).

2.8. Fluorescence measurement

The interaction between limonene and purified recombinant CDK6 was investigated using intrinsic fluorescence spectroscopy by monitoring changes in the emission spectrum of CDK6 upon ligand titration (Jameel et al., 2017). Fluorescence measurements were performed using a JASCO spectrofluorometer (Model FP-6200). Titration experiments were conducted by progressively increasing the limonene concentration, and each measurement was performed in triplicate to ensure reproducibility. Changes in fluorescence intensity were recorded and analyzed to determine binding parameters. The binding constant (Ka), number of binding sites (n), and associated thermodynamic parameters were calculated from fluorescence quenching data using the Stern–Volmer equation. Analysis of the quenching behavior enabled characterization of the binding affinity and interaction mechanism between limonene and CDK6 (Shamsi et al., 2020).

3. RESULTS

3.1. Molecular docking

Docking analysis of Conformer3D_COMPOUND_CID_22311 (limonene) against the CDK6 crystal structure (PDB ID: 3NUP) revealed a predicted binding free energy (ΔG) of −6.3 kcal mol^{-1}. The corresponding calculated pK_i value was 4.62, indicating moderate binding affinity toward the target protein. The ligand efficiency (LE) was determined to be 0.63 kcal mol^{-1} per non-hydrogen atom, suggesting favorable binding energy relative to molecular size (Figure 3).

Comprehensive interaction profiling of all generated docked conformations was performed to examine the binding orientation and molecular interactions within the 3NUP binding pocket (Figure 4). Analysis of the best-ranked pose revealed that the ligand occupies the active-site cavity and establishes stabilizing interactions with key amino acid residues, thereby contributing to its predicted binding affinity.

3.2. Protein expression and purification

Plasmid isolated using the QIAprep® Spin Miniprep Kit was checked for purity and concentration using a Nanodrop, which showed a purity of about 1.8 (A260/280), and digested bands were

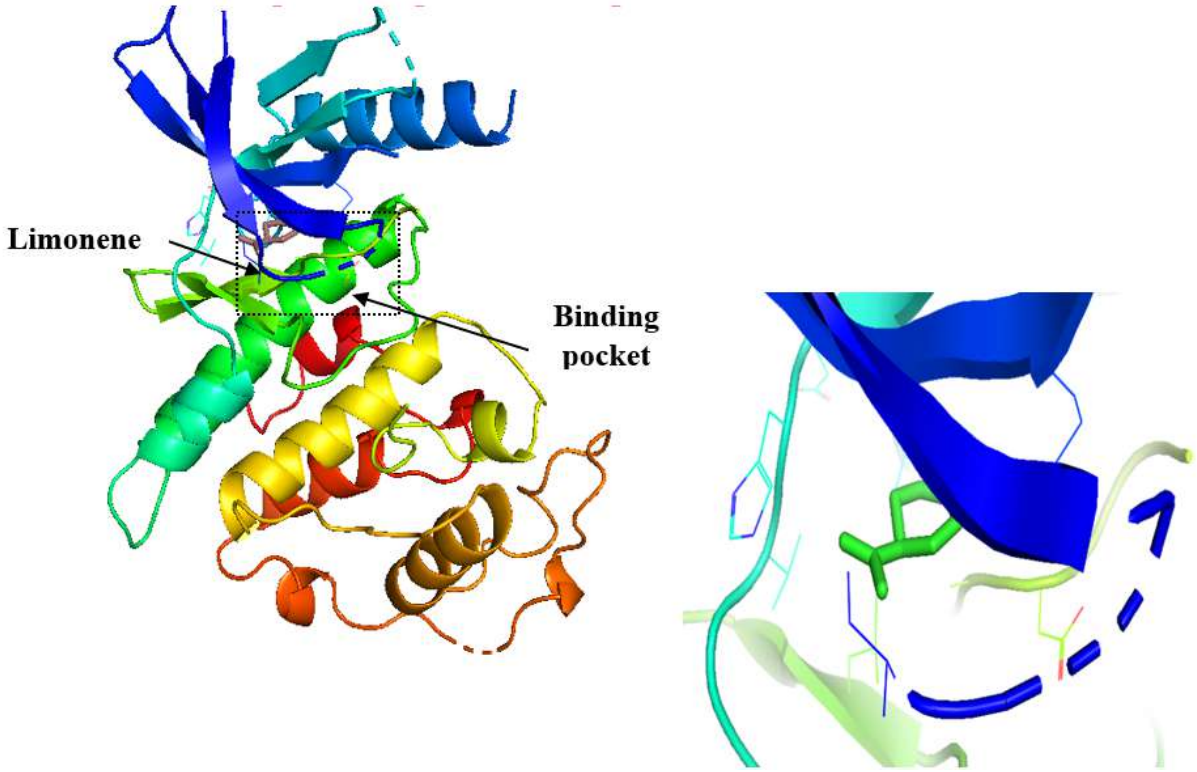

Figure 3. Representation of the binding mode of limonene to CDK6. (A) Cartoon representation of CDK6 in complex with docked limonene. (B) Zoomed interaction of limonene with the active-site pocket residues of CDK6.

visualized by running an agarose gel, which showed clear bands for nicked, linear, and supercoiled forms of DNA (Figure 5). The confirmed pET28a+ plasmid was successfully transformed into *E. coli* strain BL21 (DE3). Colonies appeared after incubation with kanamycin at 50 μg/ml for 12–16 hours at 37 °C (Figure 5). Recombinant CDK6 protein was expressed at 37 °C by inducing with 0.25 mM IPTG. Purified CDK6 protein was obtained by Ni–NTA affinity chromatography, confirmed by running SDS–PAGE (Figure 5).

3.3. Fluorescence binding studies

Intrinsic fluorescence spectroscopy was employed to evaluate the interaction between limonene and recombinant CDK6. Progressive addition of limonene led to a marked decrease in the intrinsic fluorescence intensity of CDK6, indicating effective quenching of the protein's fluorescence (Figure 6). This observation suggests that ligand binding alters the local microenvironment surrounding aromatic fluorophores (primarily tryptophan residues), thereby affecting emission characteristics. The fluorescence quenching data were analyzed using both the Stern–Volmer and modified Stern–Volmer equations to determine the quenching mechanism and binding parameters. The Stern–Volmer plot demonstrated a concentration-dependent quenching profile, supporting the formation of a protein–ligand complex.

Binding parameters derived from the modified Stern–Volmer plot revealed a binding constant (Ka) of 4.1×10^6 M^{-1}, indicative of strong affinity between limonene and CDK6. The slope of the modified plot further provided the number of binding sites (n), suggesting a defined binding interaction. The high binding constant obtained from fluorescence analysis is consistent with the molecular docking results, collectively supporting a stable and favorable interaction between limonene and CDK6.

3.4. Kinase inhibition assay

The inhibitory effect of limonene on recombinant CDK6 kinase activity was evaluated in a dose-dependent manner. Increasing concentrations of limonene resulted in a progressive reduction in CDK6 enzymatic activity. The observed decrease in kinase activity demonstrates that limonene effectively suppresses CDK6 function, supporting its potential role as a kinase inhibitor. These findings are consistent with both the molecular docking predictions and fluorescence binding studies, collectively indicating a stable interaction that translates into functional inhibition of the enzyme.

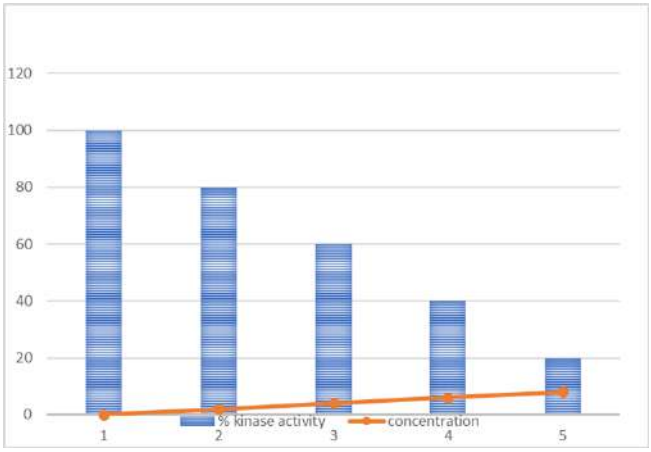

Figure 7. Graph representing enzyme activity versus ligand (limonene) concentration in the kinase inhibition assay.

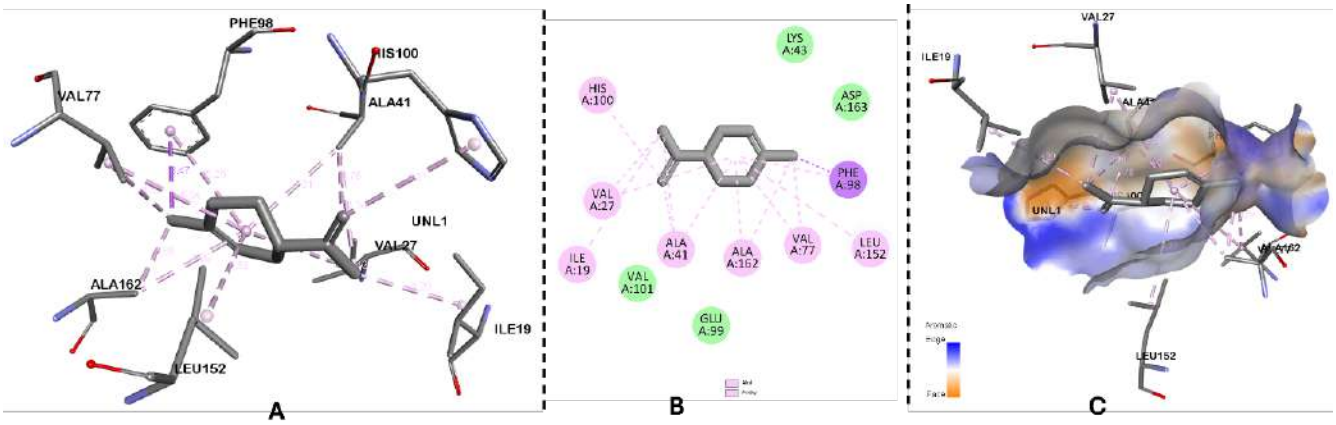

Figure 4. Molecular docking analysis of limonene with CDK6 active site. (A) Three-dimensional binding pose of limonene in the CDK6 catalytic pocket, showing key residues involved in hydrophobic stabilization. (B) Two-dimensional map showing hydrogen bonding and π–alkyl/alkyl interactions between limonene and active-site amino acids. (C) Surface representation of the CDK6 binding pocket illustrating favourable shape complementarity and hydrophobic interactions.

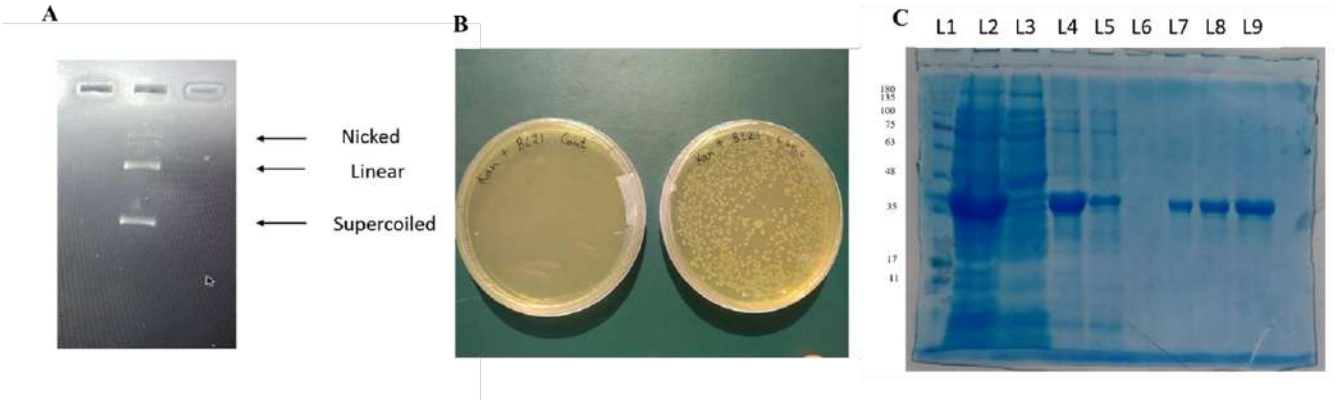

Figure 5. (A) Restriction digestion of the isolated plasmid, lane 1 showing the nicked, linear, and supercoiled forms of DNA. (B) Transformed *E. coli* BL21 colonies on an LB agar plate. (C) Purified CDK6 protein, eluted with 150 mM imidazole-containing elution buffer (lanes 7–9).

4. DISCUSSION

CDKs control cell division, which is especially important in cancer. CDK4 and CDK6 are activated by D-cyclins. As a result, CDK4 and CDK6 regulate the G1-to-S transition in the cell cycle. CDK6 plays an important part in cell-cycle regulation. CDK4/6 inhibitors suppress Rb phosphorylation and arrest the G1 phase of the cell cycle, thereby reducing tumour progression. In many cancers, the CDK4/6–INK4–Rb signalling pathway is dysregulated. CDK4/6 has emerged as a possible therapeutic target (Fontanella et al., 2022). These CDK4/6 inhibitors may become beneficial for tumour therapy after two decades of scientific development. Cyclin-dependent kinases 4 and 6 (CDK4/6) play a central role in regulating cell-cycle progression through the G1–S phase transition, making them critical drivers of uncontrolled proliferation in cancer. Pharmacological inhibition of CDK4/6 has demonstrated potent anti-proliferative effects by inducing cell-cycle arrest and promoting senescence in tumor cells. Consequently, CDK4/6 has emerged as a validated therapeutic target across multiple malignancies (Zhu & Zhu, 2023).

Several selective CDK4/6 inhibitors have received clinical approval for the treatment of hormone receptor–positive breast cancer, demonstrating favorable safety profiles and significant therapeutic benefit. Notably, agents such as Palbociclib, Ribociclib, and Abemaciclib have substantially improved progression-free survival when administered in combination with endocrine therapy. These findings underscore the clinical value of combination strategies, as CDK4/6 inhibition enhances the efficacy of other therapeutic agents and may help overcome resistance mechanisms (Illia et al., 2026).

Despite these advances, variability in patient response highlights the need for predictive biomarkers to identify individuals most likely to benefit from CDK4/6-targeted therapies. Furthermore, expanding the application of CDK4/6 inhibitors to additional tumor types requires comprehensive preclinical validation and well-designed large-scale clinical trials. Continued research focusing on combination regimens, resistance pathways, and molecular determinants of response will be essential to fully exploit the therapeutic potential of CDK4/6 inhibition in cancer management (Sun et al., 2025).

5. CONCLUSION

Elevated CDK6 expression has been strongly correlated with the progression and poor prognosis of multiple cancer types, underscoring its importance as a therapeutic target for the development of selective small-molecule inhibitors.

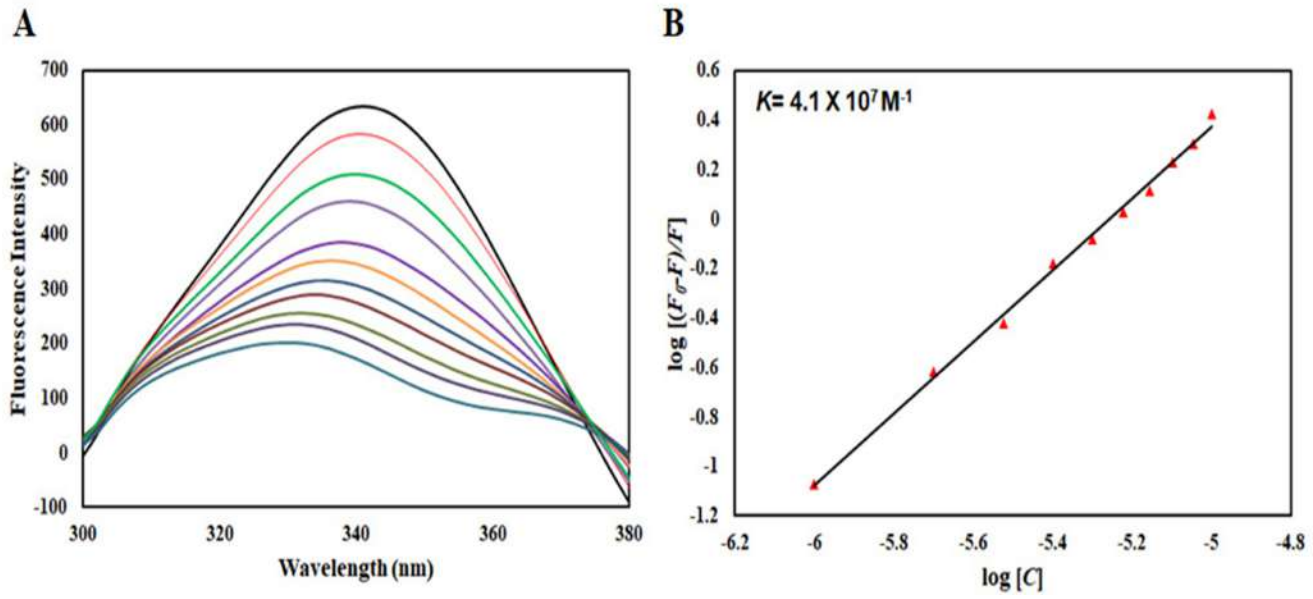

Figure 6. Fluorescence binding analysis of limonene with CDK6. (A) Intrinsic fluorescence emission spectra of CDK6 recorded in the presence of increasing concentrations of limonene. The excitation wavelength (λ_{ex}) was fixed at 280 nm, and emission spectra were collected over the range of 300–400 nm. A progressive decrease in fluorescence intensity was observed upon ligand titration, indicating quenching of CDK6 fluorescence. (B) Modified Stern–Volmer plot derived from fluorescence quenching data of CDK6 in the presence of limonene. The linear fitting of the plot was used to determine the binding constant (Ka) and the number of binding sites (n), reflecting the strength and stoichiometry of the protein–ligand interaction.

Natural compounds continue to represent a valuable reservoir of structurally diverse bioactive molecules with potential applications in oncology. In the present study, limonene was identified as a promising CDK6 inhibitor through integrated *in silico*, biophysical, and enzymatic analyses. The observed binding affinity and functional suppression of CDK6 activity suggest that limonene may contribute to its anticancer effects, at least in part, by modulating cell-cycle regulatory pathways. This mechanistic insight provides a novel perspective on the anticancer potential of limonene. Importantly, the relatively simple chemical scaffold of limonene offers opportunities for rational structural modification to enhance potency, selectivity, and drug-like properties. Structure-guided optimization strategies may facilitate the development of limonene-derived analogues as preclinical leads targeting CDK6. Further cellular and *in vivo* investigations are warranted to validate its therapeutic potential and to advance translational applications.

6. Funding

This work received no funding.

7. Acknowledgments

Zulfareen acknowledges Jamia Millia Islamia for providing the Non-NET fellowship.

8. Conflicts of Interest

The authors declare no conflicts of interest.

9. Data Availability Statement

All data generated or analyzed during this study are included in this manuscript.

10. Declaration on the Use of Artificial Intelligence (AI) Tools

The authors declare that artificial intelligence (AI) tools, specifically ChatGPT (OpenAI), were used solely to refine the language, improve grammar, and enhance the clarity of the manuscript. The AI tool was not used to generate scientific content, analyze data, interpret results, or draw scientific conclusions.

REFERENCES

Ausubel, F. M., Brent, R., Kingston, R. E., et al. 1992, Short Protocols in Molecular Biology (New York), 28764–28773

Bachs, O., Gallastegui, E., Orlando, S., et al. 2018, Oncotarget, 9, 26259

Classon, M., & Harlow, E. 2002, Nature Reviews Cancer, 2, 910

de Figueiredo, S. M., Binda, N. S., Nogueira-Machado, J. A., Vieira-Filho, S. A., & Caligiorne, R. B. 2015, Recent Patents on Endocrine, Metabolic & Immune Drug Discovery, 9, 24

Fontanella, C., Giorgi, C. A., Russo, S., et al. 2022, Critical Reviews in Oncology/Hematology, 180, 103848, doi: 10.1016/j.critrevonc.2022.103848

Garzon, R., Marcucci, G., & Croce, C. M. 2010, Nature Reviews Drug Discovery, 9, 775

Hallett, S. T., Pastok, M. W., Morgan, R. M. L., et al. 2017, Cell Reports, 21, 1386

Hanahan, D., & Weinberg, R. A. 2011, Cell, 144, 646

Hassan, N. M., Alhossary, A. A., Mu, Y., & Kwoh, C.-K. 2017, Scientific Reports, 7, 15451

Hopkins, A. L., Groom, C. R., & Alex, A. 2004, Drug Discovery Today, 9, 430

Illia, M. F., Colucci, G., Rodriguez Mignola, A., et al. 2026, Translational Breast Cancer Research, 7, 7, doi: 10.21037/tbcr-25-46

Jameel, E., Naz, H., Khan, P., et al. 2017, Chemical Biology & Drug Design, 89, 741
Kim, J.-A., Kim, D. H., Hossain, M. A., et al. 2013, International Journal of Oncology, 44, 473
Li, Y., Zhang, J., Gao, W., et al. 2015, International Journal of Molecular Sciences, 16, 9314
Malumbres, M. 2014, Genome Biology, 15, 122
Mohammad, T., Mathur, Y., & Hassan, M. I. 2021, Briefings in Bioinformatics, 22, bbaa279
Nebenfuehr, S., Kollmann, K., & Sexl, V. 2020, International Journal of Cancer, 147, 2988
O'Sullivan, C. C. 2026, Cancers, 18, doi: 10.3390/cancers18030533
Pavani, R. S., da Silva, M. S., Fernandes, C. A. H., et al. 2016, PLoS Neglected Tropical Diseases, 10, e0005181
Petroni, G., Formenti, S. C., Chen-Kiang, S., & Galluzzi, L. 2020, Nature Reviews Immunology, 20, 669
Rencuzogullari, O., Yerlikaya, P. O., Gurkan, A. C., Arisan, E. D., & Telci, D. 2020, Journal of Cellular Biochemistry, 121, 508
Russell, D. W., & Sambrook, J. 2001, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory)
Russo, A. A., Tong, L., Lee, J.-O., Jeffrey, P. D., & Pavletich, N. P. 1998, Nature, 395, 237
Sa, G., & Das, T. 2008, Cell Division, 3, 14
Shah, A., Bloomquist, E., Tang, S., et al. 2018, Clinical Cancer Research, 24, 2999
Shamsi, A., Anwar, S., Mohammad, T., et al. 2020, Biomolecules, 10, 789
Sherr, C., Beach, D., & Shapiro, G. 2016, Cancer Discovery, 6, 353
Shityakov, S., & Foerster, C. 2014, Advances and Applications in Bioinformatics and Chemistry, 1
Sielecki, T. M., Boylan, J. F., Benfield, P. A., & Trainor, G. L. 2000, Journal of Medicinal Chemistry, 43, 1
Spring, L. M., Wander, S. A., Andre, F., et al. 2020, The Lancet, 395, 817
Srivastava, N., & Saxena, A. K. 2025, Current Medicinal Chemistry, 32, 7512, doi: 10.2174/0109298673331685241205180307
Sun, J. 2007, Alternative Medicine Review, 12, 259
Sun, Y., Zhao, Y., Yu, X., Qi, Y., & Dai, X. 2025, Discover Oncology, 17, 138, doi: 10.1007/s12672-025-04269-2
Tigan, A., Bellutti, F., Kollmann, K., Tebb, G., & Sexl, V. 2016, Oncogene, 35, 3083
Trott, O., & Olson, A. J. 2010, Journal of Computational Chemistry, 31, 455
Voura, M., Khan, P., Thysiadis, S., et al. 2019, Scientific Reports, 9, 1676
Yousuf, M., Alam, M., Shamsi, A., et al. 2022, International Journal of Biological Macromolecules, 218, 394, doi: 10.1016/j.ijbiomac.2022.07.156
Yousuf, M., Shamsi, A., Khan, P., et al. 2020, International Journal of Molecular Sciences, 21, 3526
Zhu, Z., & Zhu, Q. 2023, Frontiers in Pharmacology, 14, 1212986, doi: 10.3389/fphar.2023.1212986

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 9–16, 2026
Open Access © 2026 Editory Press

Molecular Insights into Arylsulfatase B Mutation-Induced Instability in Mucopolysaccharidosis Type VI

Barka Basharat[1], Nushrat Jahan[1] and Mohammad Asim Azhar[2,*]

[1]*Department of Biotechnology, School of Chemical and Life Sciences, Jamia Hamdard, Hamdard Nagar, New Delhi, India*
[2]*Organ Transplant Center of Excellence, King Faisal Specialist Hospital and Research, Riyadh, Kingdom of Saudi Arabia*
[*]*Corresponding author: Barka Basharat, Organ Transplant Center of Excellence, King Faisal Specialist Hospital and Research, Riyadh, Kingdom of Saudi Arabia, E-mail: azharasim@gmail.com*

Abstract

Mutations in the arylsulfatase B (ARSB) gene are directly implicated in mucopolysaccharidosis type VI (MPS VI). Several non-synonymous single-nucleotide polymorphisms (nsSNPs) in ARSB have been associated with disease pathogenesis. A comprehensive evaluation of these variants is essential to understand their structural and functional consequences. In this study, a systematic in silico analysis was performed to identify deleterious nsSNPs in the ARSB gene. Initially, 430 nsSNPs were evaluated using sequence-based prediction tools, including SIFT, PolyPhen-2, FATHMM, and Mutation Assessor. Subsequently, 141 nsSNPs were subjected to structure-based stability analysis using MAESTROweb, SDM2, mCSM, and DynaMut2, of which 57 variants overlapped with previous reports. High-confidence deleterious nsSNPs were further assessed for pathogenicity using PMut and MutPred2 servers. Our integrated computational approach identified 44 highly deleterious mutations. Aggregation propensity analysis revealed that 29 of these variants exhibit increased aggregation tendencies, while one variant demonstrated progressive loss of solubility. Molecular dynamics simulations further indicated that high-confidence deleterious nsSNPs significantly disrupt ARSB structural integrity, enhance molecular flexibility, reduce structural rigidity, and promote atomic-level aggregation. Overall, this study provides mechanistic insights into how pathogenic mutations destabilize the ARSB protein and contribute to MPS VI pathogenesis, highlighting potential targets for future therapeutic investigation.

Keywords: *ARSB gene, mucopolysaccharidosis type VI, non-synonymous SNPs, pathogenic mutations, protein stability, protein aggregation, genetic variation, computational mutagenesis*

Received: April 5th, 2025 **Accepted:** August 25th, 2025 **Published:** January 5th, 2026

1. Introduction

Lysosomes function as acidic cellular hubs for macromolecule catabolism, recycling, and signaling via hydrolytic enzymes and efflux permeases. Lysosomal storage diseases (LSD), also known as dysostosis multiplex, are due to an acquired shortage of specific lysosomal enzymes (Platt et al., 2018). These diseases are frequently linked to numerous impairments of musculoskeletal growth and formation. Deficiencies in lysosomal enzymes or associated proteins cause substrate accumulation, such as sphingolipids, glycosaminoglycans (GAGs), or oligosaccharides, leading to over 41 progressive LSDs, most of which are autosomal recessive, except for X-linked Fabry disease and MPS II (Scerra et al., 2022). These disorders exhibit heterogeneous phenotypes, from neonatal lethality (e.g., non-immune hydrops fetalis) to later-onset forms (e.g., Niemann-Pick type C), with multi-organ involvement including neurological, skeletal, ophthalmic, and visceral symptoms. Pathophysiology stems from intra-lysosomal storage disrupting signaling pathways, mitochondrial function (e.g., fragmented cristae and reduced membrane potential in MPS VI, GM1 gangliosidosis), and tissue integrity; LSDs are classified by accumulated substrates, such as sphingolipidoses, mucopolysaccharidoses (MPS), and oligosaccharidosis (Gros & Muller, 2023).

Mucopolysaccharidosis type VI (MPS VI), also known as Maroteaux-Lamy syndrome, is an autosomal recessive lysosomal storage disorder caused by mutations in the arylsulfatase B (ARSB) gene, which encodes N-acetylgalactosamine-4-sulfatase (Tobacman & Bhattacharyya, 2022). This enzyme hydrolyzes sulfate groups from dermatan sulfate (DS) and chondroitin-4-sulfate, preventing their lysosomal accumulation (Rossi et al., 2025). GAGs are major components of the extracellular matrix and play critical roles in connective tissue structure, cell signaling, and tissue integrity. Proper ARSB activity ensures normal lysosomal function, cellular homeostasis, and continuous turnover of connective tissue components in skin, cartilage, tendons, and other organs. Deficient ARSB activity leads to progressive GAG buildup in multiple tissues, manifesting as dysostosis multiplex with musculoskeletal dysplasia, joint stiffness, corneal clouding, cardiac valve disease, hepatosplenomegaly, and respiratory complications (Leal et al., 2025).

Clinical severity varies widely; severe cases present symptoms by age 2–3, with loss of ambulation by age 10 and survival into

the third decades (Leal et al., 2025). Diagnosis integrates clinical evaluation, urinary GAG quantification, enzymatic assays, and ARSB genotyping. While primarily somatic, MPS VI shares skeletal and visceral features with other MPS disorders, driven by GAG-mediated connective tissue pathology, including alterations in the extracellular matrix and cardiac infiltrates (Lipiński et al., 2025). Musculoskeletal defects, joint rigidity, ocular obscuration, cardiac issues, and breathing difficulties are among the clinical symptoms of MPS VI (De Ponti et al., 2022).

Single-nucleotide polymorphisms (SNPs) represent the most common genomic variants, driving evolutionary adaptation and disease susceptibility (Sauna & Kimchi-Sarfaty, 2022). Non-synonymous single-nucleotide polymorphisms (nsSNPs) in ARSB represent a major mutational class, potentially disrupting protein stability, folding, or interactions via amino acid substitutions in conserved domains (Sinha et al., 2022). Over 200 ARSB variants have been reported, yet their structural-functional impacts remain incompletely characterized, complicating prognosis and therapy selection amid emerging options such as enzyme replacement, gene therapy, and substrate reduction (Gomez-Ospina, 2024; Tobacman & Bhattacharyya, 2022).

This study employs an integrated *in silico* pipeline comprising SIFT, PolyPhen-2, FATHMM, Mutation Assessor, MAESTROweb, SDM2, mCSM, DynaMut2, PMut, MutPred2, and molecular dynamics simulations to systematically evaluate 430 ARSB nsSNPs (Enni, 2025). Figure 1 illustrates computational methods that integrate sequence-, structure-, and dynamics-based approaches to systematically identify pathogenic variants in ARSB associated with MPS VI. We prioritize high-confidence deleterious variants, assess their aggregation propensity and changes in stability, and elucidate their mechanistic contributions to MPS VI pathogenesis, thereby informing precision diagnostics and therapeutic strategies (Sarachakov et al., 2025).

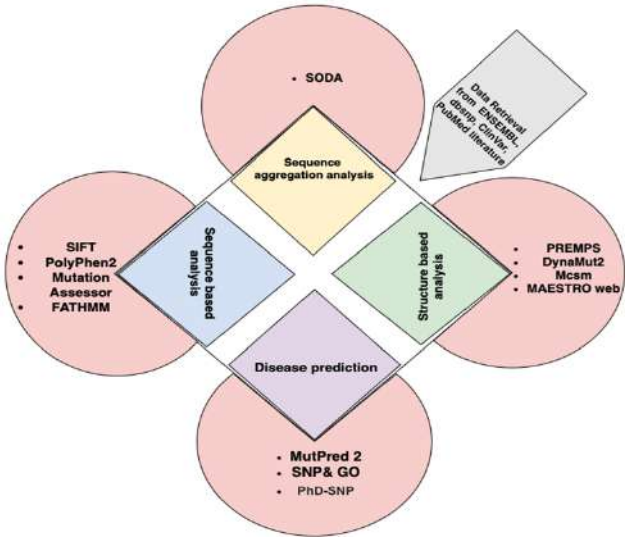

Figure 1. A description of the computational methods used to forecast the pathogenicity of the ARSB gene mutation at the structural and functional domains.

2. Materials and Methods

2.1. Data retrieval

The FASTA sequence of the human ARSB gene was obtained from the UniProt database (UniProt ID: P15848). A database of missense mutation SNPs was created using information gathered through a PubMed literature review and databases such as dbSNP, HGMD, ClinVar, and Ensemble. The list was cleared of the redundant nsSNPs. The Protein Data Bank (PDB ID: 1FSU) provided the crystal structure of human ARSB. The remaining mutations were obtained from the Ensembl databases, and all mutation types, including missense, non-synonymous, and synonymous mutations, were included (Dyer et al., 2025).

2.2. Sequence-Based Prediction of Deleterious Mutations

Sorting Intolerant From Tolerant (SIFT: http://sift.jcvi.org/) was used to predict the functional impact of amino acid substitutions based on sequence homology and evolutionary conservation (Sim et al., 2012). The underlying principle is that functionally important residues are conserved across species; therefore, substitutions at highly conserved positions are more likely to be deleterious. SIFT assigns a tolerance score ranging from 0 to 1, where scores ≤ 0.05 indicate damaging (intolerant) substitutions and scores > 0.05 suggest tolerated variants. In addition to missense variants, SIFT can also evaluate certain 3-nucleotide indels that result in amino acid insertions or deletions. These predictions are based on conservation patterns derived from multiple sequence alignments.

2.2.1. PolyPhen-2

PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/) used to predict the potential structural and functional impact of amino acid substitutions using sequence-based and structure-based features. It incorporates evolutionary conservation, physicochemical differences between residues, and structural parameters such as solvent accessibility and proximity to functional domains (Adzhubei et al., 2013). The tool calculates the difference in Position-Specific Independent Count (PSIC) scores between wild-type and mutant residues. Variants are classified as benign, possibly damaging, or probably damaging, depending on the PSIC score difference and predictive confidence.

2.2.2. Mutation Assessor

Mutation Assessor (http://mutationassessor.org/) evaluates the functional impact of amino acid substitutions based on evolutionary conservation patterns within protein families and subfamilies (Su et al., 2025). It distinguishes between general conservation across homologous proteins and functional specificity within subfamilies. Each variant is assigned a Functional Impact (FI) score and categorized as neutral, low, medium, or high impact. Variants with FI scores > 2.0 are generally considered functionally deleterious.

2.2.3. FATHMM

Functional Analysis Through Hidden Markov Models (FATHMM) predicts the pathogenicity of missense mutations using Hidden Markov Models (HMMs) and conservation-based weighting schemes (Shihab, 2021). It can incorporate disease-specific or cancer-specific models. Variants are classified as tolerated or deleterious based on a threshold score, with lower scores indicating a higher likelihood of pathogenicity.

2.3. Structure-Based Stability Prediction

2.3.1. MAESTROweb

MAESTROweb (https://pbwww.che.sbg.ac.at/maestro/web) predicts mutation-induced changes in protein stability using machine-learning methods trained on experimental thermodynamic datasets. It calculates the change in Gibbs free

energy ($\Delta\Delta G$) between wild-type and mutant proteins (Laimer et al., 2016). A negative $\Delta\Delta G$ value indicates destabilization of protein structure, whereas a positive value suggests stabilization. The tool also identifies mutation "hotspots" by scanning multiple residues.

2.3.2. mCSM

mCSM predicts the impact of mutations on protein stability and macromolecular interactions using graph-based structural signatures derived from atomic distance patterns. It estimates $\Delta\Delta G$ values, where negative scores indicate destabilizing mutations (Pires et al., 2014). Variants with significantly negative $\Delta\Delta G$ values are considered likely to disrupt protein stability and function.

2.3.3. DynaMut2

DynaMut2 (http://biosig.unimelb.edu.au/dynamut/) integrates Normal Mode Analysis (NMA) with graph-based signatures to evaluate mutation-induced changes in protein stability and dynamics (Rodrigues et al., 2021). It predicts both $\Delta\Delta G$ and changes in vibrational entropy ($\Delta\Delta S$), allowing assessment of alterations in structural flexibility. This tool is particularly useful for understanding how mutations affect conformational dynamics, well to thermodynamic stability.

2.3.4. PremPS

PremPS (https://lilab.jysw.suda.edu.cn/research/PremPS/) predicts the impact of missense mutations on protein stability using a balanced dataset of stabilizing and destabilizing mutations (Chen et al., 2020). It integrates structural and evolutionary features to estimate $\Delta\Delta G$ values and classify variants as stabilizing or destabilizing.

2.4. Pathogenicity Prediction Tools

2.4.1. MutPred2

MutPred2 (http://mutpred.mutdb.org) is a machine learning–based tool that predicts the pathogenicity of amino acid substitutions and provides mechanistic insights into possible molecular alterations (Pejaver et al., 2020). It assigns a pathogenicity probability score ranging from 0 to 1. Variants with scores > 0.5 (commonly > 0.589 for high confidence) are considered likely pathogenic. Additionally, it predicts alterations in secondary structure, post-translational modification sites, catalytic residues, and protein–protein interactions.

2.4.2. SNPs&GO

SNPs&GO integrates sequence features and Gene Ontology (GO) annotations using a Support Vector Machine (SVM) classifier to distinguish disease-associated variants from neutral polymorphisms. The inclusion of functional annotations improves predictive accuracy (Capriotti et al., 2013).

2.4.3. PhD-SNP

PhD-SNP (https://snps.biofold.org/phd-snp/phd-snp.html) is an SVM-based predictor that classifies missense mutations as disease-associated or neutral using sequence and evolutionary profile information. Variants with scores > 0.5 are predicted to be disease-causing (Calabrese et al., 2008).

2.5. Evolutionary Conservation Analysis

2.5.1. ConSurf

ConSurf (https://consurf.tau.ac.il/) evaluates evolutionary conservation of amino acid residues using multiple sequence alignment and phylogenetic analysis (Ashkenazy et al., 2016). Conservation scores range from 1 (variable) to 9 (highly conserved). Highly conserved exposed residues are often functionally important, whereas conserved buried residues are typically structural (architectural). Disease-associated mutations frequently occur at highly conserved positions.

2.6. Aggregation Propensity Analysis

2.6.1. SODA

SODA (Solubility based on Disorder and Aggregation) predicts the impact of mutations on protein solubility, secondary structure, and aggregation propensity. It helps identify variants that may promote protein misfolding or aggregation, which is relevant in disorders associated with lysosomal dysfunction (Paladin et al., 2017).

2.6.2. Arpeggio

Arpeggio analyzes interatomic interactions within protein structures by categorizing them into hydrogen bonds, van der Waals contacts, hydrophobic interactions, and ionic interactions. It accepts PDB structures and generates quantitative interaction profiles, enabling comparison between wild-type and mutant proteins to assess structural disruption (Jubb et al., 2017).

3. Results and Discussion

MPS VI, a lysosomal storage disease, is caused by ARSB deficiency (Tomanin et al., 2018). The accumulation of GAGs caused by this enzyme deficiency leads to a variety of clinical symptoms, including skeletal deformities, stiff joints, and corneal clouding. MPSVI usually has serious progression, with symptoms that start in early childhood and worsen over time (Pohl et al., 2018). Determining the molecular mechanisms driving this condition requires knowledge of the variants in the ARSB gene and their effects on protein function. In this investigation, we looked at the possibility that harmful mutations in ARSB contribute to the pathophysiology of MPSVI.

We sought to identify alterations that could have a major effect on the stability and function of the ARSB protein by employing a comprehensive strategy that combines sequence- and structure-based analyses (Leal et al., 2025). Sequence-based methods, including SIFT, PolyPhen2, FATHMM, and Mutation Assessor, were used to predict the detrimental consequences of mutations. Structure-based techniques were used to evaluate the effects of changes on protein structure and aggregating propensity, including mCSM, DynaMut2, MAESTROweb, and PremPS. Our study's key finding is that these mutations may affect protein solubility, a defining characteristic of MPS VI. We thoroughly analysed 1224 ARSB gene variants obtained from the dbSNP and Ensembl databases, as well to other mutations identified. We investigated the effects of these alterations on the structure and function impacts of variations or mutations within the ARSB gene.

Five web-based resources were used to help the sequence-based evaluation: SIFT, PolyPhen2, FATHMM, and Mutation Assessor. Simultaneously, the structure-based method assessed single-point amino acid alterations in ARSB using mCSM, DynaMut2, MAESTROweb, and PremPS. Only mutations that the analyses showed to be high-confidence variations were sent for additional testing. We used the PhD-SNP and MutPred2 websites to explain the illness characteristics associated with these high-confidence variants, as will be covered in the sections that follow.

3.1. Deleterious mutations identification of nsSNPs

Sequence-based evaluation was performed on all nsSNPs utilising five online tools: SIFT, PolyPhen2, PROVEAN, Mutation Assessor, FATHMM Using the structure-based approach, 141 nsSNPs

located in the ARSB gene were analysed (Table S1) using PremPs, MAESTROweb, DynaMut2, and mCSM. As the sole region identified through experimentation and available in the PDB database, only the highly probable nsSNPs have been selected for additional investigation. Using the PMut web server and MutPred2, high-probability nsSNPs' disease phenotypes were identified. We utilised FATHMM, SIFT, PolyPhen2, and SNPs&GO. Based on protein physical characteristics, SIFT classifies variants as either intolerant or tolerated; higher scores indicate a higher likelihood of toxicity. The study of 1224 single-point amino acid changes in ARSB using the sequence-based approach produced predictions from SIFT, PolyPhen2, Mutation Assessor, and FATHMM. These tools, particularly, highlighted the dangerous substitutions 302, 264, 238, and 167 (Figure 2). By combining these sequence-based predictive techniques, we were able to examine the possible molecular effects of mutations on functioning.

We used four different structure-based prediction technologies: DynaMut2, MAESTROweb, PremPS, and mCSM (Figure 3). By computing folding free energy, these tools assess the stability of variations by analysing atomic coordinates from the PDB file of the wild-type protein. Many of these tools use a machine-learning approach, combining different biophysics-based methodologies to predict the impact of variations on protein stabilisation. Folded (Gf) and unfolded (Gu) forms are computed in thermodynamics using the formula $\Delta G = G_u - G_f$. $\Delta\Delta G = G_m - G_w$, wherein G_m is the energy of the protein that has been altered, and G_w is the energy of the wild-type protein, is used to assess the change in protein stability and the landscape of energy. A variation that stabilises a protein is indicated by a negative $\Delta\Delta G$ value, whereas a mutation that destabilises the protein is suggested by a positive $\Delta\Delta G$ score.

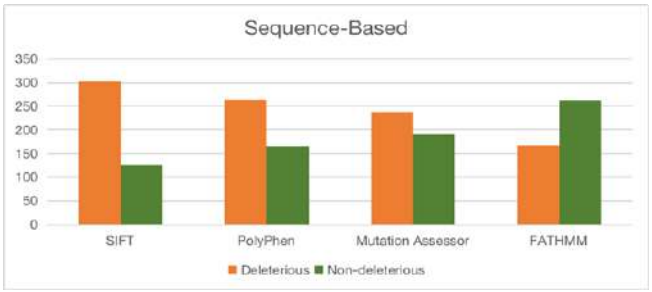

Figure 2. Deleterious mutations ARSB gene sequence-based approach. This graph illustrates the effects of mutation using a computational approach.

Concurrently, destabilising mutations were identified at 132, 130, 64, and 130 positions by structure-based predictions from mCSM, DynaMut2, MAESTROweb, and PremPS. To increase confidence in our results, we chose for further investigation only the modifications predicted by all sequence-based and structure-based methods to be detrimental. 44 amino acid changes that were considered harmful and destabilising were identified through this sorting process. These 44 changes were then analysed to investigate any possible connections to disease characteristics.

3.2. Identification of pathogenic mutations using a computational approach

We evaluated disease characteristics linked to mutations using PhD-SNP, SNAP & GO, and MutPred2 approaches. These tools classify variants based to potential disease associations and assign

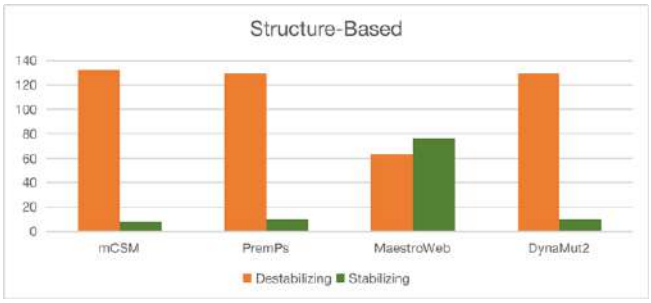

Figure 3. Deleterious mutations ARSB gene structure-based approach. This graph illustrates the effects of mutation using a computational approach.

pathogenicity levels (Figure 4). Among the 57 high-confidence variations identified through a sequence- and structure-based evaluation, PMut and MutPred predicted 44 as pathogenic. Only 44 mutations, A237D, C91R, E323K, E483D, G149R, G302R, G324V, G56D, G64R, H393P, I296N, I67N, K145E, L129P, L236P, L360P, L498P, L51P, L72P, L72R, L82P, L82R, L90P, L98P, L98R, P93R, R315P, R315Q, R327G, R327Q, R388T, T92K, V277G, V80G, W146R, W146S, W353R, W438G, Y138C, Y175D, Y210C, Y266S, Y86C, Y86N) out of 57 highly probable nsSNPs were shown to be harmful using the disease phenotype assessment algorithm (Table **??** and Table **??**)).

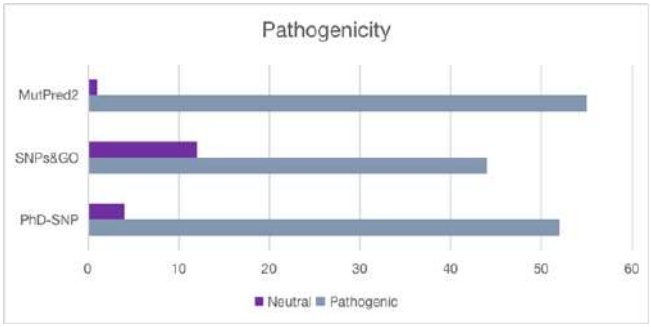

Figure 4. Pathogenic mutations predicted in the ARSB gene were predicted using structure-based tools.

3.3. Analysis of conserved residue

The ConSurf tool was utilised to analyse the human ARSB genome structure and assess residue conservation (Ashkenazy et al., 2016). According to the ConSurf study, out of 44 final mutations, these 12 (A237D, G56D, G64R, L51P, I67N, L236P, W353R, V277G, T92K, R315Q, R315P, H393P) can be highly disease-causing mutations, because they belonged in high conserved areas and mutations in these regions (Figure 5).

3.4. Analysis of aggregation propensity

The aggregation, illness, helix, and strand propensities resulting from alterations are computed using SODA48. Out of 12 nsSNPs, 2 (A237D, W353R) were shown (Table 1) to lower the protein's solubility, and 9 of them raised it, out of the 11 mutations revealed using illness phenotype prediction. The solubility of a protein significantly impacts its function. Insoluble proteins tend to congregate, which can lead to illnesses including Parkinson's, amyloidosis, and Alzheimer's. By combining solubility data with structural and sequence-based properties, SODA provides insights into protein regions susceptible to aggregation and suggests

Figure 5. Conserved residue of the ARSB region.

modifications that may improve solubility (Aggidis et al., 2024; Kumar et al., 2016; Paladin et al., 2017).

Table 1. Aggregation propensity prediction of mutant ARSB protein using SODA server.

Sequence	SODA	Remark
A237D	-0.471	Less soluble
G56D	0	More soluble
G64R	4.558	More soluble
H393P	4.84	More soluble
I67N	1.308	More soluble
L236P	1.72	More soluble
L51P	11.762	More soluble
R315P	4.59	More soluble
R315Q	4.03	More soluble
T92K	2.878	More soluble
V277G	3.236	More soluble
W353R	-0.676	Less soluble

Of the 44 alterations, 12 vnsSNPs decrease the protein's solubility. The fact that all three solubility-reducing alterations can contribute to the formation of amino acid clumps and the subsequent pathophysiology of illnesses makes this particularly concerning. It turned out that the dissolution of the amino acid was enhanced by the additional 29 nsSNPs.

Thorough conformational analyses yielded a deeper understanding of the molecular effects of these pathogenic alterations on the ARSB protein. We found that the pathogenicity associated with these changes may be driven by the addition or removal of interatomic noncovalent interactions, such as hydrophobic contacts, van der Waals forces, and hydrogen bonds (Table 2). Demonstrating structure and its interactions in Figure 6. Mutations can cause structural defects that contribute to the pathophysiology of illnesses by causing misfolding, loss of function, or enhanced agglomeration.

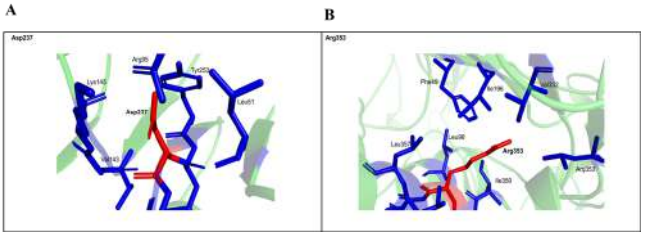

Figure 6. Mutant structure (A). Alanine mutated to Aspartic acid and (B). Tryptophan mutated to Arginine.

The significant rise in overall contacts, particularly proximal and VdW clashes, indicates increased steric hindrance and overcrowding of the altered protein. An increase in hydrogen bonds or polar interactions can counterbalance some of the destabilising effects of greater overcrowding and improve the long-term stability of the protein's structure. A denser and perhaps stressed structure is suggested by the A237D alteration, which causes a substantial rise in total acquaintances, particularly through proximal and van der Waals collisions (Table 2). This suggests that, although the A237D mutant induces structural stress, compensatory stabilising interactions may allow the protein's structure to retain its integrity (Figure 6).

Protein interactions require both hydrogen bonds and polar interactions to be specific and stable (Sheng et al., 2015). A decrease in these interactions may impair the protein's functioning

Table 2. Comparing changes with their interactions between wild type and mutant structure

Interactions	Wild type	Mutant (A237D)	Mutant (W353R)
Total number of contacts	[proximal+VdW clashinteraction] 71+3=74	[Vdw+VdW clash+proximal] 1+11+119=131	[Vdw+VdW clash+proximal] 4+7+119=130
Polar contacts	2	7	0
Weak polar contacts	2	4	4
Hydrogen bond	2	7	3
Ionic interaction	0	7	3
Carbonyl interaction	1	1	0
Hydrophobic contacts	3	3	0
Halogen bonds	0	0	4
Metal complex interaction	0	0	0
Hydrophobic contacts	3	3	2

and structural integrity. A decrease in hydrogen bonding can alter the shape of binding pockets or active sites, making it more difficult for a protein to interact with substrates, inhibitors, or other binding molecules (Chen et al., 2020; Khan et al., 2019; Shahid et al., 2015). This may result in decreased cellular activity overall, decreased binding affinity, and decreased enzyme activity in particular. The protein's internal integrity may be compromised by a significant decrease in these connections, exposing hydrophobic residues to the surrounding water. Because the hydrophobic residues try to minimise connection to the water-based environment, this contact could lead the protein to misfold. Proteins that are misfolded are more likely to aggregate, a process in which the exposed hydrophobic regions of many molecules of protein cling to one another to create insoluble clumps. Tryptophan, phenylalanine, and tyrosine are examples of aromatic residues that frequently take role in π-π stacking connections, in which the aromatic rings correspond in a parallel or edge-to-face fashion.

These mutations impair the stability of the gene overall, in addition to possibly causing illness by rupturing the structural integrity of highly conserved sections. After that, we evaluated the aggregation characteristics of these alterations using the SODA programme. It is quite probable that these two last mutations (A237D, W353R) will make the protein less soluble and contribute to the pathophysiology of the disease. This research provides a detailed account of the pathogenic nsSNPs in the ARSB gene and their possible effects. Determining whether particular mutations are harmful, unstable, and pathogenic provides important new insights into the molecular processes underlying the genesis of illness. The aggregation propensity study also identifies potential treatment targets by highlighting the significance of amino acid solubility in disease association.

To sum up, the A237D mutation results in a larger total number of connections, which causes structural strain, yet it also preserves overall integrity due to improved polar, hydrogen, and ionic interactions (Figure 6). On the other hand, the W353R mutation introduces notable changes that could disrupt local equilibrium by affecting aromatic interactions. Different structural results arise from the precise form and position of the alterations, even if both mutations bring compensatory interactions to preserve structural stability. While the W353R mutation has localised destabilising effects that affect the protein's functional regions, the A237D alteration better maintains its structural stability.

3.5. Aggregation Propensity

Treatment plans to mitigate the effects of such harmful mutations can be developed using the knowledge gained from this study (Israil et al., 2025; Nag and Tripathi, 2022; Paladin et al., 2017). Creating medications or tiny compounds that stabilise the peptide, improve its ability to dissolve, or stop it from aggregating are some possible strategies. Additionally, by understanding the structural alterations caused by these mutations, tailored therapies that restore lost or normal protein function may be developed.

In the final analysis, our thorough examination of nsSNPs in the ARSB gene not only advances our understanding of the genetic underpinnings of associated illnesses but also lays the groundwork for developing effective treatments. The integration of structural, solubility, and sequencing research offers a comprehensive knowledge of the effects of pathogenic mutations. Two of the discovered mutations are located in highly conserved regions that may affect the gene's function, according to our ConSurf analysis. Using the Arpeggio web server, we further evaluated the aggregation properties and conducted additional structural analyses (Figure 7). We discovered two changes that reduce protein solubility, suggesting a possible role in ARSB aggregation and the ensuing illness. The identification of mutations with the greatest potential to contribute to disease pathophysiology was made possible by sequence-based analysis. Simultaneously, the predictions based on structure reinforced our belief that mutations could be crucial to the onset and progression of the disease.

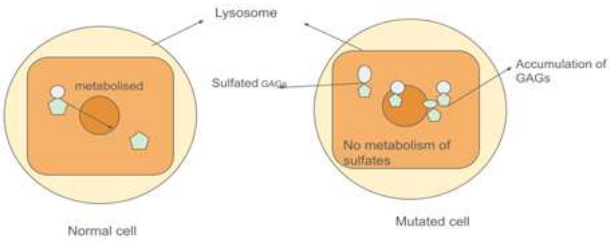

Figure 7. ARSB mutation can cause accumulation of sulfated GAG in the lysosome, which may result in a lysosomal disorder.

3.6. Molecular Dynamics Simulation Analysis

The wild-type ARSB protein, along with a few selected mutant variants, was subjected to molecular dynamics (MD) simulations

to confirm the structural consequences of high-confidence deleterious substitutions. MD simulations provide dynamic insights into the stability, flexibility, and conformational behavior of proteins under physiologically relevant conditions, in contrast to static protein structure investigations.

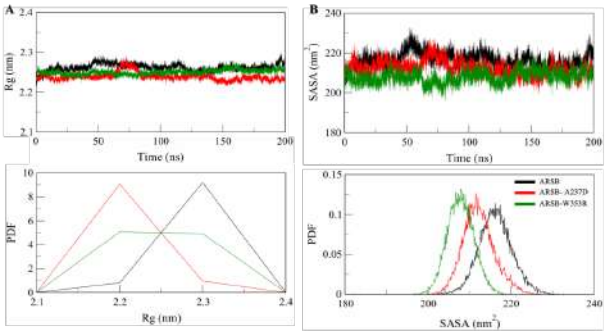

Figure 8. Radius of Gyration (Rg) analysis indicating changes in compactness of wild-type and mutant ARSB proteins over time.

RMSD, RMSF, Radius of Gyration (Rg), SASA, and hydrogen-bond evaluations were included in the trajectory evaluations (Naqvi et al., 2018). In contrast to the natural type, mutant proteins displayed delayed equilibration and higher RMSD values, suggesting decreased structural reliability. RMSF analysis revealed higher adaptability in conserved and functionally significant regions, suggesting a disturbance in catalytic activity (Figure 8). Many mutants had elevated Rg and SASA values, consistent with unfolding and increased aggregation (Figure 9). These values also indicated decreased compactness and higher solvent exposure.

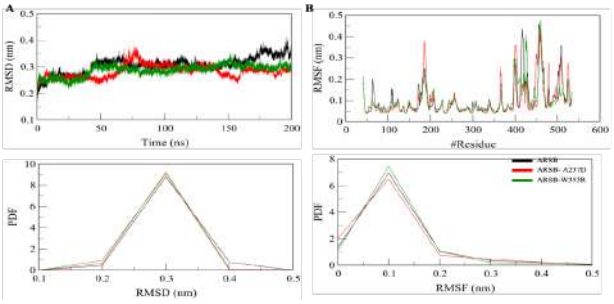

Figure 9. RMSD plot showing overall structural stability and conformational deviations of wild-type and mutant ARSB proteins during the simulation.

4. Conclusions

SNPs, or single-nucleotide polymorphisms, are thought to be among the most common genetic variations linked to several human disorders. 139 of the 429 mutations found are harmful and destabilising, according to sequence- and structure-based studies. A research investigation on pathogenicity found that 44 of the total number of variants are harmful. After consurf analysis of aggregation propensity, we found that 2 final mutations may be factors to causing disease. A thorough study of SNPs can provide insights into the mechanisms underlying disease development and inform the development of efficient therapeutic approaches. Our results indicate the importance of computational mutational analysis for understanding the genetic underpinnings of complex diseases such as MPS VI (Karageorgos et al., 2007; Prokop et al., 2022). This study lays the foundation for future research to identify targeted therapy strategies and contributes to the expanding body of knowledge on the pathophysiology of MPS VI. In the end, the study highlights the importance to using cutting-edge computational methods to better understand the molecular pathways underlying disease pathology (Dissanayake et al., 2025).

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This work received no funding.

Data availability statement

All data generated or analyzed during this study are included in this manuscript.

Declaration on the Use of AI Tools

The authors declare that ChatGPT (OpenAI) was used solely to refine the language, improve grammar, and enhance the clarity of the manuscript.

References

De Ponti, G., Donsante, S., Frigeni, M., et al. 2022, International Journal of Molecular Sciences, 23, 11168

Enni, M. A. 2025, International Journal of Scientific Interdisciplinary Research, 6, 88

Gomez-Ospina, N. 2024, Arylsulfatase A deficiency

Gros, F., & Muller, S. 2023, Nature Reviews Nephrology, 19, 366

Leal, A. F., Prieto, L. E., Pachajoa, H., & Tomatsu, S. 2025, Molecular Genetics and Metabolism, 109255

Lipiński, P., Różdżyńska-Świątkowska, A., Wiśniewska, K., et al. 2025, Biomolecules, 15, 1448

Platt, F. M., d'Azzo, A., Davidson, B. L., Neufeld, E. F., & Tifft, C. J. 2018, Nature Reviews Disease Primers, 4, 27

Rossi, A., Romano, R., Fecarotta, S., et al. 2025, Med, 6

Sarachakov, A., Yudina, A., Svekolkin, V., et al. 2025, Human Genetics, 144, 1245

Sauna, Z. E., & Kimchi-Sarfaty, C. 2022, Single nucleotide polymorphisms: human variation and a coming revolution in biology and medicine (Springer)

Scerra, G., De Pasquale, V., Scarcella, M., et al. 2022, Open Biology, 12

Sinha, A., Dinakarkumar, Y., Al-Qahtani, W. H., et al. 2022, Human Gene, 34, 201079

Tobacman, J. K., & Bhattacharyya, S. 2022, International Journal of Molecular Sciences, 23, 13146

Table 3. Disease phenotype evaluation of higher certainty nsSNPs in the ARSB gene applying PMut, MutPred, and SNAP & GO estimation methods.

Mutations	PhD-SNP	SNP & GO	MutPred2	Remark
A237D	Disease	Disease	0.9	Pathogenic
C91R	Disease	Disease	0.923	Pathogenic
E323K	Disease	Disease	0.943	Pathogenic
E483D	Disease	Disease	0.871	Pathogenic
G149R	Disease	Disease	0.947	Pathogenic
G302R	Disease	Disease	0.967	Pathogenic
G324V	Disease	Disease	0.946	Pathogenic
G527R	Disease	Neutral	0.931	Benign
G56D	Disease	Disease	0.964	Pathogenic
G64R	Disease	Disease	0.715	Pathogenic
H393P	Disease	Disease	0.889	Pathogenic
I296N	Disease	Disease	0.944	Pathogenic
I67N	Disease	Disease	0.818	Pathogenic
K145E	Disease	Disease	0.921	Pathogenic
L129P	Disease	Disease	0.937	Pathogenic
L132P	Disease	Neutral	0.735	Benign
L236P	Disease	Disease	0.925	Pathogenic
L360P	Disease	Disease	0.952	Pathogenic
L498P	Disease	Disease	0.909	Pathogenic
L51P	Disease	Disease	0.957	Pathogenic
L72P	Disease	Disease	0.939	Pathogenic
L72R	Disease	Disease	0.935	Pathogenic
L82P	Disease	Disease	0.938	Pathogenic
L82R	Disease	Disease	0.94	Pathogenic
L90P	Disease	Disease	0.836	Pathogenic
L98P	Disease	Disease	0.91	Pathogenic
L98R	Disease	Disease	0.901	Pathogenic
N84K	Neutral	Neutral	0.64	Benign
P248A	Neutral	Neutral	0.515	Benign
P531R	Neutral	Neutral	0.947	Benign
P93R	Disease	Disease	0.713	Pathogenic
R102H	Disease	Neutral	0.614	Benign
R315P	Disease	Disease	0.967	Pathogenic
R315Q	Disease	Disease	0.89	Pathogenic
R327G	Disease	Disease	0.96	Pathogenic
R327Q	Disease	Disease	0.902	Pathogenic
R388T	Disease	Disease	0.775	Pathogenic
R484G	Neutral	Neutral	0.611	Benign
R66C	Disease	Neutral	0.301	Benign
T92K	Disease	Disease	0.639	Pathogenic
V277G	Disease	Disease	0.882	Pathogenic
V332G	Disease	Neutral	0.934	Benign
V48A	Disease	Neutral	0.765	Benign
V80G	Disease	Disease	0.746	Pathogenic
W146R	Disease	Disease	0.967	Pathogenic
W146S	Disease	Disease	0.97	Pathogenic
W353R	Disease	Disease	0.914	Pathogenic
W438G	Disease	Disease	0.908	Pathogenic
W450C	Disease	Neutral	0.773	Benign
Y138C	Disease	Disease	0.946	Pathogenic
Y175D	Disease	Disease	0.957	Pathogenic
Y210C	Disease	Disease	0.837	Pathogenic
Y266S	Disease	Disease	0.861	Pathogenic
Y86C	Disease	Disease	0.922	Pathogenic
Y86N	Disease	Disease	0.93	Pathogenic

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 17–24, 2026
Open Access © 2026 Editory Press

Elucidating impacts of deleterious Missense mutations on the structure and function of β-glucuronidase: Investigating the molecular basis of pathogenesis of MPSVII

Nushrat Jahan[1], Barka Basharat[1] and Shakilur Rahman[2,*]

[1]*Department of Biotechnology, School of Chemical and Life Sciences, Jamia Hamdard, New Delhi, India*
[2]*Biology Department, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11623, Saudi Arabia*
[*]*Corresponding author: Dr. Shakilur Rahman; Biology Department, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11623, Saudi Arabia; E-mail: srahman@imamu.edu.sa*

Abstract

Single-nucleotide polymorphisms (SNPs) are among the most prevalent forms of genetic variation and are frequently associated with human diseases. In this study, we performed a comprehensive in silico analysis of non-synonymous SNPs (nsSNPs) in the GUSB gene, which encodes the lysosomal enzyme β-glucuronidase, a key regulator of glycosaminoglycan degradation. A total of 449 reported mutations were systematically evaluated using sequence- and structure-based predictive algorithms. Among these, 15 variants were identified as deleterious and structurally destabilizing. Subsequent pathogenicity assessment using SNPs&GO, MutPred, and PhD-SNP further narrowed these to eight high-confidence pathogenic mutations. Solubility and aggregation propensity analysis using the SODA tool revealed that 37.5% of these pathogenic variants showed increased aggregation or reduced solubility, suggesting a potential mechanism contributing to disease progression. Detailed structural investigation indicated that the observed destabilization likely arises from alterations in interatomic non-covalent interactions, ultimately compromising protein stability and function. Overall, this study provides a systematic and integrative characterization of pathogenic nsSNPs in the GUSB gene. The findings enhance our understanding of the structural and functional consequences of these variants and offer a foundation for future experimental validation and the development of targeted therapeutic strategies.

Keywords: *β-glucuronidase, GUSB gene, nsSNPs, missense mutation, glycosaminoglycan, MPSVII, Sly syndrome, SODA, MutPred*

Received: May 1st, 2025 **Accepted:** September 30th, 2025 **Published:** January 5th, 2026

1. Introduction

β-glucuronidase (GUSB) is a crucial enzyme widely used in molecular biology and genetic engineering projects. (Yang et al., 2017). The housekeeping enzyme GUSB plays an active role in proteoglycan synthesis in lysosomes and is expressed in most tissues. It contributes significantly to the gradual decline in dermatan and keratan sulfates by enhancing the disintegration of GAGs during the fifth cycle. GUSB promotes the dispersion of active or inert compounds from glucuronides, thereby modifying the manner in which prodrugs respond and operate (Naz et al., 2013).

GUSB deficiency leads to mucopolysaccharidosis type VII, resulting in brain lysosomal storage (Kong et al., 2022). Humans have been shown to exhibit 11 distinct types (MPS I to IIIA, B, C, and IV to IX), which are divided according to the deficient enzyme (Hytönen et al., 2012). There are seven different types of Mucopolysaccharidosis (MPS), which are categorized based on defects in a certain number of the eight specific lysosomal metabolic enzymes (Khan et al., 2017). Mutations in the IDUA gene on chromosome 4p16 cause the first kind of MPS, which is caused by a deficiency of the enzyme a-l-iduronidase. A-l-iduronidase deficiency results in the accumulation of GAGs, such as dermatan and heparan sulphates (DS and HS), in many human tissues, leading to severe organ malfunction (Nagpal et al., 2022).

MPS II is linked to lysosomal dysfunction due to a deficiency of the enzyme iduronate 2-sulphatase, which degrades heparan sulphate and dermatan sulphate (DS). A rapidly developing neurodegenerative lysosomal storage disease is MPSIII. It is caused by biallelic variants in a gene encoding an enzyme in the digestive tract that degrades heparan sulfate. Neuronal inflammation and significant activation of astrocytes and microglia are hallmarks of MPS III, a neurological disease (Seker Yilmaz et al., 2021). MPS IVA, additionally referred to as Morquio syndrome type A, constitutes one of the most significant lysosomal illnesses. The lysosomal hydrolase N-acetylglucosamine-6-sulfate sulfatase (GALNS) is absent in this autosomal-recessive genetic disorder (Sawamoto et al., 2020).

There is nothing abnormal regarding the nervous system's functions, but common features include elevated blood and urine potassium (K+) levels, a petite stature, odontoid hypoplasia, pectus carinatum, kyphoscoliosis, genu valgum, joint laxity, and corneal clouding (Tomatsu et al., 2011).

Many lysosomal storage disorders, including mucopolysaccharidosis type VII (MPS VII), commonly referred to as Sly syndrome, are frequently linked to these mutations (**?**). The integrity and activity of β-glucuronidase are dramatically affected by missense, frameshift, and nonsense mutations in the GusB gene, according to recent investigations (**?**). GusB Gene encodes a 651-amino-acid-long homotetrameric protein with three distinctive domains in every monomer (Florindo et al., 2018). One such relationship involves the lysosomal-associated membrane protein 2 (LAMP2), which supports the stability and appropriate transit of β-glucuronidase inside lysosomes (Staudt et al., 2016). Furthermore, heparan sulfate and chondroitin sulfate are broken down by β-glucuronidase, which works in tandem with other enzymes such as arylsulfatase B and heparanase (Naz et al., 2013).

According to Urayama et al. (2004), it is imperative that GUSB reach its intended intracellular location via this targeting mechanism. The lysosomal route of GAG breakdown is the main pathway in which GUSB is involved. An essential stage in the breakdown of GAGs is the cleavage of the β-D-glucuronic acid residues by GUSB. GAG buildup is caused by a deficit in β-glucuronidase activity and results in lysosomal storage disorders (Hytönen et al., 2012). GAGs (glycosaminoglycans) are broken down by many lysosomal enzymes in an intricate procedure. The remaining molecules of β-D-glucuronic acid from GAGs such as heparan sulfate, dermatan sulfate, and chondroitin sulfate are particularly susceptible to hydrolysis by GUSB (Awolade et al., 2020).

GUSB utilizes its N-terminus domain to identify and adhere to a GAG chain, its substrate. The catalytic function of the enzyme in question depends on the initial binding (Urayama et al., 2004). The site that catalyzes the breakdown of β-D-glucuronic acid residues is located in the catalytic region of GUSB (Hytönen et al., 2012). After splitting, the active region of the enzyme discharges the resulting molecules, which are shorter oligosaccharides, enabling β-Glucuronidase to interact with a different substrate (Staudt et al., 2016). Supporting autophagy and lysosomal activity requires β-glucuronidase. It supports the prevention of cellular malfunction by facilitating the breakdown of autophagic organelles. The amino acids associated with LAMP2 maintain its continued existence within lysosomes and ensure a steady supply of functional GUSB for the degradation of lysosomal proteins (Staudt et al., 2016).

GUSB is involved in the processes of extracellular matrix reorganization and cell-mediated cascade modulation, in addition to lysosomal breakdown, by controlling the accessibility of functional glycosaminoglycan fragments (Wan et al., 2020). By interacting with receptors located on the cell surface, like integrins and growth factor receptors, GAG fragments can encourage cytoskeleton remodeling and improve cell motility (Casale & Crane, 2024).

The residues of breakdown may change the extracellular matrix's mechanical features and the substance, which may affect how cells engage with their surroundings (Wang et al., 2017). Pattern recognition receptors (PRRs) on immune cell populations identify particular glycosaminoglycan components as danger-associated molecular patterns (DAMPs), which then, in turn, cause inflammatory reactions (Garantziotis & Savani, 2022). About carcinoma, the breakdown components of glycosaminoglycan might affect the activity of the cancerous cells and the stromal cells that envelop them, therefore facilitating the growth, blood vessel development, and territorial expansion of the tumor (Hua et al., 2022). Prospective treatment options for recovering β-glucuronidase function in MPS VII patients include gene therapy and enzyme replacement therapy (ERT) (Grubb et al., 2010). Furthermore, the use of small-molecule chaperones to regulate or increase the efficiency of mutant β-glucuronidase has already been investigated (Doherty et al., 2023).

Single-nucleotide polymorphisms (SNPs), particularly non-synonymous SNPs (nsSNPs), are major contributors to genetic variability and are frequently implicated in inherited metabolic disorders. Despite numerous reported variants, a systematic approach to distinguish pathogenic mutations from benign polymorphisms remains limited. Given the increasing volume of genomic data generated through high-throughput sequencing, there is a critical need to prioritize functionally significant variants using reliable approaches. Identifying deleterious nsSNPs and understanding their structural and functional consequences can provide valuable insights into disease mechanisms, genotype–phenotype correlations, and molecular instability associated with protein dysfunction. Therefore, this study was undertaken to comprehensively analyze nsSNPs in the GUSB gene using integrative sequence- and structure-based bioinformatics tools. By predicting pathogenicity, assessing structural destabilization, and evaluating aggregation propensity, the work aims to bridge the gap between genetic variation data and mechanistic understanding. Ultimately, such insights may facilitate improved diagnostic interpretation, biomarker identification, and the development of targeted therapeutic interventions.

2. Materials and methods

2.1. Data retrieval and analysis

The GUSB sequence was extracted in FASTA format from the UniProt repository (UniProt ID: P08236). A collection of non-synonymous SNPs was generated using information obtained from the PubMed literature search and databases such as dbSNP (Sherry et al., 2001), HGMD (Stenson et al., 2009), ClinVar (Landrum et al., 2014), and Ensembl (Hubbard et al., 2002). The list was cleared to eliminate redundant nsSNPs. The protein data bank provided the crystal structure of the human GUSB gene. Out of the total Data, 1305 Missense variants along with 3'UTR and 5' UTR were obtained from dbSNP and Ensembl, and are shown in Figure 1. We used a comprehensive computational approach to anticipate harmful mutations in the GUSB at both structural and functional levels, as illustrated in Figure 2.

2.2. Sorting Intolerant from Tolerant

Sorting Intolerant from Tolerant (SIFT) was used to predict how possible amino acid alterations may affect protein function. SIFT is expanded to include predictions for frame-shifting indels. Human genetic research has extensively used SIFT to assess amino acid substitutions (e.g., cancer, Mendelian disorders, and viral infections). Studies of human diseases and other study subjects are not the only areas in which SIFT is useful. The consequences of missense mutations on model animals such as rats, dogs, and Arabidopsis, as well as agricultural plants, have been studied using SIFT (Sim et al., 2012). If the SIFT score is less than or equal to 0.05, then the mutation is not tolerable (Ng & Henikoff, 2003). The

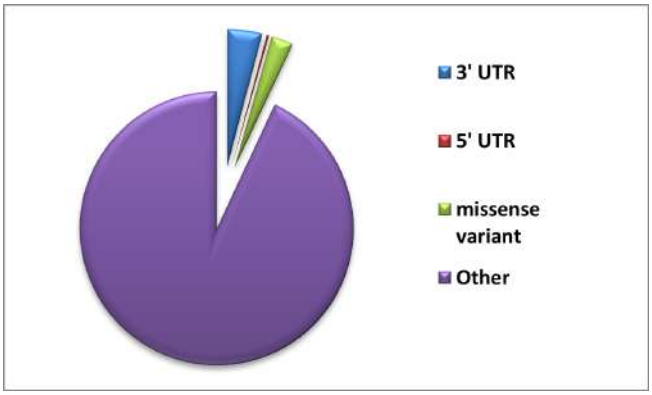

Figure 1. Number of SNPs in GusB represented using the dbSNP database.

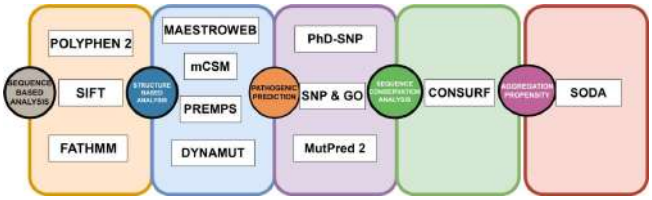

Figure 2. A recap of the computational approaches used to anticipate harmful mutations in the GUSB at both structural and functional levels.

SIFT method was utilized to forecast the impact of these nsSNPs on the protein.

2.3. Polymorphism Phenotyping v2

Polymorphism Phenotyping v2 (PolyPhen-2), a program and Web server accessible, estimates the potential effects of amino acid alterations on the stability and activity of human proteins. SNPs are functionally annotated; coding SNPs are mapped to gene transcripts; protein sequence annotations are extracted; and conservation profiles are constructed. The likelihood that the missense mutation would cause harm is then calculated using a combination of all these characteristics (Adzhubei et al., 2013). In addition to estimating the likelihood that a mutation is harmful when it is not, PolyPhen-2 also provides predictions of false-positive and true-positive rates. PolyPhen-2 computes the naive Bayes posterior probability that a particular mutation is harmful. The qualitative assessment of a mutation determines whether it is benign, perhaps harmful, or likely harmful (Adzhubei et al., 2010).

2.4. Mutation Assessor

Mutation Assessor (http://mutationassessor.org/r3/) is a sequence-based tool that predicts the functional impact of an nsSNPs on a protein. This server predicts the functional effects of missense polymorphisms and amino acid modifications observed in cancer-causing mutations in proteins. Depending on the affected amino acids and the degree of evolutionary preservation in protein homologs, the functional impact is evaluated (**?**). A sizable collection of polymorphic and disease-associated (OMIM) variants—60,000 in total has been used to verify the technique.

2.5. Function Analysis through Hidden Markov Models

Function Analysis through Hidden Markov Models (FATHMM) is a method that helps predict how both coding and non-coding variations in the human genome affect function. To determine if SNPs and other mutations tend to be damaging, hidden Markov models are employed. When it comes to genetic research and personalized therapy, FATHMM is highly beneficial, as it helps to clarify the significance of genetic variants (Shihab et al., 2013).

2.6. Mutation Cutoff Scanning Matrix

A computer method called Mutation Cutoff Scanning Matrix (mCSM) is used to predict how mutations may affect small-molecule binding, protein-protein interactions, protein-nucleic acid interactions, and protein stability. mCSM is an innovative tool for assessing non-synonymous mutations that predicts destabilizing mutations using a graph-based method. If a mutation results in a mCSM score ($\Delta\Delta G$) less than 0, it affects protein structure (Choudhury et al., 2021). It assists in forecasting the impact of mutations on drug binding, hence contributing to the development of more potent medications. It helps understand how changes in protein interactions and stabilization lead to conditions caused by mutations. Directs the creation of proteins for use in industry and medicine that have the appropriate stability and interaction characteristics (Pires et al., 2016).

2.7. MAESTRO

The free energy change on protein unfolding is estimated using MAESTRO (https://pbwww.che.sbg.ac.at/maestro/web), a multi-agent stability prediction tool. It determines how a point mutation affects the protein's stability by calculating the difference in free energy (ΔG) between the mutant and wild-type proteins. The stability of the protein is changed by a mutation if its score is less than zero (Laimer et al., 2015). Mutations may be limited to certain amino acid classes, exposed or buried residues, and user-specified areas, contingent upon the purpose of the research (Laimer et al., 2015).

2.8. PREMPS

PremPS calculates changes in the unfolding Gibbs free energy to assess the effects of single mutations on protein stability. This approach requires a protein's three-dimensional structure (Chen et al., 2020). The PREMPS software determines the mutation's corresponding change in free energy ($\Delta\Delta G$). PREMPS forecasts the impact of the mutation on protein stability based on the calculated $\Delta\Delta G$. Generally, a destabilizing mutation is indicated by a high $\Delta\Delta G$ value, whereas a stabilizing mutation is suggested by a negative $\Delta\Delta G$ (Chen et al., 2020).

2.9. DynaMut

DynaMut forecasts the effects of mutations on protein stability by considering MD simulation outcomes. The protein structure is stabilized or destabilized by a mutation based on the calculation of metrics such as changes in free energy ($\Delta\Delta G$). This prediction is essential for evaluating how mutations in biological systems will affect functionality (Rodrigues et al., 2018). PhD-SNP is a tool designed to help understand the genetic basis of illnesses. It analyzes SNPs by assessing their possible negative consequences using a variety of structural and evolutionary factors. A prediction score that indicates the probability of an SNP being deleterious is produced by PhD-SNP. A greater score denotes a higher likelihood that the SNP would result in functional alterations or increased susceptibility to illness, with a score range from 0 to 1 (Capriotti et al., 2006).

2.10. SNP & GO

A bioinformatics technique called SNP & GO (http://snps-and-go.biocomp.unibo.it/snps-and-go/) integrates data from both Gene Ontology (GO) and SNP annotations to forecast the

functional effects of SNPs. This tool facilitates understanding of genomic data in relation to disease and phenotype by illuminating the potential effects of genetic variants on protein function and biological mechanisms. The algorithms used here categorize SNPs as neutral or deleterious (likely to impair protein function) based on the attributes that were retrieved. SNP & GO produces prediction scores determined by the machine learning model that show how likely it is that an SNP will have negative effects. Higher scores indicate a greater likelihood of functional impact, which helps researchers choose SNPs for further study (Capriotti et al., 2013).

2.11. MutPred2

MutPred2 was created to predict how missense mutations in human proteins might affect protein function. It evaluates how mutations influence protein structure, function, and interactions by combining diverse features and machine learning methods, delivering significant insights into their potential roles in disease processes (Choudhury et al., 2021).

2.12. Solubility based on Disorder and Aggregation

A technique to determine the aggregation, disorder, helix, and strand propensities resulting from mutations is called Solubility based on Disorder and Aggregation (SODA) (http://protein.bio.unipd.it/soda/). The PDB format structure file or the protein sequence may be entered into this program. SODA uses Fells, ESpritz-NMR, PASTA 2.0, and other resources to forecast mutations of many types, including insertions, deletions, substitutions, and duplications. Based on how differently the WT and mutant proteins are soluble, SODA assigns a final score (Paladin et al., 2017).

2.13. Consurf Analysis

Consurf (https://consurf.tau.ac.il/) is a bioinformatics tool designed to find areas in protein sequences that have been conserved across time. It uses evolutionary links among various species and sequence homology to predict the structural and functional significance of amino acid residues. Using multiple sequence alignment, the ConSurf tool was used to assess residue conservation at a particular location (Ashkenazy et al., 2016).

3. Results and Discussion

A thorough investigation was carried out to obtain missense mutations of the GUSB gene using several well-known genetic databases, including Ensembl (http://www.ensembl.org/), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), HGMD (http://www.hgmd.cf.ac.uk), and dbSNP (http://www.ncbi.nlm.nih.gov/snp), HGMD (http://www.hgmd.cf.ac.uk), and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/). A total of 1305 missense mutations (nsSNPs) were found after an extensive search. Since certain nsSNPs may not have been indexed in these databases, a thorough PubMed literature review was conducted to ensure a more comprehensive collection. The primary objective of the present study was to conduct a comprehensive evaluation of all reported missense mutations in the GUSB gene. To understand their inheritance patterns and potential biological consequences, each variant was initially assessed at the sequence level. This approach enables a deeper understanding of how individual mutations may alter protein function and contribute to disease pathogenesis (**?**).

To ensure strong, high-confidence identification of deleterious variants, we adopted an integrative strategy combining sequence- and structure-based computational analyses. The sequence-based assessment was first performed using well-established tools, including SIFT, PolyPhen-2, Mutation Assessor (http://mutationassessor.org/r3/), and FATHMM. Because reliance on a single predictive algorithm can lead to false positives or false negatives, multiple tools were employed to improve prediction accuracy and minimize bias. SIFT classifies nsSNPs as tolerated or deleterious based on evolutionary conservation and amino acid physicochemical properties. Variants with lower tolerance index scores are predicted to have a greater functional impact on the protein (Ng & Henikoff, 2003). PolyPhen-2 evaluates the possible impact of amino acid substitutions on protein structure and function using comparative and structural features (Adzhubei et al., 2010, 2013). Mutation Assessor predicts the functional consequences of mutations by categorizing them into different impact levels (e.g., low, medium, or high) based on evolutionary conservation of affected residues (**?**). FATHMM further enhances prediction reliability by integrating sequence conservation with hidden Markov models to assess the likelihood of pathogenicity (Shihab et al., 2013).

Subsequently, structure-based tools such as MAESTROweb (https://pbwww.che.sbg.ac.at/maestro/web), PremPS, mCSM, and DynaMut were employed to evaluate the effects of mutations on protein stability, conformational dynamics, and intermolecular interactions. This combined framework provides a more reliable and comprehensive understanding of the structural and functional implications of nsSNPs in the GUSB gene (Chen et al., 2020; Pires et al., 2016; Rodrigues et al., 2018; Laimer et al., 2015).

3.1. Sequence-based predictions

Mutation Assessor categorizes variants according to their predicted functional impact. Moderate-impact mutations are expected to slightly alter protein stability or activity, whereas high-impact mutations are predicted to severely impair protein function, potentially leading to loss of activity or altered biological behavior. Similarly, FATHMM classifies variants into two principal categories based on their anticipated effects: tolerated and damaging. Tolerated variants are predicted to have little or no functional consequence and are likely benign or neutral. In contrast, damaging variants are expected to significantly disrupt protein stability, activity, or molecular interactions, thereby increasing their likelihood of being pathogenic (Shihab et al., 2015). These predictive frameworks enable researchers to prioritize high-risk variants for experimental validation and clinical investigation, particularly those identified as deleterious for disease-association and functional studies.

A total of 449 nsSNPs in the human GUSB gene were analyzed. Sequence-based predictions identified 277, 262, 160, and 192 deleterious mutations using SIFT, PolyPhen-2, FATHMM, and Mutation Assessor, respectively. Structural stability analysis revealed 13, 40, 1, and 11 destabilizing mutations as predicted by mCSM, MAESTROweb, PremPS, and DynaMut, respectively (Figure 3). Overall, this comprehensive multi-tool strategy strengthens confidence in variant prioritization and provides a reliable foundation for subsequent functional validation and disease-correlation studies (Supplementary Table S1).

3.2. Structure-based prediction

In addition to sequence-based analyses, four structure-based stability prediction tools DynaMut, PremPS, mCSM, and MAESTROweb were employed in this study. These programs estimate changes in folding free energy ($\Delta\Delta G$) using atomic

Figure 3. Distribution of deleterious and neutral nsSNPs predicted by sequence-based tools for the entire sequence of GUSB gene.

coordinates derived from the wild-type protein's PDB structure. Most of these tools integrate biophysical principles with machine-learning or neural-network–based approaches to evaluate the impact of amino acid substitutions on protein stability and conformational dynamics. To enhance prediction reliability and minimize false positives, only those variants consistently identified as deleterious by all four sequence-based tools and all four structure-based methods were selected for further analysis. This stringent integrative screening approach resulted in 15 high-confidence mutations predicted to be both deleterious and structurally destabilizing (Figure 6).

mCSM employs graph-based signatures to predict the impact of amino acid substitutions on protein stability by quantifying changes in folding free energy ($\Delta\Delta G$). Using this approach, mCSM identified 13 destabilizing mutations. MAESTROweb predicted 40 mutations as destabilizing, utilizing a combination of structural descriptors, statistical potentials, and energetic parameters to estimate the effects of mutations on protein stability and functionality. PremPS (Predictions of Protein Stability Changes upon Mutation) identified one destabilizing mutation. This tool integrates structural environmental features with probabilistic energy functions to assess the effects of single-point mutations on protein stability. DynaMut detected 11 destabilizing mutations by combining graph-based signatures with normal-mode analysis to assess their effects on both protein stability and conformational dynamics (Supplementary Table S2).

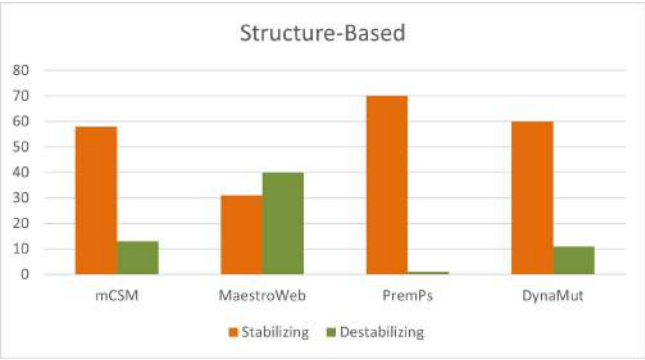

Figure 4. Distribution of destabilizing nsSNPs predicted by structure-based tools for the entire sequence of GUSB gene.

3.3. Identification of pathogenic nsSNPs

To further investigate the disease associations of the prioritized nsSNPs in the human GUSB gene, we employed three well-established pathogenicity prediction tools: PhD-SNP, SNPs&GO, and MutPred. These computational platforms classify variants as benign or pathogenic based on probability scores derived from sequence features, evolutionary conservation, and functional annotations, thereby enabling reliable assessment of disease association. From our integrative sequence- and structure-based analyses, 15 nsSNPs were identified as high-confidence deleterious and destabilizing variants. These shortlisted mutations were subsequently subjected to pathogenicity evaluation using the three prediction tools.

PhD-SNP classified 12 out of the 15 variants (80%) as pathogenic, indicating a strong likelihood of disease association. Similarly, SNPs&GO predicted 10 of the 15 variants (66.66%) to be disease-related, further supporting the potential clinical relevance of these mutations (Table 1). To enhance confidence and minimize false-positive predictions, we compared results from PhD-SNP and SNPs&GO. Eight variants were consistently identified as pathogenic by both tools, namely: L214P, L566P, N135D, P108L, R26L, T180P, Y388C, and Y399C. These eight mutations were therefore prioritized as the most probable disease-causing variants and selected for subsequent detailed analysis.

Table 1. Disease phenotype analysis of high-confidence nsSNPs in the GUSB gene.

S. No.	Mutations	PhD-SNP	SNP & GO	MutPred2
1	I499M	Neutral	Neutral	0.62
2	L121V	Neutral	Neutral	0.77
3	L214P	Disease	Disease	0.953
4	L566P	Disease	Disease	0.888
5	N135D	Disease	Disease	0.815
6	P108L	Disease	Disease	0.8
7	P196H	Neutral	Disease	0.701
8	R36L	Disease	Disease	0.943
9	R398H	Disease	Neutral	0.561
10	R56C	Disease	Neutral	0.361
11	T180P	Disease	Disease	0.843
12	T226P	Disease	Disease	0.497
13	V601G	Disease	Neutral	0.545
14	Y388C	Disease	Disease	0.785
15	Y399C	Disease	Disease	0.567

3.4. Analysis of evolutionarily conserved residues

Analysis of amino acid conservation within a protein structure provides critical insight into the functional and structural importance of specific residues and reflects the evolutionary constraints that shape them. Highly conserved residues are often essential for maintaining structural integrity, catalytic activity, or molecular interactions. In general, the likelihood of mutation tolerance is inversely related to the degree of conservation; residues that are strongly conserved across species are less likely to accommodate substitutions without functional consequences.

To evaluate evolutionary conservation in the human GUSB protein, we performed a ConSurf analysis (Figure 5). The results indicated that residues within the regions 360–370, 380–390, 415–425, and 566–571 exhibit a high degree of conservation compared to other segments of the protein. These conserved regions likely play critical roles in maintaining structural stability and functional competence. Furthermore, ConSurf classification revealed that several of these highly conserved residues are buried within the protein core, suggesting a structural role in maintaining stability, while others are exposed on the protein surface,

potentially contributing to functional interactions. These findings reinforce the biological significance of mutations occurring within conserved regions of the GUSB protein.

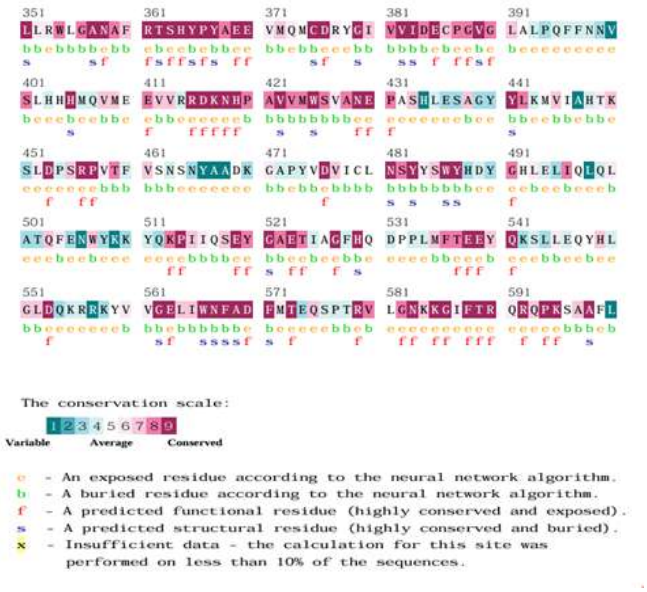

Figure 5. Conserved region of *GusB*.

3.5. Analysis of aggregation propensity

Amino acid solubility plays a critical role in maintaining proper protein folding and function (Balch et al., 2008). Previous studies have demonstrated that reduced protein solubility can promote misfolding and aggregation, which are central mechanisms underlying several neurodegenerative and aggregation-related disorders (Knowles et al., 2014). Insoluble protein regions tend to cluster, leading to pathological aggregates associated with diseases such as Parkinson's disease, Alzheimer's disease, and amyloidosis (Thal et al., 2015).

SNPs are the most common form of genetic variation and are widely associated with diverse human diseases. Comprehensive analysis of these variants provides valuable insight into the molecular mechanisms underlying disease progression and may facilitate the development of targeted therapeutic strategies. In this study, we focused on nsSNPs in the GUSB gene, a critical lysosomal hydrolase responsible for glycosaminoglycan degradation.

To evaluate the effects of prioritized variants on the solubility of the GUSB protein, we used the SODA tool. SODA provides a comprehensive assessment by estimating mutation-induced changes in aggregation propensity, intrinsic disorder, helix propensity, and strand propensity. This integrative analysis enables a deeper understanding of how specific amino acid substitutions may influence protein folding behavior and aggregation potential (Choudhury et al., 2021).

From our earlier pathogenicity assessment, eight nsSNPs were identified as disease-associated variants (Table 2). These eight mutations were subsequently analyzed using SODA to determine their impact on protein solubility. Among them, three variants were predicted to reduce protein solubility. Decreased solubility is typically associated with increased aggregation propensity, which may contribute to disease pathogenesis by forming aberrant protein assemblies. These findings suggest that these three variants could potentially promote aggregation-mediated dysfunction of the GUSB protein.

In contrast, the remaining five nsSNPs were predicted to enhance protein solubility. Increased solubility generally reflects a reduced tendency to aggregate and may help maintain a more stable, functionally competent protein conformation. Such mutations may therefore have a comparatively lower risk of inducing aggregation-related pathogenic effects. The analysis demonstrated that three of the eight pathogenic mutations (37.5%) are predicted to decrease protein solubility and increase aggregation tendency. This observation is particularly significant, as protein aggregation is a well-established contributor to several pathological conditions, including neurodegenerative and aggregation-related disorders. In contrast, the remaining variants were predicted to maintain or enhance solubility, suggesting comparatively lower aggregation risk.

Overall, the SODA analysis expands our understanding of how specific GUSB variants may influence protein solubility and aggregation behavior. This comprehensive computational analysis highlights a subset of GUSB nsSNPs with strong evidence of deleterious, destabilizing, and pathogenic effects. By integrating assessments of stability, pathogenicity, and solubility, the study provides a detailed molecular perspective on how specific variants may impair β-glucuronidase function and contribute to disease pathogenesis.

Table 2. Prediction of aggregation propensity of mutant GUSB using SODA server.

Mutations	Sequence	Helix	Strand	Aggregation	Disorder	SODA	Result
Wild type		0.281	0.296	-5.254	0.024		
L214P	1	-1.351	0.165	18.495	0.024	16.507	More soluble
L566P	2	-2.1	0.568	5.083	0.034	2.178	More soluble
N135D	3	2.394	-2.928	116	0.003	115.555	More soluble
P108L	4	0.707	-0.875	12.252	0.004	11.977	More soluble
R36L	5	-2.585	2.598	-0.668	-0.445	-2.167	Less soluble
T180P	6	-0.248	-0.335	17.006	0.149	16.553	More soluble
Y388C	7	-1.912	1.187	-0.577	-0.033	-1.992	Less soluble
Y399C	8	0.028	0.82	-22.711	-0.069	-21.467	Less soluble

To further evaluate the pathogenic potential of the prioritized variants, we employed three robust and widely validated bioinformatics tools, SNPs&GO, MutPred, and PhD-SNP, recognized for their reliability in predicting the disease association of SNPs. These analyses consistently identified eight mutations: L214P, L566P, N135D, P108L, R26L, T180P, Y388C, and Y399C as highly deleterious, suggesting significant implications for protein stability, functional integrity, and disease susceptibility.

To gain additional structural and evolutionary insight, we performed conservation analysis using the ConSurf server, which evaluates the evolutionary conservation of amino acid residues within a protein sequence (Ashkenazy et al., 2016). The results revealed that two of the eight pathogenic mutations, L566P and Y388C, are located within highly conserved regions of the protein. Such strong evolutionary conservation typically reflects critical structural or functional importance. Therefore, substitutions at these positions are likely to exert substantial effects on protein stability and activity, further supporting their potential pathogenic role.

3.6. Analyzing Structures of Specific Mutations

To gain deeper insight into the molecular consequences of the identified pathogenic variants, we performed a detailed structural and conformational analysis of the GUSB protein, with particular emphasis on residues 566 and 388. A comparative evaluation of

the wild-type and mutant structures revealed notable alterations in intermolecular interactions, highlighting the structural impact of these substitutions.

At position 566, the wild-type (Leu) protein exhibited two conventional hydrogen bonds, whereas the mutant (Pro) form retained only one, indicating a reduction in stabilizing interactions (Figure **??**A). Two weak hydrogen bonds were preserved in both forms, and neither structure displayed water-mediated hydrogen bonds. Additionally, halogen bonds, ionic interactions, and metal-coordinated contacts were absent in both the wild-type and mutant configurations.

More pronounced differences were observed in non-covalent stabilizing interactions. Aromatic interactions, which play a critical role in maintaining structural stability, were present in the wild-type structure (eight interactions) but were completely absent in the mutant. Similarly, hydrophobic interactions, essential for maintaining protein core integrity, were dramatically reduced from 17 in the wild type to only 1 in the mutant. No carbonyl interactions were detected in either form.

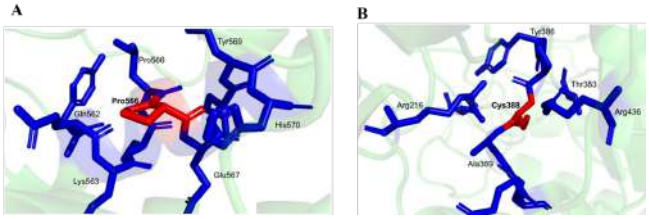

Figure 6. Analysis of Specific mutation (A). Leucine residue 566 mutated to Proline and (B). Tyrosine to Cysteine at position 388.

A detailed structural comparison was also performed for residue Tyr388Cys to evaluate the conformational consequences of the mutation. The interaction profile of the wild-type and mutant structures revealed substantial differences, highlighting the destabilizing nature of this substitution. In the wild-type protein, residue 388 formed two conventional hydrogen bonds, whereas only one hydrogen bond was retained in the mutant form, indicating a reduction in stabilizing interactions. Both structures exhibited two weak hydrogen bonds, and no water-mediated hydrogen bonds were detected in either configuration. Additionally, ionic interactions, halogen bonds, and metal-coordination interactions were absent in both the wild-type and mutant structures.

More striking differences were observed in non-covalent stabilizing interactions. The wild-type residue participated in eight aromatic interactions, all of which were completely lost in the mutant structure. Given that aromatic contacts contribute significantly to maintaining tertiary structural stability, their absence suggests a marked weakening of the local structural framework. Similarly, hydrophobic interactions were dramatically reduced from 17 in the wild type to only one in the mutant. As hydrophobic contacts are essential for preserving protein core integrity and conformational stability, this substantial reduction indicates severe structural disruption. No carbonyl interactions were observed in either form.

Overall, the mutation at residues 388 and 566 results in pronounced alterations in hydrogen bonding, aromatic stacking, and hydrophobic interactions, key forces that govern protein stability and structural coherence. These changes strongly suggest that the mutation compromises the equilibrium and functional integrity of the GUSB protein.

4. Conclusions

This study provides a comprehensive characterization of pathogenic nsSNPs in the GUSB gene and elucidates their potential molecular consequences. By systematically identifying deleterious, structurally destabilizing, and disease-associated mutations, we gain critical insight into the mechanistic basis of GUSB-related disorders. The integration of aggregation and solubility analyses further underscores the role of protein solubility in disease progression and highlights how specific amino acid substitutions can promote structural instability and functional impairment. The findings of this study may contribute to the development of targeted therapeutic strategies to mitigate the effects of harmful mutations. Potential approaches include designing small molecules or pharmacological chaperones that stabilize the mutant protein, enhance its solubility, or prevent aberrant aggregation. Moreover, understanding mutation-induced structural alterations may facilitate the rational design of precision therapeutics that restore lost enzymatic function or compensate for structural deficiencies. In conclusion, this integrative analysis of nsSNPs in the GUSB gene advances our understanding of the genetic and structural determinants of disease. By combining sequence-based prediction, structural stability assessment, pathogenicity evaluation, and solubility analysis, this work establishes a robust framework for variant prioritization and lays the foundation for future experimental validation and therapeutic development.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This work received no funding.

Data Availability Statement

All data generated or analyzed during this study are included in this manuscript.

Declaration on the Use of AI Tools

The authors declare that ChatGPT (OpenAI) was used solely to refine the language, improve grammar, and enhance the clarity of the manuscript.

References

Adzhubei, I., Jordan, D., & Sunyaev, S. 2013, Current Protocols in Human Genetics, Chapter 7, Unit7.20, doi: 10.1002/0471142905.hg0720s76

Adzhubei, I., Schmidt, S., Peshkin, L., et al. 2010, Nature Methods, 7, 248, doi: 10.1038/nmeth0410-248

Ashkenazy, H., Abadi, S., Martz, E., et al. 2016, Nucleic Acids Research, 44, W344, doi: 10.1093/nar/gkw408

Awolade, P., Cele, N., Kerru, N., et al. 2020, European Journal of Medicinal Chemistry, 187, 111921, doi: 10.1016/j.ejmech.2019.111921

Balch, W., Morimoto, R., Dillin, A., & Kelly, J. 2008, Science, 319, 916, doi: 10.1126/science.1141448

Capriotti, E., Calabrese, R., & Casadio, R. 2006, Bioinformatics, 22, 2729, doi: 10.1093/bioinformatics/btl423

Capriotti, E., Calabrese, R., Fariselli, P., et al. 2013, BMC Genomics, 14, S6, doi: 10.1186/1471-2164-14-S3-S6
Casale, J., & Crane, J. 2024, in StatPearls (Treasure Island, FL: StatPearls Publishing)
Chen, Y., Lu, H., Zhang, N., et al. 2020, PLoS Computational Biology, 16, e1008543, doi: 10.1371/journal.pcbi.1008543
Choudhury, A., Mohammad, T., Samarth, N., et al. 2021, Scientific Reports, 11, 10202, doi: 10.1038/s41598-021-89450-7
Doherty, G., Ler, G., Wimmer, N., et al. 2023, ChemBioChem, 24, e202200619, doi: 10.1002/cbic.202200619
Florindo, R., Souza, V., Mutti, H., et al. 2018, New Biotechnology, 40, 218, doi: 10.1016/j.nbt.2017.08.012
Garantziotis, S., & Savani, R. 2022, American Journal of Physiology - Cell Physiology, 323, C202, doi: 10.1152/ajpcell.00088.2022
Grubb, J., Vogler, C., & Sly, W. 2010, Rejuvenation Research, 13, 229, doi: 10.1089/rej.2009.0920
Hua, S., Viera, M., Yip, G., & Bay, B. 2022, Cancers, 15, 266, doi: 10.3390/cancers15010266
Hubbard, T., Barker, D., Birney, E., et al. 2002, Nucleic Acids Research, 30, 38, doi: 10.1093/nar/30.1.38
Hytönen, M., Arumilli, M., Lappalainen, A., et al. 2012, PLoS ONE, 7, e40281, doi: 10.1371/journal.pone.0040281
Khan, S., Peracha, H., Ballhausen, D., et al. 2017, Molecular Genetics and Metabolism, 121, 227, doi: 10.1016/j.ymgme.2017.05.016
Knowles, T., Vendruscolo, M., & Dobson, C. 2014, Nature Reviews Molecular Cell Biology, 15, 384, doi: 10.1038/nrm3810
Kong, X., Zheng, Z., Song, G., et al. 2022, Frontiers in Immunology, 13, doi: 10.3389/fimmu.2022.876048
Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., & Lackner, P. 2015, BMC Bioinformatics, 16, 116, doi: 10.1186/s12859-015-0548-6
Landrum, M., Lee, J., Riley, G., et al. 2014, Nucleic Acids Research, 42, D980, doi: 10.1093/nar/gkt1113
Nagpal, R., Goyal, R., Priyadarshini, K., et al. 2022, Indian Journal of Ophthalmology, 70, 2249, doi: 10.4103/ijo.IJO_425_22
Naz, H., Islam, A., Waheed, A., et al. 2013, Rejuvenation Research, 16, 352, doi: 10.1089/rej.2013.1407
Ng, P., & Henikoff, S. 2003, Nucleic Acids Research, 31, 3812, doi: 10.1093/nar/gkg509
Paladin, L., Piovesan, D., & Tosatto, S. 2017, Nucleic Acids Research, 45, doi: 10.1093/nar/gkx412
Pires, D., Blundell, T., & Ascher, D. 2016, Scientific Reports, 6, 29575, doi: 10.1038/srep29575
Rodrigues, C., Pires, D., & Ascher, D. 2018, Nucleic Acids Research, 46, W350, doi: 10.1093/nar/gky300
Sawamoto, K., Álvarez González, J., Piechnik, M., et al. 2020, International Journal of Molecular Sciences, 21, 1517, doi: 10.3390/ijms21041517
Seker Yilmaz, B., Davison, J., Jones, S., & Baruteau, J. 2021, Journal of Inherited Metabolic Disease, 44, 129, doi: 10.1002/jimd.12316
Sherry, S., Ward, M., Kholodov, M., et al. 2001, Nucleic Acids Research, 29, 308, doi: 10.1093/nar/29.1.308
Shihab, H., Gough, J., Cooper, D., et al. 2013, Human Mutation, 34, 57, doi: 10.1002/humu.22225
Shihab, H., Rogers, M., Gough, J., et al. 2015, Bioinformatics, 31, 1536, doi: 10.1093/bioinformatics/btv009
Sim, N.-L., Kumar, P., Hu, J., et al. 2012, Nucleic Acids Research, 40, W452, doi: 10.1093/nar/gks539
Staudt, C., Puissant, E., & Boonen, M. 2016, International Journal of Molecular Sciences, 18, 47, doi: 10.3390/ijms18010047
Stenson, P., Mort, M., Ball, E., et al. 2009, Genome Medicine, 1, 13, doi: 10.1186/gm13
Thal, D., Walter, J., Saido, T., & Fändrich, M. 2015, Acta Neuropathologica, 129, 167, doi: 10.1007/s00401-014-1375-y
Tomatsu, S., Montaño, A., Oikawa, H., et al. 2011, Current Pharmaceutical Biotechnology, 12, 931, doi: 10.2174/138920111795542615
Urayama, A., Grubb, J., Sly, W., & Banks, W. 2004, Proceedings of the National Academy of Sciences of the United States of America, 101, 12658, doi: 10.1073/pnas.0405042101
Wan, L., Zhang, S., Li, S., et al. 2020, Thrombosis and Haemostasis, 120, 647, doi: 10.1055/s-0040-1705117
Wang, M., Liu, X., Lyu, Z., et al. 2017, Colloids and Surfaces B: Biointerfaces, 150, 175, doi: 10.1016/j.colsurfb.2016.11.022
Yang, G., Ge, S., Singh, R., et al. 2017, Drug Metabolism Reviews, 49, 105, doi: 10.1080/03602532.2017.1293682

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 25–32, 2026
Open Access © 2026 Editory Press

Identification of Potent Inhibitors of Ephrin type-A receptor 1 using Virtual Screening and Molecular Dynamics Simulation Approach

Shanza Rehman[1], Sana Kauser[2], Syed Suhail Andrabi[3] and Nida Mubin[4,*]

[1]*Department of Computer Science, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India*
[2]*Department of Biosciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India*
[3]*Biomedical Engineering, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA*
[4]*Division of Haematology & Oncology, Case Western Reserve University, Cleveland, OH, USA*
[*]*Corresponding author: Nida Mubin; Division of Haematology & Oncology, Case Western Reserve University, Cleveland, OH, USA; E-mail: nxm793@case.edu*

Abstract

Ephrin type-A receptor 1 (EphA1) is a receptor tyrosine kinase implicated in tumor progression, angiogenesis, metastasis, and immune regulation, making it an attractive therapeutic target in cancer. In the present study, a structure-based virtual screening approach was employed to identify potential natural inhibitors of EphA1. A library of 90,000 natural compounds retrieved from the ZINC database was filtered using Lipinski's rule of five, yielding 32,901 drug-like molecules for docking analysis. The three-dimensional structure of EphA1 was obtained from the AlphaFold database, and molecular docking was performed using InstaDock. Based on binding affinity scores (−11.2 to −9.4 kcal/mol), the top 320 compounds were shortlisted and further subjected to PAINS filtering, physicochemical evaluation, ADMET prediction, carcinogenicity assessment, and PASS analysis. Two compounds, ZINC12660859 and ZINC12661003, demonstrated favourable drug-likeness, high gastrointestinal absorption, non-carcinogenicity, and predicted antineoplastic activity. Interaction analysis revealed that these compounds bind within the ATP-binding pocket of the kinase domain, forming key interactions with residues such as Lys656 and Asp749. Molecular dynamics simulations over 100 ns of the EphA1–ZINC12660859 complex confirmed structural stability, with stable RMSD (0.26 nm), consistent radius of gyration (1.93 nm), minimal SASA variation, and sustained hydrogen bonding throughout the simulation. Overall, ZINC12660859 exhibited strong binding affinity and dynamic stability, suggesting its potential as a lead candidate for EphA1-targeted anticancer therapeutics.

Keywords: *EphA1, receptor tyrosine kinase, virtual screening, molecular docking, molecular dynamics simulation, natural compounds, ADMET, anticancer drug discovery*

Received: June 15th, 2025 **Accepted:** October 15th, 2025 **Published:** January 5th, 2026

1. Introduction

Receptor tyrosine kinases (RTKs) are critical regulators of cellular communication and play essential roles in cell proliferation, differentiation, migration, and survival. Among RTKs, the Eph receptor family is the largest subgroup and is subdivided into EphA and EphB classes based on ligand specificity and sequence homology Wu et al. (2022). Eph receptors interact with membrane-bound ephrin ligands to initiate bidirectional signaling pathways that regulate tissue patterning, angiogenesis, immune responses, and neuronal development. Ephrin type-A receptor 1 (EphA1), encoded by the *EPHA1* gene located on chromosome 7q34, was first identified in hepatocellular carcinoma cells Wu et al. (2022). EphA1 plays diverse roles in fundamental cellular processes such as cell movement and attachment Yamazaki et al. (2009), regulation of immune response and inflammation Hjorthaug & Aasheim (2007), cancer progression, nervous system development, and angiogenesis Wu et al. (2022). During development, EphA1 is involved in blood vessel remodelling, axon guidance, the spatial organisation of diverse cell populations, and the creation of synaptic connections between neurons Lisabeth et al. (2013). Ephrin-A1 promotes directional mobility of $CD8^{+}CCR7^{+}$ T lymphocytes, a type of T cell implicated in immunological responses, and thus regulates immunological processes such as lymph node homing and antigen presentation Hjorthaug & Aasheim (2007). EphA1–ILK signalling regulates cell adhesion dynamics and cytoskeletal organisation, both of which are required for cellular functions such as migration and tissue formation Yamazaki et al. (2009).

EphA1 increases cell division and protects cells in bovine mammary epithelial cells by reducing ER stress, inhibiting inflammatory responses, and minimizing cellular damage caused by inflammation Kang et al. (2018). EphA1 also contributes to neuroinflammation through the CXCL12/CXCR4 signalling pathway and is considered an important drug target for Parkinson's disease Ma et al. (2021) and Alzheimer's disease Villegas-Llerena et al. (2016). EphA1 is differentially expressed

in different types of cancer and is therefore associated with tumour growth, tumour angiogenesis, invasion, prognosis, and metastasis Ieguchi & Maru (2019). EphA1 is overexpressed in thymoma Yu et al. (2019), prostate cancer Peng et al. (2013), hepatocellular carcinoma Wang et al. (2016), ovarian cancer Adu-Gyamfi et al. (2021), breast cancer Liang et al. (2021), clear cell renal cell carcinoma Toma et al. (2014), gastric carcinoma Wang et al. (2020), nasopharyngeal carcinoma Dai & Zhang (2020), and oesophageal squamous cell carcinoma Wang et al. (2013), while it is downregulated in colorectal cancer Wu et al. (2016) and uveal melanoma Gajdzis et al. (2020).

EphA1 has diverse functions across diseases, providing important insights into underlying mechanisms and revealing potential therapeutic targets. EphA1 has been linked to tumour development and metastasis by promoting angiogenesis, cell migration, and epithelial–mesenchymal transition Ieguchi & Maru (2019). EphA1 has also been associated with epilepsy, Parkinson's disease Ma et al. (2021), and Alzheimer's disease due to its dysregulation in synaptic plasticity and neuroinflammatory processes Ieguchi & Maru (2019). EphA1's role in angiogenesis also makes it relevant to diseases such as diabetic retinopathy and age-related macular degeneration. Disruption of EphA1 signalling has been linked to age-related macular degeneration and neovascularization in diabetic retinopathy Li et al. (2017). EphA1 regulates cytokine production, immune cell activation, and migration in immune-related disorders. Autoimmune diseases, including rheumatoid arthritis and inflammatory bowel disease, have also been linked to EphA1 dysregulation Hjorthaug & Aasheim (2007). Given its central role in tumour progression and angiogenesis, EphA1 represents an attractive therapeutic target. Inhibition of the ATP-binding pocket within the kinase domain offers a rational strategy to suppress EphA1-mediated oncogenic signalling. However, despite its biological importance, selective small-molecule inhibitors targeting EphA1 remain limited.

Structure-based drug design (SBDD) is a powerful strategy employed in drug development that uses the three-dimensional structure of a protein target to design small molecules with high affinity and specificity Yu & MacKerell (2017). With the potential to develop medicines with increased efficacy and fewer adverse effects, SBDD has become an essential approach in modern drug discovery. Virtual screening, a computational technique that efficiently identifies prospective lead compounds from large chemical libraries, is a key component of SBDD. By screening numerous compounds computationally, virtual screening allows the rapid identification of promising candidates for further experimental validation. This approach significantly streamlines the drug development process compared with traditional trial-and-error methods, thereby saving both time and cost.

In the present study, we employed an integrative computational strategy combining large-scale virtual screening, molecular docking, drug-likeness and ADMET profiling, PASS prediction, and molecular dynamics (MD) simulations to identify potential natural inhibitors of EphA1. A curated library of natural compounds from the ZINC database was screened against the EphA1 kinase domain to identify high-affinity binders targeting the ATP-binding site. The most promising candidates were further evaluated for pharmacokinetic suitability and dynamic stability to identify potential lead molecules for EphA1-targeted anticancer therapy.

2. Materials and Methods

2.1. Web Resources and the Computing Environment

The study was conducted on a Windows-based laptop equipped with a 1.19 GHz CPU, 8 GB RAM, and a 512 GB SSD. Throughout the research, a reliable power source and a high-speed internet connection were available. InstaDock Mohammad et al. (2021) and Discovery Studio were utilized for virtual screening and molecular docking, while PyMOL DeLano (2002) and Discovery Studio Visualizer Biovia (2017) were used for visualization. Various resources and servers such as NCBI, UniProt UniProt Consortium (2019), the RCSB Protein Data Bank Rose et al. (2017), AlphaFold David et al. (2022), the ZINC database Irwin & Shoichet (2005); Irwin et al. (2020), SwissADME Daina et al. (2017), pkCSM Pires et al. (2015), Way2Drug PASS Lagunin et al. (2000), and CarcinoPred-EL Zhang et al. (2017) were utilized for data retrieval, evaluation, and analysis in this study. GROMACS was used to perform molecular dynamics simulations. The various systematic steps employed in this study are depicted in Figure **??**.

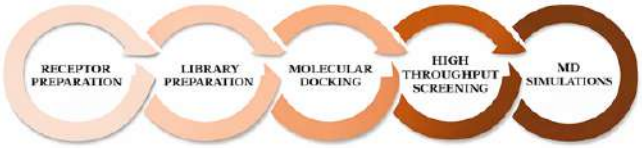

Figure 1. Steps involved in methodology

2.2. Receptor and Library Preparation

The EphA1 structure was selected and retrieved from UniProt and AlphaFold based on structural completeness, the presence of the kinase domain, mutation-related sequence information, ligand-related information, and structural purity. The structure of *EPHA1* was obtained in PDB format from the AlphaFold Protein Structure Database David et al. (2022); UniProt Consortium (2019) and further refined in PyMOL DeLano (2002) (Figure 2). Water molecules were removed during receptor preparation.

The ZINC database is a freely available public resource that compiles annotated, commercially available chemical structures Irwin & Shoichet (2005); Irwin et al. (2020). Both 2D and 3D structures of compounds can be downloaded from this database, and its interface supports rapid molecular lookup and analogue searching. ZINC maintains a catalogue of more than 230 million ready-to-dock 3D compounds for sale and also provides access to more than 750 million purchasable compounds, enabling rapid analogue searches Irwin et al. (2020). The ZINC database has expanded substantially, and ZINC20 now contains 1.4 billion chemicals from 310 catalogues and 150 suppliers Irwin et al. (2020). According to the 90/90/90 standard, 90% of catalogues are updated every 90 days, and 90% of compounds have been verified as available for purchase within the preceding three months Irwin et al. (2020).

In the present study, the Lipinski rule of five was applied to select 32,901 small molecules from the ZINC database. The Lipinski rule is a cornerstone of medicinal chemistry and provides a set of threshold-based parameters to assess the drug-likeness of chemical compounds. Drug-like molecules generally satisfy the following criteria: molecular weight below 500 Da, fewer than 5 hydrogen-bond donors, fewer than 10 hydrogen-bond acceptors, and a LogP value below 5.

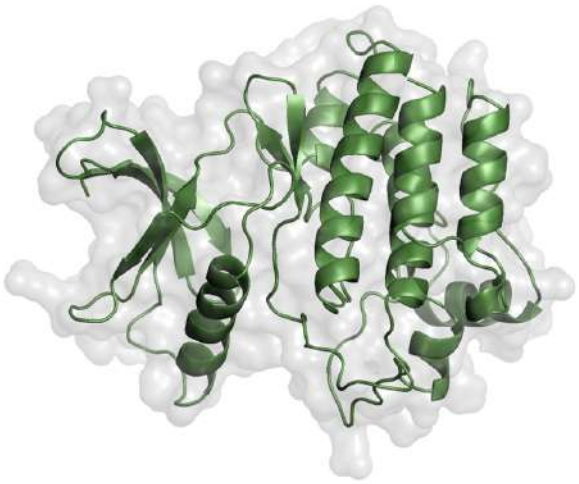

Figure 2. Kinase domain of EphA1. The structure was prepared in PyMOL.

2.3. Molecular Docking and Virtual Screening

To investigate substrate inhibitor selectivity and understand the mechanism involved, high-throughput screening was conducted using the receptor-ligand docking method for the target AF-P21709-F1. The purpose was to examine the orientation of ligands within the protein's active site cavity. A collection of natural compounds was sourced from the ZINC database, and the ligands were acquired in pdbqt format. Molecular docking of EPHA1 was conducted to examine the bond conformations and binding affinity between the ligands and EPHA1. Regarding the docking process, a defined energy grid box with the dimensions of grid box centre (x = -1.511, y = 1.922, z = -35.295) and size parameters (X=56, Y=63, and Z=75) was used to ensure maximum binding affinity and obtain the best conformational pose for the protein-ligand interactions. The docking algorithm investigated different ligand orientations within the receptor's active site. Each ligand screening used a random seed generator and a scoring function to evaluate binding affinity.

After the virtual screening was completed, the docking results were analysed by examining the output files and log files. The most appropriate docked conformation was chosen for subsequent analysis. The PyMOL software DeLano (2002) and Discovery Studio Visualizer Biovia (2017) were employed to visualize and analyse the structure of the docked complex. Using the data as a starting point, we next went on to evaluate the compounds' potential as drugs utilising the Swiss-ADME server (http://www.swissadme.ch/) for PAINs analysis, Carcinogenicity using CarcinoPred-EL. Following this, we further evaluated the physicochemical properties, absorption, distribution, metabolism, and excretion of selected compounds using PCKCSM (http://biosig.unimelb.edu.au/pkcsm). Lastly, PASS analysis was performed using way2drug (http://way2drug.com/passonline/), where ligand activities with probabilities of activation greater than 0.71 were considered.

2.4. Drug-Likeness, PAINS, and ADMET Analysis

The shortlisted compounds were subjected to pan-assay interference compounds (PAINS) filtering using the SwissADME server to eliminate false positives. Additional physicochemical filters were applied based on total polar surface area (TPSA: 65–145 Å^2) and solubility predictions. ADMET properties, including absorption, distribution, metabolism, excretion, and toxicity, were predicted using the pkCSM server. Gastrointestinal absorption, blood-brain barrier permeability, CYP2D6 inhibition, OCT2 substrate status, and AMES toxicity were evaluated. Carcinogenicity prediction was performed using CarcinoPred-EL.

2.5. PASS Analysis and Protein–Ligand Interaction Analysis

Prediction of Activity Spectra for Substances (PASS) analysis was performed using the Way2Drug PASS Online server to evaluate the potential biological activities of selected compounds. Activities with a probability of activity (Pa) greater than the probability of inactivity (Pi), particularly antineoplastic activity, were considered significant. Detailed interaction analysis of selected protein–ligand complexes was performed using Discovery Studio Visualizer. Hydrogen bonds, hydrophobic interactions, van der Waals interactions, and interactions with key catalytic residues within the ATP-binding pocket were examined. Special emphasis was placed on interactions with critical residues in the kinase domain, such as Lys656 and Asp749.

2.6. MD simulations

MD simulations were performed for 100 nanoseconds on both free EphA1 and EphA1 with the compound ZINC12660859 using GROMACS 5.1.2 at 300 Kelvin. The simulations utilized the GROMACS G54a7 force field at the molecular mechanics level. The systems were immersed in a cubic box of water (10 Å) and solvated with the spc216 water model using the gmx solvate module. Energy minimization was performed using 1500 steepest-descent steps. During the equilibration phase, the temperature gradually increased from 0 to 300 Kelvin over 100 picoseconds, while maintaining constant volume and periodic boundary conditions. Trajectory analysis employed various GROMACS utilities, including gmx rmsd for root-mean-square deviation, gmx rmsf for root-mean-square fluctuation, gmx gyrate for radius of gyration, gmx sasa for solvent-accessible surface area, and gmx hydrogen-bonds for intra- and intermolecular hydrogen bonds. Graphs and figures were generated using QtGrace.

3. Results and Discussion

3.1. Virtual Screening and Binding Affinity Analysis

We conducted virtual screening of compounds to identify a competitive inhibitor for EphA1. Log files and out-files were created, containing affinity scores and docked poses for each compound in the directory. Based on binding affinities, docking scores, and binding poses, we filtered compounds using the log and output files. Our screening identified several natural compounds with favourable binding affinity scores for EphA1's binding pocket, indicating their potential as EphA1 inhibitors. Out of 32,901 compounds that were screened from the output, 320 were found to have significant EphA1 binding affinity scores ranging from −11.2 to −9.4 (Table 1). Among the screened compounds, ZINC03845566 exhibited the highest binding affinity (−11.2 kcal/mol), while ZINC12660859 and ZINC12661003 demonstrated binding energies of −9.9 and −9.7 kcal/mol, respectively. Although several compounds showed slightly stronger docking scores, subsequent pharmacokinetic filtering and biological activity prediction narrowed the focus to ZINC12660859 and ZINC12661003 due to their favorable drug-like properties and predicted antineoplastic potential. The observed binding energies suggest strong stabilization within the EphA1 catalytic pocket, indicating potential competitive inhibition at the

Table 1. List of selected top 66 compounds based on binding affinity towards EphA1.

S. No.	Ligand ZINC ID	Binding Free Energy (kcal/mol)	S. No.	Ligand ZINC ID	Binding Free Energy (kcal/mol)
1	ZINC03845566	−11.2	34	ZINC04017374	−10.5
2	ZINC08790757	−11.1	35	ZINC08789996	−10.5
3	ZINC08791024	−11.1	36	ZINC08876667	−10.5
4	ZINC08791220	−11.1	37	ZINC12873131	−10.5
5	ZINC08791221	−11.1	38	ZINC12874710	−10.5
6	ZINC12878012	−11.0	39	ZINC12886094	−10.5
7	ZINC12902062	−11.0	40	ZINC12887931	−10.5
8	ZINC12877669	−10.9	41	ZINC12902067	−10.5
9	ZINC12902057	−10.9	42	ZINC08918297	−10.4
10	ZINC12876960	−10.8	43	ZINC02161110	−10.4
11	ZINC12881190	−10.8	44	ZINC04258868	−10.4
12	ZINC08789288	−10.8	45	ZINC08300419	−10.4
13	ZINC08790429	−10.8	46	ZINC08789048	−10.4
14	ZINC08876663	−10.8	47	ZINC12885214	−10.4
15	ZINC12883374	−10.8	48	ZINC03841970	−10.4
16	ZINC08918144	−10.8	49	ZINC08789855	−10.4
17	ZINC08791168	−10.8	50	ZINC12887928	−10.4
18	ZINC08790034	−10.7	51	ZINC08792436	−10.4
19	ZINC08789997	−10.7	52	ZINC08790023	−10.4
20	ZINC08790428	−10.7	53	ZINC03851871	−10.3
21	ZINC09033817	−10.6	54	ZINC08918422	−10.3
22	ZINC08790759	−10.6	55	ZINC08918466	−10.3
23	ZINC08790760	−10.6	56	ZINC12662629	−10.3
24	ZINC08876661	−10.6	57	ZINC03844856	−10.3
25	ZINC12879301	−10.6	58	ZINC12895934	−10.3
26	ZINC02133362	−10.6	59	ZINC04045796	−10.2
27	ZINC08876556	−10.6	60	ZINC04073739	−10.2
28	ZINC11851754	−10.6	61	ZINC12874870	−10.2
29	ZINC12662242	−10.6	62	ZINC00348369	−10.2
30	ZINC12902051	−10.6	63	ZINC02090582	−10.2
31	ZINC08791169	−10.6	64	ZINC04237100	−10.2
32	ZINC03850412	−10.5	65	ZINC12660859	−9.9
33	ZINC08876707	−10.5	66	ZINC12661003	−9.7

Table 2. The predicted druglike properties of four potentially selective inhibitors of EphA1.

Compound ID	Molecular Weight	Rotatable Bonds	H-Bond Donor	H-Bond Acceptor	LogP	PAINS
ZINC12660859	498.664	5	2	5	3.9952	0
ZINC12661003	404.459	0	0	7	2.6825	0

ATP-binding site.

3.2. Drug Likeness and ADMET Analysis

In the initial step, Discovery Studio was used to generate SMILES strings for the top hits. These SMILES IDs were subsequently subjected to PAINS analysis using SwissADME. The PAINS analysis aimed to identify pan-assay interference compounds, which are multi-target ligands that lack specificity for a single target. Eliminating such compounds is crucial because they can lead to unintended regulatory effects. Following the PAINS analysis, additional filters were applied to refine the compound selection process. Compounds violating Lipinski's rule of five were removed. Furthermore, compounds with a total polar surface area (TPSA) outside the range of 65 to 145 were excluded. Compounds exhibiting poor solubility according to Silicos-IT Solubility (mol/l), Ali Solubility (mol/l), and ESOL Solubility (mol/l) were also eliminated from consideration. These filtering criteria helped in identifying the most promising compounds for further analysis and evaluation (Table 2).

After applying the filters, 24 compounds remained for further analysis. Using pkCSM, these substances were screened to determine their ADMET parameters. Remarkably, 20 compounds successfully met the ADMET parameters, indicating favourable characteristics for drug development. The physicochemical parameters evaluated for these compounds were found to fall within the range deemed suitable for drug candidacy, as outlined in Table 3. This observation indicates that the selected compounds possess desirable properties, making them promising candidates for further investigation and drug development.

In addition, a PASS analysis was performed on the 20 selected compounds using SMILES strings via the way-2-drugs-PASS Online tool. This analysis aimed to identify compounds with anti-cancer activity, thereby offering therapeutic leads for EphA1 targeting. After careful analysis, two compounds were chosen based on their demonstrated anti-cancer properties. The activities and corresponding Pa (probability of activity) and Pi (probability of inactivity) values for these selected compounds are presented in Table 5. These compounds exhibit promising characteristics and warrant further exploration as potential anti-cancer agents targeting EphA1.

3.3. Interaction Analysis

To analyse the binding modes and interaction patterns of the two selected compounds, Discovery Studio was used. A set of 18 docked conformers was produced for each of the two natural compounds using the output files. The analysis focused on examining the interacting residues within the compounds. The Discovery Studio software was used to identify and visualize hydrogen bonding and other interactions between the compounds and EphA1. Notably, residues within the kinase domain of EphA1 were observed to play a significant role in mediating numerous interactions with the compounds. This analysis provides valuable insights into the binding mechanism and potential interactions between the selected compounds and

Table 3. Predicted ADMET characteristics of chosen hits.

ZINC ID	Absorption GI Absorption	Distribution BBB Permeation	Metabolism CYP2D6 Inhibitor	Excretion OCT2 Substrate	Toxicity AMES
ZINC12660859	94.952 (High)	−0.177	No	No	No
ZINC12661003	99.476 (High)	−0.653	No	No	No

Table 4. PASS analysis results: Activities with Pa and Pi values of 2 selected compounds.

S. No.	Compound	Pa value	Pi value	Biological Activity
1.	ZINC12660859	0,845	0,025	CYP2C12 substrate
		0,500	0,027	AR expression inhibitor
		0,464	0,083	Antineoplastic
2.	ZINC12661003	0,566	0,053	Antineoplastic
		0,649	0,027	Phosphatase inhibitor
		0,488	0,016	Myc inhibitor
		0,388	0,079	Apoptosis agonist
		0,339	0,066	Antimetastatic

EphA1. Residues of EphA1's kinase domain have been shown to offer a considerable number of interactions, including Asp749, Gly769, Gly633, Asp767, Glu673, Lys656, Arg753, Asn754, Val638, Ile630, Leu,770, Leu756, Ala709, Gly631, Ala654, Phe704, Met705, Gly708, Phe635, Glu634, Ser766, Leu686, Gly708 (Table 5). Each compound interacts with ATP binding site or is closely associated with ATP binding sites. ATP binding sites, Lys656, as of EphA1, are inhibited either by a hydrogen bond or a strong bond in each compound, and active site Asp749 of EphA1 closely forms bonds with compounds.

Compound ZINC12660859 is forming hydrogen bonds to Asp749, Gly769, Gly633 and several other interactions to656, Arg753, Asn754, Asp767, Val638, Ile630, Leu,770, Leu756, Ala709, Gly631, Ala654, Phe704, Met705, Gly708, Phe635, Glu634. Compound ZINC12661003 is interacting to GlAsp767, Glu673 via Hydrogen bonding, and Lys656, Ser766, Leu686, Ala709, Gly708, Val638, Met705, Phe704, Ala654, Leu756, Ile630 are participating in other interactions (Figure 3). A significant number of interactions, including Hydrogen bonding with the active site Asp 749 and Vander wall's interaction with ATP binding site Lys656, formed between EphA1 and ZINC12660859, demonstrating a significant affinity for binding and its further implications as a medicinal compound targeting EphA1 (Figure 4).

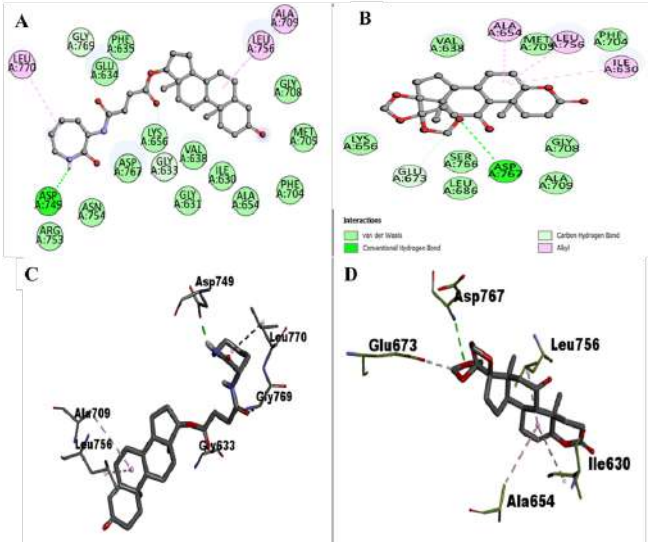

Figure 3. 2D interactions of EphA1 with (A) ZINC12660859, (B) ZINC12661003, (C) representation of amino acid residues of EphA1 interacting with ZINC12660859, and (D) ZINC12661003.

3.4. MD Simulations

An effective way to examine the atomic-level structural specifics and dynamic behaviour of protein-ligand complexes is to use the MD simulation approach. To acquire a thorough understanding of the stability and dynamics of the EphA1-ZINC12660859 complex, we conducted all-atom MD simulations for a duration of 100 ns on both the EphA1-ZINC12660859 complex and the apo EphA1 protein. We aimed to uncover numerous systematic and structural factors that affect the complex's behaviour through rigorous research into intermolecular interactions, conformational changes, hydrogen-bonding patterns, and solvent effects. Our research contributes to the field of drug development and targeted therapies by providing a deeper understanding of atomic-level protein-ligand interactions.

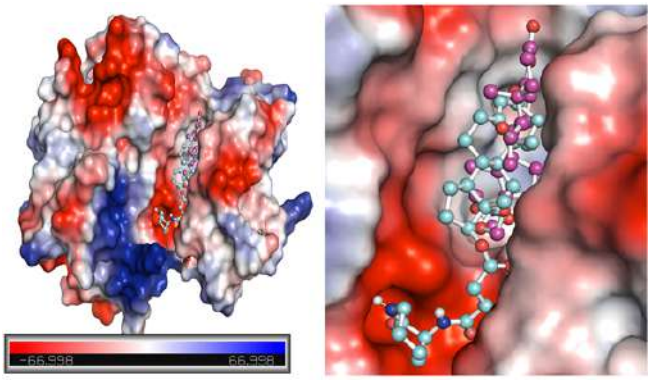

Figure 4. Potential surface representation of EphA1 complex with ZINC12660859 (cyan) and ZINC12661003 (magenta).

3.4.1. Root-Mean-Square Deviation

The root-mean-square deviation (RMSD) is a well-known metric frequently used to analyse structural heterogeneity in proteins.

Table 5. Interacting residues analyzed using Discovery Studio.

S. No.	Compound ID	Interacting Residues	
		H BONDING	Other Interactions
1.	ZINC12660859	Asp749, Gly769, Gly633	Lys656, Arg753, Asn754, Asp767, Val638, Ile630, Leu770, Leu756, Ala709, Gly631, Ala654, Phe704, Met705, Gly708, Phe635, Glu634
2.	ZINC12661003	Asp767, Glu673	Lys656, Ser766, Leu686, Ala709, Gly708, Val638, Met705, Phe704, Ala654, Leu756, Ile630

Table 6. The mean values of various MD parameters were computed based on 100 ns simulations.

System	RMSD (nm)	RMSF (nm)	R_g (nm)	SASA (nm^2)	#H-Bonds
EphA1	0.20	0.13	1.92	140.42	189
EphA1-ZINC12660859	0.26	0.12	1.93	139.62	190

It is crucial as a tool for assessing and analysing the dynamic behaviour and conformational changes of protein structures. RMSD provides vital insights into protein structural deviation and dynamics by quantifying the discrepancy between the atomic positions of a target structure and those of a reference structure. The structural dynamics of the protein EphA1 before and after ligand binding revealed low mean RMSD values of 0.20 nm and 0.26 nm for EphA1 and EphA1-ZINC12660859, respectively (Figure 5A). During the 100 ns simulation, both systems exhibited stable trajectories, with RMSD values approaching equilibrium. Analysis of the probability distribution function (PDF) revealed a significant increase in EphA1 stability, as evidenced by a pronounced peak in the distribution during compound binding. This finding indicates that the binding of ZINC12660859 positively influenced the overall stability of the complex.

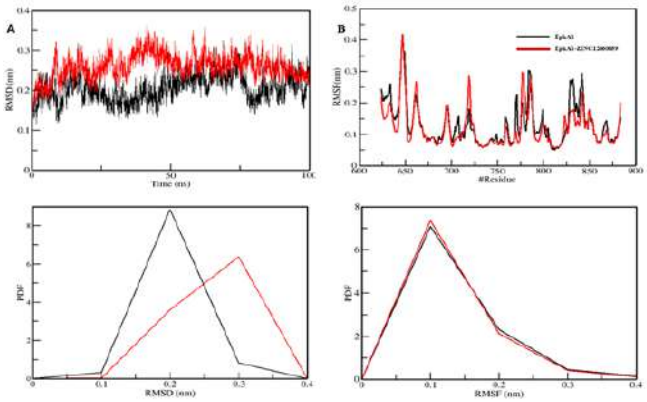

Figure 5. Structural dynamics of EphA1-ZINC12660859 complex. (A) RMSD plot. (B) RMSF plot. The lower panels show the probability distribution functions (PDFs).

3.4.2. Root-Mean Square Fluctuation

The root mean square fluctuation (RMSF) analysis has been widely employed to investigate protein dynamics during MD simulations. It offers valuable information on the influence of ligand binding on protein vibrations. Through the examination of the RMSF plot, we examined the residual dynamics of EphA1 prior to and following ligand binding. Notably, the protein-ligand complex exhibited a significant decrease in RMSF fluctuations, indicating increased stability upon binding of ZINC12660859 (Figure 5B). The RMSF analysis revealed a significant decrease in residual fluctuation within the POT1 protein. These findings are further corroborated by PDF analysis, which clearly indicates a pronounced decrease in residual flexibility following complex formation.

3.4.3. Radius of gyration

The radius of gyration (Rg) is a direct indicator of a protein's tertiary structure and overall conformation. It is commonly used to assess the compactness and folding behaviour of proteins. In our study, we investigated the stability of both EphA1 and the EphA1-ZINC12660859 complex by analysing their respective Rg values. The average Rg for EphA1 in its apo form was 1.92 nm, whereas for the EphA1-ZINC12660859 complex, it was found to be 1.93 nm (Figure 6A). The Rg plot revealed a slight increase of approximately 0.01 nm in Rg upon binding of ZINC12660859 to EphA1. This minor increase is likely due to the compound's packing effect. Remarkably, the introduction of ZINC12660859 did not result in any notable alterations to the structure of EphA1. Throughout the simulation trajectory, EphA1 maintained a constant equilibrium of Rg, indicating the enduring stability of the complex.

3.4.4. Solvent Accessible Surface Area

The solvent-accessible surface area (SASA) of a protein refers to the region on its surface that is accessible to the surrounding solvent. Analysing SASA is an important method for examining protein folding and stability. Upon analysis of the SASA plot, no notable alterations in SASA values were observed, indicating the stability of EphA1 interactions with ZINC12660859 (Figure 6B). The SASA values for the EphA1 and EphA1-ZINC12660859 complexes were 140.42 nm^2 and 139.62 nm^2, respectively. The distribution of surface-area values exhibited a consistent pattern across both systems. Notably, there was a marginal reduction in the average surface area upon binding of the compound to EphA1. This reduction suggests the formation of stable interactions or binding between the protein and ligand.

3.4.5. Hydrogen Bond Analysis

The creation of hydrogen bonds (H-bonds) is critical in the kinetics of protein folding. Protein conformational alterations are primarily governed by the disruption and formation of hydrogen bonds. In this study, we analysed the temporal progression of hydrogen bonds to assess the integrity of intramolecular bonding within the EphA1-ZINC12660859 complex. The plot we obtained indicated negligible alterations in the number of hydrogen bonds in EphA1 upon complex formation with ZINC12660859 (Figure 7). In the context of EphA1 intramolecular interactions, we observed

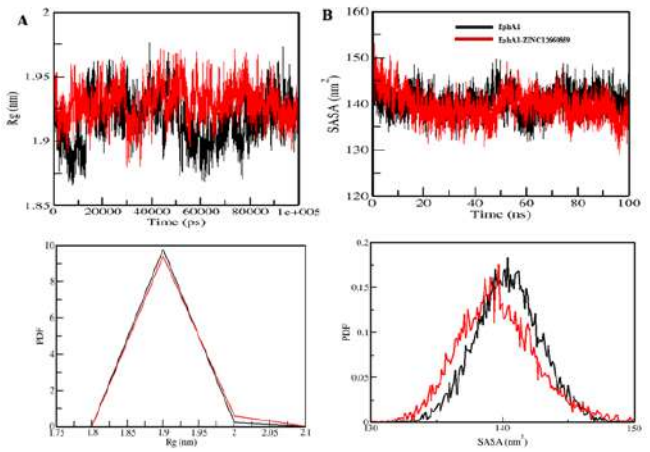

Figure 6. Analysis of the structural compactness and folding of EphA1 upon interaction with ZINC12660859. The changes in radius of gyration (Rg) and solvent-accessible surface area (SASA) are presented in plots A and B, respectively.

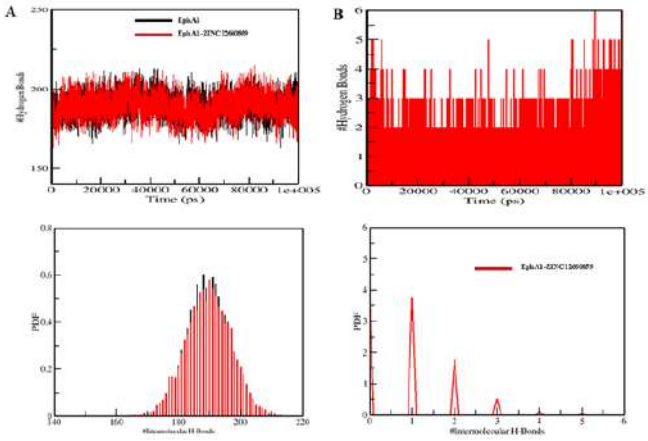

Figure 7. Analysis of hydrogen bonds. (A) Intramolecular H-bonds' temporal progression. (B) Intermolecular H-bonds established between EphA1 and ZINC12660859 within 0.35 nm. The PDF values for the hydrogen-bond distribution are shown in the lower panels. # stands for number.

that the average number of hydrogen bonds formed was 189 before binding of ZINC12660859, and this count increased to 190 after binding. A minor rise in the number of hydrogen bonds was noted, which can be attributed to the heightened compactness resulting from the binding of the ligand (Figure 7A). Analysis of the PDFs for both systems confirmed the consistent stability of intramolecular hydrogen bonds. From the generated plots, it can be inferred that the intramolecular hydrogen bonds in EphA1 remained stable throughout the simulation, even after compound binding. Throughout the simulation, ZINC12660859 was discovered to bind into the specified pocket of EphA1, creating 1–2 stable hydrogen bonds with little variation (Figure 7B). There were further occasions where ZINC12660859 showed 3–4 hydrogen bonds with larger variations. ZINC12660859 remained at its original docking site on EphA1, consistent with our predictions based on the observed intermolecular hydrogen bonding. These hydrogen bonds play an important role in the stability of the complex structure, preventing the compound from migrating or dissociating from its binding site on EphA1 during the simulation.

4. Conclusions

In the present study, an integrative structure-based computational approach was employed to identify potential natural inhibitors of the EphA1 receptor tyrosine kinase. Large-scale virtual screening of 32,901 drug-like natural compounds led to the identification of promising candidates targeting the ATP-binding pocket of the EphA1 kinase domain. Among the shortlisted molecules, ZINC12660859 demonstrated favorable binding affinity, compliance with drug-likeness criteria, acceptable ADMET properties, predicted antineoplastic activity, and stable interactions with key catalytic residues, including Lys656 and Asp749. Molecular dynamics simulations further confirmed the structural stability and sustained binding of the EphA1–ZINC12660859 complex over 100 ns without inducing conformational destabilization. Collectively, these findings suggest that ZINC12660859 represents a promising lead compound for EphA1-targeted inhibition. While the present work provides strong computational evidence supporting its therapeutic potential, further *in vitro* and *in vivo* experimental validation is required to confirm its inhibitory activity and anticancer efficacy. This study highlights the effectiveness of structure-based virtual screening and molecular dynamics simulations in accelerating the discovery of kinase-targeted therapeutic candidates.

Conflict of Interest

There is no conflict of interest to declare.

Data Availability Statement

The data supporting this study are provided in this article.

Funding

None

Acknowledgements

None

Supplementary Materials

None

Author Contributions

S.R. conceptualized and designed the study. S.R. performed data curation, virtual screening, molecular docking, ADMET analysis, PASS prediction, and molecular dynamics simulations. S.R. conducted formal analysis and interpreted the results. S.R. prepared figures and tables and wrote the original draft of the manuscript. The author reviewed and approved the final version of the manuscript.

References

Adu-Gyamfi, E., Czika, A., Liu, T., et al. 2021, Biology of Reproduction, 104, 71, doi: 10.1093/biolre/ioaa171

Biovia, D. 2017, Discovery Studio Visualizer

Dai, B., & Zhang, X. 2020, Neoplasma, 67, 794, doi: 10.4149/neo_2020_190807N724

Daina, A., Michielin, O., & Zoete, V. 2017, Scientific Reports, 7, 42717, doi: 10.1038/srep42717
David, A., Islam, S., Tankhilevich, E., & Sternberg, M. 2022, Journal of Molecular Biology, 434, 167336, doi: 10.1016/j.jmb.2021.167336
DeLano, W. 2002, CCP4 Newsletter on Protein Crystallography, 40, 82
Gajdzis, M., Theocharis, S., Gajdzis, P., et al. 2020, Life, 10, doi: 10.3390/life10100225
Hjorthaug, H., & Aasheim, H. 2007, European Journal of Immunology, 37, 2326, doi: 10.1002/eji.200737111
Ieguchi, K., & Maru, Y. 2019, Cancer Science, 110, 841, doi: 10.1111/cas.13942
Irwin, J., & Shoichet, B. 2005, Journal of Chemical Information and Modeling, 45, 177, doi: 10.1021/ci049714+
Irwin, J., Tang, K., Young, J., et al. 2020, Journal of Chemical Information and Modeling, 60, 6065, doi: 10.1021/acs.jcim.0c00675
Kang, M., Jeong, W., Bae, H., et al. 2018, Journal of Cellular Physiology, 233, 2560, doi: 10.1002/jcp.26131
Lagunin, A., Stepanchikova, A., Filimonov, D., & Poroikov, V. 2000, Bioinformatics, 16, 747, doi: 10.1093/bioinformatics/16.8.747
Li, Y., Yan, H., Wang, F., et al. 2017, Biochemical and Biophysical Research Communications, 486, 693, doi: 10.1016/j.bbrc.2017.03.100
Liang, Z., Wang, X., Dong, K., et al. 2021, BioMed Research International, 2021, 5575704, doi: 10.1155/2021/5575704
Lisabeth, E., Falivelli, G., & Pasquale, E. 2013, Cold Spring Harbor Perspectives in Biology, 5, doi: 10.1101/cshperspect.a009159
Ma, J., Wang, Z., Chen, S., et al. 2021, Molecular Neurobiology, 58, 913, doi: 10.1007/s12035-020-02122-x
Mohammad, T., Mathur, Y., & Hassan, M. 2021, Briefings in Bioinformatics, 22, doi: 10.1093/bib/bbaa279
Peng, L., Wang, H., Dong, Y., et al. 2013, International Journal of Clinical and Experimental Pathology, 6, 1854
Pires, D., Blundell, T., & Ascher, D. 2015, Journal of Medicinal Chemistry, 58, 4066, doi: 10.1021/acs.jmedchem.5b00104
Rose, P., Prlić, A., Altunkaya, A., et al. 2017, Nucleic Acids Research, 45, D271, doi: 10.1093/nar/gkw1000
Toma, M., Erdmann, K., Diezel, M., et al. 2014, PLoS One, 9, e102262, doi: 10.1371/journal.pone.0102262
UniProt Consortium. 2019, Nucleic Acids Research, 47, D506, doi: 10.1093/nar/gky1049
Villegas-Llerena, C., Phillips, A., Garcia-Reitboeck, P., Hardy, J., & Pocock, J. 2016, Current Opinion in Neurobiology, 36, 74, doi: 10.1016/j.conb.2015.10.004
Wang, J., Ma, J., Dong, Y., et al. 2013, APMIS, 121, 30, doi: 10.1111/j.1600-0463.2012.02941.x
Wang, Y., Dai, Y., Xu, G., et al. 2020, Medical Science Monitor, 26, e923409, doi: 10.12659/msm.923409
Wang, Y., Yu, H., Shan, Y., et al. 2016, Journal of Experimental & Clinical Cancer Research, 35, 65, doi: 10.1186/s13046-016-0339-6
Wu, B., Jiang, W., Zhou, D., & Cui, Y. 2016, Anticancer Research, 36, 1211
Wu, Y., Du, Z., Mou, J., et al. 2022, Current Medicinal Chemistry, doi: 10.2174/0929867329666220820125638
Yamazaki, T., Masuda, J., Omori, T., et al. 2009, Journal of Cell Science, 122, 243, doi: 10.1242/jcs.036467
Yu, L., Ke, J., Du, X., Yu, Z., & Gao, D. 2019, Scientific Reports, 9, 2369, doi: 10.1038/s41598-019-38878-z
Yu, W., & MacKerell, A.D., J. 2017, Methods in Molecular Biology, 1520, 85, doi: 10.1007/978-1-4939-6634-9_5
Zhang, L., Ai, H., Chen, W., et al. 2017, Scientific Reports, 7, 2118, doi: 10.1038/s41598-017-02365-0

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 33–38, 2026
Open Access © 2026 Editory Press

Impact of Single Nucleotide Polymorphisms in TERF1-Interacting Nuclear Factor 2 and Their Association with Dyskeratosis Congenita Autosomal Dominant-3

Sarfraz Ahmed[1,*], **Lalita Thangwal**[2] **and Kim Ji Hoe**[1,*]

[1]*Department of Medical Biotechnology and Research Institute of Cell Culture, Yeungnam University, Gyeongsan 38541, Republic of Korea*
[2]*Department of Computer Science, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India*
[*]*Corresponding authors: Sarfraz Ahmed and Kim Ji Hoe; Department of Medical Biotechnology and Research Institute of Cell Culture, Yeungnam University, Gyeongsan 38541, Republic of Korea; E-mails: sarfrazahmed6954@yu.ac.kr, kimjihoe@ynu.ac.kr*

Abstract

Dyskeratosis congenita (DC) is a rare hereditary telomere biology disorder characterized by bone marrow failure, nail dystrophy, mucosal leukoplakia, and abnormal skin pigmentation. Telomere shortening resulting from defects in telomere maintenance genes is a central pathogenic mechanism of the disease. TERF1-interacting nuclear factor 2 (TIN2), encoded by the *TINF2* gene, is a critical component of the shelterin complex that safeguards telomere integrity by coordinating interactions among TRF1, TRF2, and TPP1/POT1. Heterozygous missense mutations in *TINF2* are known to cause autosomal dominant dyskeratosis congenita type 3. In this study, a comprehensive *in silico* analysis was performed to identify deleterious and pathogenic non-synonymous single-nucleotide polymorphisms (nsSNPs) in the TIN2 protein. A total of 271 nsSNPs were subjected to sequence-based prediction tools (SIFT, PROVEAN, PolyPhen-2, Mutation Assessor, and SAAFEC-SEQ). Structure-based stability analyses of 134 variants located within the resolved crystal structure (PDB ID: 5XYF) were conducted using SDM2, DUET, mCSM, and MUpro. High-confidence deleterious and destabilizing mutations were further evaluated for disease association using PMut, PhD-SNP, and Rhapsody. Twelve missense mutations (A16D, V22G, V34G, L38P, V49G, R52P, R56G, E109G, F114S, L121P, Y139C, and L150P) were consistently predicted as pathogenic. Detailed structural characterization revealed significant alterations in protein stability, packing density, accessible surface area, aggregation propensity, and intermolecular noncovalent interactions. Most of these mutations showed increased aggregation tendency or reduced solubility, suggesting disruption of shelterin complex stability. This study provides structural and functional insights into pathogenic *TINF2* variants and highlights potential molecular mechanisms underlying autosomal dominant dyskeratosis congenita.

Keywords: *Dyskeratosis congenita, TINF2, TIN2, shelterin complex, telomere biology, nsSNP, in silico analysis*

Received: July 1st, 2025 **Accepted:** November 1st, 2025 **Published:** January 5th, 2026

1. Introduction

Telomeres are nucleoprotein complexes located at the linear ends of chromosomes that maintain genome stability and integrity. Human telomeres consist of approximately 5–15 kb of TTAGGG tandem repeats, forming double-stranded DNA with a 3' single-stranded overhang bound by the shelterin complex (Pike et al., 2019). The shelterin complex is composed of six subunits: TRF1, TRF2, POT1, TIN2, TPP1, and RAP1 (de Lange, 2005). This complex plays a critical role in protecting chromosome ends and regulating telomerase activity to maintain telomere length.

TIN2, encoded by the *TINF2* gene, is a central component of the shelterin complex, facilitating its assembly by bridging the double-stranded telomeric DNA-binding proteins TRF1 and TRF2 with the single-stranded DNA-binding complex TPP1/POT1. Structurally, TIN2 contains a telomeric repeat factor homology (TRFH) domain that mediates essential protein–protein interactions within the complex. The N-terminal region of TIN2 exhibits structural similarity to the TRFH domains of TERF1 and TRF2, suggesting an evolutionary relationship, although these proteins have diverged functionally.

The TIN2 protein consists of 354 amino acids. Within the shelterin complex, TIN2 interacts with TPP1 through its N-terminal domain (approximately 200 amino acids) and contains a TRF1-binding motif spanning approximately 20 amino acids (Yang et al., 2011). Due to its central scaffolding role, TIN2 is essential for telomere length regulation and chromosome end protection. It contributes to TRF1-dependent telomere length control and stabilizes TRF1 through multiple mechanisms. Notably, TIN2 protects TRF1 from tankyrase 1-induced poly(ADP-ribosyl)ation, thereby preventing its dissociation from telomeres.

Additionally, TIN2 competes with SCF^{FBX4} for TRF1 binding, protecting it from ubiquitin-dependent proteolysis (Frescas & de Lange, 2014).

Defects in telomere maintenance result in progressive telomere shortening and genomic instability, which are hallmarks of a group of inherited disorders known as telomeropathies. Dyskeratosis congenita (DC) is a rare hereditary bone marrow failure syndrome characterized by the classical triad of nail dystrophy, mucosal leukoplakia, and abnormal skin pigmentation (Savage et al., 2008). Patients with DC often develop additional complications, including aplastic anemia, pulmonary fibrosis, myelodysplastic syndrome, and leukemia. The disorder exhibits genetic heterogeneity and can be inherited in X-linked recessive, autosomal recessive, or autosomal dominant patterns. Mutations in several telomere-associated genes, including *DKC1*, *TERC*, *TERT*, *NOP10*, *NHP2*, *TCAB1*, *RTEL1*, and *TINF2*, have been implicated in DC. Notably, heterozygous missense mutations in *TINF2* are responsible for autosomal dominant dyskeratosis congenita type 3 and are frequently associated with severe clinical manifestations and markedly shortened telomeres.

Several studies have identified mutation hotspots within residues 269–298 of TIN2, particularly near the TRF1-binding domain. These mutations are believed to disrupt shelterin complex assembly and compromise telomere stability. Given the critical scaffolding function of TIN2, even minor amino acid substitutions may significantly affect protein stability, intermolecular interactions, and overall telomere function. Non-synonymous single-nucleotide polymorphisms (nsSNPs) can alter protein structure and function, contributing to disease pathogenesis. Computational prediction tools provide a cost-effective and systematic approach to identify and prioritize potentially deleterious variants. Integrating sequence-based, structure-based, and pathogenicity prediction methods enables the identification of high-confidence disease-associated mutations and facilitates understanding of their molecular mechanisms.

In this study, we performed a comprehensive *in silico* analysis of nsSNPs in the *TINF2* gene to identify deleterious, destabilizing, and pathogenic variants associated with dyskeratosis congenita. By integrating multiple predictive algorithms and structural analyses, we aimed to elucidate the molecular effects of selected mutations on TIN2 stability and function, thereby providing insights into the structural basis of autosomal dominant dyskeratosis congenita.

2. Materials and Methods

2.1. Data Retrieval

The FASTA format sequence of TINF2 protein was extracted from the UniProt database using UniProt ID (Q9BSI4). Using different databases such as dbSNP (Sherry et al., 2001), HGMD (Stenson et al., 2009), Clinvar (Landrum et al., 2014), and Ensembl (Hubbard et al., 2002) make a list of missense mutations and remove the duplicate nsSNPs. Approximately 900 mutations are obtained from the Ensembl database, of which 271 are non-synonymous mutations. The 3-D crystal structure obtained from the PDB (Berman et al., 2002) database of TINF2 protein using (PDB ID: 5XYF). This structure was used for structural-based analysis.

2.2. Prediction of deleterious mutations using sequence-based tools

2.2.1. SIFT

The Sorting Intolerant from Tolerant (SIFT) program (http://sift.jcvi.org/) uses physical properties of amino acids to predict whether amino acid substitutions are deleterious or neutral. SIFT also estimates a protein's sequence homology. If the SIFT score is more than 0.05, then the mutation is tolerable or neutral; if the score is equal to or less than 0.05, then it is not tolerable or deleterious (Kumar et al., 2009). All 271 nsSNPs mutations of the TINF2 protein are tested using the SIFT tool.

2.2.2. PolyPhen-2

Polymorphism phenotyping (PolyPhen-2) (http://genetics.bwh.harvard.edu/pph2/) is a sequence-based tool that finds the damaging probability of amino acid substitution in any protein on the basis of its physical and comparative properties (Ramensky et al., 2002). It takes a FASTA format sequence as input and estimates the PSIC score between 0 to 1. PolyPhen-2 scores range from 0.0 to 1.0. Variants closer to 1.0 are more likely to be damaging and are classified as benign, possibly damaging, or probably damaging based on predefined thresholds.

2.2.3. PROVEAN

Protein variation effect analyzer (PROVEAN) (http://provean.jcvi.org/) is a sequence-based tool take FASTA format sequence as an input and predicts the damaging nsSNP of protein on the basis of the functionality of the protein (Choi & Chan, 2015). If the PROVEAN score is greater than -2.5, the mutation is predicted to be neutral; scores less than -2.5 are considered deleterious.

2.2.4. Mutation Assessor

Mutation Assessor (http://mutationassessor.org/r3/) is a sequence-based tool that takes a UniProt ID or NCBI RefSeq ID as input for a protein. The mutation assessor classifies mutations as medium, low, or neutral in terms of deleteriousness based on multiple sequence alignment and evolutionary conservation (Reva et al., 2011). This tool predicts the functional impact of a missense mutation on a protein and computes the FI score. If the FI score is less than 2.00, the mutation is neutral; if the FI score is greater than 2.00, the mutation is considered deleterious.

2.2.5. SAAFEC-SEQ

Single Amino Acid Folding free Energy Changes (SAAFEC-SEQ) (http://compbio.clemson.edu/SAAFEC-SEQ/index.php). SAAFSEC-SEQ is a sequence-based tool used to determine the change in folding free energy caused by amino acid substitution (Getov et al., 2016). It takes a protein FASTA sequence as input and predicts changes in free energy based on physicochemical properties and sequence features.

2.3. Prediction of the destabilizing mutation using structure-based tools

2.3.1. SDM2

Site-Directed Mutator (SDM2) (http://marid.bioc.cam.ac.uk/sdm2) is structure based tool that takes a PDB ID or a PDB file as input and uses environment-specific amino acid substitution tables to calculate the stability of an nsSNP in a protein (Pandurangan et al., 2017). It estimates the change in stability between the wild-type and mutant nsSNP of a protein using $\Delta\Delta G$. SDM2 estimates the change in stability ($\Delta\Delta G$) between the wild-type and mutant protein. Negative $\Delta\Delta G$ values indicate destabilizing mutations, whereas positive values indicate stabilizing effects. All 134 nsSNPs

located within the resolved crystal structure were analyzed using SDM2.

2.3.2. mCSM

It is a structure-based tool to estimate nsSNP. mCSM take PDB file or a FASTA sequence as input and predict the mutation is destabilizing or not on the basis of graph based approach and calculates $\Delta\Delta G$ (http://structure.bioc.cam.ac.uk/mcsm) (Pires et al., 2013). For a range of proteins, mCSM provides better comparisons of mutations linked to disease. If the score is less than 0, the mutation reverses its effect.

2.3.3. DUET

DUET (http://biosig.unimelb.edu.au/duet/) is a structure-based tool that takes a PDB file and an nsSNP mutation as input and calculates the DUET score, along with mCSM and SDM scores, as outputs. The DUET tool employs SVM and combines the results of mCSM and SDM to compute $\Delta\Delta G$. DUET integrates residue-level RSA (BY SDM) and secondary-structure and pharmacophore (by mCSM) features, and combines them with a supervised learning program (Pires et al., 2014).

2.3.4. MUpro

MUpro (http://mupro.proteomics.ics.uci.edu/mutation_intro.html/) is a machine-learning technique used to determine how a single amino acid mutation affects protein stability. It employs two distinct machine-learning models to compute energy change. These programs are SVM and neural networks (Cheng et al., 2006). It computes a confidence score between -1 and 1 and the change in energy to quantify confidence. If the MUpro score is greater than 0, the mutation is predicted to increase protein stability; if the score is less than 0, it is predicted to decrease protein stability.

2.4. Identification of Pathogenic nsSNPs

2.4.1. Pmut

PMut (http://mmb.irbbarcelona.org/PMut) is structure based tools which is used to check whether a non-synonymous mutation is related to disease or not (López-Ferrando et al., 2017). Updated version of pmut determine. Pathogenicity of mutations, but it also provides an opportunity to train one's own prediction model on a particular dataset.

2.4.2. Phd-SNP

Phd-SNP (http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi/) is a sequence or structure-based tool that is used to differentiate disease-related mutations regulated by the local environment of substitution. It takes a PDB file or a FASTA sequence as input.

2.4.3. RHAPSODY

Rhapsody (http://rhapsody.csb.pitt.edu/) is a computational tool primarily used to determine the Pathogenicity of missense mutations based on sequence- and structure-based properties of the mutant protein (Ponzoni et al., 2020). It outperforms EV mutation and Polyphen-2 in analyzing residue-average pathogenicity.

2.5. Analysis of packing density and accessible surface area by using SDM2

The updated version of sdm2 also calculates the RSA, OSP, and residue depth for both mutant and wild-type proteins. For RSA, OSP, and residue-depth calculations, an environment-specific amino acid table is employed. RSA can calculate by the Lee and Richard technique. For the Structure stability analysis, OSP and residue depth are extended properties of the protein (DeDecker et al., 1996; Richards & Lim, 1993).

2.6. Analysis of noncovalent interactions

2.6.1. Arpeggio

Arpeggio server is an structure based tool that takes a PDB ID or a PDB format file as input and estimates all the interatomic interactions of a protein. The Arpeggio server estimated mainly 15 different types of interatomic interactions, which can also be seen by this server (http://structure.bioc.cam.ac.uk/arpeggio/) (Jubb et al., 2017).

2.7. Analysis of aggregation propensity

2.7.1. SODA

Solubility based on Disorder and Aggregation (SODA) (http://protein.bio.unipd.it/soda/) is a sequence or structure-based tool used to estimate the aggregation of protein, disorder, helix, and strand propensity that emerge because of mutations. Use the FASTA protein sequence or the PDB structure file as input. Insertion, deletion, substitution, and duplication (mutation type) were estimated by soda using PASTA 2.0, ESpritz-NMR, and Fells Soda score estimated according to the solubility difference between the mutant and WT of protein (Paladin et al., 2017).

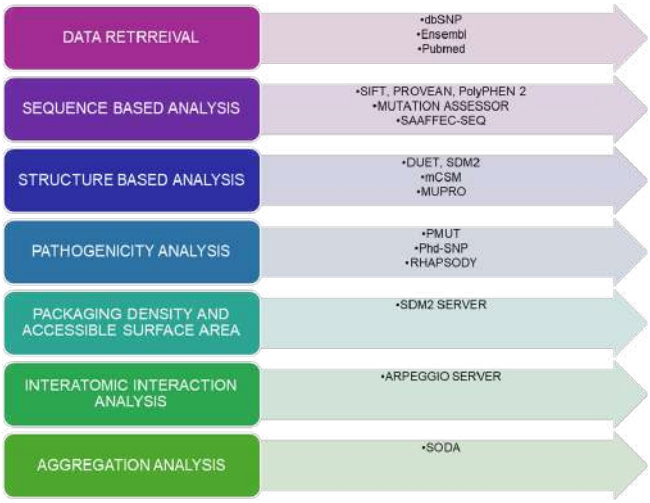

Figure 1. Overview of the computational aspects to predict the pathogenic mutation in TINF2.

3. Results and Discussion

3.1. Identification of Non-Synonymous Variants in TINF2

A list of 900 mutations was obtained from Ensembl, of which only 271 were non-synonymous. Out of 271 nsSNPs, only the structure of the N-terminal region (residue 2-202) is found in the PDB structure (PDB ID: 5XYF) of the TIN2 protein in the RCSB database. This analysis focuses on the sequence- and structure-based study of all nsSNPs. A multilevel approach was used to determine the functional and structural importance of the mutation on the TINF2 protein. To determine high-confidence deleterious and destabilizing non-synonymous mutations, sequence- and structure-based analyses were performed. All 271 non-synonymous mutations were subjected to sequence-based analysis using five tools: SIFT, PROVEAN, PolyPhen-2, Mutation Assessor, and SAAFEC-SEQ. Structure-based analysis is done by four different web-based tools, which are SDM2, DUET, mCSM, and MUpro, to estimate the 134 non-synonymous mutations present within the resolved

crystal structure of the TINF2 protein (PDB ID 5XYF). For further studies, only high-confidence nsSNPs were included. Pathological analysis of high-confidence nsSNPs was performed using two web tools: PMut and PhD SNP.

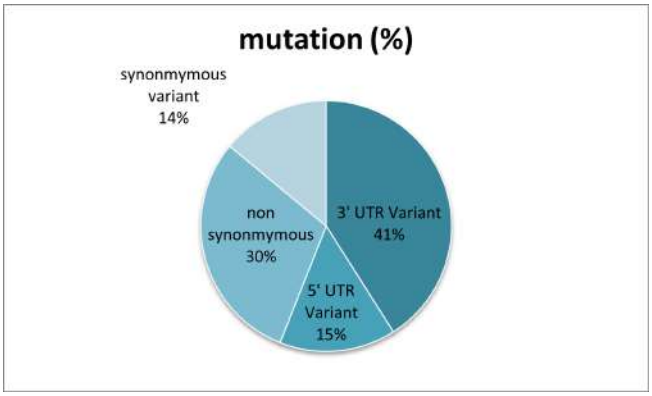

Figure 2. Representation of the number of SNPs in TINF2 using the Ensembl database.

3.2. Sequence-Based Prediction of Deleterious Mutations

All 271 nsSNPs were analyzed using five independent sequence-based prediction tools. The proportion of variants predicted as deleterious varied among tools, reflecting differences in underlying algorithms. SIFT predicted 142 variants (52.3%) as damaging, PolyPhen-2 identified 96 (35.4%) as possibly or probably damaging, PROVEAN classified 41 (15.1%) as deleterious, Mutation Assessor detected 80 (29.5%) as functionally impactful, and SAAFEC-SEQ predicted 264 (96.3%) variants to reduce folding free energy (Table S1). Although individual tools yielded variable results, consensus analysis reduced the risk of false positives (Figure 3). Variants predicted as deleterious by at least four of the five tools were considered high-confidence functionally damaging mutations. This approach improved reliability by integrating evolutionary conservation, physicochemical changes, and predicted changes in folding energy. The high proportion of predicted deleterious variants suggests that many substitutions in TIN2 may disrupt its structural integrity or protein–protein interactions within the shelterin complex.

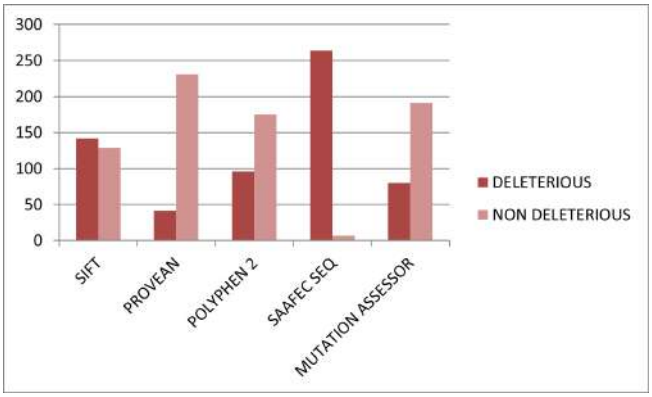

Figure 3. Graphical representation of deleterious nsSNPs predicted by sequence-based tools.

3.3. Structure-Based Stability Analysis

Structure-based prediction was performed on 134 nsSNPs located within the resolved N-terminal structure of TIN2. Stability changes were assessed using SDM2, DUET, mCSM, and MUpro. Among the analyzed variants, SDM2 predicted 58 (43.2%) as destabilizing, DUET identified 103 (76.8%), mCSM detected 120 (89.5%), and MUpro predicted 128 (95.5%) variants as decreasing protein stability (Figure 4). The predominance of destabilizing predictions indicates that many substitutions in the N-terminal domain may impair structural stability. Destabilization of TIN2 is particularly significant because the N-terminal region mediates high-affinity binding with TRF2 and contributes to scaffold formation within the shelterin complex. Structural perturbations in this domain may weaken intermolecular interactions, leading to defective telomere protection and compromised telomere length regulation. To increase prediction accuracy, only variants classified as destabilizing by at least three of four structure-based tools and as deleterious by at least four of five sequence-based tools were retained. This stringent filtering yielded 35 high-confidence deleterious and destabilizing nsSNPs.

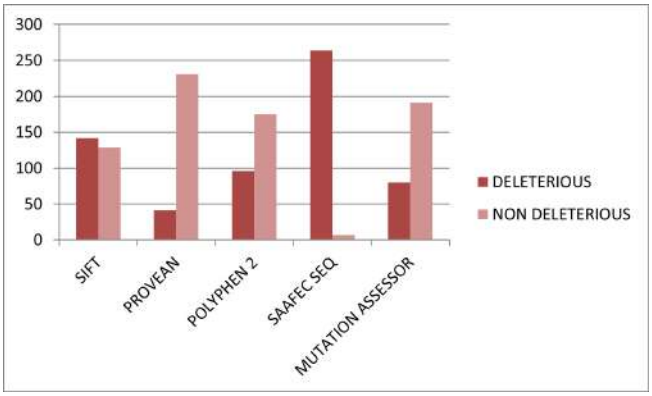

Figure 4. Graphical representation of deleterious and neutral nsSNPs predicted by structure-based tools.

3.4. Identification of pathogenic nsSNPs

The 35 high-confidence variants were further evaluated for disease association using PMut, PhD-SNP, and Rhapsody. PMut predicted 25 (71.4%) variants as pathogenic, PhD-SNP identified 24 (68.5%), and Rhapsody classified 11 (31.4%) as disease-associated. Twelve mutations, A16D, V22G, V34G, L38P, V49G, R52P, R56G, E109G, F114S, L121P, Y139C, and L150P, were consistently predicted as pathogenic by at least two of the three tools. These variants were selected for detailed structural investigation. Interestingly, several of these mutations are located in highly conserved residues within the N-terminal region, supporting their functional importance. Many involve substitution of small or hydrophobic residues with charged or bulky amino acids, which may significantly alter local structural packing.

3.4.1. Changes in Stability and Packing Density

Analysis of relative solvent accessibility (RSA), occluded surface packing (OSP), and residue depth revealed that most of the selected mutations induced noticeable alterations in local packing density. Mutations such as V22G, V34G, V49G, and L150P replace hydrophobic residues with glycine or proline, thereby disrupting secondary-structure elements and reducing hydrophobic-core stability. Proline substitutions (L38P, L121P, L150P) are particularly disruptive due to conformational rigidity

Table 1. Prediction of RSA, residue depth, and OSP for the wild-type and mutant TINF2 proteins using the SDM2 server.

S. No.	Mutation	WT RSA (%)	WT Depth (Å)	WT OSP	MT RSA (%)	MT Depth (Å)	MT OSP	Outcome
1	A16D	0.0	8.1	0.60	0.0	8.2	0.65	Increased Solubility
2	V22G	9.9	4.6	0.54	13.6	5.2	0.44	Reduced Solubility
3	V34G	0.4	7.5	0.53	17.0	7.0	0.47	Increased Solubility
4	L38P	0.3	8.6	0.52	6.2	8.1	0.51	Increased Solubility
5	V49G	7.1	5.1	0.42	34.4	4.1	0.26	Increased Solubility
6	R52P	37.7	3.9	0.31	13.3	4.8	0.42	Reduced Solubility
7	R56G	44.7	3.6	0.29	70.6	4.0	0.34	Increased Solubility
8	E109G	85.2	3.3	0.22	94.3	3.5	0.34	Increased Solubility
9	F114S	7.2	6.1	0.49	21.6	5.2	0.37	Increased Solubility
10	L121P	5.6	5.9	0.52	13.8	5.5	0.51	Increased Solubility
11	Y139C	12.9	4.8	0.49	9.1	5.3	0.45	Reduced Solubility
12	L150P	4.4	5.3	0.53	5.5	5.7	0.56	Increased Solubility

Table 2. Arpeggio server estimates of interatomic interactions for WT and mutant protein structures.

S. No.	Variants	Vander Waals	Hydrogen	Ionic	Aromatics	Hydrophobic
1	Wild Type	175	267	17	18	471
2	A16D	176	267	17	18	471
3	V22G	174	267	17	18	464
4	V34G	175	267	17	18	464
5	L38P	175	263	17	18	463
6	V49G	174	267	17	18	459
7	R52P	176	267	17	18	472
8	R56G	175	267	17	18	470
9	E109G	175	267	17	18	471
10	F114S	173	266	17	14	446
11	L121P	173	263	17	18	465
12	Y139C	172	268	17	13	451
13	L150P	175	268	17	18	472

and their tendency to break α-helices. These changes likely perturb TIN2's folding and structural integrity (Table 1).

3.4.2. Alterations in Noncovalent Interactions

Interatomic interaction analysis using Arpeggio demonstrated that several pathogenic mutations reduced hydrogen bonds, hydrophobic contacts, and ionic interactions compared to the wild-type structure (Table 2). Loss or rearrangement of these interactions can destabilize the local microenvironment and impair protein–protein interaction interfaces. For example, substitutions involving charged residues (R52P, R56G, E109G) may abolish electrostatic interactions critically for maintaining tertiary structure or mediating binding with TRF1 and TRF2.

3.4.3. Aggregation Propensity and Solubility Changes

Aggregation analysis using SODA indicated that nine of the twelve selected mutations showed increased aggregation propensity, whereas the remaining three showed reduced solubility. Increased aggregation propensity suggests a higher likelihood of misfolding and potential formation of non-functional protein assemblies (Table 3). Protein misfolding or aggregation of TIN2 could disrupt shelterin complex assembly, thereby impairing telomere capping and exposing chromosome ends to DNA damage response pathways.

Table 3. Prediction of aggregation propensity of mutant TINF2 protein using the SODA server.

S. No.	Mutation	SODA	Remark
1	A16D	-2.9	Less Soluble
2	V22G	20.6	More Soluble
3	V34G	13.4	More Soluble
4	L38P	5.9	More Soluble
5	V49G	18.9	More Soluble
6	R52P	13.9	More Soluble
7	R56G	7.74	More Soluble
8	E109G	1.59	More Soluble
9	F114S	9.37	More Soluble
10	L121P	1.3	More Soluble
11	Y139C	-6.5	Less Soluble
12	L150P	-8.1	Less Soluble

4. Conclusions

In this study, a comprehensive *in silico* framework was employed to systematically evaluate non-synonymous single-nucleotide polymorphisms in the *TINF2* gene and to identify variants with potential pathogenic relevance in dyskeratosis congenita. By integrating multiple sequence-based, structure-based, and pathogenicity prediction tools, we prioritized high-confidence deleterious and destabilizing mutations affecting the TIN2 protein. Twelve missense variants (A16D, V22G, V34G, L38P, V49G, R52P, R56G, E109G, F114S, L121P, Y139C, and L150P) were consistently predicted to compromise protein stability and structural integrity. Detailed structural analyses indicated that these substitutions alter packing density, disrupt noncovalent interactions, and increase aggregation propensity, potentially impairing shelterin complex assembly and telomere protection. Given TIN2's central scaffolding role within the shelterin complex, these destabilizing effects may potentially contribute to telomere dysfunction associated with autosomal-dominant dyskeratosis congenita. Although the findings provide mechanistic insights into the structural consequences of *TINF2* mutations, experimental validation is required to confirm their biological impact. The prioritized variants identified in this study may be suitable candidates for functional assays and genetic screening strategies. Overall, this work enhances understanding of the molecular basis of telomere-associated disorders and offers a systematic computational approach for investigating disease-related mutations.

Conflict of Interest

There is no conflict of interest to declare.

Data Availability Statement

The data supporting this study are provided in this article.

Funding

None

Acknowledgements

None

Supplementary Materials

None

Author Contributions

L.T. conceptualized and designed the study. L.T. performed data retrieval, computational SNP analysis, structural stability analysis, pathogenicity prediction, and aggregation propensity analysis. L.T. conducted formal analysis, prepared figures and tables, and wrote the manuscript. The author approved the final version of the manuscript.

References

Berman, H. M., Battistuz, T., Bhat, T. N., et al. 2002, Acta Crystallographica Section D

Cheng, J., Randall, A., & Baldi, P. 2006, Proteins

Choi, Y., & Chan, A. P. 2015, Bioinformatics

de Lange, T. 2005, Genes & Development

DeDecker, B. S., O'Brien, R., Fleming, P. J., et al. 1996, Journal of Molecular Biology

Frescas, D., & de Lange, T. 2014, Journal of Biological Chemistry

Getov, I., Petukh, M., & Alexov, E. 2016, Bioinformatics

Hubbard, T., Barker, D., Birney, E., et al. 2002, Nucleic Acids Research

Jubb, H. C., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., & Blundell, T. L. 2017, Journal of Molecular Biology

Kumar, P., Henikoff, S., & Ng, P. C. 2009, Nature Protocols

Landrum, M. J., Lee, J. M., Riley, G. R., et al. 2014, Nucleic Acids Research

López-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M., & Gelpí, J. L. 2017, Nucleic Acids Research

Paladin, L., Piovesan, D., & Tosatto, S. C. E. 2017, Nucleic Acids Research

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., & Blundell, T. L. 2017, Nucleic Acids Research

Pike, A. M., Strong, M. A., Ouyang, J. P. T., & Greider, C. W. 2019, Molecular Cell

Pires, D. E. V., Ascher, D. B., & Blundell, T. L. 2013, Bioinformatics

—. 2014, Nucleic Acids Research

Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N., & Bahar, I. 2020, Bioinformatics

Ramensky, V., Bork, P., & Sunyaev, S. 2002, Nucleic Acids Research

Reva, B., Antipin, Y., & Sander, C. 2011, Nucleic Acids Research

Richards, F. M., & Lim, W. A. 1993, Quarterly Reviews of Biophysics

Savage, S. A., Giri, N., Baerlocher, G. M., et al. 2008, Nature Genetics

Sherry, S. T., Ward, M. H., Kholodov, M., et al. 2001, Nucleic Acids Research

Stenson, P. D., Mort, M., Ball, E. V., et al. 2009, Genome Medicine

Yang, D., He, Q., Kim, H., Ma, W., & Songyang, Z. 2011, Journal of Biological Chemistry

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 39–43, 2026
Open Access

Investigation of inhibitory potential of resveratrol to the microtubule affinity regulating kinase 4: Implications in anticancer therapy

Khuzin Dinislam[1], Minhal Abidi[2] and Pratibha Prasad[3,*]

[1] *Bashkir State Medical University, Department of General Chemistry, Ufa, Republic of Bashkortostan, Russia. E-mail: Dinislamkhuzin@mail.ru*
[2] *Department of Biochemistry, Jawaharlal Nehru Medical College, Aligarh Muslim University, Aligarh, India*
[3] *Department of Basic Medical Sciences, Ajman University, United Arab Emirates*
[*] *Corresponding author: Pratibha Prasad; Department of Basic Medical Sciences, Ajman University, United Arab Emirates; E-mail: p.prasad@ajman.ac.ae*

Abstract

Resveratrol shows efficacy against a wide range of cancers and neurodegeneration by targeting specific molecular pathways involved in cancer progression. It exhibits neuroprotective properties, which are crucial for reducing neurodegenerative diseases like Alzheimer's disease (AD). It works by reducing neuroinflammation, protecting neurons from oxidative damage, and controlling important signaling pathways that are essential for synaptic plasticity and neuronal survival. One of the main targets in cancer and neurodegenerative diseases, the kinase MARK4 is linked to the formation of tauopathies like AD. Resveratrol is an excellent antioxidant, anti-cancer and a neuroprotectant used as an inhibitor against the kinase MARK4. To comprehend the binding and structural changes in the protein's structure upon ligand binding, *in silico* investigations were conducted. After the *in silico* experiments, the pure protein was studied *in vitro*. With an IC_{50} value of 7 µM, the natural substance resveratrol effectively inhibited the protein. Additionally, a significant binding affinity for MARK4 was demonstrated by the fluorescence binding experiment ($K = 10^4\ M^{-1}$). The findings demonstrated that the ligand attaches to the protein's binding pocket and blocks MARK4's enzymatic activity, indicating that the substance is a strong binding agent against MARK4.

Keywords: *Resveratrol, Antioxidant, Anti-cancer, Neuroprotectant, Drug discovery*

Received: August 1st, 2025 **Accepted:** November 25th, 2025 **Published:** January 5th, 2026

1. Introduction

Protein kinases form one of the largest and functionally most diverse enzyme groups of eukaryotic biology, which coordinate cellular signalling by reversible phosphorylation of serine, threonine, or tyrosine residues on substrate proteins Mohammad et al. (2021); BIOVIA, Dassault Systèmes (2017). It is a post-translational modification that functions as a molecular switch regulating protein activity, localisation, stability, and intermolecular interactions, and thus determining cellular fate choices DeLano (2002). Over 500 kinases constitute the human kinome, which mediates the combination of extracellular signals with intracellular events and controls cell proliferation, cell metabolism, cell cytoskeleton, apoptosis, and stress-adaptation Atiya et al. (2023). Considering this pivotal regulatory position, disruptions in kinase network signaling have a close relationship with disease pathogenesis, especially cancer, in which disrupted phosphorylation cascades result in malignant transformation and resistance to therapy Alrouji et al. (2023).

The development of cancer is defined by the constant activation of signaling pathways that maintain the proliferation, inhibit cell death, and induce invasion and metastasis Dahiya et al. (2019). Although it has been widely studied, there is growing interest in kinases that control cellular architecture and polarity; many of these have been neglected due to their lack of association with the classical oncogenic kinases, including receptor tyrosine kinases and components of the mitogen-activated protein kinase MAPK pathway Fan et al. (2019). These kinases not only provide structural support, but they are also involved in plasticity signalling, cell cycle and metastatic dissemination signalling. The microtubule affinity-regulating kinase (MARK) plays a role in cell cycle progression and cytoskeletal dynamics. Shamsi et al. (2020).

MARK kinases are a subfamily of AMP-activated protein kinases (AMPK), and it comprises of four isoforms (MARK1-4) which regulate microtubule stability by phosphorylating microtubule-associated proteins (MAPs). MARK kinases regulate cytoskeletal reorganisation, vesicle trafficking, polarity, and organisation of mitotic spindles by regulating the separation of MAPs from microtubules Anwar et al. (2022). MARK4 isoform that has received specific attention among them is its rampant tissue distribution and up-and-coming functions in the pathological conditions like those associated with neurodegenerative illnesses, metabolic imbalance, and malignancy. MARK4 is structurally a conserved catalytic kinase molecule with a regulatory region on either side of a conserved

catalytic domain that enables it to participate in the coordination of structural and signaling pathways Shamsi et al. (2020).

Research suggests that MARK4 is overexpressed in multiple types of cancers, and it leads to tumour progression in a variety of ways. Increased levels of MARK4 have been linked to increased cell proliferation, migration, invasion, and epithelial-to-mesenchymal transition (EMT)Valeur & Berberan-Santos (2013). Mechanistically, MARK4 controls cytoskeletal dynamics, controls mitotic spindle activity, and is linked with signalling cascades that regulate cellular survival and metabolic adjustment. This dualism structural organization MARK4 and signal transduction modulation makes this kinase a central non-canonical oncogenic kinase Eftink (1997); Lakowicz & Weber (1973). Notably, the MARK4 activity seems to support metastatic phenotypes, such as epithelial polarity loss, motility, and stress-resistance, which can be regarded as its therapeutic importance.

Specific inhibition of kinases has significantly improved the therapeutic intervention of cancer, but the difficulties to overcome, including the development of resistance, off-target toxicity, and redundancy of pathways, remain. This means that the discovery of new targets of kinases and the development of other forms of inhibition methods are still priorities Lakowicz (2006). The reason why MARK4 is a promising candidate is that its blocking disrupts both cytoskeletal structure and oncogenic signalling pathways. Empirical evidence has shown that MARK4 blockade activates cell-cycle arrest, inhibits malignant cell proliferation, and induces apoptosis, which indicates that MARK4 blockers could provide multifunctional therapeutic effects Vivian & Callis (2001).

Along with the growth of the kinase-based drug discovery, natural products have resurfaced as a leading modulator of cellular signalling pathways Taraska & Zagotta (2010). Natural products offer structurally diversified chemical scaffolds that may interact with protein kinases in an ATP-competitive, substrate-competitive, or allosteric-mechanism [16]. The idea of natural molecules is also translational, with a number of anticancer drugs being based on the molecules that can be found in nature. Phytochemicals like flavonoids, phenolic acids, and terpenoids have been reported to control phosphorylation networks, dampen oncogenic signaling and increase cellular responses to stress in recent years [17]. Their pleiotropic property enables them to modulate many pathways at once, a property that is highly important in complex diseases such as cancer.

Natural compounds have been singled out by emerging research to have a specific MARK4 modulatory effect [18]. The research by Anwar et al. revealed that rosmarinic acid has anticancer effects due to its mechanism by which it directly interacts with MARK4 and consequently inhibits the proliferation of tumour cells and triggers apoptosis [19]. This study was a mechanistic understanding of the ability of phytochemical scaffolds to disrupt cytoskeletal kinase activity. Later studies indicated that the flavonoid myricetin suppresses breast and lung cancer cell growth by suppressing MARK4, which supports the use of the kinase as a therapeutic target [20]. Notably, these results indicated that structurally different natural molecules accumulate around MARK4, thereby giving the indication that the kinase has structural properties that can be utilized in the design of phytochemical-based drugs [21]. In addition to the sphere of oncology, MARK4 inhibition with the help of natural compounds has been shown in neurodegenerative conditions[22]. Bacopaside II was also discovered to mediate with MARK4 and regulate disease-related signalling, hence providing momentum on the broader biological importance of MARK4 targeting [18]. These studies, in general, define MARK4 as a key molecular hub involved in a variety of pathologies and, therefore, support the hypothesis that natural compounds could be universal modulators of this kinase. The targeting of cross-diseases in this way is especially appealing, as the control of cytoskeleton is central to cellular homeostasis [18].

Resveratrol (RV), a non-flavonoid polyphenol is a prevalent phytochemical under intense research in the field of bioactive plant-based compounds that demonstrates a comprehensive therapeutic capability. It is a derivative of stilbene that is found in grapes, berries, and peanuts having antioxidant, anti-inflammatory, cardioprotective, neuroprotective, and anticancer properties. [23]. On the molecular level, resveratrol disrupts a wide variety of signalling cascades, such as PI3K/Akt, MAPK, NF-κB signalling, p53, and sirtuin signalling. Such interactions regulate the major cell functions, including proliferation, apoptosis, angiogenesis and metabolic adaptation. All these interactions are shown to modulate cell proliferation, apoptosis, angiogenesis and metabolic adaptation processes which are the key processes in cancer biology [24].

RV has been demonstrated to cause tumor growth, cell cycle arrest, apoptosis and metastasis inhibition in cancer models. Its mechanism of action can often be interfered with by kinase-mediated phosphorylation cascades, and this implies that kinase modulation is a fundamental part of its anticancer effect [25]. It is also interesting to note that resveratrol can influence cytoskeletal organization and microtubule dynamics, which are closely linked to the MARK4 activity [22]. Resveratrol is a logical extension of the same mechanistic paradigm, which has been shown to be inhibited by related polyphenols, including myricetin and rosmarinic acid. It interacts with kinase catalytic domains via its polyphenolic structure and its pleiotropic signaling effects could increase its therapeutic activity by producing a coexisting effect on multiple pathways [26].

MARK4 is a relevant model of a kinase with its inhibition having the potential to modulate structural and signalling elements of tumour progression, and natural products, including RV, represent chemically versatile templates that can implement such changes [22]. The convergence highlights that phytochemical-derived kinase inhibitors will be useful as complements or adjuncts to current targeted therapies [30]. Recent research achievement has shown that it is possible to inhibit MARK4 and reduce the growth of cancer cells using natural compounds, which provide a mechanistic basis to phytochemical therapeutic interventions [31]. A prospective candidate in this emerging structure is resveratrol with its thoroughly established anticancer action and ability to regulate kinase signaling. The decoding of the interaction between MARK4 signalling and RV-mediated regulation could provide new opportunities for multi-target cancer therapy and precision medicine. The further convergence of kinase biology, structural pharmacology and natural product research is set to play a critical role in turning these findings into clinically relevant interventions.

2. Materials and Methods

2.1. Materials

Difco LB broth (Becton, Dickinson and Company, Sparks, MD, USA) was autoclaved at 121 °C for 15 min and stored at room temperature. IPTG (Sigma-Aldrich, Bengaluru, India; now Merck KGaA, Darmstadt, Germany) was prepared as a 1 M stock in

Milli-Q water and stored at −20 °C. Kanamycin (50 mg/mL) and ampicillin (100 mg/mL) stock solutions were prepared in Milli-Q water and stored at −20 °C. CAPS buffer was prepared as a 1 M stock in Milli-Q water, and pH was adjusted as required. NaCl (5 M) and glycerol (100% v/v) stock solutions were prepared in Milli-Q water and stored at room temperature. Resveratrol (Sigma-Aldrich) was prepared as stock solutions in an appropriate solvent (e.g., DMSO) and stored according to the manufacturer's instructions.

2.2. Molecular Docking

The three-dimensional crystal structure of human MARK4 was retrieved from the RCSB Protein Data Bank (PDB ID: 5ES1) at a resolution of 2.8 Å. Prior to docking, the protein structure was prepared by removing co-crystallised ligands and water molecules, followed by the addition of hydrogen atoms and assignment of Kollman charges using MGL Tools to ensure proper geometry and charge distribution for molecular docking analysis Mohammad et al. (2021). Molecular docking of MARK4 with resveratrol was carried out using InstaDock. Blind docking was carried out to explore the surface of the entire protein, applying exhaustiveness of 8. The resultant poses were further analysed on the basis of binding affinity and pattern of interaction. All the interactions between the protein and ligand, including hydrogen bonds and hydrophobic interactions, were studied with Discovery Studio Visualizer and PyMOL BIOVIA, Dassault Systèmes (2017); DeLano (2002).

2.3. Protein expression and purification

The protein MARK4 was expressed in M15 *E.coli* bacterial cells. The M15 strain contains the pREP4 plasmid, which has the lac repressor (lacI) gene, allowing regulation of protein expression under the T5 promoter in pQE vectors. The plasmid vector selected was pQE30 which is compatible with M15 and carried N-terminal polyHistidine tag, facilitating as a tag in purification and high yield. The recombinant plasmid with the protein of interest was transformed into *E.coli*-M15 cells, and protein purification was carried out with the previously established protocols Atiya et al. (2023); Alrouji et al. (2023).

2.4. Enzyme inhibition assay

Enzyme here is a kinase, MARK4, which works by using ATP as a substrate for its activity, converting ATP to ADP, and the released phosphate is used by the kinase to phosphorylate its target. For the assay mentioned here, the principle is that the protein kinase is considered 100 per cent active by measuring the amount of phosphates that are released upon interaction with ATP. Keeping the concentration of protein constant, different concentrations of ligand are added to the protein and incubated, allowing the ligand to bind with the protein. Afterwards, ATP is added to the control, i.e. protein without ligand and protein with increasing concentration of ligand, and the released phosphate is estimated with a reagent known as BIOMOL green reagent. MARK4 and ATP stock solutions were prepared, and the reaction was set in a 96-well plate by mixing protein, ATP, buffer and ligands to obtain a final MARK4 concentration of 5 µM and ATP of 200 µM. The reaction mixtures were first set up without ATP and incubated with increasing concentrations of ligands for 30 minutes at 37°C. Further, ATP was added and incubated at 37°C for approximately 30 minutes. Following incubation, Biomol Green reagent was added according to the manufacturer's instructions to terminate the reaction and enable detection of released inorganic phosphate. The mixtures were thoroughly mixed and incubated at room temperature in the dark to allow full color development. Absorbance was measured at 620 nm using a microplate reader.

2.5. Fluorescence binding studies

The binding studies of RV with MARK4 were performed on the Jasco spectrofluorometer at 25°C. The compound was prepared by dissolving it in DMSO, giving the desired 1 mM working concentration in the working buffer (pH 8.0). The binding studies were performed with a fixed concentration of MARK4 (4 µM), and the ligand was added gradually in increasing concentration till saturation. The addition of ligand was mixed thoroughly and gently to ensure interaction between protein and ligand before measurements. The intrinsic fluorescence of protein, mainly contributed by the aromatic amino acids and predominantly Tryptophan, was measured by excitation at 280 nm, which generated an emission spectrum at 300–400 nm. The fluorescence quenching observed in the emission spectra was analysed using the modified Stern-Volmer equation Dahiya et al. (2019).

In the given equation, F0 and F denote the fluorescence intensity of protein and protein ligand complex, respectively. K is the binding constant; n denotes number of binding sites and C for concentration of RV.

$$\frac{F_0}{F} = 1 + K_{sv}[C] \tag{1}$$

$$\log\left(\frac{F_0 - F}{F}\right) = \log K + n \log[C] \tag{2}$$

2.6. Cell proliferation assay

Cancer cells H1299 are a widely used human non-small cell lung carcinoma (NSCLC) cell line, specifically identified as a large cell carcinoma subtype. The cells were procured from NCCS, Pune, India. A culture medium was prepared for H1299 cells with antibiotics and heat-inactivated FBS. Using a 96-well culture plate, approximately 5×10^3 cells were plated per well following passages till passage number 20, with a confluency of 85%, and all the treatments were given. A compound stock solution was made with DMSO, with a maximum limit of 0.2% per treatment. The MTT stock was prepared as 5 mg/ml, and 10 μL of the stock was incubated in all the treated wells at 37°C with 5% CO_2 for a few hours. The 96-well plate was read spectrophotometrically at 570 nm using a 96-well plate reader.

3. Results and Discussion

3.1. Molecular docking analysis

To understand the binding of RV with MARK4, molecular docking studies were carried out as shown in Figure 1. According to the docking research, RV interacts with Ile62 and Glu133 by typical hydrogen bonds, which are crucial for anchoring the ligand inside the binding cavity. The stability of ligand-protein complexes is greatly aided by hydrogen bonds, which are important factors in determining ligand specificity. RV also forms multiple hydrophobic and van der Waals interactions with several residues lining the MARK4 binding pocket, including Met132, Val116, Tyr134, Gly138, Glu139, Asp142, Ala135, and Leu185. Together, these interactions help keep the ligand inside the kinase's catalytic domain and stabilize the structure of the MARK4–RV complex. Moreover, π–alkyl interactions with Val70 and Ala83 were noted, suggesting stronger hydrophobic interactions between RV's aromatic ring and the nonpolar residues in the active site. Docking

studies reveal RV occupies the catalytic pocket of MARK4 with multiple interactions Fan et al. (2019); Shamsi et al. (2020); Anwar et al. (2022).

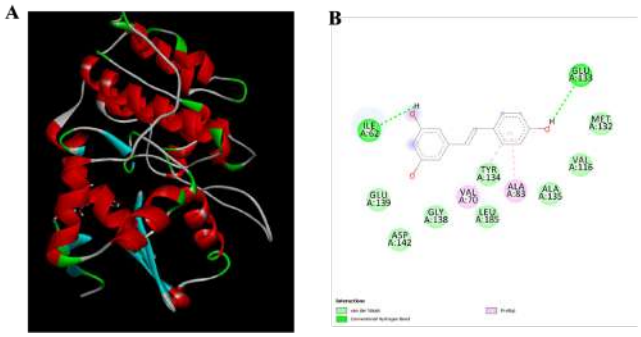

Figure 1. Molecular docking of RV with MARK4; A. 3D interaction of RV and MARK4; B. 2D interactions of RV and MARK4.

3.2. Fluorescence-based binding studies

The aromatic amino acids in the protein structure are responsible for the intrinsic fluorescence of the protein. The major contributor amongst all aromatic amino acids is tryptophan (Trp), followed by tyrosine and phenylalanine. Trp has a high quantum yield, resulting in high protein fluorescence at 300–350 nm when excited at 280 nm. The intensity and spectral position of this emitted light provide information about the structural and environmental status of the protein Valeur & Berberan-Santos (2013); Eftink (1997); Lakowicz & Weber (1973). Trp fluorescence intensity and emission wavelength can be dramatically changed by changes in solvent polarity, conformational shifts, or ligand binding, which makes it an effective probe for researching protein structure, folding, and intermolecular interactions Lakowicz (2006); Vivian & Callis (2001). As tryptophan fluorescence is highly sensitive to conformational rearrangements and binding-induced perturbations, it is widely used to investigate protein-ligand interactions, conformational changes, and to estimate binding constants Taraska & Zagotta (2010). In the current investigation, an increase in the ligand concentration resulted in decreased fluorescence intensity of the protein, as shown in Figure 2A. The quenching in the fluorescence of the protein can likely be due to the changes in the microenvironment near the tryptophan residues in the presence of RV. The binding constant of the interaction was estimated to be $7.0 \times 10^4\ M^{-1}$ using the modified Stern-Volmer equation (Figure 2B).

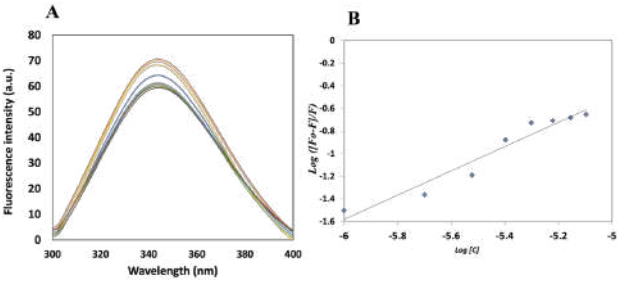

Figure 2. (A) Fluorescence emission spectra of native MARK4 protein and increasing concentration of RV (0–8 µM). (B) Modified Stern-Volmer plot of MARK4–RV system.

3.3. Enzyme inhibition assay

The kinase MARK4 is an important therapeutic target due to its association with various diseases, including cancer and neurodegeneration. MARK4 has a strong correlation with Alzheimer's disease, as MARK4 is associated with pathological phosphorylation of tau protein, contributing to dementia and early onset of AD. MARK4 is upregulated in various cancers, including breast cancer, lung cancer and hepatocellular carcinoma Mohammad et al. (2021); BIOVIA, Dassault Systèmes (2017); DeLano (2002). Therefore, MARK4 has emerged as an interesting therapeutic target in cancer, neurodegeneration and other diseases. RV is a naturally occurring compound with documented anti-inflammatory, anti-oxidant, anti-ageing, anti-cancer and neuroprotective activities. In order to study the influence of RV on kinase activity and to complement molecular docking and fluorescence-based binding studies, an enzyme inhibition assay was carried out.

A decrease in kinase activity of MARK4 was observed with increasing ligand concentration (Figure 4), indicating inhibition of the kinase activity of the protein, which can be due to the formation of protein-ligand association, as evident from docking studies. AAT Bioquest IC_{50} calculator was used to determine the half-maximal inhibitory concentration (IC_{50}), which is the inhibitor concentration needed to decrease enzyme activity by 50% as shown in Figure 3. In line with earlier findings, the IC_{50} value for RV was found to be ≈ 7 µM. Dahiya et al. (2019); Fan et al. (2019). These findings validate RV's therapeutic potential and confirm its status as a strong MARK4 inhibitor. RV might be a promising addition to the expanding family of small-molecule kinase inhibitors that target MARK4, considering the pivotal role that kinases play in a variety of pathological situations, such as cancer and neurological diseases.

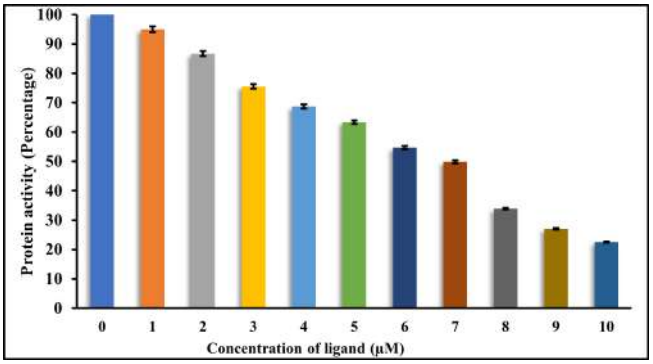

Figure 3. ATPase activity assay of MARK4 with varying RV concentration (0–10 µM).

3.4. Cell proliferation assay

H1299 lung cancer cell lines were used to examine the cytotoxicity of RV against them. We performed a cell survival assay to verify the anti-cancerous activity of the compounds. The cell growth was observed for 48 hours following the treatment with increasing concentrations of RV. Compared to the DMSO vehicle control, RV showed significant cytotoxicity in the H1299 cell lines. The IC_{50} values of RV against H1299 were ≈ 114 µM, as shown in Figure 4.

4. Conclusion

To conclude, the current paper presents an in-depth analysis of resveratrol as an antimicrotubular affinity-regulating

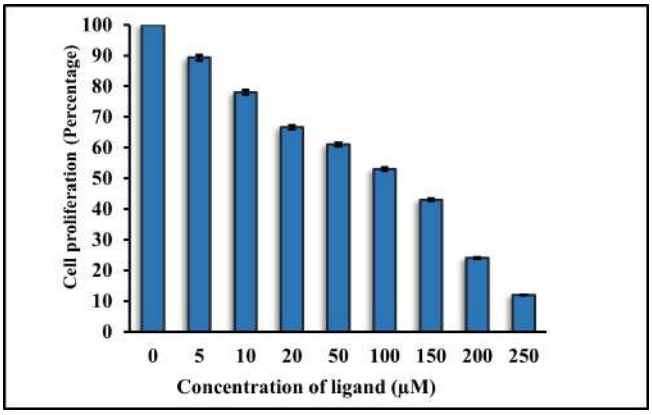

Figure 4. Cell proliferation assay of RV against cell line H1299.

kinase 4 (MARK4) inhibitors, a kinase that is gradually gaining prominence as the cause of cancer advancement and neurodegenerative diseases. Combining both computational and experimental methods, we can show that resveratrol forms a strong interaction with MARK4 and that it can effectively regulate the enzyme activity. The molecular docking analysis exhibited positive binding of resveratrol in the catalytic site of MARK4, which is an indication of structural compatibility and may dislodge the kinase activity. The binding studies conducted using fluorescence also supported the presence of a direct interaction between resveratrol and MARK4 as demonstrated by the quenching of intrinsic protein fluorescence with rising resveratrol concentration, and calculation of a strong binding constant, which indicates a high affinity. Enzyme inhibition assays of functional validation showed that resveratrol inhibits the activity of the MARK4 kinase dose-dependently with an IC_{50} of the order of micromolar, which is an indicator of a good small-molecule inhibitor.

In combination, these data indicate resveratrol as a promising phytochemical scaffold that could be used to target MARK4 and disrupt the kinase-mediated signalling pathways related to cytoskeletal regulation, tumour progression and neurodegeneration. The additional multi-target signalling capabilities of resveratrol enhance the future of resveratrol as a systems-level therapeutic agent. Nevertheless, more research in cellular models, structural dynamics and pharmacokinetic optimization is required to convert these results into clinical applications. Although these encouraging results exist, there are still translation issues. Naturally occurring substances tend to have low bioavailability, short metabolism, and situational potency. These limitations are being overcome using medicinal chemistry, structural biology, and delivery systems based on nanotechnology to optimise the scaffold of phytochemicals and enhance pharmacokinetic activities. In-depth knowledge of the interactions of MARK4 with ligands, such as binding site structure and conformational dynamics, will be fundamental to rational drug design [29].

Altogether, this article contributes to the comprehension of MARK4 modulation by natural compounds and offers a mechanistic concept of phytochemical-based kinase inhibitors evolution. The MARK4 resveratrol-interaction is also a good beginning in rational drug design to multi-pathway therapeutic intervention of cancer and other related diseases.

Declaration of Competing Interest

The authors report no conflicts of interest.

Data availability statement

All data generated or analyzed during this study are included in this article.

References

Alrouji, M., et al. 2023, International Journal of Biological Macromolecules, 235, 123831
Anwar, S., et al. 2022, Journal of Molecular Liquids, 355, 118928
Atiya, A., et al. 2023, Journal of Biomolecular Structure and Dynamics, 41, 8824
BIOVIA, Dassault Systèmes. 2017, Discovery Studio Visualizer
Dahiya, R., et al. 2019, RSC Advances, 9, 23302
DeLano, W. L. 2002, CCP4 Newsletter on Protein Crystallography, 40, 82
Eftink, M. R. 1997, Methods in Enzymology, 278, 221
Fan, J., Fu, A., & Zhang, L. 2019, Quantitative Biology, 7, 83
Lakowicz, J. R. 2006, Principles of Fluorescence Spectroscopy (Springer)
Lakowicz, J. R., & Weber, G. 1973, Biochemistry, 12, 4161
Mohammad, T., Mathur, Y., & Hassan, M. I. 2021, Briefings in Bioinformatics, 22
Shamsi, A., et al. 2020, Biomolecules, 10, 789
Taraska, J. W., & Zagotta, W. N. 2010, Neuron, 66, 170
Valeur, B., & Berberan-Santos, M. N. 2013, Molecular Fluorescence: Principles and Applications (John Wiley & Sons)
Vivian, J. T., & Callis, P. R. 2001, Biophysical Journal, 80, 2093

Clinical & Molecular Biomedicine, Vol. 1, No. 1, pp. 44–55, 2026
Open Access © 2026 Editory Press

Genome-wide annotation and structural modeling of hypothetical proteins in *Listeria monocytogenes*

Shama Khan[1,*]

[1] *South African Medical Research Council, Vaccine and Infectious Diseases Analytics Research Unit (VIDA), Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa*
[*] *Corresponding author: Shama Khan, PhD , Wits Vaccines and Infectious Diseases Analytics (VIDA) Research Unit , Faculty of Health Sciences, University of the Witwatersrand , Email: Shama.Khan@wits-vida.org; Shama.Khan@wits.ac.za , ORCID: 0000-0002-0874-5029*

Abstract

The rapid expansion of genome sequencing projects has resulted in the identification of numerous hypothetical proteins whose functions remain uncharacterized. In *Listeria monocytogenes* serotype 4b, a major food-borne pathogen associated with high mortality rates, several predicted proteins lack functional annotation despite their potential role in pathogenicity and survival. In the present study, a comprehensive *in silico* approach was employed to functionally annotate 92 hypothetical proteins identified from the genome of *L. monocytogenes*. Protein sequences were retrieved and analyzed using sequence similarity searches, conserved domain identification, motif analysis, and multiple sequence alignment. Functional classification was performed based on BLAST, Pfam, and InterPro analyses. Structural prediction was performed via homology modeling via the SWISS-MODEL server, followed by structural validation and comparative analysis using PyMOL and DALI-Lite. Functional inference from structural models was supported by conserved-residue mapping and ProFunc analysis. Sequence-based analysis enabled classification of most hypothetical proteins into functional groups, including hydrolases, transferases, transporters, kinases, stress-response proteins, membrane proteins, DNA-binding proteins, and ATP-binding proteins. Structure-based modeling of selected proteins further confirmed the predicted catalytic residues, metal-binding sites, ligand-interaction sites, and conserved functional motifs. Several proteins were predicted to be involved in enzymatic activity, nucleotide metabolism, membrane transport, transcriptional regulation, and stress adaptation. This integrative computational analysis provides functional insights into previously uncharacterized proteins of *L. monocytogenes*. The findings enhance genome annotation quality and identify potential targets for further experimental validation, contributing to a better understanding of bacterial physiology and pathogenesis.

Keywords: *Listeria monocytogenes, hypothetical proteins, functional annotation, homology modeling, genome analysis, structural prediction, protein function prediction*

Received: September 1st, 2025 **Accepted:** December 10th, 2025 **Published:** January 5th, 2026

1. Introduction

Due to the speed and cost-effectiveness of genome sequencing, many small bacterial, archaeal, and eukaryotic genomes have been sequenced, and larger eukaryotic genomes are expected to be fully sequenced soon. In such a case, annotation becomes problematic when genes are predicted faster than annotation can keep pace. Despite ongoing improvements in genome annotation, a substantial proportion of open reading frames remain classified as "conserved hypothetical proteins." Hypothetical proteins are predicted from nucleotide sequences but lacking direct experimental validation at the protein level (**?**). In many instances, these predicted proteins show limited sequence similarity to previously characterized and annotated proteins. Conserved hypothetical proteins constitute a significant fraction of genes in sequenced genomes; they encode proteins identified across multiple phylogenetic lineages but remain uncharacterized with respect to biological function or biochemical properties, sometimes accounting for a large portion of the predicted proteome (**?**).

Determining the functions of hypothetical proteins is essential for improving genomic and proteomic annotations. The identification of new hypothetical proteins can reveal previously unknown structural folds and biological activities (**?**). As more structures are characterized, novel domains, motifs, and conformational arrangements are likely to emerge, contributing to a clearer understanding of structure–function relationships in proteins (**?**). Functional characterization may also uncover new molecular pathways and regulatory cascades. Furthermore, hypothetical proteins may serve as biomarkers or therapeutic targets. Their identification can facilitate the discovery of previously unrecognized or computationally predicted genes (**?**). Ultimately, elucidating the roles of these proteins will deepen our understanding of biological systems and may contribute to the development of improved therapeutic interventions (**?**Rao, 1989).

The biological roles of many predicted proteins remain unknown because their existence has been inferred solely from genomic sequences (**?**). The rapid completion of genome sequencing projects has generated extensive biological datasets, enabling the association of genes with their corresponding protein products, which are fundamental to cellular function.

Advances in genome and cDNA sequencing technologies allow the prediction of numerous proteins for which experimental evidence is lacking (Gury et al., 2004). Functional insights into hypothetical proteins are often derived from sequence comparisons with proteins of known function in model organisms. Since similarities in sequence or conserved domains frequently indicate shared functional properties, such analyses can suggest potential biological roles (**?**). Moreover, if a hypothetical protein exhibits significant structural homology with characterized proteins, it is reasonable to infer comparable molecular features (**?**).

Traditional biochemical and molecular biology approaches can accurately assign functions to genes; however, these experimental methods are labor-intensive and costly. Consequently, even in fully sequenced genomes, only about 50–60% of genes have been functionally annotated (**?**). Automated genome analysis and computational annotation strategies provide alternative means to interpret genomic information. Therefore, determining protein function remains a major challenge in the post-genomic era. This has increased reliance on bioinformatics tools to predict the roles of uncharacterized protein sequences, and several contemporary computational approaches have been developed to address this issue (**?**).

Listeria monocytogenes is the causative agent of listeriosis. It is a facultative anaerobic bacterium capable of surviving under both aerobic and anaerobic conditions (**?**). The organism can invade and proliferate within host cells and is recognized as one of the most severe foodborne pathogens, with mortality rates of 20–30% in clinical cases (**?**). In the United States, listeriosis is responsible for approximately 2,500 reported cases and 500 deaths annually, making it one of the leading causes of death among foodborne bacterial infections (**?**). Although relatively rare, the disease is associated with a high case-fatality rate. Its emergence as a public health concern has been attributed to changing dietary habits, advances in food preservation technologies that extend shelf life, and the bacterium's ability to survive and grow at refrigeration temperatures.

Disease caused by *Listeria monocytogenes* is called Listeriosis (a bacterial infection). In most cases, these infections occur in immunocompromised individuals (**?**), such as pregnant women (they are about 20 times more likely to get infected than healthy adults), Newborns, persons with diabetes, cancer, and kidney disease, persons with AIDS, and persons who take glucocorticosteroid medications. Listeriosis is fatal in at least 1 in 5 infected individuals. *Listeria monocytogenes* causes flu-like symptoms and meningitis if it spreads to the central nervous system. The first reported case of Listeriosis was in 1981; since then, numerous listeriosis outbreaks have been detected and investigated. Subsequent studies have confirmed that the *Listeria monocytogenes* serovar primarily responsible for Listeria infection is serotype 4b. So, of the 13 serovars of *Listeria monocytogens*, *Listeria monocytogenes* 4b is the most prominent one (**?**). *Listeria monocytogenes* has been reported to cause miscarriages, stillbirths, or even premature pregnancies among pregnant women with underlying diseases. The chances of a fetus being affected by an affected mother are high.

The *Listeria monocytogenes* genome is approximately 3 Mb. Its genome consists of one chromosome. Total number of nucleotides is 2912690, total number of protein genes found to be 2766, and total number of RNA genes is 85. Many stress proteins are expressed in *Listeria monocytogenes* under stressful conditions, helping to bacterium survive harsh environments for extended periods. There are several important facts about *Listeria monocytogenes* genome available. Genome analysis suggests that the *Listeria monocytogenes* 4 b genome contains several hypothetical proteins. There are 92 hypothetical proteins in *Listeria monocytogenes* 4 b. There is little information available on these hypothetical proteins from *Listeria monocytogenes* genome. Here, our aim is to annotate the hypothetical proteins from *Listeria monocytogenes* by using recent bioinformatics tools. After annotating a genome, we grouped genes into two types: the first comprises known genes with functional characterization; the second comprises conserved hypothetical genes conserved across many organisms. Generally, a newly sequenced genome contains approximately 30% of genes that are poorly or incorrectly annotated (**?**). Hence, these poorly characterized hypothetical genes might play an important role in our understanding of the pathogen's pathogenesis and virulence. Given that so little is known about these genes, we have attempted to assign plausible functions to hypothetical proteins. The functional evidence suggests that these hypothetical proteins may have specific functions, but further detailed analysis to needed to confirm their functions.

2. Materials and Methods

2.1. Data mining

We have analyzed the *Listeria monocytogenes*'s genome and the hypothetical proteins that it contains by retrieving the information at the P1R website. We have identified and downloaded 92 hypothetical protein sequences of *Listeria monocytogenes*. The sequences of all proteins analyzed with their primary accession number, length, molecular mass, and proposed functions (Table 1–3). Protein sequences are saved in FASTA format for further processing. A systematic pipeline used in this study is illustrated in Figure 1.

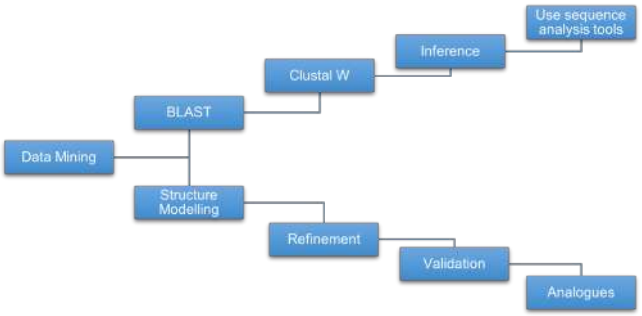

Figure 1. Outline of structure and function assignment to the hypothetical proteins.

2.2. Sequence similarity search

We have derived the function of hypothetical proteins from sequence similarity to a well-characterized homolog in GenBank. Standard protein-protein BLAST (BLASTP) is used to identify a query amino acid sequence and to find similar sequences in protein databases. The results of a BLAST query against a public database provide insight into the functional properties of related sequences. Hence, BLASTP revealed several related searches for our hypothetical proteins. With the help of multiple sequence alignments (ClustalW) of the related selected sequences, we have been able to analyze the conserved residues present in our

hypothetical protein by comparing with the proteins of known function, and we have related the functions of the conserved positions of some residues. A multiple sequence alignment was performed using ClustalW to study conserved sequence patterns and ancestral relationships among the organisms. The next step was to predict the functions using structural analysis.

2.3. Structure Prediction

The three-dimensional structures of hypothetical proteins were examined to infer their potential biological functions. Structural prediction was carried out using homology modeling, which generally involves multiple sequential steps. First, a suitable template protein with an experimentally determined structure, typically belonging to the same or a closely related family, is identified. Second, appropriate sequence alignment tools are employed to obtain the best possible alignment between the target and template proteins. Third, the target sequence is divided into smaller segments to facilitate accurate modeling. Fourth, corresponding fragments are searched in structural databases based on sequence similarity and the spatial conformation of the selected template. Fifth, the spatial coordinates of these matched segments are assembled and adjusted to construct the full-length target structure, ensuring that all atomic positions are properly incorporated. These procedures are repeated multiple times to generate several models, from which an averaged structure is obtained, followed by comprehensive energy minimization to refine the final predicted model. The 3-D structures were retrieved using the SWISS-MODEL server, an automated comparative modeling system for predicting protein 3-D structures. The entire homology modeling workflow can be performed in the 'project mode' of PyMOL, an integrated sequence-structure workbench. Using the SWISS-MODEL server, we obtained suitable 3D structures for thirteen hypothetical proteins selected for detailed structural analysis, and the resulting structural models were downloaded in PDB format for further analysis.

2.4. Function prediction

Clustering of gene expression profiles is widely used for functional prediction, based on the principle that genes involved in related biological processes tend to exhibit coordinated expression patterns. Functional inference can also be derived from protein–protein interaction networks, where the role of an uncharacterized protein is estimated from the known functions of its interacting partners. Schwikowski and co-workers introduced a neighbor-counting strategy in which the probable function of an unknown protein is determined according to the frequency of functional annotations among its interaction partners. This strategy was later improved by Hishigaki et al. (2001) through the incorporation of a chi-square (χ^2) statistical framework to enhance prediction reliability. In both methods, all functional contributions from neighboring proteins are treated with equal significance. Several publicly accessible signature and motif databases support functional annotation and are frequently integrated with sequence clustering and domain-based resources. These include PROSITE, PRINTS, Pfam, ProDom, Blocks, SMART, and InterPro; in the present study, Pfam and InterPro were specifically utilized. Structural classification of proteins into homologous families, superfamilies, and folds can be performed using databases such as SCOP, CATH, FSSP, CAMPASS, and HOMSTRAD. Additionally, the ProFunc server facilitates structure-based functional prediction by comparing modeled protein structures against multiple databases and reporting significant matches for functional interpretation.

2.5. Visualization

We carried out structural analysis using PyMOL for superposition of molecules, secondary-structure-based alignment, and calculation of RMSD values. PyMOL is also used to depict structures, as well as to analyze potential residue positions and functions. We visualized the template residues and the model after superimposing the model onto the template. Each residue was visualized individually, and residue conservation was assessed.

2.6. Structure analysis

Structure-based comparison approaches are highly effective for detecting remote evolutionary relationships that may not be apparent from sequence similarity alone. Structural information also enables the recognition of functional sites that have arisen independently during evolution. Because protein function is inherently dependent on three-dimensional conformation, structural analysis provides direct insight into the molecular mechanisms underlying biological activity. Local structural alignment focuses on identifying conserved three-dimensional arrangements of specific residues, even among proteins that differ substantially in overall fold. Several computational tools are available for such analyses, including DaliLite, which utilizes Cα–Cα distance matrix comparisons; TOPS; SSM; VAST; GRATH, which represents secondary structure elements as graph nodes connected by spatial relationships; and SSAP, which evaluates structural similarity based on Cα comparisons and additional scoring features. Other methods such as LSQMAN, CE (Combinatorial Extension), MATRAS, and LOCK2 also employ atom-based or secondary structure vector representations for alignment. In the present study, DALI-Light was selected for structural comparison and analysis.

3. Results and Discussion

3.1. Sequence-based Function Prediction

Information derived from BLAST, ClustalW, and other online servers for sequence analysis revealed close similarities of hypothetical proteins to proteins of known function. Based on the proposed functions, we classified all 92 hypothetical proteins into functional classes. All the Hypothetical proteins have been classified into the following categories.

3.1.1. Hydrolase

The enomic proteins of *Listeria monocytogenes* most abundantly showed similarity with proteins having hydrolase function, and hence, by comparing the conserved residues, we can assume the proteins to have the same function as well. The sequence of C1L0U3, C1L0R8, C1KZQ3, C1KZE1, C1KYZ2, C1KYY1, C1KY45, C1KY43, and C1KXZ5 has been predicted to have hydrolase function following a precise analysis that compared its sequence similarity with that of known, characterized proteins. Hydrolases are enzymes that facilitate hydrolytic reactions, in which a water molecule is utilized to break chemical bonds within a substrate. During this process, the hydrogen (H^+) and hydroxyl (OH^-) components of water are incorporated into the reacting molecule, resulting in its cleavage into two or more smaller products. The term "hydrolase" serves as the systematic designation for enzymes classified under Enzyme Commission (EC) class 3, which encompasses all enzymes that catalyze hydrolysis reactions.

For the C1L0U3 gene, 65 matching sequences were found by FASTA search. 8 neighboring genes found as Q8Y970, D2PAB8,

D2NZC4, C1L0U3, E1UE28, B8DIC3, Q92DZ3, D3UKJ4 having the percentage identity of 100%, 100%, 100%, 100%, 99.6%, 99.6%, 99.3%, 91.5% respectively. 10 sequence motifs were identified when the sequence was searched against the InterPro database. There are 57 matching sequences found by FASTA search for C1KZQ3 gene and 7 neighboring genes were found as E1UBT1, B8DAU2, Q8Y3Y0, D2P7B6, D2NVU7, Q927E2, D3USV7 with identities as 99.3%, 99.3%, 97.8%, 96.8%, 96.8%, 95.3% and 83.5% respectively. 9 sequence motifs were found by using InterPro database. For the C1KY43 gene, we identified 47 matching sequences via a FASTA search and 9 sequence motifs via InterProScan. E1UAS5, B8DDF3, Q8Y4N4, D2P959, D2NY72, Q928N2, and D3UR85 are the 7 neighbouring genes, with percentage identities of 98%, 97.6%, 97.2%, 97.2%, 97.2%, 95.6%, and 94%, respectively.

3.1.2. Transferase

Transferases are enzymes that catalyze the movement of specific chemical groups, such as methyl, glycosyl, acyl, or phosphate groups, from one molecule to another. In these reactions, one compound acts as the donor of the functional group, while another serves as the acceptor. The nomenclature and classification of transferases follow the format "donor:acceptor group transferase," reflecting the nature of the transferred group and the participating substrates. Proteins belonging to this group are C1L2K8, C1L1Z7, C1L0P3, C1L0L0, C1L024, C1KYI4, C1KXY4, C1KVW1, and C1KVA0. The protein sequence has been found to be similar to that of characterized proteins in this category. This group is the second most abundant in *Listeria monocytogenes*, after hydrolases. For the gene C1L0L0, 65 matching sequences were identified by a FASTA search, and 8 motifs were matched by InterProScan, which searches the InterPro database. The 7 neighbouring genes are as follows D2PA32, D2NYL3, Q8Y9E8, E1UDU6, B8DA44, Q92E70 and D3USH9 with percentage identity as 99.5%, 99.5%, 99.5%, 96.7%, 96.7%, 92.8% and 89.4% respectively. There are 29 matching sequences found by FASTA search for the C1KXY4 gene and 6 motifs matched while running it against InterPro database. There are 6 neighboring genes found which are E1UCN5, B8DGQ6, Q8YAG1, D2PAN3, D2NZA9 and Q92F98 with percentage identity as 97.1%, 97.1%, 95.5%, 95.1%, 95.1% and 95.9% respectively.

3.1.3. Transporter

Transporter proteins are proteins that move substances within an organism. Transport proteins are vital to the growth and life of all living things. There are several different kinds of transport proteins. The proteins found in *Listeria monocytogenes* included in this category are C1L300, C1L1R1, C1L164, C1L0K2, C1KZ81, C1KZ17, C1KYZ6, C1KY61, and C1KY46. For the C1L1R1 gene, 4 matching sequences were found by FASTA search, and 6 motifs were found by InterPro scan. There were 7 neighboring genes found which are E1UF08, B8DEE5, Q8Y8B8, D2P3J7, D2P0R1, Q92D31, and D3ULQ9 with percentage identity as follows 97.9%, 97.9%, 96.8%, 96.6%, 96.6%, 93.4%, and 87.4%, respectively.

3.1.4. RNA Binding

The following residues belong to this group: C1L2S2, C1L1U8, and C1L1G8. RNA-binding proteins are proteins that bind to RNA recognition motifs (RRMs) of double or single-stranded RNA in cells and participate in forming ribonucleoprotein complexes. They are cytoplasmic and nuclear proteins. For the C1L2S2 gene, 15 matching sequences were identified by FASTA search, and 11 sequence motifs were identified in InterPro. The 7 neighbouring genes are Q8Y7C0, D2P4S6, D2P1Z0, E1UG75, B8DFW2, Q92BY9 and D3UMV0 with sequence identity as 97.4%, 97.4%, 95.2%, 95.2%, 94.9% and 94.9% respectively. Whereas for the C1L1U8 gene, 51 matching sequences were found, and 11 sequence motifs matched in InterPro. There are 7 neighbouring genes which are as follows Q92CZ5, Q7AP12, E1UF45, D2P3N4, D2P0U8, B8DEA8 and D3ULU6 with similarities as 100%, 100%, 100%, 100%, 100%, 100%, and 99.5% respectively.

3.1.5. Hydrolase acting on carbon-nitrogen but not on peptide

Following mentioned are the proteins C1L161, C1KYM2, and C1KXZ0, which belong to this category. This family contains hydrolases that break carbon-nitrogen bonds. This sub family of hydrolase is numbered as EC 3.5. There were 35 matching sequences found by FASTA search and 6 sequence motifs matched in the InterPro scan. There were 6 neighbouring genes found which are E1UD07, D2P8J8, D2NWP2, B8DEW3, Q8YA77, and Q92EZ8 with percentage identity as 99.2%, 99.2%, 99.2%, 99.2%, 98.1% and 95.3% respectively.

3.1.6. Kinase

Kinases are enzymes that catalyze the transfer of phosphate groups from high-energy donor molecules, typically adenosine triphosphate (ATP), to specific target substrates in a reaction known as phosphorylation. Through this process, kinases regulate numerous cellular activities, including signal transduction, metabolism, and cell cycle progression. They belong to the broader class of phosphotransferases. In the present study, proteins C1L143, C1KZU1, and C1KYX3 were identified as members of this enzyme category.

3.1.7. Response to stress

Cells exposed to stressful conditions often accumulate abnormal, misfolded, or damaged proteins as a consequence of environmental or physiological challenges. To mitigate these effects, cells increase the production of stress-related proteins that assist in stabilizing, refolding, or degrading altered proteins. Rather than solely enabling survival under extreme or lethal stress, these stress proteins also play a crucial role in cellular recovery following stress exposure, thereby helping restore normal cellular homeostasis. C1LO28, C1KZM7, and C1KW03 are the proteins that belong to this category. A FASTA search of the C1KZM7 gene identified 21 matching sequences, and an InterPro scan identified 4 sequence motifs. Q927G8, E1UBQ4, D2P7S2, D2NVX3, B8DAW9, Q8Y405 and D3USS9 are the 7 neighbouring genes found with 99.3%, 99.3%, 99.3%, 99.3%, 99.3%, 98.6% and 98.6% identity respectively.

3.1.8. Integral to membrane

An integral membrane protein is a protein molecule, or protein complex, that is permanently embedded within the plasma membrane through hydrophobic regions that interact strongly with the lipid bilayer. These proteins are tightly associated with membrane phospholipids and cannot be easily removed without disrupting the membrane structure. Integral membrane proteins are broadly divided into two principal categories: (1) transmembrane proteins and (2) integral monotropic proteins. Transmembrane proteins represent the more prevalent group and extend completely across the lipid bilayer. Such proteins perform diverse biological functions, including acting as transporters, ion channels, receptors, enzymes, structural anchors, components involved in energy storage and signal transduction, and mediators of cell–cell adhesion. Representative examples include integrins, cadherins, the insulin receptor, neural cell adhesion molecules (NCAM), selectins, glycophorin, and rhodopsin. C1L1R3, C1L1R2, and C1KY64 are the proteins that belong to this group.

For the gene C1KY64, we found 39 matching sequences by FASTA search and 5 motifs matched in InterPro scan. 6 neighbouring genes found were E1UAU6, B8DDD2, Q8Y4L1, D2P980, D2NY93, and Q928L2 with percentage identity as 98.3%, 98.3%, 96.6%, 96.6%, 96.6%, and 97.4%, respectively.

3.1.9. DNA Binding

DNA-binding proteins are proteins that contain a DNA-binding domain, which in turn is responsible for binding to single or double-stranded DNA molecules. C1L0T5 and C1KWG4 are the proteins that have this domain and are therefore classified in this category.

3.1.10. Phosphoesterase

Phosphoesterase is a family of hydrolase with acts on ester bonds. Phosphoesterases are involved in cleaving ester bonds. Proteins belonging to this category are C1L2V7 and C1L2E5. For gene C1L2E5 there are 24 matching sequences found by FASTA search and 6 sequence motifs matched in InterPro. There are 7 neighbouring genes which are Q8Y7N4, E1UFM2, B8DHZ2, D2P485, D2P1E9, Q92CG9 and D3UME0 with percentage identity as 100%, 100%, 100%, 99.4%, 99.4%, 97.6%, and 95.3% respectively.

3.1.11. ATP Binding

An ATP-binding protein is a protein that specifically interacts with adenosine 5′-triphosphate (ATP), a ribonucleotide composed of the purine base adenine attached to the sugar D-ribofuranose and linked to three phosphate groups. ATP serves as a primary carrier of chemical energy and phosphate groups within the cell. Proteins that bind ATP typically utilize this interaction to drive biochemical reactions, regulate cellular processes, or facilitate molecular transport, thereby playing essential roles in energy-dependent cellular functions. C1KYL6 and C1KV71 are the proteins in this category. For the C1KYL6 gene, we identified 69 matching sequences via a FASTA search. D2P8J2, D2NWN6, Q8YA83, Q92F06, D3URG2, E1UCZ8, and B8DEX2 are the 7 neighboring genes found with percentage identities as 98.9%, 98.9%, 98.5%, 90.3%, 89.2%, 89.6%, and 89.6%, respectively. Additionally, 10 sequence motifs were matched in the InterPro scan. For the C1KV71 gene, 102 matching sequences were identified by a FASTA search, and 11 sequence motifs were identified by an InterPro scan. 7 neighboring genes found were E1U8I5, B8DAN0, Q8YAT3, D2PBF4, D2P030, Q92FS4 and D2P8J2 with percentage identity as 96.6%, 96.6%, 96.2%, 96.2%, 96.2%, 95.8% and 50.2% respectively.

3.1.12. Other Proteins

Proteins in this category are proteins that are one of their kind, i.e., no other protein in *Listeria monocytogenes* shares the same function as they do. Proteins like C1L1B3 whose function is mRNA cleavage, C1L165 whose function is to bind to polyisoprenoid, C1L158 which is responsible for phenazine biosynthesis, C1L0M8 which is involved in negative regulation of phenolic acid metabolism, C1K0G9 whose function is to catabolize myo-inositol, C1L075 which is polyphosphogluconolactonase, C1KZV6 which is responsible for nucleotide binding, C1KYZ4 which is monoxygenase, C1KYX9 which is involved in proteolysis, C1KYT3 is a GTP binding protein, C1KY51 is iron binding protein, C1KY50 has iron-sulphur cluster assembly, C1KY30 has chloride chemical activity, C1KXN7 is an oxidoreductase, C1KWC7 has a role in stationary phase survival, C1KVW8 has catalytic activity. The proteins are among those that don't share any function with any other protein. This is why they are being classified in this category. For the C1L165 gene, 11 matching sequences were identified by a FASTA search, and 4 sequence motifs were identified by an InterPro scan. There were 7 neighbouring genes found which are as follows Q8Y8U6, E1UEF3, D2PB35, D2NZR0, B8DGH0, Q92DM4 and D3UL61 with percentage identities of 99.4%, 99.4%, 99.4%, 99.4%, 99.4%, 98.3%, and 95.9%. 25 matching sequences were found by a FASTA search for the C1L0M8 gene. Additionally, 4 sequence motifs were identified by InterProScan. 7 neighbouring genes found are Q92E53, Q7AP25, D2PA50, D2NYN1, E1UDW6, B8DA24 and 3UKC6 with percentage identity as 100%, 100%, 100%, 99%, 99%, 99%, and 99% respectively.

3.1.13. Unknown

Proteins included in this category are the proteins whose function could not be predicted due to a lack of any known characterized protein hits. Due to the unavailability of good hits, we classified them in this category, as no function can be assigned to them until we obtain hits of characterized proteins whose function is known to us, instead of hypothetical, uncharacterized, or putative proteins, which we obtained for proteins of this category.

3.2. Structure-based Function Prediction

We have sought to develop a template for building models of hypothetical proteins. Fortunately, we identified templates for 13 proteins and subsequently predicted their structures. To predict the function of hypothetical proteins, we have extensively analyzed the atomic coordinates of models and proposed corresponding functions based on structural similarity, fold, domain, motif, and analysis of structurally conserved residues.

3.2.1. C1L1U8

The protein C1L1U8 is 555 residues long, and the model was successfully built for residues 6–555. The overall structure of this protein is shown in Figure 2A. Our model shares 81.09% sequence identity with the template structure and has an RMSD of 0.1 Å. Furthermore, there were 15 α helices and 23 β sheets in our model. The active site of the template structure 3ZQ4 contains two Zn^{2+} ions positioned approximately 3 Å apart within an octahedral coordination geometry, situated near the cleft separating the β-lactamase and β-CASP domains. The first zinc ion is coordinated by the side chains of Asp78, His79, Asp164, and His390, whereas the second zinc ion interacts with His74, His76, His142, and Asp164 (**?**). As anticipated, residues involved in metal coordination are highly conserved. Structural superimposition demonstrated that these key residues are preserved in our modeled protein. Additionally, Asp195 and His368 have been implicated in acid–base catalysis. In the template structure, dimerization is stabilized by a Ca^{2+} ion located at the interface, coordinated by Gly49 and Asp51 from one monomer and Asp443 and Glu464 from the opposing chain. A salt bridge between Asp449 and Arg544 further reinforces the dimer interface. Hydrogen-bonding interactions involve His364, Gly367, Asp78, and Asp164 (**?**). The residue Ser366 is essential for catalytic activity, and its mutation abolishes enzymatic function. Two conserved residues positioned near the catalytic center, His368 and Asp195, adopt conformations consistent with a classical catalytic dyad. ProFunc analysis identified 20 significant ligand-binding templates among 94,569 entries and one notable DNA-binding template from a dataset of 4,194. The phosphate group of nucleotide 3 forms hydrogen bonds with the side chain of Ser233 and the backbone nitrogen of Glu77. The nucleotide base is oriented to allow potential π–π stacking interactions with Phe42, although optimal stacking requires an alternative rotamer conformation of this residue. Notably,

substitution with a pyrimidine base at this position enhances structural compatibility and enables additional hydrogen bonding with the main-chain carbonyl of Asp51 and the hydroxyl group of Tyr52; the structural model was adjusted accordingly to optimize these interactions. Based on the pronounced structural similarity to the template and conservation of catalytic residues, C1L1U8 is proposed to function as a nuclease, possibly exhibiting 5′–3′ exonuclease activity. Nevertheless, experimental studies are necessary to validate its exact enzymatic function.

3.2.2. C1KYM2

The protein C1KYM2 comprises 259 residues, and the model was successfully built for residues 2–259. The overall structure of this protein is shown in Figure 2. Our model shares 37.31% sequence identity with the template and has an RMSD of 1.359 Å. Furthermore, our model contained 6 α-helices and 11 β-sheets. The catalytic triad residues Glu43, Lys109, and Cys143 have been shown to be conserved in our model. Lys 109 and Tyr 144 are responsible for the formation of hydrogen bonds (Chin et al., 2007). In the protein–ligand complex, the two methyl groups of cacodylate establish favorable C–H$\cdots\pi$ interactions with nearby aromatic residues. One methyl group interacts with Phe49, while the other engages Phe113 and Trp175. Such C–H$\cdots\pi$ contacts are recognized as stabilizing forces that can significantly enhance the structural integrity of proteins and protein–ligand assemblies. The binding cavity is largely composed of hydrophobic amino acids, including Phe49, Phe113, Val142, Trp175, and Pro176, which together with residues of the catalytic triad contribute to shaping the active-site pocket (Chin et al., 2007). ProFunc analysis identified 3 statistically significant ligand-binding templates from a total of 94,569 entries examined. Additionally, 2 significant matches were detected among 4,194 DNA-binding templates, suggesting possible nucleic acid interaction capability.

3.2.3. C1KYY1

The protein C1KYY1 comprises 440 residues, and its model was successfully built for residues 1–439. The overall structure of this protein is shown in Figure 2. Our model shares 48.23% sequence identity with the template structure and has an RMSD of 0.7 Å. Furthermore, there were 21 α helices and 9 β sheets in our model. The residues His-129–Glu-122 form a putative catalytic dyad in the template structure and are among the most important for its activity (Vorontsov et al., 2011). His 66, His 101, Asp 111, and Asp 183 are important for the formation of a bowl-shaped structure where active binding of the ligand takes place. Residues Lys-14, Asn-36, Gln-41, Arg-44, Phe-64, Arg-326, and Lys-330 are in direct contact with the ligand and make this site highly specific to dGTP (Vorontsov et al., 2011). All residues except Asn-36 are highly conserved and play important roles in ligand binding. The His-129, Glu-122 couple has been shown to function as a catalytic dyad for the generation of the nucleophilic OH^- hydroxyanion, and a nearby His-114 residue may also assist in the proper positioning of the attacking group. Tyr 239 and Arg 63 form a hydrogen bond (Vorontsov et al., 2011). Leu49–Tyr368 are involved in van der Waals interactions. Because most of the important residues are conserved in our model, we propose that C1KYY1 functions as a deoxynucleotide triphosphate triphosphohydrolase (dNTPase). Furthermore, during ProFunc analysis, 20 significant ligand binding templates were found out of 94569 ligand binding templates.

3.2.4. C1L0M8

The protein C1L0M8 is 110 residues long, and the model was successfully built for residues 5–105. The overall structure of this protein is shown in Figure 2. Our model shares 40.78% sequence identity with the template and has an RMSD of 0.5 Å. Furthermore, our model contains 4 α-helices and 2 β-sheets. Tyr46 and Arg51 are the DNA-binding residues in the template and help in attaching to DNA. The residues Gly25, Tyr26, Glu42, Arg71, and Lys72 contribute to the strengthening of the binding between DNA and the protein (Fibriansah et al., 2012). Based on the conserved nature of important residues, it can be proposed that the function of C1L0M8 is a transcriptional regulator of phenolic acid.

3.2.5. C1L0R8

The protein C1L0R8 comprises 606 residues, and the model was successfully built for residues 201–605. The overall structure of this protein is shown in Figure 2. Our model shares 30.07% sequence identity with the template and has an RMSD of 0.5 Å. Furthermore, there are 11 α helices and 9 β sheets in our model. It has been shown that LTA (Lipoteichoic Acid) is required for divalent cation homeostasis and that its absence has severe effects on cell morphogenesis and division. Disruption of LTA synthesis severely affects cell morphology and viability. The residue Thr 297 has been shown to be essential for LTA synthesis and is conserved in our model. Sequence examination of residues directly interacting with the phosphothreonine at position 297, namely His412, Glu253, and Trp350, indicates that these amino acids are highly conserved across LtaS homologs. Their strong conservation among orthologous proteins suggests a critical functional role, likely contributing significantly to lipoteichoic acid (LTA) biosynthesis (**?**). Glu253, Asp471, and His472, which surround the residue Thr297, have been found to coordinate with the magnesium ion in the active site. His343, Asn345, Arg352, Glu253, and Thr408 are involved in hydrogen-bond formation with water (**?**).

3.2.6. C1L165

The protein C1L165 is 176 residues long, and the model was successfully built for residues 5–176. The overall structure of this protein is shown in Figure 2. Our model shares 41.28% sequence identity with the template and has an RMSD of 0.2 Å. Furthermore, our model contains 3 α-helices and 9 β-sheets. His18, Arg62, His65, and Try146 residues are responsible for binding to poly-isoprenoid (Handa et al., 2004). In the ProFunc analysis, 1 significant DNA-binding template was identified among 4194 DNA-binding templates.

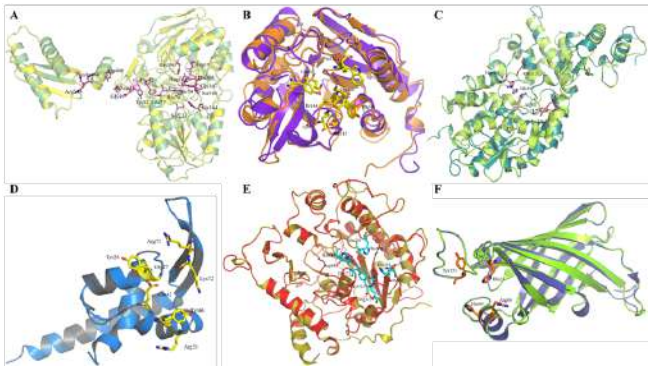

Figure 2. Structural superimposition of modeled proteins C1L1U8, C1KYM2, C1KYY1, C1L0M8, C1L0R8, and C1L165 with their respective template structures. Cartoon representations showing the three-dimensional models superimposed over their corresponding template structures.

3.2.7. C1L2E5

The protein C1L2E5 comprises 169 residues, and the model was successfully built for residues 1–158. The overall structure of this protein is shown in Figure 3. Our model shares 22.50% sequence identity with the template and has an RMSD of 3.919 Å. Furthermore, our model contains 5 α-helices and 10 β-sheets. Residues presumably involved in metal binding include Asp-8, His-10, Asp-36, Asn-59, Asn-60, His-97, His-120, Thr-121, and His-122 (**?**). All are present except Thr-121, which is replaced by Ser; however, because it is like Thr, the presence of conserved metal-binding residues suggests possible phosphodiesterase activity; however, the relatively low sequence identity (22.50%) limits confidence in this prediction.

3.2.8. C1KZM7

The protein C1KZM7 comprises 156 residues, and the model was successfully built for residues 3–142. The overall structure of this protein is shown in Figure 3. Our model shares 30.41% sequence identity with the template and has an RMSD of 0.085 Å. Furthermore, our model contains 5 α-helices and 5 β-sheets. It has been found that the residues Arg135, Ser131, Val38, Pro8, Gly117, Gln119, Ala133, Val132, Gly120, Gly123, and Asn122 are responsible for ATP binding (**?**). Some residues are conserved in our model, while some are not. We can tentatively conclude that this protein may function as a stress protein and as an ectoine producer, which helps the bacteria to survive osmotic stress. But it cannot be said with certainty. Additionally, during the ProFunc analysis of the model, 1 significant ligand-binding template was identified out of 94569 ligand-binding templates.

3.2.9. C1KYZ6

The protein C1KYZ6 is 362 residues long, and the model was successfully built for residues 42–196. The overall structure of this protein is shown in Figure 3. Our model shares 12.35% sequence identity with the template and has an RMSD of 5.687 Å. Furthermore, our model contains 4 α-helices and 7 β-sheets. The residues Ile 342, Asn 346, Tyr 465, and Leu 469 are responsible for the formation of the crevice region where many bound water molecules were observed (Xu et al., 2009). This crevice can virtually divide the domain into two parts: the upper and lower subdomains. These are the only residues of importance observed in the template. Due to very low sequence identity (12.35%) and high RMSD (5.687 Å), the structural model lacks sufficient reliability for functional inference.

3.2.10. C1KYT3

The protein C1KYT3 is 211 residues long, and the model was successfully built for residues 5–201. The overall structure of this protein is shown in Figure 3. Our model shares 33.65% sequence identity with the template and has an RMSD of 3.75 Å. Furthermore, our model contains 5 α-helices and 8 β-sheets. The residues Ser23, His54, Glu77, and Arg104 have been shown to participate in catalysis (**?**). The residues Lys22, Arg24, Phe25, Asp75, Gly76, Pro78, Ala82, Tyr105, Tyr106, Gly107, Leu111, Leu116, Tyr120, Asp74, and Thr81 probably contribute to substrate binding mechanism of the protein. In the ProFunc analysis, 1 significant DNA-binding template was identified among 4,194 available DNA-binding templates. Because most residues are conserved in our model, the protein's function may involve differentiation, development, and regulation of cell proliferation.

3.2.11. C1KY51

The protein C1KY51 comprises 147 residues, and the model was successfully built for residues 6–145. The overall structure of this protein is shown in Figure 3. Our model shares 46.00% sequence identity with the template and has an RMSD of 3.3 Å. Furthermore, our model contains 5 α-helices and 3 β-sheets. In the template structure, the zinc-binding region is positioned at the apex of the structural core, which consists of three antiparallel β-strands flanked by two α-helices. This site is exposed on the protein surface and remains accessible to solvent molecules. The coordinated zinc ion interacts with three conserved cysteine residues, Cys40, Cys65, and Cys127, as well as an additional conserved residue, Asp42 (Liu et al., 2005). The geometry of coordination between the zinc ion and the sulfur atoms of the cysteines, along with the OD1 atom of Asp42, is tetrahedral, with an average bond distance of approximately 2.5 Å. Within the conserved PXCGD motif, Cys40 represents one of the essential cysteines required for activity. Asp42 is another invariant residue in this motif and has been shown to significantly stabilize IscU–$[Fe_2S_2]^{2-}$ complexes in both *Thermotoga maritima* and human IscU proteins (Liu et al., 2005).

Electron density corresponding to the side chains of Asn36 and Asn37 in Loop 2 is absent, and the density is discontinuous between Pro38 and Thr39. In contrast, Cys40, Gly41, and Asp42 are clearly resolved. Loop 3 comprises Gly64 and Cys65, residues that are highly conserved among members of the IscU/NifU protein families. Cys65 (equivalent to Cys63 in *E. coli*) is another critical cysteine known to participate in disulfide bond formation with an active-site cysteine of IscS in *E. coli*. The third essential cysteine, Cys127, is situated on the upper segment of helix α6 and is well defined in the Sp_IscU structure. Arg124, located near the zinc-binding region at the N-terminus of helix α6, is also clearly resolved and highly conserved across IscU/NifU proteins (Liu et al., 2005). Due to its proximity to Cys65, Arg124 may contribute to stabilization of the S_2 intermediate during sulfur transfer from IscS to IscU. Surface analysis of the template further revealed a conserved negatively charged region formed by acidic residues Asp17, Asp28, Asp57, Asp77, and Glu56. ProFunc analysis identified one statistically significant ligand-binding template among 94,569 screened entries. Considering the conservation pattern of key residues and structural features, C1KY51 is likely involved in processes such as electron transfer, regulatory mechanisms, environmental sensing, and substrate activation, although experimental validation is necessary to confirm these proposed functions.

3.2.12. C1KYL6

The protein C1KYL6 is 273 residues long, and the model was successfully built for residues 4–271. The overall structure of this protein is shown in Figure 3. Our model shares 25.19% sequence identity with the template and has an RMSD of 0.091 Å. Furthermore, our model contains 8 α-helices and 12 β-sheets. The residues Arg45, Asp10, Thr43, Asp8, Lys185, Asn211, and Asp208 have been identified to constitute the template's active site (Lu et al., 2011), and these residues are conserved in our protein as well. Hence, the function of our protein can be predicted as the same as that of the HAD superfamily phosphatases, which are involved in phosphoryl transfer reactions and hydrolytic dephosphorylation.

3.2.13. C1KY50

The protein C1KY50 comprises 464 residues, and the model was successfully built for residues 33–462. The overall structure of this protein is shown in Figure 3. Our model shares 17.20%

sequence identity with the template and has an RMSD of 7.538 Å. Furthermore, our model contains 8 α-helices and 18 β-sheets. The residues Phe373, Leu375, Ile380, Met388, Ile389, Ala392, and Ala395 are involved in hydrophobic interactions (Wada et al., 2009). Iron-sulfur (Fe-S) proteins that contain a Fe-S cluster as a prosthetic group, such as our template, are widely utilized in organisms for a variety of cellular processes, including respiratory and photosynthetic electron transport, and in the regulation of gene expression (Wada et al., 2009). Given the low sequence identity (17.20%) and high RMSD (7.538 Å), the structural model of C1KY50 should be interpreted with caution. Although conserved residues suggest a possible iron–sulfur cluster–related role, experimental validation is necessary.

This comprehensive *in silico* analysis enabled functional annotation of 92 hypothetical proteins from *Listeria monocytogenes*. Sequence-based classification identified multiple functional classes, while structural modeling provided additional mechanistic insight for selected proteins. Although several predictions were strongly supported by conserved catalytic residues and structural similarity, others require cautious interpretation due to low sequence identity or model reliability. The study improves genome annotation quality and identifies promising candidates for subsequent biochemical and functional validation.

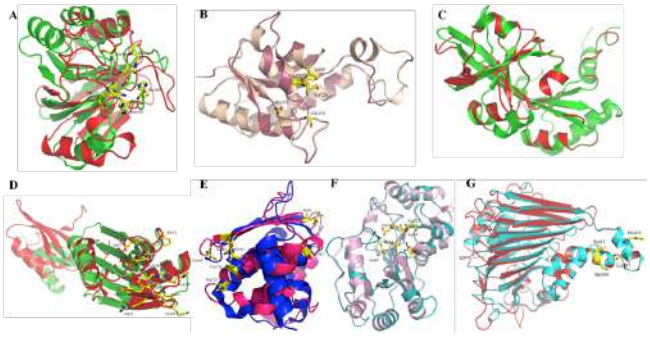

Figure 3. Structural superimposition of modeled proteins C1L2E5–C1KY50 with their respective template structures. Cartoon representations depicting the structural alignment of modeled proteins with their corresponding templates.

4. Conclusions

We have conducted an extensive analysis of the HPs of *Listeria monocytogenes*. Following genome analysis of pathogens, we observed numerous proteins whose functions remain uncharacterized. Comparative genomics indicates that these proteins are present across organisms but have not yet been functionally characterized to date. The first part of this study comprises proteins for which biochemical activity can be predicted with reasonable confidence using genomic sequence-based approaches. We listed the proposed functions of most HPs using sequence analysis tools. The second part includes structure determination and the development of a structure-function relationship. We have successfully determined the structure-function relationship for 13 HPs from *Listeria monocytogenes*. This structural genomics project enables us to better understanding of pathogenesis of this microorganism. We have finally achieved our goal of assigning structure-based biochemical functions to most hypothetical proteins from *Listeria monocytogenes*. However, further biochemical work is needed to assign functions with accurate precision.

5. Conflict of Interest:

There is no conflict of interest to declare.

6. Data Availability Statement:

The data supporting this study are provided in this article.

7. Funding:

None

8. Acknowledgements:

None

9. Supplementary materials:

None

10. Author Contributions:

A.N. conceptualized and designed the study. A.N. performed data curation, virtual screening, molecular docking, ADMET analysis, PASS prediction, and molecular dynamics simulations. A.N. conducted formal analysis and interpreted the results. A.N. prepared figures and tables and wrote the original draft of the manuscript. The author reviewed and approved the final version of the manuscript.

References

Chin, Ko-Hsin, Tsai, et al. 2007
Fibriansah, G., Kovacs, A., Pool, T., et al. 2012, PLoS ONE, 7, e48015
Gury, J., Barthelmebs, L., Tran, N., Divies, C., & Cavin, J. 2004, Appl Environ Microbiol, 70, 2146
Handa, N., Terada, T., Doi-Katayama, Y., et al. 2004, Protein Sci., 14, 1004
Liu, Jinyu, Oganesyan, et al. 2005, Proteins, 59, 875
Lu, Zhibing, Dunaway-Mariano, et al. 2011, thetaiotaomicron. Proteins, 79, 3099
Rao, N. 1989, J. Thorac. Cardiovasc. Surg., 98, 303
Vorontsov, I., I., Minasov, et al. 2011, J. Biol. Chem., 286, 33158
Wada, K., Sumi, N., Nagai, R., et al. 2009, J. Mol. Biol., 387, 245
Xu, Yongbin, Sim, et al. 2009, Biochemistry, 48, 5218

Table 1. List of hypothetical proteins in *Listeria monocytogenes* (Part 1 of 3).

S. No.	Name of Hypothetical Protein	pI	Length	Pfam	Proposed Function
1	C1L300/ C1L300_LISMC	9.69	447	MatE (PF01554)	antiporter activity, drug transmembrane transporter activity
2	C1L2X5/ C1L2X5_LISMC	7.85	345	UPF0118 (PF01594)	NA
3	C1L2V8/ C1L2V8_LISMC	5.02	120	DUF964 (PF06133)	NA
4	C1L2V7/ C1L2V7_LISMC	5.45	267	YmdB (PF13277)	putative phosphoesterases
5	C1L2S2/ C1L2S2_LISMC	6.41	274	FtsJ (PF01728), S4 (PF01479)	methyltransferase involved in viral RNA capping, RNA binding
6	C1L2Q9/ C1L2Q9_LISMC	6.47	321	DUF1385 (PF07136)	NA
7	C1L2P0/ C1L2P0_LISMC	4.78	118	No domains	NA
8	C1L2N2/ C1L2N2_LISMC	5.44	92	DUF503 (PF04456)	NA
9	C1L2K8/ C1L2K8_LISMC	5.38	174	Acetyltransf_3 (PF13302)	transfers acetyl group
10	C1L2J0/ C1L2J0_LISMC	7.84	103	NA	NA
11	C1L2E5/ C1L2E5_LISMC	5.13	174	Metallophos_2 (PF12850)	Phosphoesterase (hydrolase activity, acting on ester bonds)
12	C1L1Z7/ C1L1Z7_LISMC	6.42	774	Glycos_transf_2 (PF00535), Glyphos_transf (PF04464)	glycerophosphotransferase (sequential transfer of glycerol-phosphate units) (CDP-glycerol glycerophosphotransferase activity)
13	C1L1V8/ C1L1V8_LISMC	9.43	267	DUF817 (PF05675)	NA
14	C1L1U8/ C1L1U8_LISMC	6.18	555	Lactamase_B (PF00753), RMMBL_DRMBL (CL0398)	RNA binding, hydrolase activity, acting on ester bonds, metal ion binding
15	C1L1S5/ C1L1S5_LISMC	10.1	344	DUF218 (PF02698)	NA
16	C1L1R3/ C1L1R3_LISMC	9.3	255	TerC (PF03741)	integral to membrane
17	C1L1R2/ C1L1R2_LISMC	9.36	255	TerC (PF03741)	integral to membrane
18	C1L1R1/ C1L1R1_LISMC	9.09	453	MatE (PF01554)	antiporter activity, drug transmembrane transporter activity
19	C1L1K4/ C1L1K4_LISMC	9.81	201	SNARE_assoc (PF09335)	NA
20	C1L1G8/ C1L1G8_LISMC	5.59	725	Tex_N (PF09371), HHH_3 (PF12836), S1 (PF00575)	RNA binding, hydrolase activity- acting on ester bonds
21	C1L1B3/ C1L1B3_LISMC	4.78	125	Ribonuc_L-PSP (PF01042)	Inhibits protein synthesis by cleaving the mRNA
22	C1L190/ C1L190_LISMC	5.09	282	NA	NA
23	C1L165/ C1L165_LISMC	4.69	176	YceI (PF04264)	binds to polyisoprenoid
24	C1L164/ C1L164_LISMC	9.54	306	EamA (PF00892)	transporter activity
25	C1L162/ C1L162_LISMC	9.27	233	DUF554 (PF04474)	NA
26	C1L161/ C1L161_LISMC	4.79	296	CN_hydrolase (PF00795)	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
27	C1L158/ C1L158_LISMC	5.68	282	PhzC-PhzF (PF02567)	Phenazine biosynthes
28	C1L143/ C1L143_LISMC	5.33	309	DAGK_cat (PF00781)	NAD+ kinase activity, diacylglycerol kinase activity
29	C1L0U3/ C1L0U3_LISMC	4.83	288	Hydrolase_3 (PF08282)	hydrolase activity
30	C1L0T5/ C1L0T5_LISMC	5.42	209	HhH-GPD (PF00730)	DNA binding, endonuclease activity
31	C1L0T0/ C1L0T0_LISMC	9.82	187	DUF420 (PF04238)	NA

Table 2. List of hypothetical proteins in *Listeria monocytogenes* (Part 2 of 3).

S. No.	Name of Hypothetical Protein	pI	Length	Pfam	Proposed Function
32	C1L0R8/ C1L0R8_LISMC	5.34	606	Sulfatase (PF00884)	sulfuric ester hydrolase activity
33	C1L0Q0/ C1L0Q0_LISMC	9.71	246	TauE (PF01925)	NA
34	C1L0P3/ C1L0P3_LISMC	4.87	172	Acetyltransf_1 (PF00583)	acetyltransferase activity
35	C1L0M8/ C1L0M8_LISMC	4.78	110	PadR (PF03551)	involved in negative regulation of phenolic acid metabolism
36	C1L0L0/ C1L0L0_LISMC	6.08	394	Methyltrans_SAM (PF10672)	methyltransferase activity
37	C1L0K2/ C1L0K2_LISMC	7.7	431	Xan_ur_permease (PF00860)	transporter activity
38	C1L0G9/ C1L0G9_LISMC	4.82	264	AP_endonuc_2 (PF01261)	involved in the myo-inositol catabolism (isomerase activity)
39	C1L075/ C1L075_LISMC	5.18	346	Lactonase (PF10282)	6-phosphogluconolactonase activity
40	C1L028/ C1L028_LISMC	5.72	143	Usp (PF00582)	response to stress
41	C1L024/ C1L024_LISMC	5.26	226	GATase (PF00117)	transferase activity
42	C1KZV6/ C1KZV6_LISMC	4.89	281	NmrA (PF05368)	nucleotide binding (negative transcriptional regulator involved in the post-translational modification of the transcription factor AreA.)
43	C1KZU1/ C1KZU1_LISMC	5.41	377	Gly_kinase (PF02595)	glycerate kinase activity
44	C1KZQ3/ C1KZQ3_LISMC	4.64	279	Hydrolase_3 (PF08282)	hydrolase activity
45	C1KZM7/ C1KZM7_LISMC	8.93	156	Usp (PF00582)	response to stress
46	C1KZM4/ C1KZM4_LISMC	6.15	120	YjbR (PF04237)	NA
47	C1KZE1/ C1KZE1_LISMC	4.83	270	Hydrolase_3 (PF08282)	hydrolase activity
48	C1KZ86/ C1KZ86_LISMC	5.18	111	PhnA_Zn_Ribbon (PF08274), PhnA (PF03831)	NA
49	C1KZ81/ C1KZ81_LISMC	5.97	494	FTR1 (PF03239)	transmembrane transport
50	C1KZ17/ C1KZ17_LISMC	8.78	220	MgtC (PF02308)	transport of Mg2+
51	C1KZ02/ C1KZ02_LISMC	5.31	280	SAM_adeno_trans (PF01887)	NA
52	C1KYZ6/ C1KYZ6_LISMC	8.7	362	MacB_PCD (PF12704), FtsX (PF02687)	transport lipids targeted to the outer membrane across the inner membrane
53	C1KYZ4/ C1KYZ4_LISMC	5.44	97	ABM (PF03992)	monooxygenase activity
54	C1KYZ2/ C1KYZ2_LISMC	4.62	275	Hydrolase_3 (PF08282)	hydrolase activity
55	C1KYY1/ C1KYY1_LISMC	5.4	440	HD (PF01966)	metal ion binding, phosphoric diester hydrolase activity
56	C1KYX9/ C1KYX9_LISMC	6.83	217	Peptidase_M50 (PF02163)	proteolysis, metalloendopeptidase activity
57	C1KYX3/ C1KYX3_LISMC	9.03	306	DAGK_cat (PF00781)	kinase, transferase
58	C1KYW9/ C1KYW9_LISMC	10.1	357	UPF0104 (PF03706)	NA
59	C1KYT3/ C1KYT3_LISMC	5.78	211	UPF0029 (PF01205), DUF1949 (PF09186)	GTP binding
60	C1KYS5/ C1KYS5_LISMC	6.43	287	DUF161 (PF02588), DUF2179 (PF10035)	NA
61	C1KYP0/ C1KYP0_LISMC	4.97	322	UPF0052 (PF01933)	NA
62	C1KYM2/ C1KYM2_LISMC	5.12	259	CN_hydrolase (PF00795)	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds

Table 3. List of hypothetical proteins in *Listeria monocytogenes* (Part 3 of 3).

S. No.	Name of Hypothetical Protein	pI	Length	Pfam	Proposed Function
63	C1KYL6/ C1KYL6_LISMC	5.39	273	Hydrolase_3 (PF08282)	ATP binding, ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism
64	C1KYI4/ C1KYI4_LISMC	5.98	201	MTS (PF05175)	16S rRNA (guanine(966)-N(2))-methyltransferase activity
65	C1KY64/ C1KY64_LISMC	5.76	117	ArsC (PF03960)	regulate the transcription of multiple genes in response to disulfide stress
66	C1KY61/ C1KY61_LISMC	5.27	291	Cation_efflux (PF01545)	cation transmembrane transporter activity
67	C1KY51/ C1KY51_LISMC	5.94	147	NifU_N (PF01592)	iron ion binding, iron-sulfur cluster binding
68	C1KY50/ C1KY50_LISMC	4.87	464	UPF0051 (PF01458)	iron-sulfur cluster assembly
69	C1KY46/ C1KY46_LISMC	9.02	279	TauE (PF01925)	involved in the transport of anions across the cytoplasmic membrane during taurine metabolism as an exporter of sulfoacetate
70	C1KY45/ C1KY45_LISMC	5.08	462	Metallophos (PF00149), 5_nucleotid_C (PF02872)	hydrolase activity (5_nucleotid_C)
71	C1KY43/ C1KY43_LISMC	4.94	255	Hydrolase_like (PF13242), Hydrolase_6 (PF13344)	hydrolase activity
72	C1KY41/ C1KY41_LISMC	4.82	436	DUF21 (PF01595), CBS (PF00571), CorC_HlyC (PF03471)	flavin adenine dinucleotide binding (CBS)
73	C1KY30/ C1KY30_LISMC	9.43	406	Voltage_CLC (PF00654)	voltage-gated chloride channel activity
74	C1KY03/ C1KY03_LISMC	5.29	156	Rrf2 (PF02082)	NA
75	C1KXZ5/ C1KXZ5_LISMC	4.65	281	Hydrolase_3 (PF08282)	hydrolase activity
76	C1KXZ0/ C1KXZ0_LISMC	4.76	532	Amidohydro_3 (PF07969)	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides
77	C1KXY4/ C1KXY4_LISMC	6.26	257	Methyltransf_26 (PF13659)	methyltransferase activity
78	C1KXX7/ C1KXX7_LISMC	4.83	270	Hydrolase_3 (PF08282)	hydrolase activity
79	C1KXV5/ C1KXV5_LISMC	5.52	249	EAL (PF00563)	NA
80	C1KXN7/ C1KXN7_LISMC	6.53	331	Bac_luciferase (PF00296)	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
81	C1KXN1/ C1KXN1_LISMC	6.58	147	DUF523 (PF04463)	NA
82	C1KWJ0/ C1KWJ0_LISMC	6.74	121	DUF1798 (PF08807)	NA
83	C1KWG8/ C1KWG8_LISMC	6.52	327	DUF939 (PF06081), DUF939_C (PF11728)	NA
84	C1KWG7/ C1KWG7_LISMC	4.39	125	Glyoxalase_2 (PF12681)	NA
85	C1KWG4/ C1KWG4_LISMC	6.52	205	HTH_11 (PF08279), CBS (PF00571)	sequence-specific DNA binding transcription factor activity (CBS)
86	C1KWC7/ C1KWC7_LISMC	5.29	291	YicC_N (PF03755), DUF1732 (PF08340), DUF1732 (PF08340)	play a role in the stationary phase survival (YicC)
87	C1KW03/ C1KW03_LISMC	6.11	126	OsmC (PF02566)	has a novel pattern of oxidative stress regulation
88	C1KVW1/ C1KVW1_LISMC	6.22	192	rRNA_methylase (PF06962)	methyltransferase activity
89	C1KVW0/ C1KVW0_LISMC	5.48	321	Radical_SAM (PF04055)	catalytic activity, iron-sulfur cluster binding
90	C1KVA0/ C1KVA0_LISMC	6.02	234	DUF633 (PF04816)	tRNA (adenine-N1-)-methyltransferase activity
91	C1KV99/ C1KV99_LISMC	5.5	373	NIF3 (PF01784)	NIF3 interacts with the yeast transcriptional coactivator NGG1p, which is part of the ADA complex
92	C1KV71/ C1KV71_LISMC	4.92	269	Hydrolase_3 (PF08282)	cation transport, ATP binding, ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism

Table 4. Major Classes of Hypothetical Proteins in *Listeria monocytogenes*

S. No.	Proposed function of Hypothetical protein	Primary Accession number
1	Hydrolase	C1L0U3, C1L0R8, C1KZQ3, C1KZE1, C1KYZ2, C1KYY1, C1KY45, C1KY43, C1KXZ5
2	Transferase	C1L2K8, C1L1Z7, C1L0P3, C1L0L0, C1L024, C1KYI4, C1KXY4, C1KVW1, C1KVA0
3	Transporter	C1L300, C1L1R1, C1L164, C1L0K2, C1KZ81, C1KZ17, C1KYZ6, C1KY61, C1KY46
4	RNA binding	C1L2S2, C1L1U8, C1L1G8
5	Hydrolase acting on carbon-nitrogen but not on peptide	C1L161, C1KYM2, C1KXZ0
6	Kinase	C1L143, C1KZU1, C1KYX3
7	Response to stress	C1LO28, C1KZM7, C1KW03
8	Integral membrane	C1L1R3, C1L1R2, C1KY64
9	DNA binding	C1L0T5, C1KWG4
10	Phosphoesterase	C1L2V7, C1L2E5
11	ATP Binding	C1KYL6, C1KV71
12	Others	C1L1B3, C1L165, C1L158, C1L0M8, C1L0G9, C1L075, C1KZV6, C1KYZ4, C1KYX9, C1KYT3, C1KY51, C1KY50, C1KY41, C1KY30, C1KXN7, C1KWC7, C1KVW8, C1KV99
13	Unknown	C1L2X5, C1L2V8, C1L2Q9, C1L2P0, C1L2J0, C1L1S5, C1L1K4, C1L198, C1L162, C1L0T0, C1L0Q0, C1KZM4, C1KZ86, C1KZ02, C1KYW9, C1KYS5, C1KYP0, C1KY03, C1KYN7, C1KXN1, C1KWJ0, C1KWG8, C1KWG7

Table 5. Superimposed residues between template and modeled proteins C1L1U8–C1L165. Residue pairs identified through structural superimposition of each modeled protein with its corresponding template structure. Amino acid residues from the template are shown on the left, followed by the aligned residues in the model (Template → Model). Identical residue matches and positional shifts are indicated according to their residue number in the respective structures.

Protein ID	Superimposed Residues (Template → Model)
C1L1U8	Asp78→Asp78, His79→His79, Asp164→Asp164, His390→His390, His74→His74, His76→His76, His142→His142, Asp195→Asp195, His368→His368, Gly49→Gly49, Asp51→Asp51, Asp443→Asp443, Glu464→Glu464, Asp449→Asp449, Arg544→Arg544, Ser366→Ser366, His364→His364, Gly367→Gly367, Ser233→Ser233, Glu77→Glu77, Phe42→Phe42, Tyr52→Tyr52
C1KYM2	Glu43→Glu42, Lys109→Lys111, Cys143→Cys145, Tyr144→Tyr146, Phe49→Tyr48, Phe113→Phe115, Trp175→Trp171, Val142→Ile144, Pro176→Pro172
C1KYY1	His129→His120, Glu122→Glu113, His66→His64, His110→His110, Asp111→Asp102, Asp183→Asp173, Lys14→Lys12, Asn36→Ala34, Gln41→Gln39, Arg326→Arg317, Lys330→Arg321, His114→His105, Tyr239→Tyr231, Arg63→Arg61, Leu49→Leu47, His119→His110, Tyr187→Tyr177, Tyr243→Tyr235, Tyr368→Tyr358
C1L0M8	Gly25→Gly27, Tyr26→Tyr28, Glu42→Glu42, Arg71→Arg71, Lys72→Lys72, Tyr46→Tyr46, Arg51→Arg51
C1L0R8	Thr297→Thr272, His412→His388, Glu253→Glu230, Trp350→Tyr325, His472→His449, Asp471→Asp448, His343→His318, Asn345→Asn320, Arg352→Arg327, Thr408→Ser384
C1L165	His18→His21, Arg62→Arg66, His65→His69, Trp146→Tyr151

Table 6. Superimposed residues between template and modeled proteins C1L2E5–C1KY50. Residue correspondences obtained from structural alignment between each template and its respective modeled protein. Residues are presented as Template → Model to indicate positional and amino acid conservation or substitution observed upon superimposition.

Protein ID	Superimposed Residues (Template → Model)
C1L2E5	Asp8→Asp8, His10→His10, Asp36→Ser36, Asn59→Asn54, Asn60→Cys55, His97→His78, His120→His107, Thr121→Ser108, His122→His109
C1KZM7	Arg135→Tyr130, Ser131→Ser126, Val38→Val40, Pro8→Gly10, Gly117→Gly112, Gln119→Thr114, Ala133→Ser128, Val132→Val127, Gly120→Gly115, Gly123→Ala118, Asn122→Ser117
C1KYZ6	Leu469→Lys186, Tyr465→Trp182, Asn346→Ala79, Ile342→Thr75
C1KYT3	Ser23→Ser21, His54→His52, Glu77→Glu74, Arg104→Arg100, Lys22→Lys20, Arg24→Arg22, Phe25→Phe23, Asp75→Asp71, Gly76→Gly72, Pro78→Pro74, Ala82→Ala78, Tyr105→Tyr101, Tyr106→Phe102, Gly107→Gly103, Leu111→Leu107, Leu116→Leu112, Tyr120→Tyr116, Asp74→Asp70, Thr81→Thr77
C1KY51	Cys40→Cys41, Asp42→Asp43, Cys65→Cys66, Arg124→Arg125, Cys127→Cys128, Gly64→Gly65, Asp57→Ala58, Asp77→Gln78, Glu56→Val57
C1KYL6	Trp171→Ser187, Phe175→Asn191, Arg45→Arg46, Asp1→Asp12, Thr43→Thr44, Asp8→Asp10, Lys185→Lys201, Asn211→Asn227, Asp208→Asp22
C1KY50	Phe373→Phe415, Leu375→Leu417, Ile380→Leu422, Met338→Met430, Ile389→Ile431, Ala392→Gly434, Ala395→Glu437

www.editorypress.uz
Volume 1, Issue 1
June 2026