

Three-Dimensional Structure Prediction and Validation of Simran z1 Protein using Homology Modelling

Deeksha Gupta¹, Tarun², Aakanksha Gupta³, Hari Om Sharma⁴, Akshay Kumar Vats⁵, Pradeep Kumar⁶,
Puran Singh Khati⁷, Shashank Rana⁸ and Nikunaj Bhardwaj⁹

^{1,8}Department of Microbiology, Chaudhary Charan Singh University, Meerut (Uttar Pradesh)

²Department of Chemistry, Hariom Saraswati PG College, Dhanuri, Haridwar (Uttarakhand)

^{3,4}Department of Chemistry, CCR (PG) College, Muzaffarnagar (UP)

⁵Department of Chemistry, Motherhood University, Roorkee (Uttarakhand)

⁶Department of Chemistry, HVM (PG) College, Raisi, Haridwar (Uttarakhand)

⁷Department of Chemistry, S. D. M. Government (P. G.) College, Doiwala, (Uttarakhand)

⁹Department of Zoology, MS College, Saharanpur (UP)

Corresponding author: Nikunaj Bhardwaj / bhardwajnikunaj8@gmail.com

Received: 14th June, 2026; Revised: 20th July 2026; Accepted: 1st October 2026; Available Online: 05th October 2026

ABSTRACT

Many hypothetical proteins with unknown functions, especially in mycobacteriophages, have been discovered due to the explosive growth of genomic data. Using homology modeling techniques, this study conducted an in silico structural and functional analysis of the putative protein AQT25855.1 from Mycobacterium phage SimranZ1. Find appropriate structural templates; the amino acid sequence was obtained from NCBI and analysed using BLASTp against the Protein Data Bank. Based on sequence similarity, coverage, and functional relevance, the template (PDB ID: 5HXY) was chosen. The best model was chosen from several three-dimensional models created with MODELLER, using the lowest DOPE score as the criterion. Loop modeling was used to refine the structure and improve conformational accuracy. Ramachandran plot analysis, which showed that most residues were in energetically favourable areas, supporting a high degree of stereochemical dependability, was used to evaluate the predicted model's quality. Domain and structural analysis revealed similarities with members of the tyrosine recombinase/integrase family of proteins, pointing to a function in DNA recombination events. In general, this work demonstrates the utility of computational modeling for forecasting the structure and function of hypothetical proteins and lays the groundwork for future experimental validation and functional investigations.

Keywords: Homology modelling, Hypothetical protein, Mycobacterium phage SimranZ1, MODELLER, DOPE score, Ramachandran plot, Loop refinement, Protein structure prediction, Tyrosine recombinase, In silico analysis

How to cite this article: Deeksha Gupta, Tarun, Aakanksha Gupta, Hari Om Sharma, Akshay Kumar Vats, Pradeep Kumar, Puran Singh Khati, Shashank Rana, Nikunaj Bhardwaj, Three-Dimensional Structure Prediction and Validation of Simran z1 Protein using Homology Modelling Int J Drug Deliv Technol. 2026;16(85s): 1832-1942; DOI: 10.25258/ijddt.16.85s.191

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

The rapid expansion of whole-genome sequencing has revealed a strikingly persistent class of *hypothetical proteins*, open reading frames for which no experimental evidence of expression or function yet exists. In both bacterial and phage genomes, these enigmatic proteins routinely comprise 20 – 40% of the predicted proteome, creating a significant annotation gap that obscures our understanding of microbial physiology, pathogenesis, and evolutionary innovation.

Functional inference by conventional sequence comparison often fails because many hypothetical proteins lack detectable homology to characterized families, underscoring the need for structure-guided approaches to elucidate their biological roles.

Mycobacteriophages, viruses infecting Mycobacterium species, provide a compelling testbed for such efforts because hypothetical proteins overwhelmingly populate their extraordinarily diverse genomes. More than 2,100 mycobacteriophage genomes are now available, yet only about 15% of their predicted open reading frames have assigned functions, leaving the vast remainder uncharacterized. The cluster F1 phage SimranZ1, isolated from New Delhi soil, typifies this challenge: of its 103 annotated genes, the majority encode proteins of unknown function despite clear conservation across related F1 phages. Decoding these proteins is essential not only for understanding phage biology and evolution but also for exploiting phage as diagnostic or therapeutic tools against mycobacterial diseases. [1]

*Author for Correspondence: bhardwajnikunaj8@gmail.com

Three-dimensional (3-D) structure determination offers a powerful route to function prediction because protein folds and active-site geometries are more conserved than primary sequences. In practice, however, experimental techniques such as X-ray crystallography and cryo-electron microscopy cannot keep pace with the deluge of genomic data, making *in silico* structure prediction indispensable. Homology (comparative) modelling fills this gap by threading an uncharacterized sequence onto the backbone of a structurally resolved homolog, thereby generating an atomic model that often approaches experimental accuracy when template identity exceeds 30%.

Recent advances in automated servers, exemplified by SWISS-MODEL, I-TASSER, and Phyre2, now integrate profile–profile alignments, template ensembles, and physics-based refinement, enabling reliable modelling even for remote homologs. These computational pipelines have been successfully applied to bacterial hypothetical proteins, revealing catalytic motifs, ligand-binding pockets, and plausible interaction interfaces that were invisible from sequence alone. [2]

Despite these advances, few studies have leveraged homology modelling to investigate mycobacteriophage proteins. Notably, Das and colleagues recently combined modelling with biochemical validation to characterize Gp65, a DNA-dependent AAA ATPase from SimranZ1, underscoring the utility of structure prediction in phage genomics.

Yet the overwhelming majority of SimranZ1 hypothetical proteins remain unexplored, representing an untapped reservoir of novel folds and functions. Addressing this knowledge gap is crucial for developing a comprehensive systems-level understanding of phage infection strategies and for identifying novel molecular tools for biotechnology.

The present study focuses on the *in-silico* elucidation of one such protein, herein designated Hp-SimZ1, an uncharacterized gene product conserved across cluster F1 phage but absent from bacterial genomes. By integrating state-of-the-art homology modelling, rigorous stereochemical validation, and structure-based functional annotation, we aim to:

- i) generate a high-confidence 3-D model of Hp-SimZ1
- ii) identify putative active-site residues and ligand-binding pockets.
- iii) propose plausible biological functions testable by subsequent experimentation.

Through this work, we seek to demonstrate how computational structural biology can systematically illuminate the dark proteome of mycobacteriophages, thereby accelerating functional discovery and expanding the palette of phage-encoded biomolecules available for therapeutic engineering.

3. MATERIALS AND METHODS

3.1. databases:

3.1.1. NCBI (National Centre for Biotechnology Information)

NCBI is a comprehensive database that provides information on several types of biological databases. It offers access to biomedical and genomic information. NCBI is a part of the United States National Library of Medicine (NLM). GenBank is a genomic sequence database, and PubMed is a literature database; the database used (AQT25855.1) was retrieved from NCBI. Link: <http://www.ncbi.nlm.nih.gov/>. [3, 4]

3.1.2 PDB (Protein Data Bank)

PDB is the global, freely accessible, and primary repository for 3D structural data of biological macromolecules, including proteins, DNA, and RNA. It enables research in medicine, technology, and biology, with structures determined by X-ray crystallography, NMR spectroscopy, and electron microscopy. The 3D structure of the retrieved protein was obtained from the PDB. Link: <https://www.rcsb.org/> [3, 5]

3.2. Tools & Software:

3.2.1. BLAST (Basic Local Alignment Search Tool)

Basic Local Alignment Search Tool refers to a suite of programs utilized to produce alignments between a nucleotide or protein sequence, referred to as a “query,” and nucleotide or protein sequences within a database, referred to as “subject” sequences. (“BLAST QuickStart | Springer Nature Link”) It operates online at the National Centre for Biotechnology Information (NCBI) website. BLAST is also available on other sites. BLAST can fast align and evaluate a query DNA and amino acid sequences with a database of sequences, which create it a critical tool in ongoing genomic and proteomic research. (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) [6, 7]

3.2.2 MODELLER

MODELLER is a computer program for comparative protein structure modeling of protein tertiary structures as well as quaternary structures (rarer). In the simplest case, the input is an alignment of a sequence to be modelled by the template structures, the atomic coordinates of the templates, and a simple script file. [8]

3.2.3 ModLoop

ModLoop is a web server for automated protein loop modeling, predicting conformations by satisfying spatial restraints without relying on known structure databases. (“ModLoop: automated modeling of loops in protein structures”) Developed by the Sali Lab, it uses the MODELLER package to build 300 models, returning the best one based on the lowest molpdf score. It is used for refining protein structures, particularly loops missing in PDB files. The system processes the model, and users receive the results, with low molecular probability density functions (molpdf), via the web interface or email.[9]

3.2.4 Ramachandra Plot

A Ramachandran plot or map, developed by Gopala Samudra Narayanan Ramachandran and V. Sasisekharan, is a way to visualize the dihedral angle ψ against ϕ of amino acid residues in protein structure.

The Ramachandran plot shows the statistical distribution of the combinations of the backbone dihedral angles ϕ and ψ . ("Ramachandran Plot - an overview | ScienceDirect Topics") The Ramachandran plot provides a way to view the distribution of torsion angles in a protein structure and shows that the torsion angles corresponding to the two major secondary structure elements (α -helices and β -sheets) are clearly clustered within separate regions.[10, 11]

3.2.5. S-PDB viewer

Swiss-PDB viewer (Deep View) is a 32-bit application designed for analysing, superimposing, and visualizing protein structures. It helps determine structural alignments, active sites, and H-bonds. It is often used with AlphaFold to model sidechains and mend atomic nomenclature.

Offers capabilities for modeling side chains, assessing stability, and amending amino acid nomenclature.

3.3 Steps/ Method

3.3.1. Sequence Retrieval

The amino acid sequence of the hypothetical protein **AQT25855.1** (372 amino acids) from *Mycobacterium phage SimranZ1* was retrieved from the National Centre for Biotechnology Information. The sequence was obtained in FASTA format and served as the primary input for subsequent structural and computational analyses. The selection of this protein was based on its annotation as a hypothetical protein, necessitating structural prediction for functional insights.

3.3.2. Template Identification

Identify structurally homologous templates; the retrieved sequence was subjected to similarity search using BLASTp against the Protein Data Bank. The search aimed to detect proteins with known three-dimensional structures exhibiting sequence similarity to the target. Templates were selected based on multiple criteria, including sequence identity, query coverage, and structural resolution. The most suitable template structure was downloaded in PDB format for use in comparative modeling.

3.3.3. Sequence Alignment Preparation

Accurate alignment between the target sequence and the selected template is critical for reliable model construction. The alignment was prepared and formatted into **PIR (.ali) format**, ensuring compatibility with downstream modeling software. Special attention was given to maintaining correct residue numbering and alignment integrity, particularly in regions corresponding to insertions or deletions, which are typically associated with loop regions.

3.3.4. Homology Modeling

Three-dimensional structural models of the target protein were generated using MODELLER (version 10.6), which employs a satisfaction of spatial restraints approach. A Python-based script was constructed to automate the modeling process, allowing the generation of multiple candidate structures. These models were produced to sample conformational space and to account for variability in structurally ambiguous regions. The modeling process was executed via the command-line interface within a Windows-based computational environment.

3.3.5. Model Selection and Evaluation

The generated models were evaluated using the **Discrete Optimized Protein Energy (DOPE) score**, a statistical potential implemented within MODELLER. The DOPE score provides an estimate of the energetic favourability of the predicted structure. Among the generated models, the structure exhibiting the lowest (most negative) DOPE score was selected as the most reliable representation of the target protein. This model was carried forward for further structural inspection and refinement.

3.3.6. Structural Visualization and Secondary Structure Analysis

The selected model was visualized and analysed using Swiss-PdbViewer. Structural inspection focused on the identification of secondary structure elements and overall fold organization. Residues were further examined based on their backbone dihedral angles (ϕ and ψ), which define the conformational state of the protein backbone. Regions exhibiting deviations from energetically favourable ϕ/ψ distributions, particularly those inconsistent with expected secondary structure patterns, were identified as candidates for refinement.

3.3.7. Loop Modeling and Refinement

Regions exhibiting improper backbone conformation, typically corresponding to loop segments, were subjected to refinement. Loop modeling was performed using ModLoop and/or loop refinement protocols within MODELLER. Continuous residue segments were defined to allow sufficient flexibility for conformational adjustment. Multiple loop conformations were generated, and the most energetically favourable structures were selected. This step aimed to improve backbone geometry and reposition residues into conformations consistent with stable structural arrangements.

3.3.8. Post-Refinement Validation

The refined models were subjected to further validation to assess structural quality. Backbone geometry was re-evaluated through analysis of ϕ and ψ angle distributions, ensuring that residues occupied energetically allowed regions. Structural integrity was also assessed by inspecting steric clashes and overall fold consistency using Swiss-PdbViewer. Improvements in loop regions and a reduction in conformational strain indicated successful refinement.

3.3.8.1. Ramachandran Plot Analysis (<https://saves.mbi.ucla.edu/>) [12]

The stereochemical quality of the predicted and refined protein structures was further evaluated using Ramachandran plot analysis. This analysis assesses the distribution of backbone dihedral angles (ϕ and ψ) of amino acid residues and provides a quantitative measure of structural reliability.

The Ramachandran plot was generated to examine whether residues occupy energetically favourable conformational regions corresponding to well-defined secondary structures. In the initial modelled structure, certain residues were observed outside the favoured regions, indicating conformational strain, particularly within flexible loop segments.

Following loop refinement, a marked improvement in the distribution of residues was observed. Most residues were found within the **most favoured regions**, while additional residues were in **allowed regions**, and only a minimal number of residues appeared in **disallowed regions**. This shift in distribution reflects improved backbone geometry and reduced steric strain.

The reduction of outlier residues after refinement confirms the effectiveness of the loop modeling approach in

enhancing structural quality. Overall, the Ramachandran plot analysis supports the reliability of the final refined model for further structural and functional interpretation.

3.3.9. Software and Computational Environment

All computational analyses were performed using a Windows-based system. The primary tools utilized in this study included MODELLER (version 10.6) for homology modeling, Swiss-PdbViewer for structural analysis, BLASTp for template identification, and ModLoop for loop refinement. Standard text editors were employed for preparing input files such as FASTA sequences, alignment files, and Python scripts.

RESULT

Sequence Retrieval and FASTA Representation

The sequence (AQT25855.1) was retrieved from the NCBI, annotated as a putative integrase from *Mycobacterium phage SimranZ1*. The protein comprises 372 amino acids, suggesting a large enzyme potentially involved in recombination-related functions.

The retrieved FASTA sequences used are as shown in fig. 1.

FASTA ▾

putative integrase [Mycobacterium phage SimranZ1]

GenBank: AQT25855.1

[GenPept](#) [Identical Proteins](#) [Graphics](#)

```
>AQT25855.1 putative integrase [Mycobacterium phage SimranZ1]
MASVRFRRDGTAYLVSYRFGGRGSAQGALTDDRKAADAFAAVDAHGAARALEMHGIDPTPRGSKS
ELTVAEWIRHHIDHLTGVEQYTIKYEQYLANDIAPNLGDIPLSKLSEEDIARWVKVMTTGGRDGNHGA
PKTLRNKYGFLSGALNAAVPRYLSTNPAAGRRLPRGDAEDDDIRMLTHAEFDRLRDAVTPHWKLMVQFM
VSTGLRWGEVSALQPKHVDLETSTIRVRQAWKYSSAGYVLGPPKTKRSRRTVDVPARLLERLDSHEFVF
VNTDGGPVRYPGFLRRVWNPAAVEKAGLVPRPTPHDLRHTYASWQLTGGETVTVSRQLGHESIQTVDTY
TDVDRSSRVAAEFMDGLLGDF
```

Figure 1: FASTA seq of the hypothetical protein AQT25855.1

BLAST Analysis and Hit Selection

Identify the structure of a homologous protein. The selected protein sequence was subjected to BLASTp

against the Protein Data Bank database. The BLAST result gives some information, as in the figure. 2.

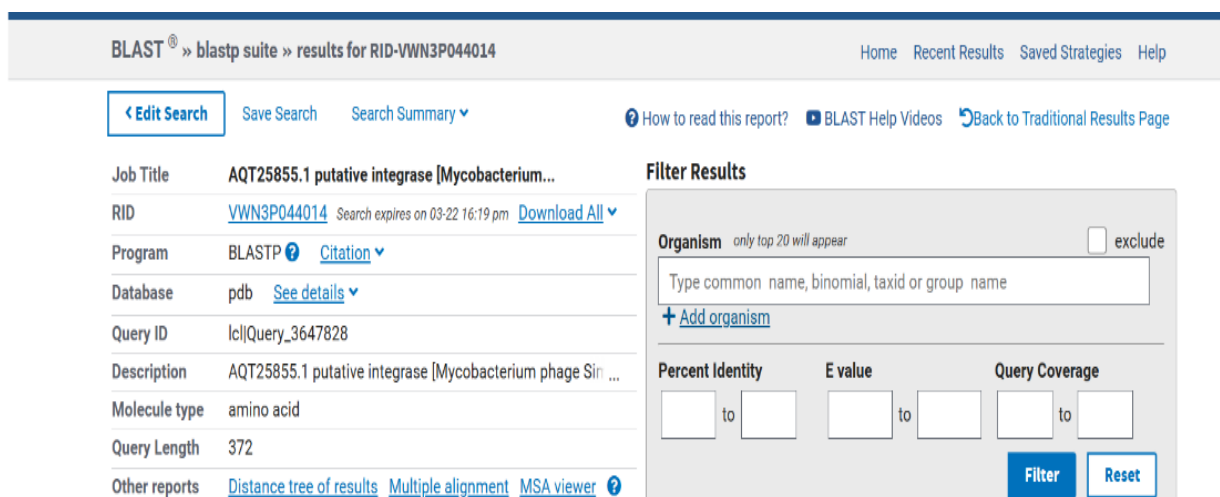


Figure 2: BLASTp output showing homologous protein hits against the Protein Data Bank, indicating similarity with recombinase/integrase family proteins.

Along with this information, it also yields significant hits corresponding to tyrosine recombinases and integrase-like proteins. Among these Hits, as shown in fig. 3, we selected

the first Hit for modelling, which was 5HXY (Chain A), which was the top-ranked among all.

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download

Select columns

Show

100

☒ select all

8 sequences selected

[GenPept](#)
[Graphics](#)
[Distance tree of results](#)
[Multiple alignment](#)
[MSA Viewer](#)

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc Len	Accession
☒ Chain A, Tyrosine recombinase XerA [Thermoplasma acidophilum DSM 1728]	Thermoplasma acidophilum DSM 1728	55.1	55.1	76%	9e-08	25.61%	317	5HXY_A
☒ Chain 14, Ribosome-associated translation inhibitor RaiA [Porphyromonas gingivalis]	Porphyromonas gingivalis	50.4	50.4	44%	3e-06	26.38%	401	95V_14
☒ Chain A, Integrase [Peduovirus P2]	Peduovirus P2	46.6	46.6	65%	4e-05	24.00%	292	5C6K_A
☒ Chain A, site-specific recombinase Int4 [Vibrio cholerae O1 biovar El Tor str. N16961]	Vibrio cholerae O1 biovar El Tor str. N16961	43.1	43.1	57%	6e-04	26.00%	320	2A3V_A
☒ Chain A, HP1 INTEGRASE [Hpnavirus HP1]	Hpnavirus HP1	40.4	40.4	45%	0.002	25.88%	170	1AH_A
☒ Chain A, Integrase [Peduovirus P2]	Peduovirus P2	39.7	39.7	43%	0.004	26.38%	176	5DOR_A
☒ Chain A, PROBABLE TYROSINE RECOMBINASE XERC-LIKE [Pyrococcus abyssi]	Pyrococcus abyssi	38.5	38.5	16%	0.017	32.26%	292	4A8F_A
☒ Chain A, Tyrosine recombinase XerH [Helicobacter pylori 26695]	Helicobacter pylori 26695	38.1	38.1	56%	0.024	22.38%	363	5LIV_A

FASTA

Chain A, Tyrosine recombinase XerA

PDB: 5HXY_A

[GenPept](#) [Identical Proteins](#) [Graphics](#)

```
>pdb|5HXY|A Chain A, Tyrosine recombinase XerA
XGSSHHHHHSSGLVPRGSHXASXTGGQXGRGSXPAETNEYLSRFVEYXTGERKSRYTIKEYRFLVDQF
LSFXNKKPDEITPXDIERYNKFLAVKKRYSKTSQYLAIAKAVKLFYKALDLRVPINLTPPKRPSHXPVYLS
EDEAKRLIEAASSDTRXYAIVSVLAYTGVRVGELCNLKISVDLQESIINVRSGKGDKDRIVIXAECEVK
ALGSYDLRLSXDTNDNYLFVSNRRVRFDSTIERXIRDLGKKAGIQKKVTPHVL RHTFATSVLRNGGDI
RFIIQILGHASVATTQIYTHLNDLSALREXYTQHRPRY
```

Figure 3: Selected template structure (PDB ID: 5HXY, Chain A) used for homology modelling.

The Hit was selected based on:

- significant E-value indicating strong homology,
- substantial query coverage,
- and functional similarity (recombinase/integrase family).

Additionally, conserved domain analysis revealed the presence of an **XerD-like recombinase domain**, supporting the functional relevance of the selected Hit.

Homology Modeling and Model Generation

Homology modelling was performed using MODELLER10.6. Multiple (x10) structural models were generated based on the alignment between the target and template sequences, several candidate structures (e.g., 5hxy.B99990001.pdb to 5hxy.B99990010.pdb) representing different conformational possibilities of the protein.

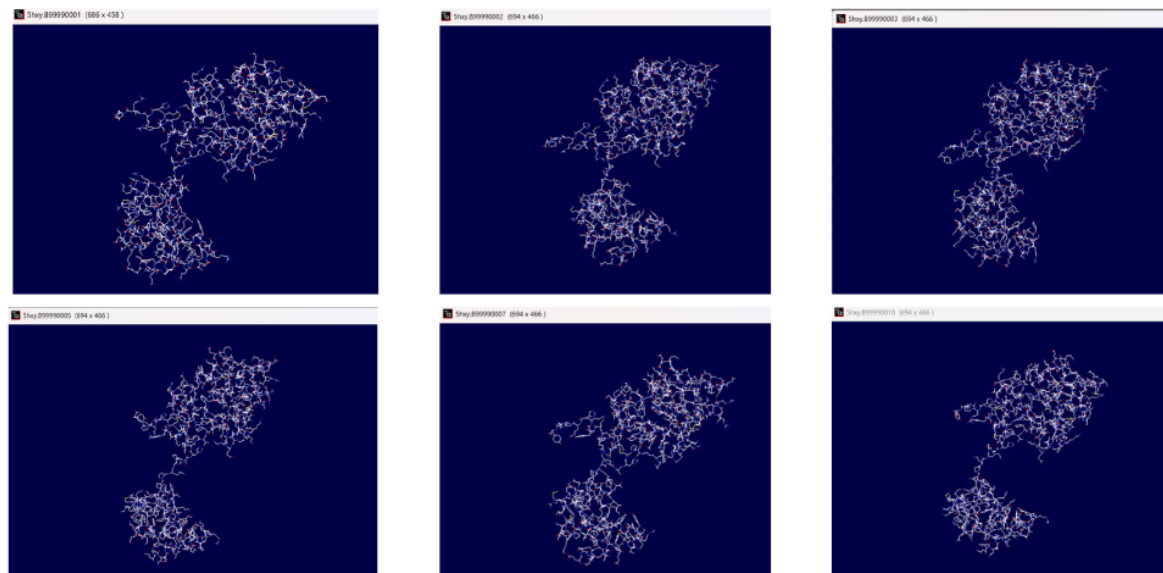


Figure 4: Representative homology models generated using MODELLER show different predicted conformations.

Some of the structures generated are as above in fig.4, and we selected the 5hxy.B99990003.pdb structure to proceed with based on the DOPE score value or DOPE score-based model selection, the model with the lowest (most negative) DOPE score, as it indicates the most energetically favourable conformation. This selected

model was used for further structural analysis and refinement.

Structural Visualization and Identification of Problematic Regions

The selected model was visualised using Swiss-Pdb Viewer.

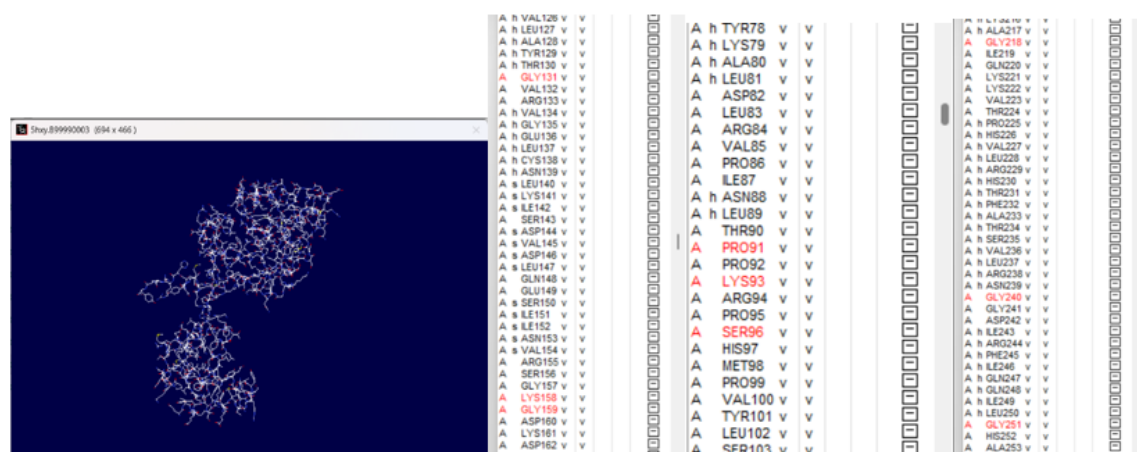


Figure 5: Structural visualization of the selected homology model (5hxy.B99990003.pdb) of AQT25855.1 showing initial loop irregularities, which were corrected through refinement, resulting in improved stability and structural compactness.

Analysis of the backbone geometry, in fig. 5, focusing on ϕ and ψ angles, identified specific regions, especially

within loop segments, which exhibited unfavourable conformations (red regions). These areas were deemed structurally inconsistent and needed refinement.

Refinement and loop modeling of the structure

We need to improve the structural quality of the model selected. Loop refinement was performed on the selected residue segment, which had an improper backbone

geometry. The refinement process was done using the modloop web server.

Following refinement, previously distorted regions were repositioned into energetically favourable conformations, improving the overall structural integrity of the model. As in Table 1, the highlighted model was selected.

Table 1: Comparative evaluation of homology models generated using MODELLER for the hypothetical protein AQT25855.1, including MolPDF, DOPE score, and GA341 score. Model 3 (5hxy.B99990003.pdb) was selected as the optimal structure based on the lowest DOPE score, indicating the most energetically favourable conformation.

Model No.	Filename	MolPDF Score	DOPE Score	GA341 Score
Model 1	5hxy.B99990001.pdb	1218.38538	-33449.27734	1.00000
Model 2	5hxy.B99990002.pdb	1450.81482	-33567.09375	1.00000
Model 3	5hxy.B99990003.pdb	1239.22949	-33894.66016	1.00000
Model 4	5hxy.B99990004.pdb	1320.98096	-33340.36328	1.00000
Model 5	5hxy.B99990005.pdb	1491.06360	-33429.04297	1.00000
Model 6	5hxy.B99990006.pdb	1233.99683	-33443.73047	1.00000
Model 7	5hxy.B99990007.pdb	1178.48071	-33556.08984	1.00000
Model 8	5hxy.B99990008.pdb	1224.84363	-33783.40234	1.00000
Model 9	5hxy.B99990009.pdb	1134.95581	-33317.35547	1.00000
Model 10	5hxy.B99990010.pdb	1255.39014	-33613.33594	1.00000

Ramachandra plot analysis

The stereochemical quality of the final refined model was assessed using Ramachandran plot analysis.

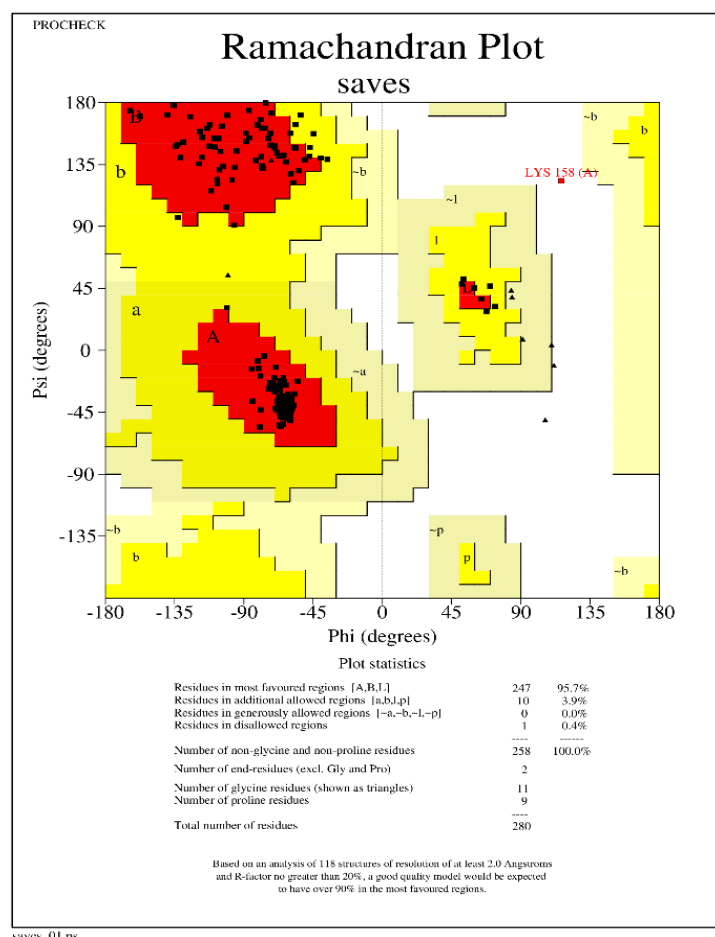


Figure 6: Ramachandran plot of the model before refinement, indicating residue distribution across conformational regions.

This Ramachandra plot was generated before the looping refinement of the structure it shows Residues in most favoured regions [A, B, L] 247 95.7% Residues in additional allowed regions [a,b,l,p] 10 3.9% Residues in generously allowed regions [-a,-b,-l,-p] 0 0.0% Residues in disallowed regions 1 0.4%, Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions as shown in fig. 6.

The Ramachandra plot generated after the refinement of the model shows A total of 249 residues.

Analysis of the model showed that 96.5% of residues were in the most favoured regions, indicating energetically stable conformations typically associated with α -helices

and β -sheets. The remaining 3.5% of residues were found in additionally allowed regions, which are acceptable and often represent flexible or loop regions of the protein. No residues appeared in generously allowed or disallowed regions, suggesting the absence of steric clashes and unfavourable backbone conformations.

As in Fig. 7, this reflects a high degree of stereochemical correctness in the predicted structure. Overall, the Ramachandran analysis indicates that the final protein model possesses **excellent structural quality and reliability**, making it suitable for further functional and structural interpretation. The obtained stereochemical parameters are consistent with high-quality protein structures, where over 90% of residues are expected to occupy favoured regions.

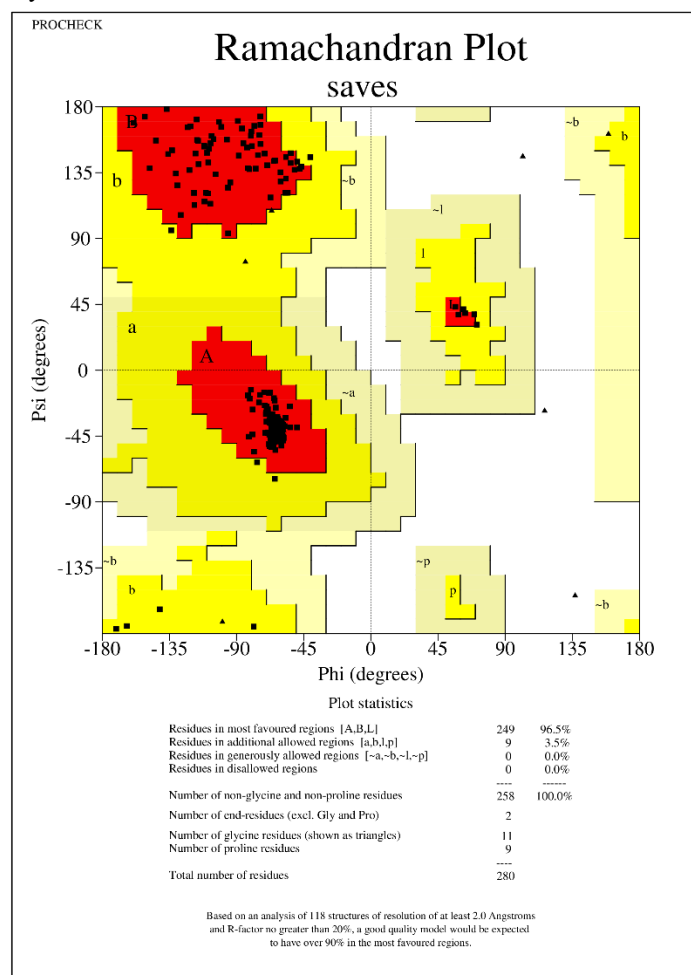


Figure 7: Ramachandran plot after refinement, demonstrating improved stereochemical quality with residues predominantly in favoured regions.

The final structure of the model in the SPDB viewer after the loop refinement is as shown in fig. 8:

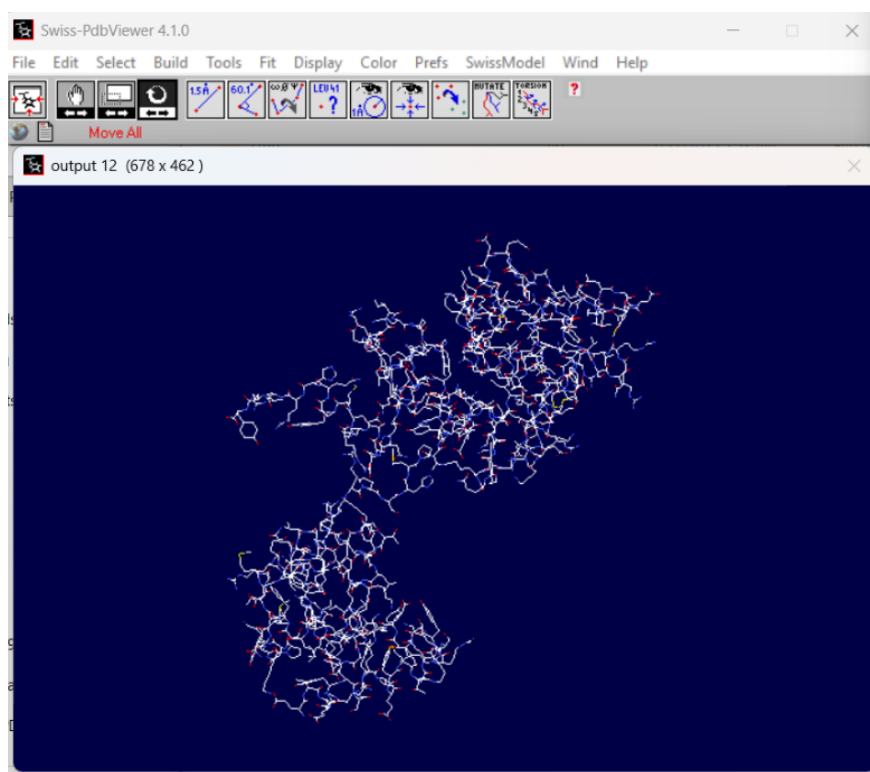


Figure 8: after modloop refinement model and in the Swiss model viewer, the model before(a) and after(b) loop refinement are as follows:

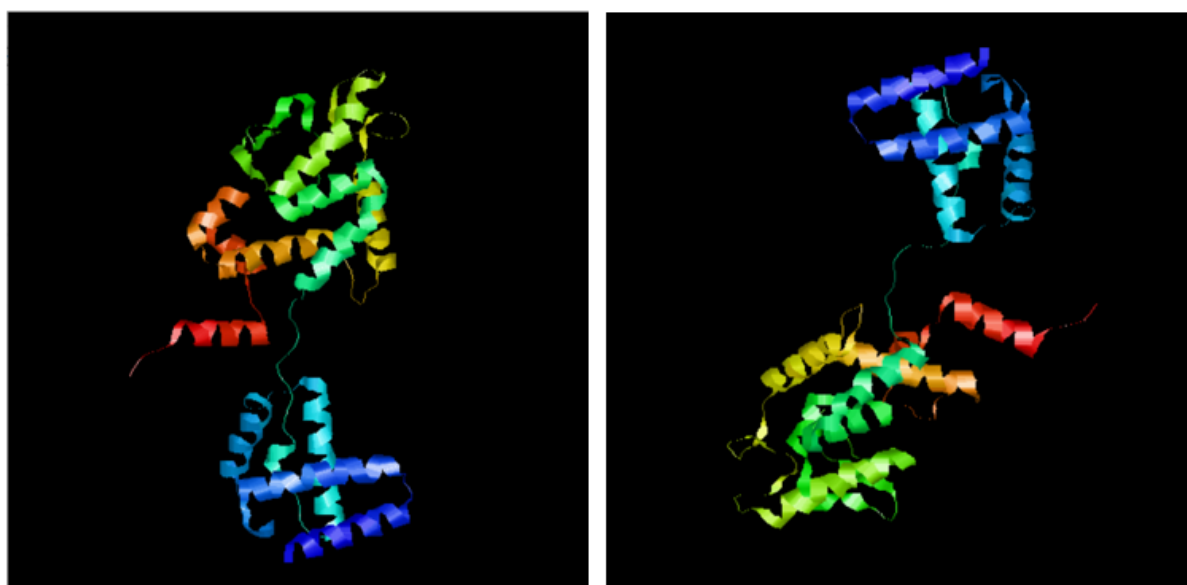


Figure 9: Comparative structural analysis of the protein model before (a) and after (b) loop refinement, showing enhanced structural organization.

Structural Comparison Before and After Loop Refinement

A comparative analysis of the structure of the protein before and after the refinement of the loop, as shown in fig. 9, indicates a significant improvement in the structure of the protein. Before the refinement, the structure of the protein was found to have an extended and irregular

structure, resulting in a disjointed structure. As a result, the structure lacked integration of various structural domains.

After the refinement of the loop, the structure of the protein was found to have significantly improved. The structure of the protein was compact and coherent. As a result, the structure of the protein was stable. The structure

of the protein was found to be well-organized. As a result, the helices were well-organized, and the structure was compact.

The refinement of the structure of the protein is consistent with the results of the analysis of the structure of the protein using the Ramachandran plot. As a result, the analysis of the structure of the protein using the Ramachandran plot was found to have an optimal structure. Therefore, the refinement of the structure of the protein was found to have a stable structure.

DISCUSSION

The objective is to elucidate the three-dimensional organization and potential functional characteristics of a hypothetical protein, AQT25855.1, derived from the *Mycobacterium* phage SimranZ1. By constructing an idealized model, we can get a more insightful idea regarding what functional role this newly detected protein might perform.

When we compare the hypothetical protein model with the modeling template, we can notice some similarities as well as dissimilarities. The template contains a tyrosine recombinase, XerA. Even though there is a mere 25% sequence similarity, the overall topology of the hypothetical protein model AQT25855.1 resembles a recombinase family protein, particularly regarding the alpha-helical regions. This indicates that AQT25855.1 might belong to a larger group of site-specific recombinases and integrases, which play a significant role in DNA recombination and integration, particularly in phages.

However, there are also differences. In the structure of AQT25855.1, there are some regions that have an irregular shape, which could be important for the stability of the structure. After refining these regions, the structure becomes much more compact and stable. This is important because it is possible that these regions could be important for the function of the enzyme. Thus, they could be behaving differently in the two proteins.

The above analysis of the structure through the BLAST tool and the conserved domains also shows that there is a possible XerD-like recombinase domain. Thus, the protein may be used for recombination-related functions. From the above analysis, it is also possible to understand the evolutionary history of the protein. The structure of the protein is conserved over time, whereas the sequences have diverged. This is possible because the structure of the protein could have evolved quickly so that it could be fine-tuned for the function of the bacteriophages. This is also possible because it is known that bacteriophages have a quick evolution rate, so that they could be fine-tuned for the function of the host.

The quality of the stereochemical structure of the model was checked through the Ramachandran plot. From the above plot, it is possible to see that the structure of the protein is of high quality. Thus, the structure of the protein is not only of high quality from a geometric point of view

but also satisfies the characteristics of a well-established protein structure.

CONCLUSION

In this study, a comprehensive in silico approach was employed to elucidate the structural and functional characteristics of the hypothetical protein AQT25855.1 from *Mycobacterium* phage SimranZ1. Through homology modelling using MODELLER, followed by systematic refinement and validation, a high-confidence three-dimensional model of the protein was successfully generated. The selection of an appropriate template (PDB ID: 5HXY) and the use of DOPE score-based model evaluation ensured the reliability of the predicted structure.

Subsequent structural refinement, particularly through loop modelling, significantly improved the geometric and conformational integrity of the model. This improvement was quantitatively supported by Ramachandran plot analysis, which demonstrated that most residues occupied energetically favourable regions, confirming the stereochemical quality and stability of the final model.

Structural comparison and domain analysis revealed that, despite relatively low sequence identity, the protein shares notable topological similarities with tyrosine recombinase/integrase family proteins. The presence of a putative XerD-like recombinase domain suggests that AQT25855.1 may play a role in site-specific DNA recombination or integration processes, which are critical for phage genome maintenance and host interaction. These findings highlight the evolutionary conservation of protein structure over sequence and reinforce the utility of structure-based annotation in characterizing hypothetical proteins.

Overall, this study demonstrates the effectiveness of computational structural biology tools in bridging the annotation gap associated with hypothetical proteins. The predicted model provides a valuable framework for future experimental validation and functional studies. Furthermore, it contributes to a deeper understanding of mycobacteriophage biology and opens avenues for exploring novel phage-derived biomolecules with potential applications in biotechnology and therapeutics.

REFERENCES

1. Das, R. and U. Bajpai, Functional characterization of a DNA-dependent AAA ATPase in SimranZ1, a F cluster mycobacteriophage. *bioRxiv* 2021.
2. Bin Hafeez, A., et al., In Silico Functional Annotation and Structural Characterization of Hypothetical Proteins in *Bacillus paralicheniformis* and *Bacillus subtilis* Isolated from Honey. *ACS Omega*, 2025. 10(9): p. 8993-9006.
3. a, S.R., et al., Modeling Structure Design Of *Mycobacterium tuberculosis* KU- Protein Using Modeler As Well Study Interaction Between KU-

- Protein and DNA Polymerase. Uttar Pradesh Journal of Zoology, 2021. 2021 - Volume 42
4. NCBI <http://www.ncbi.nlm.nih.gov/>
5. RCSB PDB
6. Aanchal Tyagi, S.R., Nikunaj Bhardwaj and Arvind Kumar Tiwari, Whole genome sequencing Lactobacillus rhamnosus GG genome : Reference for a complex bacterial genome. 2021. Volume: 21.
7. BLAST <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
8. Marti-Renom, M.A., et al., Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct, 2000. 29: p. 291-325.
9. Fiser, A., R.K. Do, and A. Sali, Modeling of loops in protein structures. Protein Sci, 2000. 9(9): p. 1753-73.
10. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, Stereochemistry of polypeptide chain configurations. J Mol Biol, 1963. 7: p. 95-9.
11. Ramachandran plot, in Ramachandran plot.
12. Colovos, C. and T.O. Yeates, Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci, 1993. 2(9): p. 1511-9.