In-silico Structure Prediction and Validation of Human Interleukin-34 Using AlphaFold ColabFold, and GalaxyRefine

Project Aim:

This project focuses on practicing protein structure prediction using a hypothetical protein with no experimentally resolved structure available. A stepwise computational approach involving structure prediction, refinement, and validation was employed to obtain a reliable three-dimensional model for further analysis.

Protein Selection

Human Interleukin-34 (IL-34) was selected as the target protein for this study.

Rationale for Protein Selection

Interleukin-34 (IL-34) is a cytokine involved in immune regulation and inflammatory signaling pathways. Despite its biological significance, a complete experimentally determined 3D structure of IL-34 is not available in the Protein Data Bank (PDB), making it a suitable candidate for computational structure prediction. The relatively small size of the protein enables efficient and rapid structure prediction using AlphaFold, making it ideal for practice and methodological demonstration. Additionally, the availability of an AlphaFold Database prediction allows comparative analysis to assess prediction consistency and model confidence.

Protein Information

- **Protein name:** Interleukin-34
- **Gene symbol:** *IL34*
- **Organism:** *Homo sapiens*
- **UniProt accession ID:** Q8WWI3
- **Protein length:** 222 amino acids

PDB Structure Search

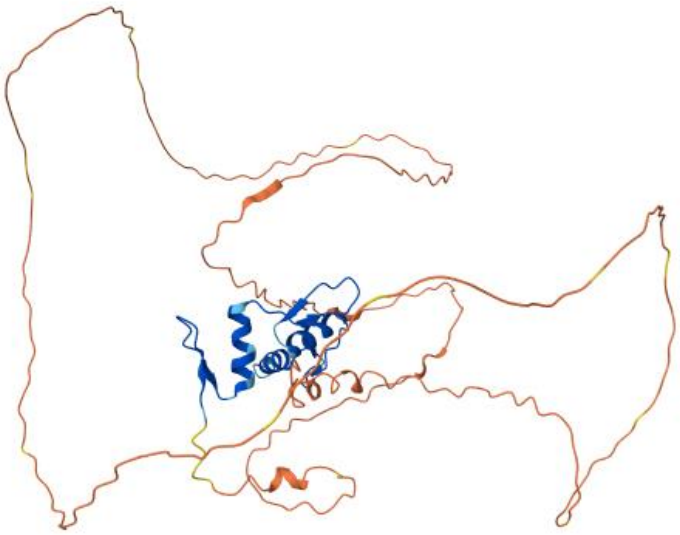
Before proceeding with structure prediction, an extensive search was performed to identify any experimentally resolved three-dimensional structure of the selected protein. The **Protein Data Bank (PDB)** was queried using the UniProt accession ID **Q8WWI3** and the protein name *Interleukin-34*. The search revealed that no complete experimentally determined 3D structure of human IL-34 is available in the PDB. Only partial or indirect structural information was found, which was insufficient for detailed structural analysis. This confirmed the absence of an experimentally solved structure and justified the use of computational structure prediction approaches.

AlphaFold Database Predicted Structure and Confidence Assessment

Following the initial PDB search, the **AlphaFold Protein Structure Database** was explored to determine whether a predicted structure of human IL-34 (UniProt ID: Q8WWI3) was already available. The already predicted model retrieved revealed a partial structural prediction; however, the overall confidence of the model was relatively low.

The **predicted local distance difference test (pLDDT) scores** indicated that most regions of the protein fell within the **low- to medium-confidence range**, corresponding to the **orange-colored segments** observed in the 3D structure visualization. These regions primarily included loops and terminal residues, suggesting increased flexibility or disorder in these parts of the protein. Only a few secondary structure elements, such as short α-helices, displayed moderately higher confidence (blue regions).

Additionally, the **Predicted Aligned Error (PAE) plot** showed elevated error estimates across most residue pairs, indicating uncertainty in the relative positioning of structural elements. This suggests that, while AlphaFold provides a tentative model, experimental or further computational refinement would be required for reliable structural interpretation.



Model Confidence

Very high (pLDDT > 90)

High (90 > pLDDT > 70)

Low (70 > pLDDT > 50)

Very low (pLDDT < 50)

pLDDT is a per-residue measure of local confidence.

Domains (0)

This protein does not currently have any structural domains identified. Domain annotations will appear here if data becomes available in future updates.

Fig.1 : Already predicted structure of human IL-34 (UniProt ID: Q8WWI3)

Result :Average pLDDT 47.47 (Very Low)

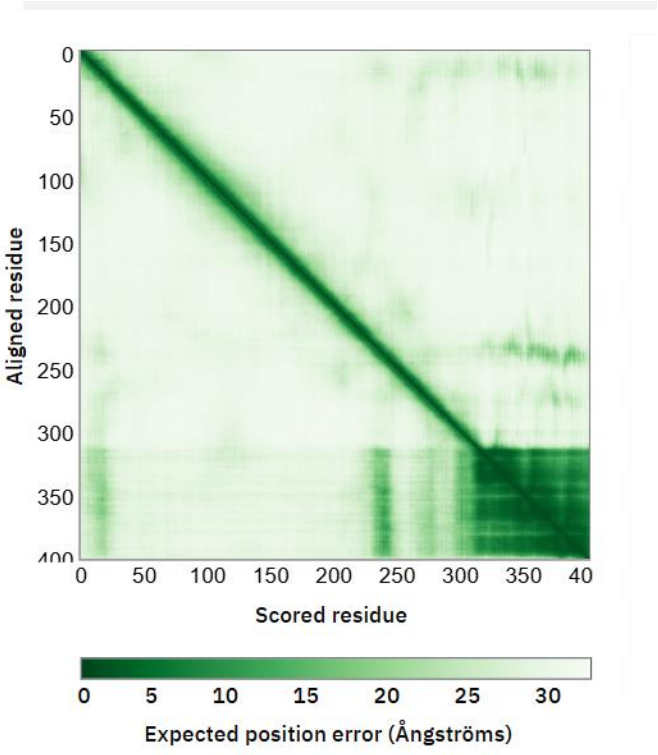


Fig.2 :PAE plot

The **PAE plot** is **not** an inter-residue distance map or a contact map. Instead, the shade of green indicates the expected distance error in Ångströms (Å), ranging from 0 Å to an arbitrary cut-off of 31 Å. The colour at (x, y) corresponds to the expected distance error in the residue x's position when the predicted and the true structures are aligned on residue y.

A dark green tile corresponds to a good prediction (low error), whereas a light green tile indicates poor prediction (high error).

AP-2 repressor, Sequence length 402 human IL-34 (UniProt ID: Q8WWI3)

MNIHMNRKTIKNINTFENRMLMLDGMPAVRVKTELLESEQGSPNVHNYPDMEAVPLLLNNVKG
EPPEDSLSVDHFQTQTEPVDLSINKARTSPTAVSSSPVSMTASASSPSSTSTSSSSSSRLASSPTVITS
VSSASSSSTVLTPGPLVASASGVGGQQFLHIIHPVPPSSPMNLQSNKLSHVHRIPVVVQSVPVVYT
AVRSPGNVNNTIVVPLLEDGRGHGKAQMDPRGLSPRQSKSDSDDDDLPNVTLDSVNETGSTALS
IARAVQEVHPSPVSRVRGNRMNNQKFPCSISPFSIESTRRQRRSESPDSRKRRIHRCDFEGCNKVY
TKSSHLKAHRRTHTGEKPYKCTWEGCTWKFARSDELTRHYRKHTGVKPFKCADCDRSFSRSDH
LALHRRRHMLV

AlphaFold-Predicted Structure Confidence and Validation Analysis

The AlphaFold Protein Structure Database provides a predicted three-dimensional structure for human Interleukin-34 (UniProt ID: Q8WWI3). However, confidence assessment of the model revealed that the majority of the protein exhibited **low prediction confidence**, as indicated by the predominance of **orange-colored regions** in the pLDDT-based visualization. These regions correspond mainly to flexible loops and terminal segments, while only a limited central region showed moderate confidence.

Visualization of the AlphaFold-predicted structure using **PyMOL** and **Discovery Studio** further confirmed the presence of extensive disordered and poorly defined regions, suggesting structural instability and uncertainty in residue positioning. The **Predicted Aligned Error (PAE) plot** showed elevated error values across many residue pairs, indicating low confidence in the relative spatial arrangement of structural elements throughout the protein.

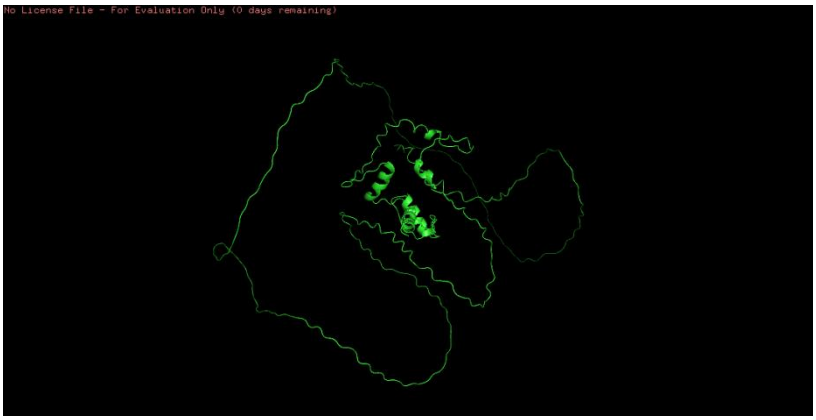


Fig.3 : Visualization of the AlphaFold-predicted structure using **PyMOL**

Structure Validation Results_ (Already predicted)

To assess the reliability of the AlphaFold-predicted structure, multiple structure validation tools were employed.

ERRAT Analysis

The ERRAT evaluation yielded an **overall quality factor of 93.51**, which generally indicates acceptable non-bonded atomic interactions. Despite this high score, ERRAT alone does not assess stereochemical correctness or residue environment compatibility, necessitating additional validation .

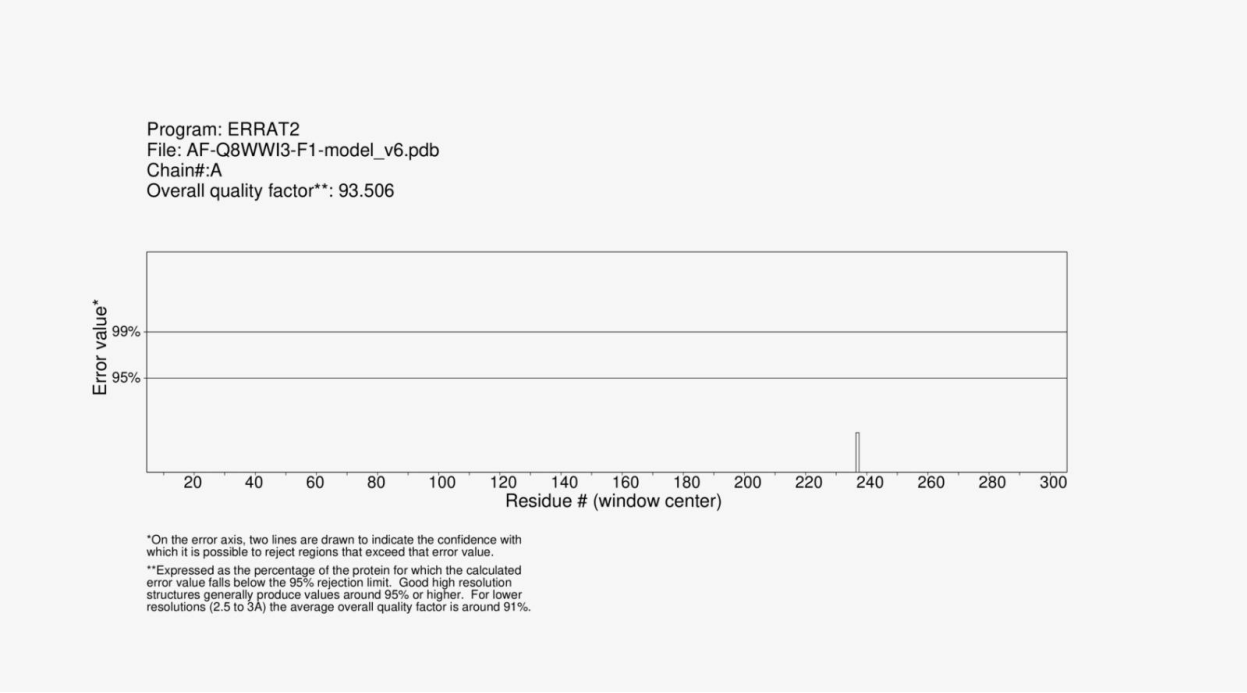


Fig.4 : ERRAT plot

Verify3D Analysis

The Verify3D assessment revealed that only **24.88% of residues achieved an averaged 3D–1D score ≥ 0.1**, which falls well below the recommended threshold of 80%. Consequently, the model **failed the Verify3D validation**, indicating poor compatibility between the predicted three-dimensional structure and its corresponding amino acid sequence environment.

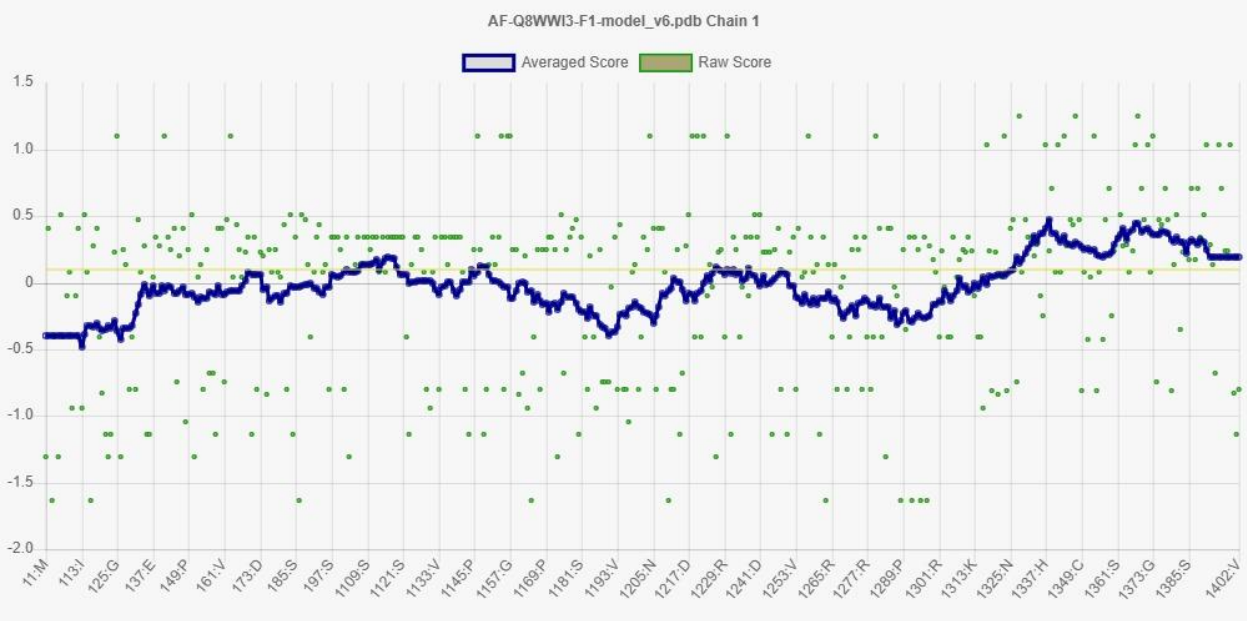
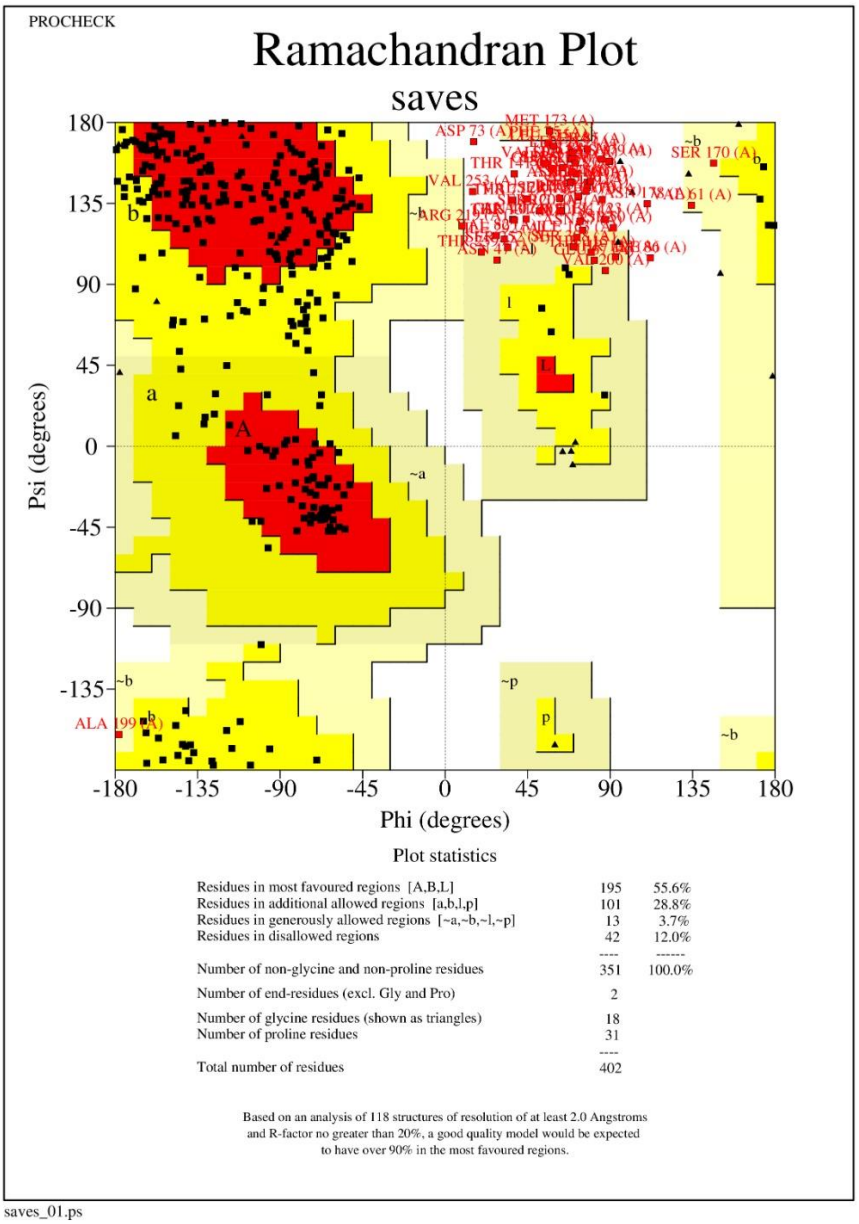


Fig.5 : Verify3D plot

PROCHECK Analysis

PROCHECK stereochemical evaluation further highlighted significant structural deficiencies. Ramachandran plot analysis showed that only **55.6% of residues were located in core (most favored) regions**, while **12.0% of residues occupied disallowed regions**, which exceeds acceptable limits for a reliable protein model. In total, **7 errors and 2 warnings** were reported across nine evaluation criteria, with no evaluation passing without issues. A high number of labeled residues in Ramachandran and side-chain conformational plots further indicated deviations from ideal geometry.



saves_01.ps

Fig.6: PROCHECK plot

Overall Interpretation_ (Already predicted)

Although the AlphaFold-predicted model of IL-34 demonstrated a high ERRAT score, the **low pLDDT confidence**, **elevated PAE values**, **failure of Verify3D**, and **poor Ramachandran statistics** collectively indicate that the structure is **not sufficiently reliable for detailed structural or functional analysis** without further refinement. These results emphasize the importance of cautious interpretation of AlphaFold predictions, particularly for proteins with extensive flexible or disordered regions.

AlphaFold-Based Structure Prediction Using ColabFold

To improve upon the low-confidence model obtained from the AlphaFold Protein Structure Database, an independent structure prediction of human Interleukin-34 (UniProt ID: Q8WWI3) was performed using the **ColabFold implementation of AlphaFold**.

Model Selection and Parameter Settings

The full-length amino acid sequence of IL-34 was retrieved from UniProt and submitted to the ColabFold notebook for de novo structure prediction. The following parameters were applied:

- **Prediction engine:** AlphaFold2
- **MSA generation:** MMseqs2
- **Template usage:** Disabled
- **Number of models generated:** Five (rank_1 to rank_5)
- **Recycles:** Default
- **Amber relaxation:** Enabled

Template-free modeling was selected to ensure unbiased prediction based solely on evolutionary information. Multiple models were generated to allow selection of the most reliable structure based on confidence metrics.

The final model was selected based on overall confidence scores and Predicted Aligned Error (PAE) analysis, with the **rank_1 model** exhibiting the most favorable quality parameters.

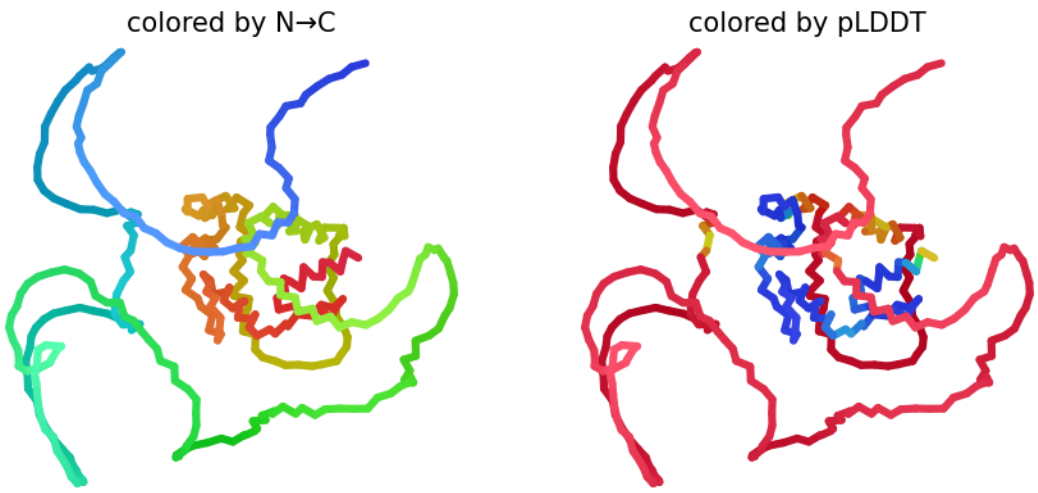


Fig.7 : Rank_1 model ColabFold implementation of AlphaFold**.**

Prediction Results and Confidence Assessment

Structural prediction of the IL-34 protein was performed using ColabFold, generating five ranked models (rank_1 to rank_5). Model confidence was evaluated using predicted Local **Distance Difference Test (pLDDT) scores, Predicted Aligned Error (PAE) maps, and Multiple Sequence Alignment (MSA) coverage.**

The per-residue pLDDT profiles showed consistent confidence patterns across all five models. Overall, the N-terminal and central regions (approximately residues 1–280) exhibited low to moderate pLDDT values (≈25–55), suggesting structural flexibility or partial disorder. In contrast, a pronounced increase in pLDDT scores was observed in the C-terminal region (approximately residues 300–390), where values exceeded 85–90, indicating high local structural reliability. Minor fluctuations were present at the extreme C-terminus, but overall this region demonstrated strong prediction confidence across all ranked models. Among the five models, rank_1 displayed slightly higher and more uniform pLDDT values, supporting its selection for further analysis.

PAE heatmap analysis revealed similar confidence distributions among all models. A distinct low-error diagonal pattern was observed, reflecting reliable local residue positioning. However, extensive high-error regions were present away from the diagonal, indicating uncertainty in long-range inter-residue relationships. Notably, the C-terminal segment (residues ~300–400) showed comparatively lower PAE values, consistent with its elevated pLDDT scores and suggesting enhanced structural stability. In contrast, the N-terminal and central regions exhibited higher PAE values, indicative of conformational flexibility and reduced global confidence.

MSA coverage analysis further supported these observations. The N-terminal region (residues 1–200) showed sparse sequence coverage, reflecting limited evolutionary conservation and weak homologous support. Conversely, the region spanning residues ~250–350 demonstrated substantially higher MSA depth, suggesting strong evolutionary conservation. This increased alignment depth corresponded with improved pLDDT scores and reduced PAE values, underscoring the influence of evolutionary information on prediction accuracy. Terminal regions displayed reduced coverage, contributing to their lower confidence scores.

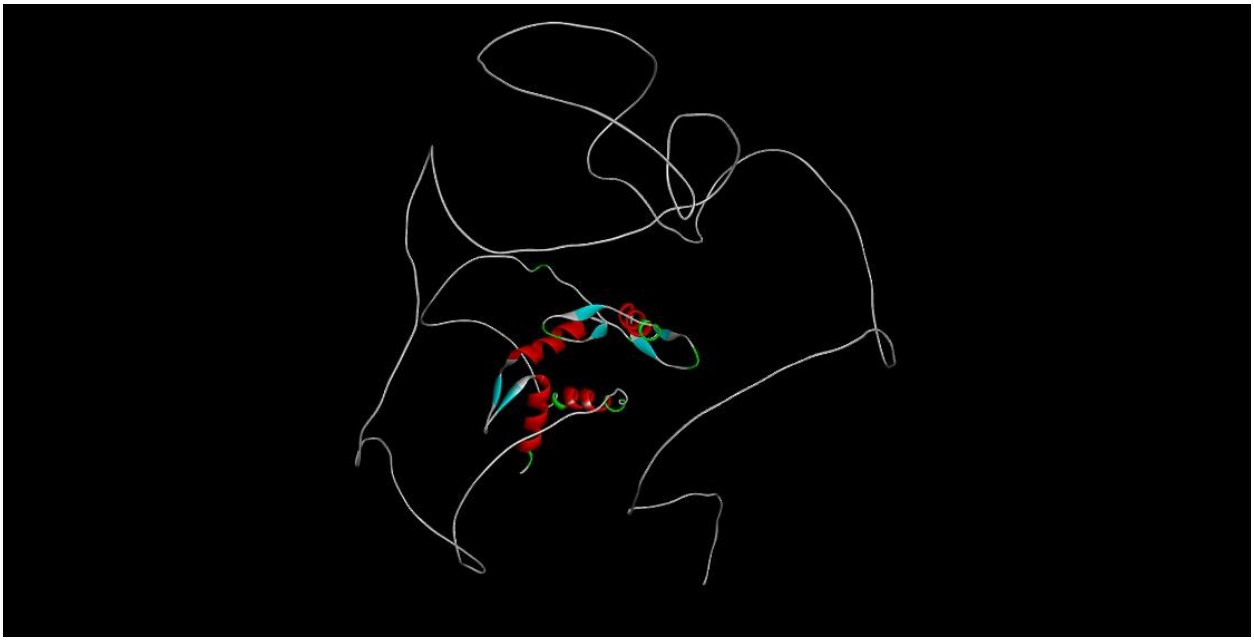


Fig.8 : Pymol visuliazation of Alphafold_Collab of Rank model-1

Comparative assessment with the AlphaFold Database model indicated that the ColabFold-predicted structure exhibited improved secondary structure definition and enhanced domain coherence, particularly within conserved regions. Despite these improvements, several low-confidence and potentially disordered segments persisted, especially in regions with poor MSA coverage

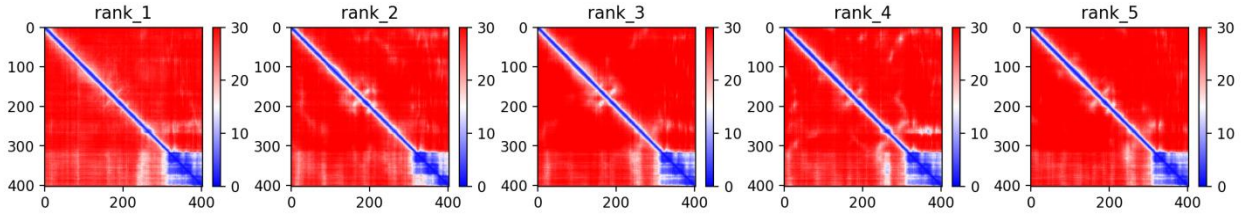


Fig.9: PAE heatmap of ranked models (rank_1 to rank_5)

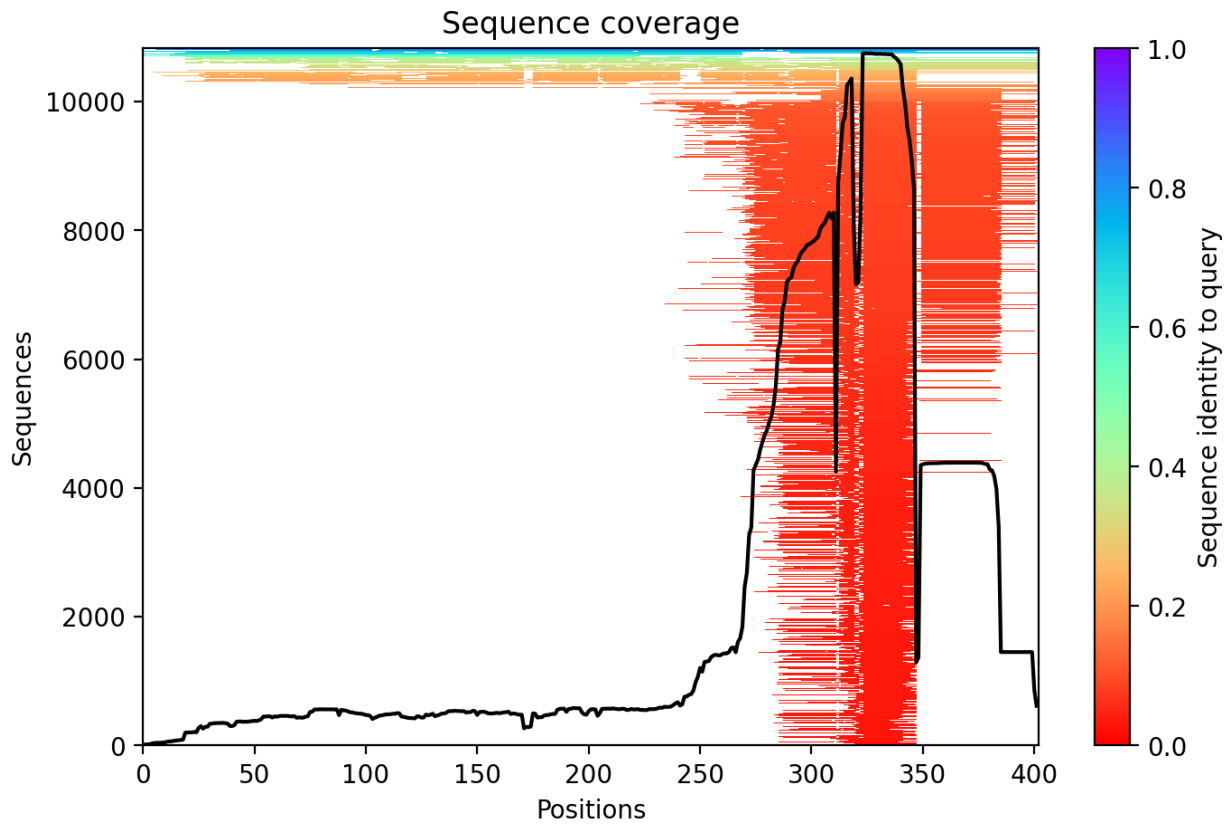


Fig.10: MSA coverage analysis plot

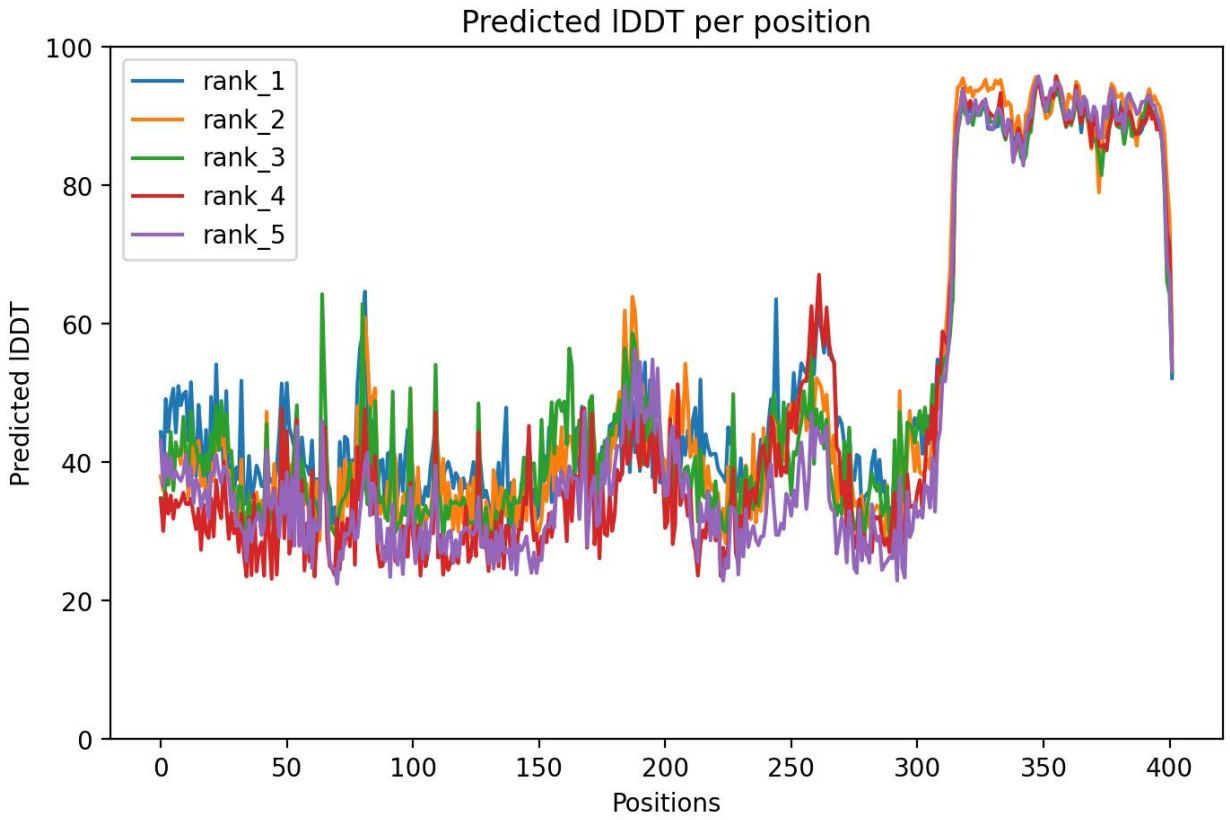


Fig.9: pLDDT plot of ranked models (rank_1 to rank_5)

Overall, the integrated analysis of pLDDT, PAE, and MSA coverage suggests that while the IL-34 model is reliable in conserved C-terminal regions, the N-terminal and central segments likely possess intrinsic flexibility or disorder. Therefore, the predicted structure is suitable for preliminary structural and functional insights; however, experimental validation and further refinement are required for high-resolution structural interpretation.

Structure Prediction and Refinement of IL-34

The initial AlphaFold-predicted model of IL-34, while demonstrating a high ERRAT score, exhibited low pLDDT confidence, elevated PAE values, failure in Verify3D assessment, and suboptimal Ramachandran statistics. Collectively, these indicators suggested that the raw AlphaFold structure was not sufficiently reliable for detailed structural or functional analysis, highlighting the need for further refinement, particularly for proteins with flexible or disordered regions.

To improve model quality, the structure was refined using **GalaxyRefine**, which applies iterative side-chain rebuilding and overall relaxation to enhance stereochemical parameters. Five refined models were generated and assessed based on multiple structural validation metrics, including GDT-HA, RMSD, MolProbity score, clash score, poor rotamers, and Ramachandran favored regions (Table 1).

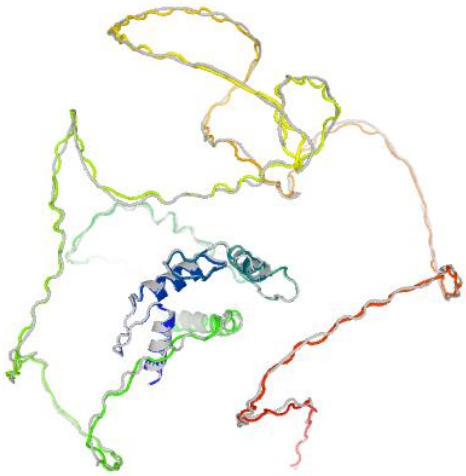


Fig.10: Galaxy Refine Model-4 UniProt ID: Q8WWI3 ranked models (rank_1 to rank_5)

Structure Information

Model	GDT-HA	RMSD	MolProbity	Clash score	Poor rotamers	Rama favored
Initial	1.0000	0.000	4.182	70.3	12.6	37.5
MODEL 1	0.7363	1.038	1.547	3.0	0.3	92.8
MODEL 2	0.7357	1.060	1.491	3.0	0.0	94.0
MODEL 3	0.7481	1.021	1.566	3.6	0.8	93.8
MODEL 4	0.7425	1.039	1.527	3.8	0.3	94.8
MODEL 5	0.7544	1.011	1.585	3.5	0.3	93.0

combination of **high Ramachandran favored regions (94.8%)**, low clash score (3.8), minimal poor rotamers (0.3%), and overall favorable MolProbity score (1.527). This refined model provides a reliable framework for subsequent structural analyses, docking studies, or functional investigations of IL-34.

Results and Conclusion

In this study, the three-dimensional structure of Human Interleukin-34 (IL-34) was investigated using a systematic computational modeling and validation workflow. Initial assessment of the AlphaFold Database model revealed very low prediction confidence, with an average pLDDT score of 47.47. Furthermore, structural validation using ERRAT, Verify3D, and PROCHECK indicated poor stereochemical quality and limited compatibility between the predicted structure and its amino acid sequence. These results demonstrated that the pre-predicted AlphaFold model was not sufficiently reliable for detailed structural or functional interpretation.

To overcome these limitations, a new structure prediction was performed using the ColabFold implementation of AlphaFold. Multiple ranked models were generated, showing improved local confidence, particularly within conserved regions of the protein. However, several flexible and poorly conserved segments continued to exhibit low confidence, highlighting the intrinsically disordered nature of certain regions of IL-34.

Subsequently, the predicted structures were refined using GalaxyRefine to enhance structural stability and stereochemical accuracy. Five refined models were produced and evaluated based on GDT-HA, RMSD, MolProbity score, clash score, poor rotamer frequency, and Ramachandran plot statistics. Among these, Model 4 demonstrated the most favorable validation profile, with 94.8% residues in favored Ramachandran regions, low clash score (3.8), minimal poor rotamers (0.3%), and a satisfactory MolProbity score (1.527). These improvements reflect substantial enhancement in geometric quality and overall structural reliability.

Overall, this study highlights the importance of comprehensive validation and refinement of AlphaFold-predicted models, particularly for proteins lacking experimentally determined structures and containing flexible regions. The final refined IL-34 model obtained in this work provides a reliable structural framework for future molecular docking, functional analysis, and drug discovery studies. Nevertheless, experimental validation remains essential to confirm the accuracy of the predicted structure at high resolution.