Report on designing an expression strategy for target sequence using a Baculovirus Expression Vector System (BEVS) in Sf9/Sf21insect cells

Author : Shreya Adhya

Date: 09.02.26

Target Sequence

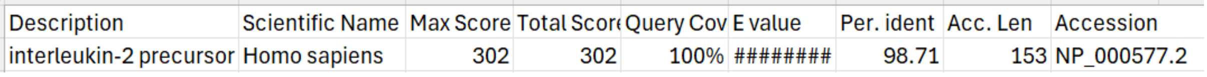
>Target_X = Human IL-2 Precursor

MYRMQLLSCIALSLALVTNSAPTSSSTKKTQLQLEHLLLDLQMVILNGINNYKNPKLTRMLTFKFYMPKKA
TELKHLQCLEEELKPLEEVLNLAQSKNFHLRPRDLISNINVIVLELKGSETTFMCEYADEKTATIVEFLNRWI
TFCQSIISTLT

Part 1: Expression & Engineering Report

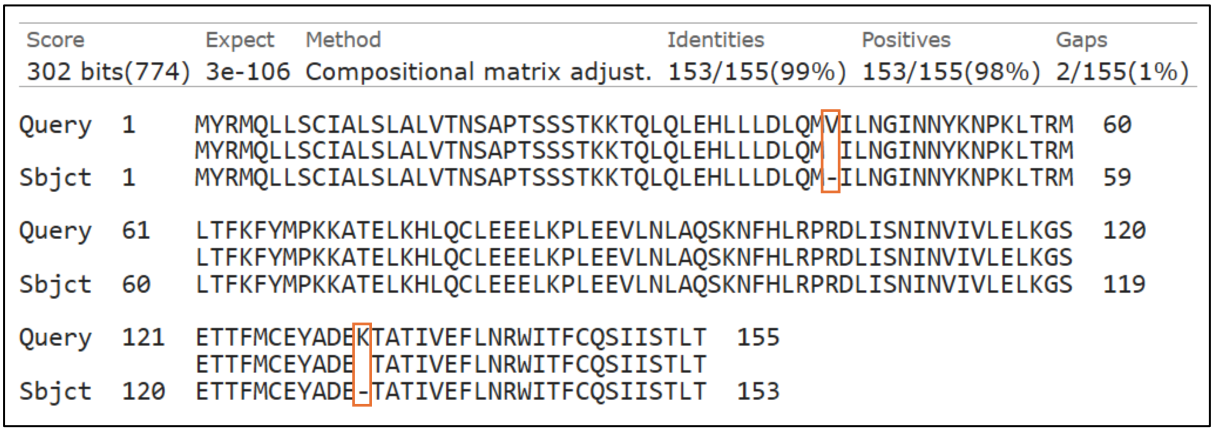
- **Evaluation of sequence data:**

On running a **BLASTp** analysis against the NCBI nr database identified the sequence was human interleukin-2 precursor **(NP_000577.2, corresponding UniProt - P60568)**



Description	Scientific Name	Max Score	Total Scor	Query Cov	E value	Per. ident	Acc. Len	Accession
interleukin-2 precursor	Homo sapiens	302	302	100%	########	98.71	153	NP_000577.2

BLAST ouput



```
Score           Expect  Method                        Identities      Positives       Gaps
302 bits(774)  3e-106  Compositional matrix adjust.  153/155(99%)  153/155(98%)  2/155(1%)

Query  1    MYRMQLLSCIALSLALVTNSAPTSSSTKKTQLQLEHLLLDLQMVILNGINNYKNPKLTRM  60
            MYRMQLLSCIALSLALVTNSAPTSSSTKKTQLQLEHLLLDLQM ILNGINNYKNPKLTRM
Sbjct  1    MYRMQLLSCIALSLALVTNSAPTSSSTKKTQLQLEHLLLDLQM-ILNGINNYKNPKLTRM  59

Query  61   LTFKFYMPKKATELKHLQCLEEELKPLEEVLNLAQSKNFHLRPRDLISNINVIVLELKGS  120
            LTFKFYMPKKATELKHLQCLEEELKPLEEVLNLAQSKNFHLRPRDLISNINVIVLELKGS
Sbjct  60   LTFKFYMPKKATELKHLQCLEEELKPLEEVLNLAQSKNFHLRPRDLISNINVIVLELKGS  119

Query  121  ETTFMCEYADEKTATIVEFLNRWITFCQSIISTLT  155
            ETTFMCEYADE TATIVEFLNRWITFCQSIISTLT
Sbjct  120  ETTFMCEYADE-TATIVEFLNRWITFCQSIISTLT  153
```

The % identity was <100% - because at two positions two amino acids were introduced in provided sequence.

- **Codon Optimization**

Step 1 : Optimisation by tool

Tool – **IDT Codon Optimization** (link)

Parameters set –

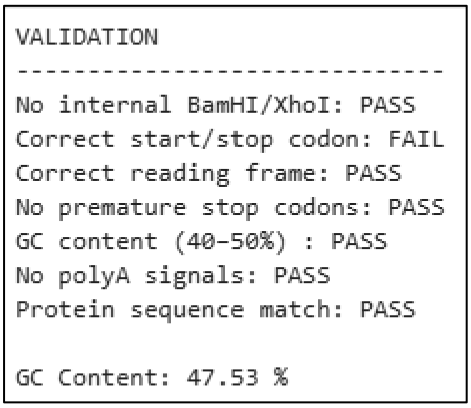
- host - *Spodoptera frugiperda* cells
- restriction sites excluded – BamHI and XhoI

Output sequence –

ATG TAC AGG ATG CAG CTC CTC AGT TGT ATA GCC TTG TCA CTG GCC CTC GTG ACG AAC TCG GCT CCA ACG TCT TCC AGT ACG AAG AAA ACC CAG CTC CAA TTG GAG CAC CTT CTC CTT GAT CTG CAA ATG GTG ATC CTT AAT GGT ATC AAC AAC TAC AAG AAC CCG AAA CTG ACC CGT ATG TTG ACC TTC AAA TTT TAC ATG CCT AAA AAG GCC ACA GAG CTG AAA CAC CTG CAA TGC CTC GAA GAA GAA CTG AAA CCA CTG GAG GAA GTC CTT AAC CTG GCT CAG TCG AAA AAT TTC CAC CTT AGG CCC CGT GAC CTC ATC TCT AAC ATT AAC GTG ATT GTC CTC GAA CTG AAG GGC AGC GAG ACA ACC TTC ATG TGT GAA TAT GCC GAC GAA AAG ACT GCT ACG ATA GTG GAG TTT CTG AAT CGC TGG ATT ACG TTC TGC CAA TCA ATC ATC AGC ACC CTG ACC

Step 2 : Validate the codon sequence

A Python code was written to validate the optimized codon sequence for the following fields(attached in a .zip file)



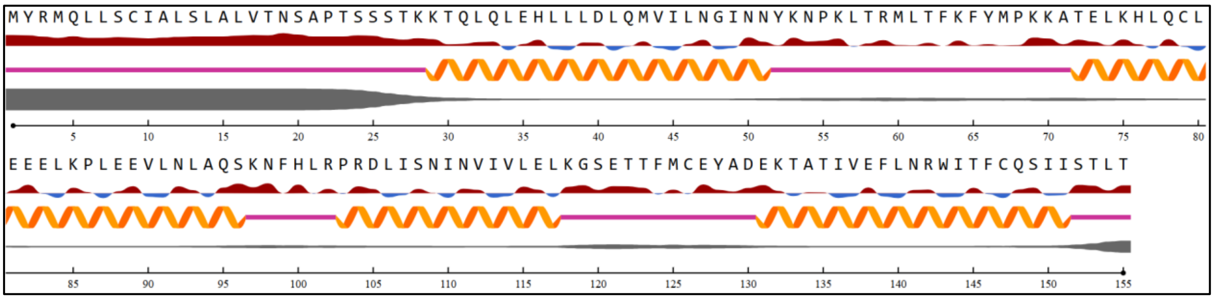
```
VALIDATION
------------------------------
No internal BamHI/XhoI: PASS
Correct start/stop codon: FAIL
Correct reading frame: PASS
No premature stop codons: PASS
GC content (40-50%) : PASS
No polyA signals: PASS
Protein sequence match: PASS

GC Content: 47.53 %
```

There was a case of FAIL : the stop codons TAA, TAG or TGA were absent.

- **Rational Design:**

Tool - NetSurfP - 3.0 tool

This tool predicts the surface accessibility, secondary structure, disorder, and phi / psi dihedral angles of amino acids in an amino acid sequence.



Relative Surface Accessibility: Red is exposed and blue is buried, thresholded at 25%.
Secondary Structure: Helix, ➡ Strand, — Coil.
Disorder: Thickness of line equals probability of disordered residue.

Inference – The image result below indicates the regions where the protein is disordered in grey and it is **highly disordered towards N-terminal** at positions 1 to 30, as the hydrophobic residues are exposed. This region brings instability in the protein and are likely to cause stability or folding issues.

Suggested point mutation to improve the solubility of the recombinant product –

L → Q - Reduces surface hydrophobicity, improves solubility while preserving fold

I → T - Increases local polarity, lowers aggregation tendency

L → E - Introduces negative charge, strongly disrupts hydrophobic clustering

Y → Q - Removes aromatic hydrophobicity, enhances secretion and solubility

- **Vector Construction**

1. Insect-cell-specific leader sequence

A **gp67 signal peptide** is recommended as the insect-cell-specific leader sequence for IL-2 expression in BEV system.

gp67 signal peptide is **optimized for insect cells**(Sf9 and Sf21), it ensures proper protein folding and efficient processing along the insect secretory pathway. This makes it particularly **effective for secreted, soluble proteins** such as cytokines.

Its reliability and high secretion efficiency have made gp67 the **standard choice** for producing recombinant human proteins in insect cell expression systems.

Comparison with Other Signal Peptides

- **gp67**: Provides the most consistent and high-level secretion of **soluble recombinant proteins** in Sf9/Sf21 cells and is therefore preferred for cytokines such as IL-2.
- **gp64**: Derived from a baculoviral membrane fusion protein; although efficient for ER targeting, it often leads to partial membrane association, making it less suitable for fully secreted soluble proteins.
- **Honeybee melittin**: Can promote strong secretion and, in some cases, perform comparably to gp67; however, its efficiency is more protein-dependent, whereas gp67 offers greater reliability across targets.

References

Scholz, J., Suppmann, S. A new single-step protocol for rapid baculovirus-driven protein production in insect cells. *BMC Biotechnol* **17**, 83 (2017). https://doi.org/10.1186/s12896-017-0400-3

Chakraborty S, Trihemasava K, Xu G. Modifying Baculovirus Expression Vectors to Produce Secreted Plant Proteins in Insect Cells. J Vis Exp. 2018 Aug 20;(138):58283. doi: 10.3791/58283. PMID: 30176019; PMCID: PMC6128212.

Titus Kretzschmar, Laurent Aoustin, Otto Zingel, Marcello Marangi, Bénédicte Vonach, Harry Towbin, Martin Geiser, High-level expression in insect cells and purification of secreted monomeric single-chain Fv antibodies, Journal of Immunological Methods, h ttps://doi.org/10.1016/0022-1759(96)00093-2.

C.I. Murphy, J.R. Mcintire, D.V. Davis, H. Hodgdon, J.R. Seals, E. Young,Enhanced Expression, Secretion, and Large-Scale Purification of Recombinant HIV-1 gp120 in Insect Cells Using the Baculovirus egt and p67 Signal Peptides,Protein Expression and Purification, https://doi.org/10.1006/prep.1993.1046.

2. Fusion tags necessary for purification

Fusion tags are added for two purposes : better purification or better solubility of the expressed protein.
Since it's mentioned about purity these following fusion tags can be used based on our current design :

- **His_6 tag** – small size, less interference with protein structure and function and has low-cost purification.
- **Strep-tag II –** small size and used when sensitive proteins or when avoiding metal ions is important
- Other than these there's FLAG, MBP, CBP, Fc

References

Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system.Frontiers in Microbiology,2014,https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2014.00063,10.3389/fmicb.2014.00063

Kosobokova, E.N., Skrypnik, K.A. & Kosorukov, V.S. Overview of fusion tags for recombinant proteins. *Biochemistry Moscow* **81**, 187–200 (2016). https://doi.org/10.1134/S0006297916030019

https://pef.facility.uq.edu.au/fusion-tags-protein-purification

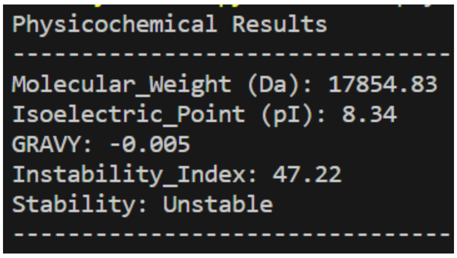
Part 2: Computational Analysis

All the codes are zipped into a separate file named by the corresponding section

1. Physicochemical Profiling: Calculate Molecular Weight (MW), Isoelectric Point (pI), Grand Average of Hydropathicity (GRAVY), and Instability Index

Code logic : the code used the in-built function to calculate the above physiochemical properties using **ProteinAnalysis** class from SeqUtils package. It is based on ProtParam tools on the Expasy Proteomics Server.

Output :



```
Physicochemical Results
-------------------------------
Molecular_Weight (Da): 17854.83
Isoelectric_Point (pI): 8.34
GRAVY: -0.005
Instability_Index: 47.22
Stability: Unstable
-------------------------------
```

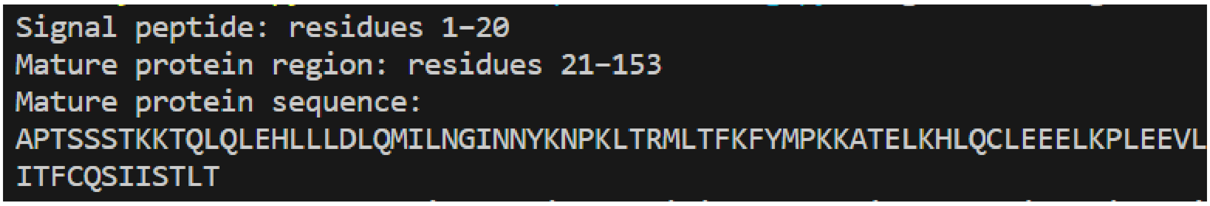
This indicates the protein is unstable as the instability index is above 40, which is coming from the exposed hydrophobic patch in the signal peptide.

2. Sequence Parsing: Identify the native human signal peptide (if present) and output the mature protein sequence.

Step 1. Using a Python code to find signal peptide and mature protein from UniProt annotation

Code logic : Using BLASTp result protein sequence, the UniProt data was extracted for that protein P60568 using API call > The sequence was parsed using Biopython and the features ‘Signal’ and ‘Chain’ was searched from it. > ‘Signal’ represents the peptide signal and ‘Chain’ represents the mature protein.

Output :



```
Signal peptide: residues 1-20
Mature protein region: residues 21-153
Mature protein sequence:
APTSSSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEELKPLEEVL
ITFCQSIISTLT
```

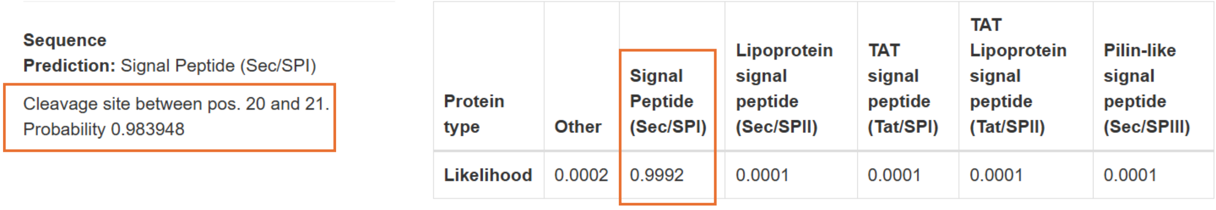
The mature protein is :

APTSSSTKKTQLQLEHLLLDLQMILNGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEELKPLEEVLNLAQSKNFHLRPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT

Step 2. Using SignalP tool to validate the Python output.

Tool : https://services.healthtech.dtu.dk/services/SignalP-6.0/

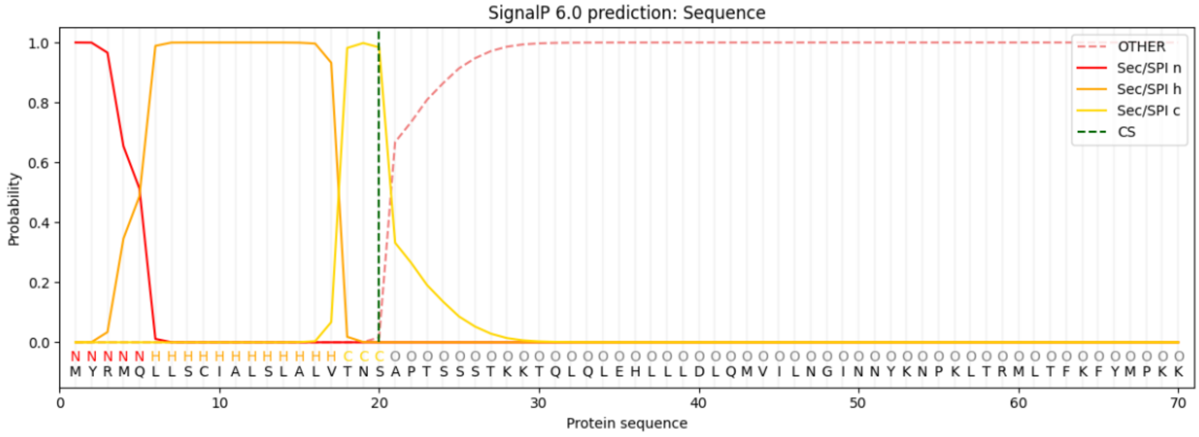
Output :



Sequence
Prediction: Signal Peptide (Sec/SPI)

Cleavage site between pos. 20 and 21. Probability 0.983948

Protein type	Other	Signal Peptide (Sec/SPI)	Lipoprotein signal peptide (Sec/SPII)	TAT signal peptide (Tat/SPI)	TAT Lipoprotein signal peptide (Tat/SPII)	Pilin-like signal peptide (Sec/SPIII)
Likelihood	0.0002	0.9992	0.0001	0.0001	0.0001	0.0001



Inference :

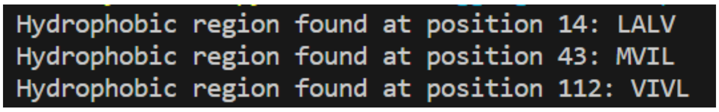
UniProt annotates the protein as a precursor with an N-terminal **signal peptide spanning residues 1–20**; the **mature protein starts from residue 21 till 153**. SignalP 6.0 predicts a classical Sec-dependent signal peptide with a **cleavage site between residues 20,21 (probability 0.9839)**.

The UniProt annotation is consistent with SignalP 6.0 predictions, confirming that the identified signal peptide is the native human signal peptide.

3. **Aggregation Propensity:Implement a function to flag regions with high local hydrophobicity (e.g., >4 consecutive hydrophobic residues).**

Code logic : The code finds >4 consecutive hydrophobic residues in the provided sequence.

Output :



```
Hydrophobic region found at position 14: LALV
Hydrophobic region found at position 43: MVIL
Hydrophobic region found at position 112: VIVL
```

Inference :

The sequence has three regions of >4 consecutive hydrophobic regions that will promote intermolecular aggregation.
The hydrophobic stretch corresponds to the N-terminal signal peptide(at 14) and is expected to be cleaved during secretion. Any additional hydrophobic clusters(at 43,112) within the mature protein would represent potential aggregation hotspots and can be handled with rational mutations to improve solubility and recombinant expression.

4. **Structural Evaluation: Provide a predicted 3D model (using AlphaFold2 or homology modeling). Identify surface-exposed hydrophobic patches that may contribute to aggregation**.

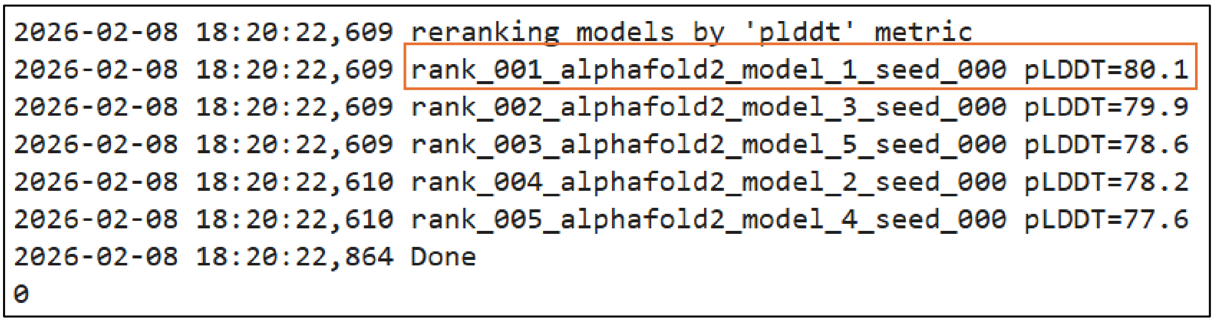
For 3D model prediction, both AlphaFold2 and Modeller were used and the final outputs were superimposed.

- Part 1 : Predicted 3D model from **AlphaFold2**

Step 1: The 3D model of the provided sequence was predicted using the AlphaFold2 ran on ColabFold. (parameters mentioned in attached code)

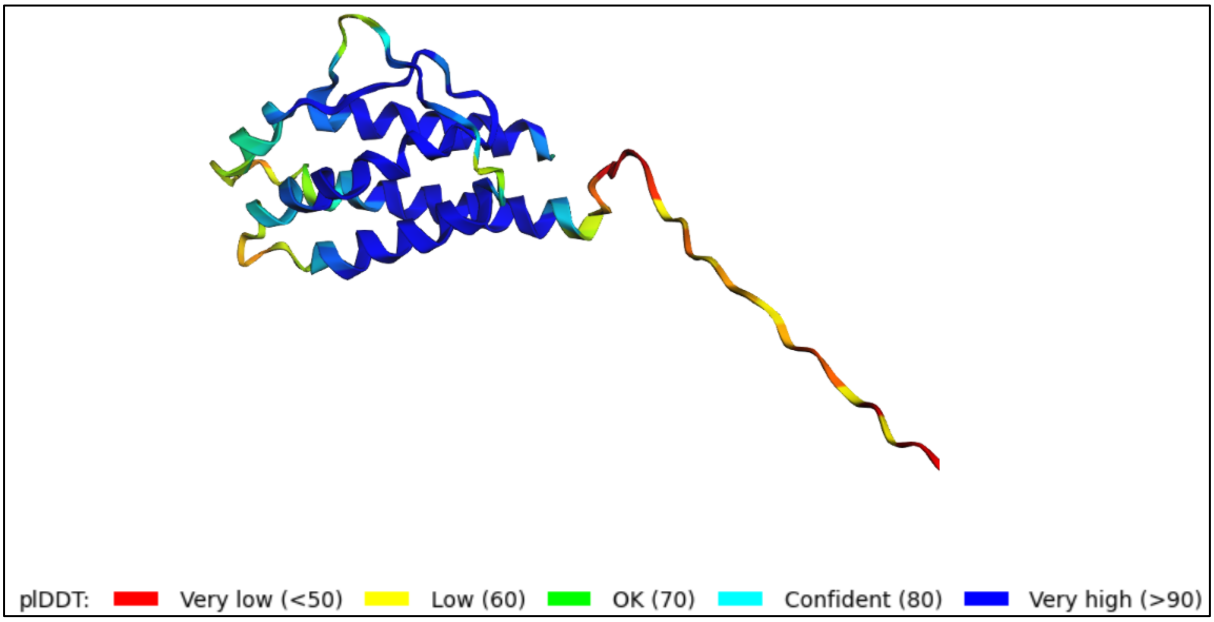
Output

The tool generated 5 models out of which the model 1 had highest residue-level confidence score.



```
2026-02-08 18:20:22,609 reranking models by 'plddt' metric
2026-02-08 18:20:22,609 rank_001_alphafold2_model_1_seed_000 pLDDT=80.1
2026-02-08 18:20:22,609 rank_002_alphafold2_model_3_seed_000 pLDDT=79.9
2026-02-08 18:20:22,609 rank_003_alphafold2_model_5_seed_000 pLDDT=78.6
2026-02-08 18:20:22,610 rank_004_alphafold2_model_2_seed_000 pLDDT=78.2
2026-02-08 18:20:22,610 rank_005_alphafold2_model_4_seed_000 pLDDT=77.6
2026-02-08 18:20:22,864 Done
0
```

Reporting the **Model 1 having 80.1%** pLDDT score.



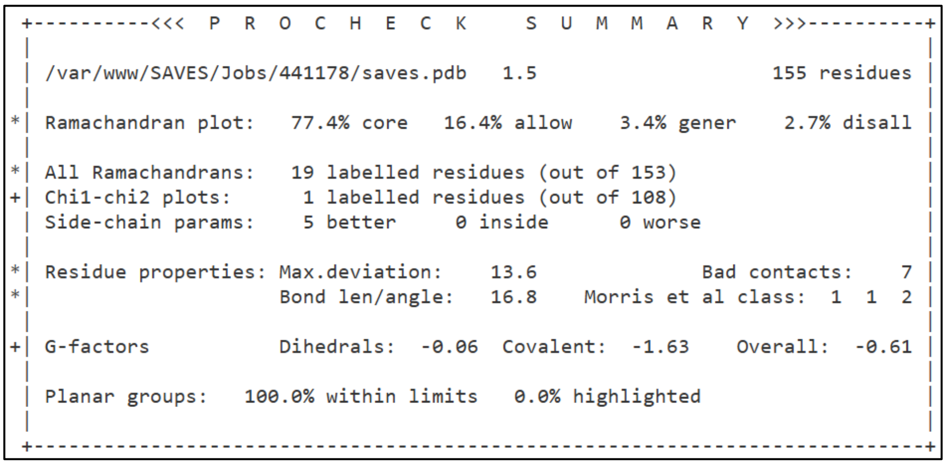
Model 1 predicted structure

Inference - The predicted model provided residue-level confidence scores (pLDDT), which were used to assess structural reliability and identify surface-exposed regions which was **80% which indicates a correct backbone prediction** with misplacement of some side chains. (source)

(pLDDT is a per-residue confidence score from AlphaFold2 that tells how confident the model is about the local structure around each residue.)

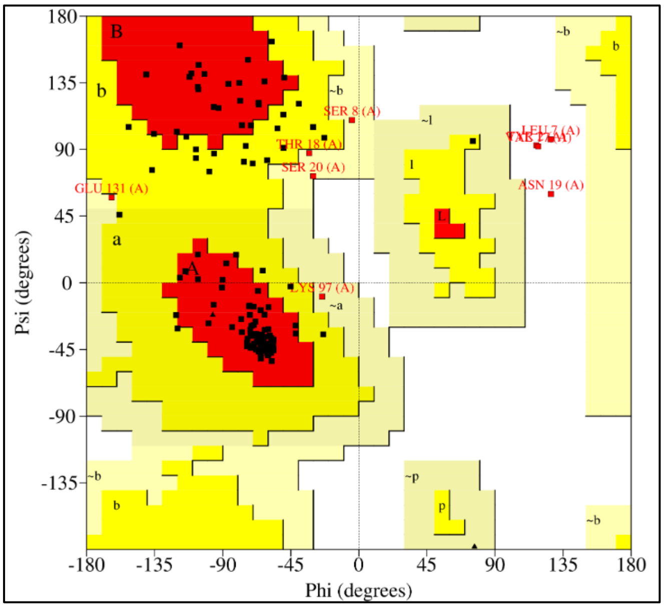
Step 2 : Checking stability of the AlphaFold2 predicted structure using PROCHECK tool

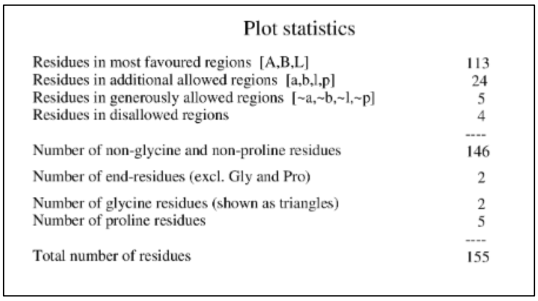
Overall statistics



```
 +----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
 |                                                                             |
 | /var/www/SAVES/Jobs/441178/saves.pdb   1.5                    155 residues |
 |                                                                             |
*| Ramachandran plot:   77.4% core   16.4% allow    3.4% gener    2.7% disall |
 |                                                                             |
*| All Ramachandrans:   19 labelled residues (out of 153)                     |
+| Chi1-chi2 plots:      1 labelled residues (out of 108)                     |
 | Side-chain params:    5 better      0 inside      0 worse                  |
 |                                                                             |
*| Residue properties: Max.deviation:    13.6            Bad contacts:    7   |
*|                     Bond len/angle:   16.8    Morris et al class:  1  1  2 |
 |                                                                             |
+| G-factors           Dihedrals:  -0.06  Covalent:  -1.63    Overall:  -0.61 |
 |                                                                             |
 | Planar groups:   100.0% within limits   0.0% highlighted                   |
 |                                                                             |
 +-----------------------------------------------------------------------------+
```

Ramachandran Plot





Plot statistics

Residues in most favoured regions [A,B,L]	113
Residues in additional allowed regions [a,b,l,p]	24
Residues in generously allowed regions [~a,~b,~l,~p]	5
Residues in disallowed regions	4

Number of non-glycine and non-proline residues	146
Number of end-residues (excl. Gly and Pro)	2
Number of glycine residues (shown as triangles)	2
Number of proline residues	5

Total number of residues	155

Inference

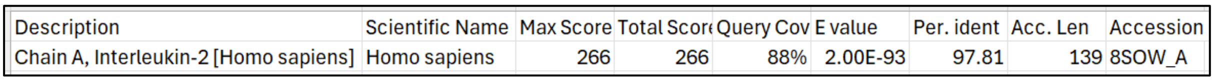
- Ramachandran plot shows that most residues occupy favoured or allowed regions, with only a few disallowed residues likely arising from flexible regions and the overall G-factor (−0.61) is within an acceptable range for a computational model, supporting the structure’s suitability for downstream analysis while highlighting limited regions for potential refinement.

- Part 2 : Predicted 3D model from **homology modelling using Modeller**

Tool used - https://salilab.org/modeller/

Steps 1 : Running BLASTp using the provided sequence on PDB database and aligning with the sequences after that.

Result :



Description	Scientific Name	Max Score	Total Scor	Query Cov	E value	Per. ident	Acc. Len	Accession
Chain A, Interleukin-2 [Homo sapiens]	Homo sapiens	266	266	88%	2.00E-93	97.81	139	8SOW_A

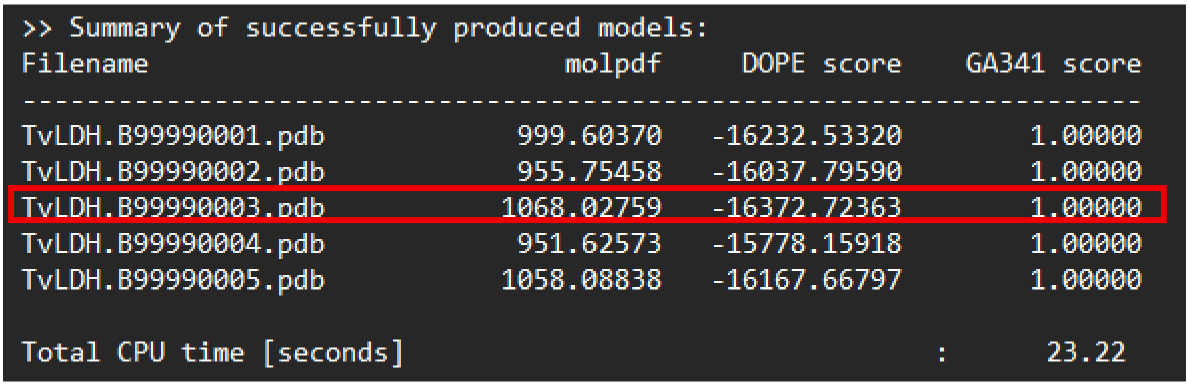
The chain A of the PDB accession 8SOW should the highest statistical match with the input sequence, it belongs to Chain A, Interleukin-2 [Homo sapiens].

The sequences were then aligned using the align2d.py code.(Log and code are attached)

Step2. Modelling the sequences

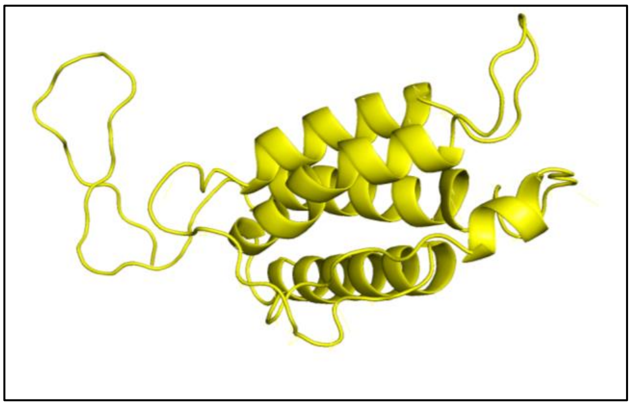
Output

The tool generated 5 models out of which the model3 had highest negative DOPE score.



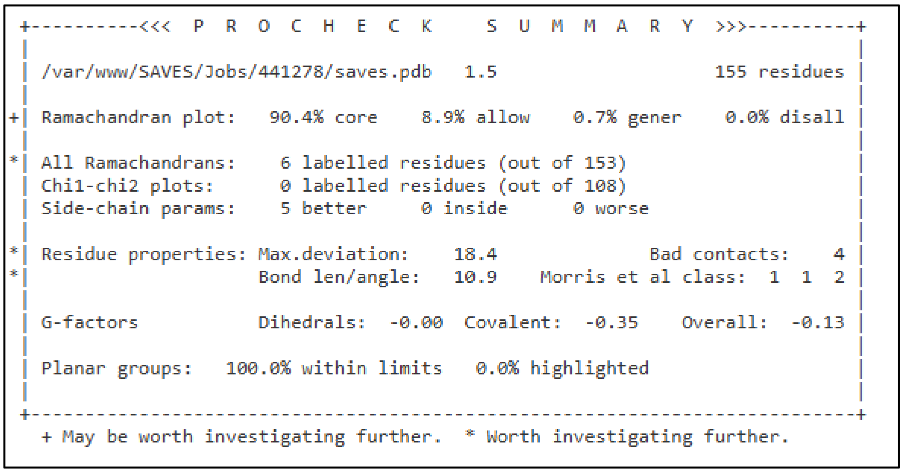
```
>> Summary of successfully produced models:
Filename                          molpdf     DOPE score    GA341 score
----------------------------------------------------------------------
TvLDH.B99990001.pdb            999.60370   -16232.53320        1.00000
TvLDH.B99990002.pdb            955.75458   -16037.79590        1.00000
TvLDH.B99990003.pdb           1068.02759   -16372.72363        1.00000
TvLDH.B99990004.pdb            951.62573   -15778.15918        1.00000
TvLDH.B99990005.pdb           1058.08838   -16167.66797        1.00000

Total CPU time [seconds]                                 :       23.22
```

Reporting model 3 image below :



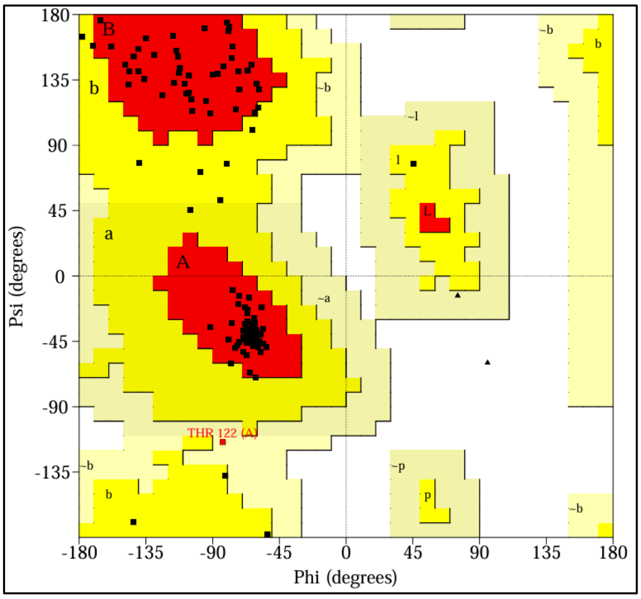
Step 2 : Checking stability of the AlphaFold2 predicted structure using PROCHECK tool

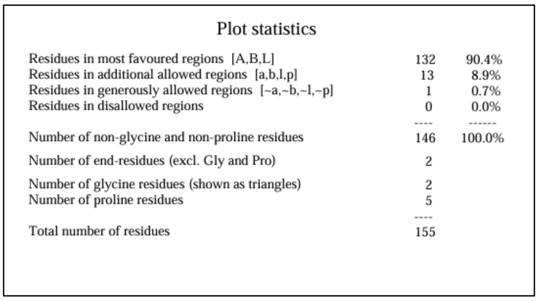
Overall statistics



```
+----------<<<  P  R  O  C  H  E  C  K     S  U  M  M  A  R  Y  >>>----------+
 |                                                                             |
 | /var/www/SAVES/Jobs/441278/saves.pdb   1.5                   155 residues |
 |                                                                             |
+| Ramachandran plot:   90.4% core    8.9% allow    0.7% gener    0.0% disall |
 |                                                                             |
*| All Ramachandrans:    6 labelled residues (out of 153)                     |
 | Chi1-chi2 plots:      0 labelled residues (out of 108)                     |
 | Side-chain params:    5 better      0 inside       0 worse                 |
 |                                                                             |
*| Residue properties: Max.deviation:    18.4             Bad contacts:     4 |
*|                     Bond len/angle:   10.9    Morris et al class:  1  1  2 |
 |                                                                             |
 | G-factors           Dihedrals:  -0.00  Covalent:  -0.35     Overall:  -0.13 |
 |                                                                             |
 | Planar groups:   100.0% within limits   0.0% highlighted                   |
 |                                                                             |
+-----------------------------------------------------------------------------+
   + May be worth investigating further.  * Worth investigating further.
```

Ramachandran Plot



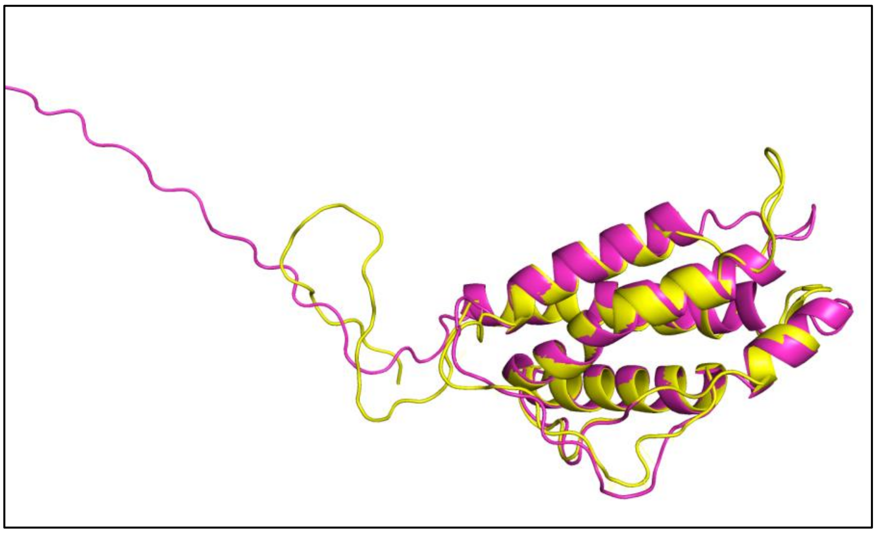


```
                              Plot statistics

Residues in most favoured regions  [A,B,L]              132      90.4%
Residues in additional allowed regions  [a,b,l,p]        13       8.9%
Residues in generously allowed regions  [~a,~b,~l,~p]     1       0.7%
Residues in disallowed regions                            0       0.0%
                                                        ----    ------
Number of non-glycine and non-proline residues          146     100.0%

Number of end-residues (excl. Gly and Pro)                2

Number of glycine residues (shown as triangles)           2
Number of proline residues                                5
                                                        ----
Total number of residues                                155
```

- Ramachandran plot shows that 90% residues occupy favoured or allowed regions and the overall G-factor (−0.13) is within an acceptable range for a computational model, supporting the structure’s suitability for downstream analysis while highlighting limited regions for potential refinement.

Part 3 : Super-imposing the Alphafold2 and Modeller predicted models

Superimposed model image



Calculated RMSD : **0.862** Å (from PyMol)

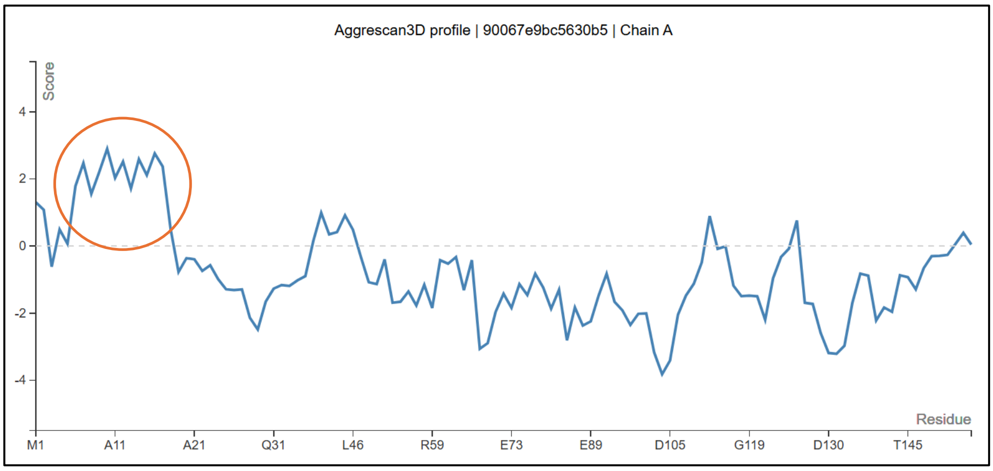
RMSD (Root Mean Square Deviation) measures the average distance between the atoms of two superimposed proteins.

Inference : An RMSD value below 1.0 Å indicates excellent structural agreement. The close alignment between the MODELLER and AlphaFold2 models (**RMSD = 0.82 Å**) demonstrates near-identical conformations, providing strong confidence in the predicted structure and confirming the reliability of the modelling results.

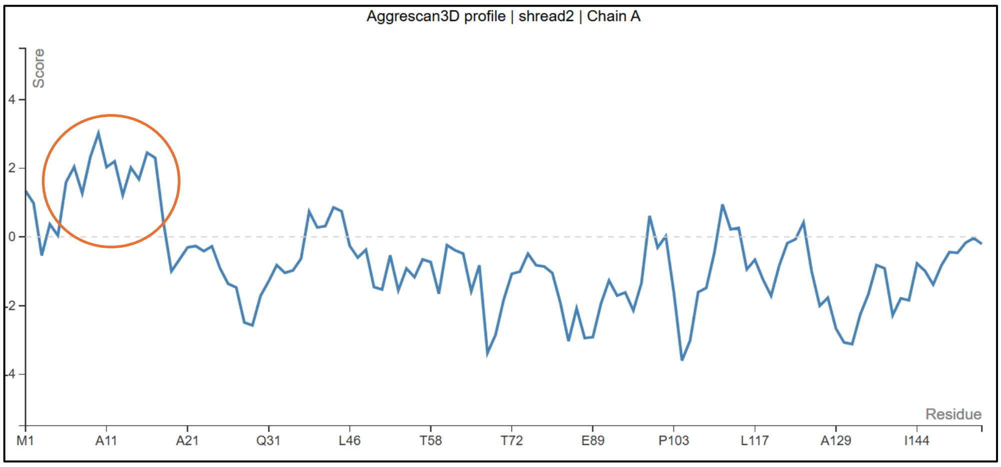
5. **Identification of surface-exposed hydrophobic patches that may contribute to aggregation.**

- **Tool 1 – Aggrescan3D**
  It predicts of aggregation propensity in predicted protein structures (3D input)

AlphaFold2 predicted model



Homology modelling predicted model(Modeller)



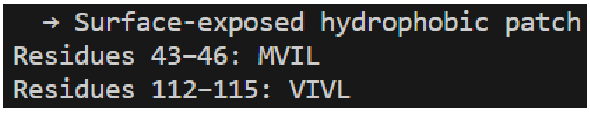
Inference - Aggrescan3D analysis revealed **a strong aggregation signal at the N-terminus corresponding to the signal peptide**, which will be removed during secretion, along with a few moderate aggregation peaks downstream in the mature sequence. These regions are largely surface-exposed and limited in extent.

Also, the aggregation profiles are highly similar for both predicted structural models, adding confidence to the robustness and consistency of the prediction.

- **Tool 2 – NetSurf + Python code for Aggregation Propensity**

Code logic - finds hydrophobic patches on sequence (>4 consecutive hydrophobic residues) -> checks which ones are surface-exposed using NetSurfP RSA (previously calculated) -> Final flagged regions => aggregation-prone surface patches

Code output



```
  → Surface-exposed hydrophobic patch
Residues 43-46: MVIL
Residues 112-115: VIVL
```

Inference - Hydrophobic regions were identified using a sliding-window sequence analysis (>4 consecutive hydrophobic residues) and subsequently filtered using NetSurfP-predicted relative solvent accessibility (RSA ≥ 0.25 in 3 out 4 residues) to identify surface-exposed aggregation-prone patches.

6. **Purification Recommendation: Suggest a suitable buffer (including pH) and a chromatography method for purification**.

Since the suggested fusion tags for an efficient purification for BEV of human protein in insect cells are His$_6$ tag and Strep-tag II, here are their suitable buffers.

7. **His$_6$ tag** purification: needs IMAC (Ni^{2+}/Co^{2+} resin) with 20–50 mM phosphate buffer, pH 7.5–8.0, 300 mM NaCl; elute with 250–300 mM imidazole.
8. **Strep-tag II** purification: Use Strep-Tactin affinity chromatography with 100 mM Tris-HCl, pH 7.5–8.0, 150 mM NaCl; elute gently with 2.5 mM desthiobiotin

Reference

Bornhorst & Falke, *Methods in Enzymology* (2000); Kost et al., *Nature Biotechnology* (2005); Schmidt et al., *Analytical Biochemistry* (2013); Hitchman et al., *Biotechnology Advances* (2010)

OVERALL CONCLUSION

The target sequence was validated against reference databases, confirming a reliable starting point for engineering. Codon optimization for *Spodoptera frugiperda* cells improved translational efficiency while avoiding rare codons and unfavorable sequence features.

Rational analysis identified limited aggregation-prone regions, enabling conservative point mutations to enhance solubility without affecting structure or function.
An insect-cell expression strategy was designed using a gp67 signal peptide for efficient secretion and affinity tags for robust purification. Python-based computational pipelines automated physicochemical profiling, signal peptide identification and mature sequence extraction, aggregation hotspot detection, and structural evaluation using AlphaFold2.

Structural reliability is supported by AlphaFold2 and homology modelling, with acceptable stereochemical quality from PROCHECK and limited aggregation-prone surface regions identified by NetSurfP and Aggrescan.

Overall, the structure is suitable for downstream analysis and rational engineering for expressing the Target-X using a Baculovirus Expression Vector System (BEVS) in Sf9/Sf21insect cells, with only minor local regions potentially benefiting from refinement or solubility-enhancing mutations.