CHIMERA_AA: Feature Extraction Guide

CHIMERA_AA feature extraction utilizes the Biopython library for extracting features of protein chains. The workflow of the CHIMERA_AA feature extraction Google Colab is represented in **Figure 1**. The link to Google Colab for feature extraction is (https://colab.research.google.com/drive/1UWqETo1c1eX6uXWxUn3eIY5Q0Gj2UsoJ?usp=sharing). The list of features included in the feature extraction table is described below.

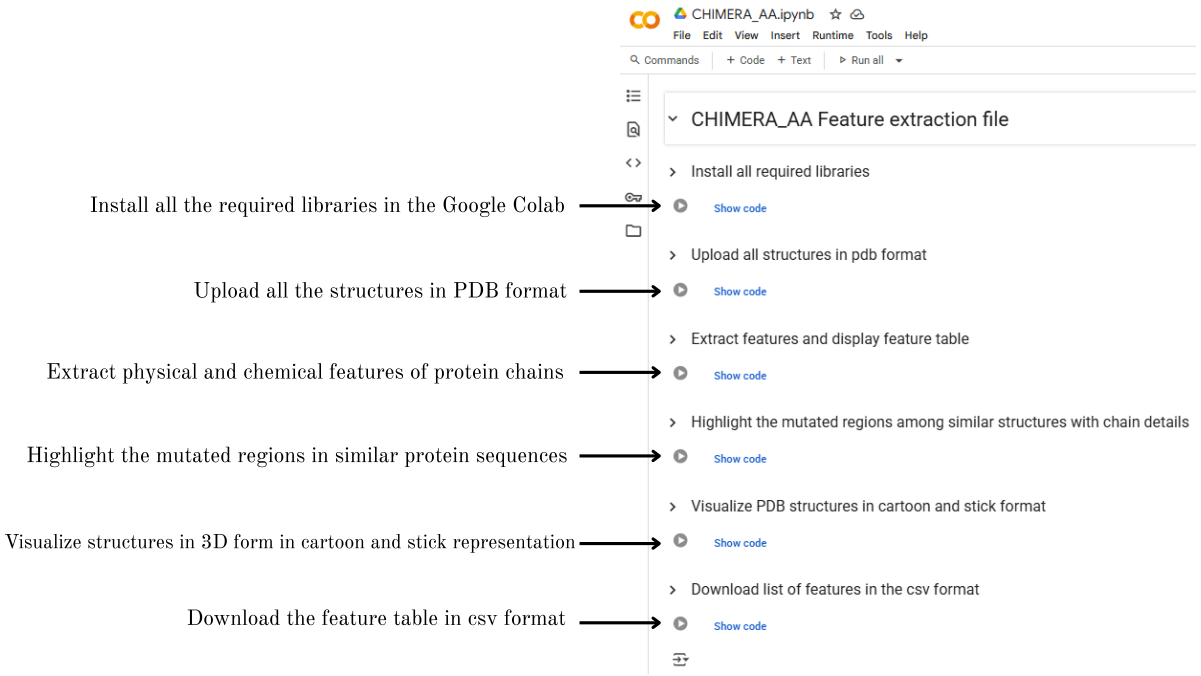


Figure 1. Representation of Google Colab for protein feature extraction

1. Protein Sequence

The primary sequence of a protein refers to its specific linear arrangement of amino acids. This sequence ultimately dictates the 3D structure of the protein through the process of folding.

2. Sequence Length

The total number of amino acids present in the protein chain.

3. Isoelectric Point (pI)

The isoelectric point is the pH at which the protein neutralizes with no net electric charge. This affects solubility, as proteins tend to aggregate or precipitate at their pI.

4. Hydropathy Index (GRAVY Score)

The hydropathy index indicates the average hydrophobicity or hydrophilicity of the amino acids in the protein. A positive GRAVY score suggests a hydrophobic (likely membrane-bound) protein, whereas a negative score suggests a soluble, hydrophilic protein.

5. Molecular Weight

Molecular weight is the sum of the atomic masses of all atoms in the protein

6. Average B-Factor

The B-factor, or temperature factor, is derived from X-ray crystallography and reflects atomic displacement or flexibility. A high average B-factor suggests that atoms within the protein have more movement, which could indicate intrinsic disorder or structural flexibility.

7. Standard Deviation of B-Factor

Standard deviation of the B-factor measures the variability in atomic motion across the protein. A high standard deviation implies that some regions are highly flexible while others are rigid. Such variation may be important for proteins that undergo conformational changes.

8. Fraction of Helices

This feature refers to the percentage of residues forming α-helices, a common secondary structural motif. Helices provide rigidity and often span membranes or stabilize the core of globular proteins.

9. Fraction of Sheets

This corresponds to the proportion of β-sheet content in the protein chains. β-sheets provide structural integrity and are common in the protein core.

10. Fraction of Turns

Turns and loops connect secondary structure elements and often reside on the protein surface, participating in molecular recognition. Their flexibility allows conformational changes.

11. Percentage of Hydrophobic Residues

Hydrophobic residues like leucine and valine usually reside in the interior of the protein, helping maintain the folded structure via hydrophobic interactions. A high percentage of hydrophobic residues often correlates with a stable core. However, hydrophobic patches on the surface may promote aggregation.

12. Percentage of Polar Residues

Polar residues such as serine and threonine are typically exposed on the surface and involved in hydrogen bonding or solvent interactions. They contribute to solubility and structural specificity.

13. Percentage of Aromatic Residues

Aromatic residues like tryptophan, tyrosine, and phenylalanine are involved in π-π stacking, stabilization of hydrophobic cores, and recognition interfaces. These residues are important in enzyme active sites and ligand binding.

14. Percentage of Charged (Acidic + Basic) Residues

Charged residues (Asp, Glu, Lys, Arg) form salt bridges and contribute to electrostatic interactions, which are vital for stability and protein–protein binding. A balanced charged residue content is important for solubility and pH-dependent behavior.

Demonstration

For demonstration of the feature extraction Colab, we utilized the PDB structure of the thermostable subdomain from the chicken villin headpiece, NMR, and minimized average structure (PDB ID: 1VII). Using CHIMERA_AA, the PDB structure was mutated at 2 locations, residue IDs 41 and 42 of chain A, with alanine and valine, respectively. The mutated structure file was created in the PDB format (ALAL41VAL42.pdb). The mutated structure was then minimized using the *minimize.sh* shell script and saved in PDB format (ALAL41VAL42_min.pdb). The feature extraction Colab was used to compare and obtain features of all three PDB structures.

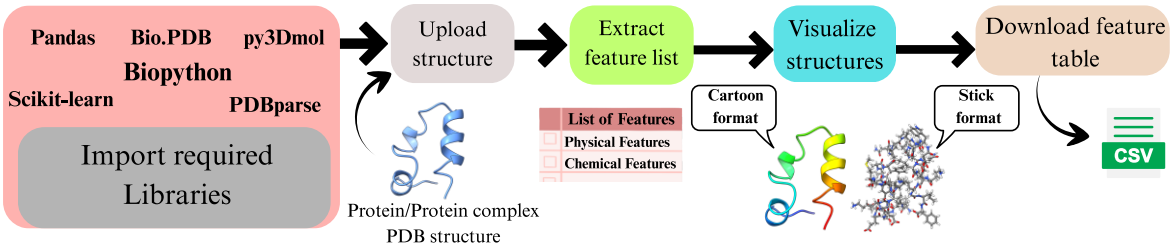


Figure 2. Schematic representation of feature extraction Google Colab

Flowchart

This section demonstrates the use of Google Colab stepwise to obtain results. The user needs to execute each tab individually in the series and obtain results. The Colab snippets, along with the output, are represented further.

1. Install all required libraries and dependencies.

> Install all required libraries

[1] Show code

```
Collecting biopython
  Downloading biopython-1.85-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (13 kB)
Requirement already satisfied: termcolor in /usr/local/lib/python3.11/dist-packages (3.1.0)
Collecting py3Dmol
  Downloading py3dmol-2.5.1-py2.py3-none-any.whl.metadata (2.1 kB)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from biopython) (2.0.2)
Downloading biopython-1.85-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.3/3.3 MB 55.3 MB/s eta 0:00:00
Downloading py3dmol-2.5.1-py2.py3-none-any.whl (7.2 kB)
Installing collected packages: py3Dmol, biopython
Successfully installed biopython-1.85 py3Dmol-2.5.1
Requirement already satisfied: biopython in /usr/local/lib/python3.11/dist-packages (1.85)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (1.6.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from biopython) (2.0.2)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.15.3)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.5.1)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.6.0)
```

2. Upload the reference and mutated structures for feature extraction

› Upload all structures in pdb format

[2] Show code

Choose Files 3 files

- **1vii.pdb**(n/a) - 62694 bytes, last modified: 7/15/2025 - 100% done
- **ALA41VAL42.pdb**(n/a) - 60051 bytes, last modified: 7/15/2025 - 100% done
- **ALA41VAL42_min.pdb**(n/a) - 61251 bytes, last modified: 7/15/2025 - 100% done

```
Saving 1vii.pdb to 1vii.pdb
Saving ALA41VAL42.pdb to ALA41VAL42.pdb
Saving ALA41VAL42_min.pdb to ALA41VAL42_min.pdb
```

3. Extract features and create a dynamic table of all the features

› Extract features and display feature table

Show code

```
Processed 1vii.pdb
Processed ALA41VAL42.pdb
Processed ALA41VAL42_min.pdb
```

	Filename	Chain_ID	Sequence	Sequence_Length	Isoelectric_Point	Hydropa
0	1vii.pdb	A	MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF	36	9.4	
1	ALA41VAL42.pdb	A	AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF	36	9.4	
2	ALA41VAL42_min.pdb	A	AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF	36	9.4	

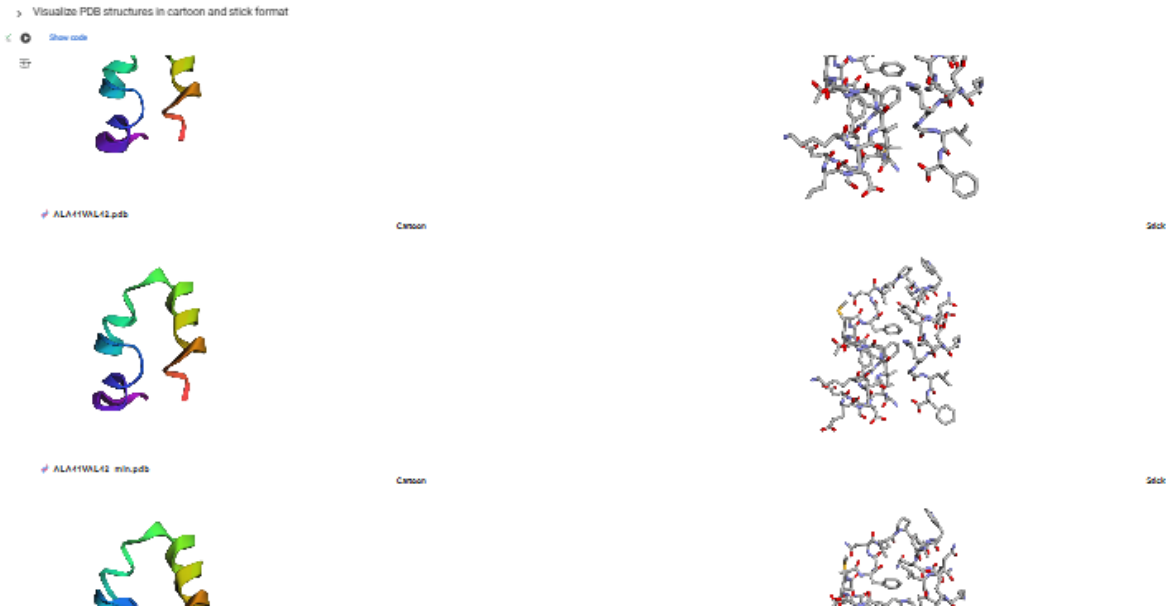
4. Mutated residues in the structures were highlighted to understand the change of residues.

› Highlight the mutated regions among similar structures with chain details

[4] Show code

```
1vii.pdb [Chain: A]           MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF
ALA41VAL42.pdb [Chain: A]     AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF
ALA41VAL42_min.pdb [Chain: A] AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF
```

5. Visualize structures in 3D form in cartoon and stick format



6. Download the feature table in CSV format.
   After complete execution, the following CSV format file will be downloaded. **Table 1** represents the complete generated feature table with features for all uploaded structures.

Table 1. List of different physical and chemical features of initial protein structure (1VII), mutated structure (ALA41VAL42), and minimized structure (ALA41VAL42_min) from feature extraction Google Colab notebook.

Filename	**1vii**	**ALA41VAL42**	**ALA41VAL42_min**
Chain_ID	A	A	A
Sequence	MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF	AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF	AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF
Sequence_Length	36	36	36
Isoelectric_Point	9.4	9.4	9.4
Hydropathy_Index	-0.39	-0.38	-0.38
Molecular_Weight	4191.93	4102.67	4117.79
Average_B_Factor	1.39	1.34	1.3
Std_B_Factor	1.04	1.03	1.04
Helix_Fraction	0.47	0.44	0.44
Sheet_Fraction	0.25	0.25	0.25
Turn_Fraction	0.33	0.33	0.33
Hydrophobic_Residues_Percent	38.89	38.89	38.89
Polar_Residues_Percent	19.44	19.44	19.44
Aromatic_Residues_Percent	13.89	13.89	13.89
Acidic_Residues_Percent	11.11	11.11	11.11
Basic_Residues_Percent	16.67	16.67	16.67

Analysis

With the generated feature table (Table 1), a detailed comparative analysis of the initial (1vii), mutated (ALA41VAL42), and minimized (ALA41VAL42_min) structures of the thermostable subdomain from the chicken villin headpiece may be conducted. These features are helpful to deduce a change of features, a single protein change on introduction of a mutation, and minimization of the mutated structure.

1. Sequence Analysis

Initial Sequence: MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF

Mutated & Minimized Sequence:
AVSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF

Observation: The first two residues changed from M (methionine) and L (leucine) to A (alanine) and V (valine). These substitutions replace bulkier and more hydrophobic residues with smaller ones. This might affect N-terminal interactions and folding dynamics slightly.

2. Sequence Length

All Structures: 36 residues

Observation: Since there is no change in the number of residues, this signifies that there is no insertion or deletion of any residue in the protein structure.

3. Isoelectric Point (pI)

All Structures: 9.4

Observation: The overall charge distribution remains unchanged throughout the process of mutation and minimization. This signifies that no charged amino acids were involved during the process of mutation.

4. Hydropathy Index (GRAVY)

Initial structure: -0.39

Mutated structure: -0.38

Minimized structure: -0.38

Interpretation: The structure shows slightly less hydrophilicity after mutation due to substitution of methionine (GRAVY: 1.9) and leucine (3.8) with alanine (1.8) and valine (4.2). The change is minor, which suggests that the overall solubility and hydrophobicity are preserved.

5. Molecular Weight (Da)

Initial structure: 4191.93 Da

Mutated structure: 4102.67 Da

Minimized structure: 4117.79 Da

Interpretation: The decrease in molecular weight is consistent with substituting heavier methionine and leucine (149.21 + 131.17) with lighter alanine and valine (89.09 + 117.15).

The slight increase after minimization (4117.79) suggests the addition of missing atoms or coordinate adjustments, possibly from hydrogen addition during structure optimization.

6. Average B-Factor

Initial structure: 1.39

Mutated structure: 1.34

Minimized structure: 1.30

Interpretation: B-factor reflects atomic flexibility. The slight decrease after mutation and minimization suggests that the protein has become slightly more rigid, likely due to less flexible side chains at the N-terminal or improved packing from minimization.

7. Standard Deviation of B-Factor

All Structures: ~1.03–1.04

No significant difference in structures implies that the overall flexibility across the protein has not changed, even after mutations.

8. Secondary Structure Fractions

Helix Fraction:

Initial: 0.47

Mutated/Minimized: 0.44

Sheet Fraction:

All: 0.25

Turn Fraction:

All: 0.33

Interpretation: A slight reduction in helical content after mutation may be due to N-terminal substitution disrupting local helical initiation. Since the sheets and turns remain constant, it indicates that the core structure is largely unaffected.

9. Hydrophobic Residue Percent

All: 38.89%

Observation: Since we know that the substitution occurred from 2 hydrophobic amino acids with the other two hydrophobic residues and the overall content of hydrophobic residues remains unchanged despite the substitution, this confirms that the change was within the hydrophobic class.

10. Polar Residue Percent

All: 19.44%

Observation: Since there is no change in the overall content of polar residue percentage, which suggests that the mutation did not introduce/remove polar residues.

11. Aromatic Residue Percent

All: 13.89%

Observation: Since there is no change in the overall content of aromatic residue percentage, which suggests that the mutation did not introduce/remove aromatic residues.

12. Acidic & Basic Residue Percent

Acidic: 11.11%

Basic: 16.67%

Observation: Since all variants are constant, this indicates that electrostatic properties were preserved even after substitution.

Implication for Structural Stability

The slight decrease in B-factors and reduction in helix fraction suggest that the N-terminal mutation slightly alters flexibility and helical tendency, but not dramatically. The molecular weight and hydropathy are moderately altered, and electrostatic and structural features remain stable, implying that the mutation is structurally conservative.

References

1. Scott, L. R., & Fernández, A. (2017). Protein Basics. In *A Mathematical Approach to Protein Biophysics* (pp. 47-64). Cham: Springer International Publishing.
2. Schulz, G. E., & Schirmer, R. H. (2013). *Principles of protein structure*. Springer Science & Business Media.
3. Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, *20*(2), 15-19.
4. Sillero, A., & Ribeiro, J. M. (1989). Isoelectric points of proteins: theoretical determination. *Analytical biochemistry*, *179*(2), 319-325.
5. Yuan, Z., Bailey, T. L., & Teasdale, R. D. (2005). Prediction of protein B-factor profiles. *Proteins: Structure, Function, and Bioinformatics*, *58*(4), 905-912.
6. McKnight, C. J., Matsudaira, P. T., & Kim, P. S. (1997). NMR structure of the 35-residue villin headpiece subdomain. *Nature structural biology*, *4*(3), 180-184.