

## Task Description

Proteins are large, complex molecules composed of one or more chains of amino acids, known as residues. There are 20 standard types of residues (amino acids) that make up proteins, each with unique chemical properties influencing the protein's structure and function. Proteins can vary in size, containing anywhere from a few dozen to several thousand residues. A typical protein sequence can be represented as a string of single-letter codes for each residue, such as:

Example protein sequence: “MKTFFVFTTQRYDEQELFFQN” (M, K, T, ... are the protein's residues)

Residue-residue contacts refer to pairs of residues that are spatially close within the 3D structure of the protein. In this task, two residues are considered in contact if the distance between their C $\alpha$  atoms is below 8 Å (angstroms). Identifying these contacts is crucial for understanding protein folding and interactions, as they provide insights into the overall structure and function of the molecule.

You are tasked with developing a machine learning model for predicting residue-residue contacts within protein sequences. This model will be implemented in Python and trained using an open-source protein database that provides sequence and 3D structural data. The goal is to predict which residues in a protein sequence come into contact, without explicitly predicting the full 3D configuration.

This model should build upon the ESM2 architecture, which already provides a contact prediction head but relies solely on input sequences. The main goal of this task is to re-use ESM2 embeddings and additionally incorporate structural data from proteins with sequences similar to the input sequence, enhancing the prediction performance.

## Task Details

**Objective:** Create a new residue contact prediction model that leverages both the ESM2 embeddings of the input sequence residues and structural data from similar sequences.

### Implementation:

- Implement your model in Python, modifying or extending the ESM2 model to incorporate additional structural information. The input of the model is a single sequence, the output is a binary contact map (1 - if there is a contact between the residue pair, 0 - if there is no contact).
- Use an open-source protein database, such as the Protein Data Bank (PDB), to access sequences and their 3D structural data for training and evaluation.
- Ensure that the dataset is split into training, validation, and test sets, with a proper handling of sequence similarity to avoid train-test data leakage.

**Deliverables:**

- Submit your code as a ZIP file or upload it to a public GitHub repository.
- Include a concise paper that outlines:
  - A detailed description of your method and how it extends ESM2.
  - The data split strategy used for training, validation, and testing.
  - Hyperparameters and other relevant details of the ML setup.
  - Analysis of the results, including performance metrics and insights.

**Note:** You are encouraged to leverage open-source resources and appropriately cite any external code or libraries used. Ensure originality in your approach and final implementation.

**Evaluation Criteria**

- Code quality
- ML training and evaluation setup, data handling
- Model performance
- Results analysis
- Method description