**Deep Learning Scientist Assignment for Basecamp Research Selective Process**

Introduction

Basecamp research is a biotech start-up aimed at providing biodiversity guardians with a fair and equitable way to share their environmental genomic data with biotechnology innovators.

The main objective of our team is to ethically build a global collection of unique information that can be used to drive new product innovation. We do this by accessing the most diverse set of novel sequences for a vast array of protein families and use this insight to steer complex function predictions with graph learning and AI methods identifying sequences most relevant for substrate class and reaction coordinates.

Data

You are provided with the following data:

**Sequences.fasta** - contains 1000 protein sequences with a header preceding each sequence which denotes the number of the sequence and the group it belongs to.

**Predictions.fasta** - contains protein sequences you will need to make a prediction on.

Goal

**Make a prediction for each sequence in the Predictions.fasta file predicting which group each sequence belongs to.**

You are required to explore the data, state your hypothesis and make a prediction on the dataset provided. Present your findings in a clear and concise report. What are the limitations of your analysis?