# Research Engineer Intern - Computational Biology Technical Test

As a research engineer intern you will have the opportunity to work on biology related projects. Proteins are ubiquitous in our day-to-day work, hence this technical test aims at introducing you to this type of data.

We recommend that you spend no more than 5 hours on this test.

## Problem Setting

The goal of this test is to build a protein classifier: for each protein you have to assign the corresponding Pfam family (i.e. protein family). You can find more information regarding the Pfam family here.

## Data

The dataset to use is hosted on Kaggle: Pfam seed random split - Using Deep learning to Annotate the Protein Universe

## Deliverable

Your deliverable should be divided into 2 main parts:

1. Dataset Analysis
2. Classifier

The format of your deliverable is up to you (script, jupyter, pdf report, private GitHub repository, ...). If you choose a pdf report you must also provide the code used for the project.

You are free to use any machine learning / deep learning framework.

Requirements:

- Python 3.6+
- Easily reproducible on a laptop with 16GB of RAM + 4GB GPU

## Compute

In case you need more compute power than locally available on your computer, the following resources provide interesting amounts of computing power for free:

- Google Colab: access to one GPU or one TPU, time limit of 12 hours (kernels are shut down after 12 hours)
- Kaggle notebooks: access to one GPU (NVIDIA P100), time limit of 6 hours

## Evaluation

- You won’t be evaluated on the final performance of your classifier but rather on the methodology you used to tackle this task, so make sure to explain each step.
- We will pay attention to the code quality and the documentation.
- You will be evaluated on your capacity to communicate the results of your work both verbally and in writing to a technical audience.

Hope you have fun!

Please feel free to contact us if you have any questions, we’ll be happy to help.