Learning the Language of Proteins

## Supporting Org Information

 

| Company | Instadeep |
| --- | --- |
| Point(s) of Contact  (name(s), title(s), email(s)) | * [Thomas Pierrot](https://www.linkedin.com/in/thomas-pierrot-120a43128/), Research Lead, t.pierrot@instadeep.com * [Alain-Sam Cohen](https://www.linkedin.com/in/a2c/), Engineering Lead, as.cohen@instadeep.com |
| Location (city/time zone) | France & UK |

## 

[**Supporting Org Information**](#_1uamq249imw4) **1**

[**AI Product/Capstone Project Description**](#_4ha20x76zzk) **2**

[Problem](#_6vdkmoqff0lo) 2

[Why](#_oomvm8feyy9c) 2

[Success](#_zta80kbm5v7) 3

[Audience](#_h6pw71d4a1o4) 3

[What](#_d1mkjxch3qo0) 3

[Final Deliverables](#_u4s3d51gryb9) 4

[Anything Else?](#_qjw6h3i3ymk) 4

## AI Product/Capstone Project Description

### Problem

[Write a succinct statement of the problem that you're trying to solve (<50 words)]

| Drug development and discovery is a time and labor intensive process that has the potential to be enhanced and improved by next-generation protein sequencing techniques. |
| --- |

### Why

Write about why this is a problem worth solving. What is the business value hypothesis that connects to what success looks like and for whom? (~50-250 words)

| Recent breakthroughs in Natural Language Processing (NLP) and training of large transformer models have paved the way for new types of domain-specific deep language models. In the biological domain, development of general-purpose protein language models that are capable of predicting specific protein types could, among other things, pave the way to more effective, safer cancer treatments.  The total business value of AI-inspired protein-sequencing breakthroughs like this in the long term is not yet calculable. |
| --- |

### Success

Write about what success looks like. What is the Key Performance Indicator (or couple of KPIs)? How might they connect to a relevant ML model accuracy metric? (<50 words)

| This is a *research*-focused project focused on *natural language processing* and *binary classification*.  Successful projects will produce a generic Python code capable of assessing the performance of an protein language embedding generation technique on a series of three binary classification tasks, where pairs of inputs and labels are provided.  As a modeling-focused project, *the key performance indicator for this work is the ML model accuracy metric chosen for classification accuracy*. |
| --- |

### Audience

Specify exactly which users/customers this AI/ML product is being built for. What is the customer's pain or need that connects back to the problem? (<50 words)

| A functional AI product, once developed, would be used by pharmaceutical companies to accelerate drug development and improve R&D efficiency. |
| --- |

### What

Now describe what the ML looks like. This includes a discussion of data and sources, potential/likely models, a choice of an accuracy metric to optimize for and a defense of your choice. How does your accuracy metric connect back to the KPI(s) named above?

| Data   * Three datasets, all relatively small, will be provided as .csv files of amino acid sequences (of variable length) with labeled pairs of inputs and outputs. The outputs are 0 or 1. * Upon request, pre-computed hand-engineered features that have been selected by bio-experts are also available.   Modeling   * Implement one (or many) deep language models that have been *previously pre-trained on large datasets of proteins*. Note that the goal is not to pre-train models but rather to exploit existing ones and assess their performance on each task. * Comparison of classical approaches such as hand-engineering representations based on biological/statistical/chemical properties of the peptides is also welcomed. |
| --- |

You may also find it helpful to fill out an [MLOps Stack Canvas](https://ml-ops.org/content/mlops-stack-canvas) or [MLOps Stack Template](https://ml-ops.org/content/state-of-mlops). These tools really help to clarify tech stack requirements for our students.

### Final Deliverables

At the conclusion of a cohort, students are expected to deliver:

* Deployed Demo, where the provided solution satisfies the specified problem setting
* 10-minute Presentation
* GitHub Repo ( description in README + code )
  + The models used and links towards associated papers and repositories
  + The pooling technique used
  + The different classification techniques tested
  + Any baseline based on hand-engineered features implemented

For more detailed information on student capstone projects, you can check out the guide that we provide MLE students [here](https://docs.google.com/document/d/1lFRKgc9darivZaNaGoGni9Gch3hIuwnkN8AVql-kreU/edit?usp=sharing).

### Anything Else?

Please provide any additional information on key activities, technologies, datasets, expected learning outcomes, potential mentorship or employment opportunities, or anything else not listed above!. And thanks for supporting our students!

| **Tool Recommendations:**   * Object-oriented Python * Manipulating large quantities of data * Scikit-learn, TensorFlow, Keras, Jax, or PyTorch for ML/DL modeling * Experiment monitoring * Data visualization * [pytest](https://docs.pytest.org/en/7.1.x/) for testing * [BioPython](https://biopython.org/) to manipulate bio data   **Recommended Course of Action**   * This project has two parts. * First, we would like you to explore the literature and identify from 1 up to 3 deep protein language models that have been pre-trained and open-sourced and use them to compute embeddings for the provided datasets in the three tasks. * Second, we would like you to compare different classifiers that would take the pre-computed embeddings as inputs. Notably, we would like you to test non-supervised techniques such as clustering and dimensionality reduction techniques as well as supervised techniques from classical ML methods such as XGBoost to more modern and involved methods that would leverage neural networks. Any other original method, properly justified, would be very welcomed as well.   **Technical details:**   * You should identify a few pre-trained language models candidates for which the inference code and weights have been made open-source. A few good references to start with might be [Tristan Bepler and Bonnie Berger, 2021](https://www.cell.com/cell-systems/pdf/S2405-4712(21)00203-9.pdf), and [Rives et al., 2021](https://www.pnas.org/doi/10.1073/pnas.2016239118). * As these models are computationally expensive, you'll have to batch computations and use accelerators (GPUs/TPUs). We also expect you to identify the proper data format to store the embeddings of all the peptides present in the datasets for the three tasks. * You have the freedom to test any classification method you like. We would like to test a variety of different methods and discuss for each one, the motivation behind as well as the performance and limitations of your results.   **Key research questions and technological constraints that the project will answer:**   1. Are the existing pre-trained models from the literature general enough to provide good representations in these three very different tasks? 2. Do representations that emerge from unsupervised learning outperform systematically hand-engineered representations coming from expert knowledge? 3. Which existing model is the more appropriate? Which activation of the model should be used to extract the embeddings, the final or an intermediate one? What is the best technique to transform a vector of embeddings per position into a single embedding that represents the input sequence, average pooling? Max pooling? Any learned pooling? 4. Are the learned representations powerful enough that any very basic ML technique such as logistic regression performs well or do we need more involved techniques?   **Expected learning outcomes**   1. Identifying relevant models in the literature and being able to use provided open-source code 2. Writing an inference code that exploits batching over accelerators 3. Storing data in an appropriate format for efficient saving and loading in the disk and in the RAM 4. Testing different classification approaches and efficiently presenting the results in a visual way to compare them   **Background information on Proteins, Peptides, and Classification Tasks**  Proteins are large, complex molecules that play many critical roles in the human body. They do most of the work in cells and are required for the structure, function, and regulation of the body’s tissues and organs.  Proteins are made up of hundreds or thousands of smaller units called amino  acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein.  The sequence of amino acids determines each protein’s unique 3-dimensional structure and its specific function. Amino acids are coded by combinations of three DNA building blocks (nucleotides), determined by the sequence of genes.  Neural Language Processing (NLP) techniques demonstrated several breakthroughs in the past decade, enabling deep neural networks to learn in an unsupervised fashion the underlying structure of any natural language such as English. The learned representations by these models can then be leveraged to solve varieties of tasks, often called downstream tasks, even though very little data is available. In the same way, a sentence in English can be modeled as a sequence of tokens that represent one or several words, a protein can be represented by a sequence of tokens where each token is an amino acid. This observation led to the development in the past 3 years of a deep language model trained on large datasets of proteins leading to breakthroughs in several domains of biology.  In this project, we propose to leverage pre-trained transformers over proteins to compute representations, also called embedding vectors, of proteins to solve a variety of bio tasks that are very significant for the industry even though little data is available. We will consider three tasks in particular.  **Classification Tasks**  Note: in these tasks, we will mostly talk about peptides which are proteins of relatively small lengths.  *Task 1. - ACP Design*  Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. Conventional cancer treatment relies on radiotherapy and chemotherapy, but both methods bring serious side effects to patients, as these therapies not only attack cancer cells but also damage normal cells. Anticancer peptides (ACPs), as a new type of therapeutic agent, have gained more and more attention since they have been selected as a safe drug. Therefore, it’s necessary to develop an efficient and accurate method to predict ACPs. In this task, we want to develop a system that can predict, given a sequence of amino acids, if a peptide is an ACP or not.  *Task 2. - DNA-Binding proteins*  As a part of the protein family, DNA-binding proteins play an important role in DNA replication, DNA methylation, gene expression, and other biological processes. Due to the importance of DBPs, it is highly desirable to develop effective methods to identify DBPs. At present, some experimental techniques, such as filter binding assays, X-ray crystallography [1] genetic analysis [2], etc., are developed for identifying DBPs. However, experimental methods are both costly and time-consuming. Meanwhile, more and more protein sequences have exploded with efficient next-generation sequencing techniques. Therefore, it is an important  research topic to develop fast and effective computational methods to handle such large-scale protein sequence data. In this task, we want to develop a system that can predict, given a sequence of amino acids, if a peptide is a DNA-binding protein or not.  *Task 3. - Antimicrobial peptide recognition*  Antimicrobial peptides (AMPs) are a class of small peptides that widely exist in nature and they are an important part of the innate immune system of different organisms. AMPs have a wide range of inhibitory effects against bacteria, fungi, parasites, and viruses. The emergence of antibiotic-resistant microorganisms  and the increasing concerns about the use of antibiotics resulted in the development of AMPs, which have good application prospects in medicine, food, animal husbandry, agriculture, and aquaculture. Faced with this reality, Machine Learning methods are now commonly adopted by wet-laboratory researchers to screen for promising candidates in less time. In this task, we want to develop a system that can predict, given a sequence of amino acids, if a peptide is an AMP or not. |
| --- |

**About FourthBrain**

FourthBrain trains aspiring Machine Learning engineers in the technical and practical skills necessary to contribute immediately to an AI team. Our remote, online program is designed to be flexible and accessible for anyone with software experience. We infuse values of collaboration, communication, empathy, and equity throughout the program.

We are part of the AI Fund, founded by Andrew Ng.