Secondary structure prediction project 2025

You are tasked with applying what you have learned about Deep Learning to make a secondary structure predictor. It is up to you to decide how to design the model (many tutorials/papers are available for you to study - do remember to refer to them correctly in your written report.

Training data can be found here. The data is available both as PSSM and sequences, your model should be able to handle both. The expected output is per residue class H(helix), E(strand), C(coil).

Please provide your notebook that runs the code as a link to a public colab notebook.

You will be evaluated on reproducibility, readability, functionality, and adherence to best practices.

Please upload a single PDF written as a short research paper.

- You should write a full lab report on your project.

  Including:

  - Introduction, materials/methods, results, conclusion/discussion, supplementary material if needed.
  - Max 1000 words (including supplementary).
  - The report should include the following points in the results section:
    - What is the performance of your predictor? (use metrics such as accuracy,mcc, etc). Don’t forget to include loss curves (validation and training).
    - Some ablation studies, e.g. how is the performance affected by the details of the network (regularization, dropouts, input features, number of layers, etc.)
  - This is not an exam – please feel free to look up relevant papers, collaborate with your peers, look at old notes, labs, etc to help you with the assignment. Of course, do not plagiarize. (We will check each report with plagiarism software.)
  - Please submit your report as a PDF.

Grading of the Report (total 50 points):

- **Format and coherence - 10%**
  - Establish a well-defined report structure by including separate sections for the introduction, results, discussion, and methods.
  - Do not exceed 1000 words in the report.
- **Answered all required elements - 30%**
  - You can utilize figures to illustrate your findings, but make sure you also provide a written explanation for each point.
- **Critical thinking/discussion - 30%**
  - Demonstrate your capability to reason about your results regardless of the performance
- **Figures - 15%**
  - Spend the time to make concise figures with proper references
  - Ensure that all figures are legible and easily understood.
  - The caption should be written in the form of a complete sentence and should provide a detailed explanation of all the elements depicted in the figure.
  - No stand-alone figures! Each figure needs to be referenced within your main body text.
- **Overall impression - 15%**
  - Your report's overall impression, understanding of the assignment, and effort put into the report will be taken into account here.
  - Answering additional questions beyond the required ones may positively impact the overall impression of the report.

Grading of the Code (total 50 points):

You have to present your code (no slides needed) on March 19th, information on how to reserve a time will be uploaded on March 17th.

The presentation will be of ~15 minutes:

- **10 minutes** for **code presentation** (focus on the deep learning part, model architecture and training procedure)
- **5 minutes** for **questions** from the TAs

Evaluation will be based on:

- **Reproducibility 30%**
  - We should be able to rerun your notebook and get the same results as you present in your report (make sure to use fixed random seeds)
- **Documentation 20%**
  - Comments throughout the code and explanation of methods/functions
- **Explanation of the Code 50%**
  - During the presentation, you will go through your code and explain what the different parts do
  - We might ask follow-up questions and you will be graded on the ability to answer these (there might not be a straight-forward answer to the questions, it's more about testing your ability to reason about your code)