Heinrich-Heine-Universität Düsseldorf
Computational Cell Biology
Alexander Kroll

Applications of Transformer Networks in Bio- and Cheminformatics

Worksheet 7

Submission Deadline:	10. June 2024, 11:59 pm
Discussion of solutions:	18. June 2024, 4:30 - 6:00 pm

Submission Instructions:

- Upload a Jupyter notebook with your solutions to the exercises that you solved on your own PC. Upload the Python scripts and log files for any code that could not be executed in Jupyter Notebooks or was executed on the HPC.
- If you are submitting multiple files, zip them together.
- Submit your solutions by uploading the file to https://uni-duesseldorf.sciebo.de/s/hCt1rTP23EeWmUC. The uploaded file should have the following filename: "lastname_studentID_worksheet7".

Excercise 7.1 *Predicting substrates for enzymes* (50 Points)

In this exercise, we will predict potential substrates for enzymes. More specifically, we will build a model that takes an enzyme and a small molecule as input and predicts whether the small molecule is a substrate for the enzyme. For this exercise, you will need to download the files located at https://uni-duesseldorf.sciebo.de/s/g4WTvXi49oSyBOU. In particular, we need the training, validation and test files stored in *data/input_files*. The "protein sequence" column stores the amino acid sequence of the enzyme, the "SMILES" column stores the SMILES string of the small molecule, and the "output" column stores whether or not the small molecule is a substrate for the enzyme.

(a) For this prediction task, we need to store information about the protein and the small molecule in numerical vectors. What do you think are the best ways to obtain numerical protein and small molecule representations for this task? Explain your answer.

(b) Compute Transformer Network-based numerical representations for all enzymes and compute Transformer Network-based numerical representations for all small molecules in the datasets. Consider that many small molecules and many enzymes occur multiple times in the datasets, and you only need to compute a numerical representation once for each unique molecule. This reduces computation time.

(c) Perform a hyperparameter optimization for a gradient boosting model for the binary prediction task described above by training on the training set and selecting the set of hyperparameters that yields the highest Matthews Correlation Coefficient (MCC) on the validation set. Use the same hyperparameter ranges as defined in exercise 5.1(b).

(d) Select the set of hyperparameters that yielded the best MCC on the validation set in (c). Train a gradient boosting model with this set of hyperparameters on the training set and evaluate it on the test set, i.e., compute the MCC, accuracy, and ROC-AUC score for the test set.

Excercise 7.2 *Multimodal Transformer Network* (50 Points)

The data folder downloaded in Exercise 7.1 stores a multimodal transform Network, ProSmith, that has been trained for the task of predicting whether or not a small molecule is an enzyme. The file "Get ProSmith Representations.ipynb" stores code to use the trained model to generate a representation of the updated classification token embedding. For each enzyme-small molecule pair provided as model input, this numerical representation stores task-specific information about the small molecule and the enzyme.

(a) Run the code in 'Get ProSmith Representations.ipynb' to compute numerical representations of the first 100 entries of the full dataset (consisting of the training, validation, and test dataset loaded in Exercise 7.1).

(b) Provide a detailed description of the forward process of the ProSmith Transformer Network from the provided code (What is the model input? How is the input processed? What is the architecture of the ProSmith model? etc...)

(c) Does the ProSmith model use positional encodings? Can you guess why/why not?

(d) **Bonus exercise (30 points):** Extract the ProSmith embeddings for all data points in the datasets and perform hyperparameter optimization of a gradient boosting model using the resulting numerical representations.

Excercise 7.3 *Choose research project* (0 Points)

Until Saturday you will receive project descriptions of several research projects via the rocketchat. Please rank at least 3 projects, starting with the one you would prefer to work on.

Excercise 7.4 *Use of LLMs* (0 Points)

State for which exercise you have used LLMs (large language models) such as ChatGPT or GitHub Copilot. State which tools you have used and for which steps. This answer does not influence how many points you receive for your submission.