Compound-protein interaction (CPI) prediction

Rodrigo Castellano Ontiveros
roco@kth.se

***Abstract*—The main idea of the project is to create a model that can predict compound-protein interaction, which plays an important role in drug-discovery. There are several models that can be applied by using different machine learning (ML) learning approaches. For this project, the focus will be on the model applied in [2], based on transformers. The authors have shown that their model improves previous work done in this area.**

I. Introduction

Predicting the interaction between compounds and proteins can be useful when it comes to discovering and developing new drugs. It can help to detect negative effects of drugs, as well as new potential uses. ML approaches are applied given that biological experiments are time and resources consuming, and some times virtual screening cannot be applied.

CPI have been widely studied and several ML techniques have been applied, such as recurrent neural networks [3], graph neural networks [1] or transformers. The authors of [2] have applied a model based on transformers and self-attention mechanism, in which they work with datasets representing the protein and the ligand. In this case the transformers will try to predict CPI by having proteins and ligands as two sort of sequences. The self-attention mechanism will be used to explain the interaction features.

II. Method

The plan is the following:

- Study the relevant literature on CPI and get familiar with the topic.
- Study and review literature about transformers, why they are applied to this problem and give better results than other techniques such as RNN and how to apply them in python with pytorch.
- Learn how to implement transformers in pytorch.
- Implement the model with the dataset GPCR used in [2] and try to reproduce the results. Compare with a baseline model.
- Try different datasets and compare the results obtained.
- Give an explanation of the behaviour of the model.
- If possible, try to find an alternative to this model that gives good results.

III. Results

The expected results of the project and the course are:

- Learn a introduction to machine learning applied to computational biology and get some insights in this field. Learn the role machine learning plays in drug discovery.
- Learn about state of the art architectures such as transformers that can be applied not only to computational biology but to other fields.
- Try to obtain similar results to the authors of [2]. If it's not possible due to computational resources, an approach could be to train the model with a fraction of the dataset or to find smaller datasets. Some parts such hyperparameter optimization or k-fold cross-validation would be then feasible.
- Find a way to explain the obtained results, such as by visualizing the attention weights.

IV. Time plan

Table I presents the different tasks that compose the project and time estimation for each of them. Final time estimation may vary for each task.

V. Risks

Some of the risks of undertaking this particular project include:

- Struggle to understand the topic, specially when there is no previous experience on biology-related courses/projects.
- Not being able to implement different datasets.
- Not implementing the model as the authors do.
- Not being able to reproduce the results.
- Not being able to find a proper interpretation of the behaviour of the model.

References

[1] Ermal Elbasani et al. Gcrnn: graph convolutional recurrent neural network for compound–protein interaction prediction. In *International Conference on Biomedical Engineering Innovation 2019 Kaohsiung, Taiwan*, 19 November 2019.

[2] Lifan Chen et al. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. In *Bioinformatics, 36(16), 2020, 4406–4414*, 19 May 2020.

[3] Zhangyang Wang Mostafa Karimi, Di Wu and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. In *Bioinformatics, Volume 35, Issue 18, 15 September 2019, Pages 3329–3338*, 15 February 2019.

Task	Hours
Study the relevant literature on CPI and find a suitable model to implement	18
Find information about literature on transformers. Study possible alternatives to implement	25
Learn how to implement transformers in pytorch	25
Implement the model	40
Training, hyperparameter optimization, try different datasets	19
Evaluation and interpretation of results	18
Writing the project report	15

TABLE I

ESTIMATION OF THE TIME NECESSARY TO COMPLETE EACH RESPONSIBILITY.