Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Matyáš Brabec
Název práce	Using combined sequence and structural features to predict protein-ligand binding sites
Rok odevzdání	2024
Studijní program	Informatika **Studijní obor** Softwarové a datové inženýrství

Autor posudku	Petr Škoda **Role** oponent
Pracoviště	Katedra softwarového inženýrství

Text posudku:

The thesis has two objectives motivated by the usage of protein language models for binding site prediction. The first objective is a fine-tuning of a protein language model. The second objective is a utilization of 3-dimensional, structure-related, features into the model.

Altogether, the objectives bridge domains of large language models and ligand binding site prediction. Thus, to tackle both objectives, the student must familiarize himself with two quite complicated domains.

The author demonstrates a general awareness of both domains in the introduction section of the thesis. In this section, the author introduces selected concepts used later in the thesis. Unfortunately, the level of detail does vary. For example:

- The author introduces mmCIF format, just to use PDB instead. Without additional explanation.
- Basic topics like sensitivity and negative predictive value are quite well explained, but not used.
- The author explains "forward propagation" and later states that we "execute the forward pass" as the first step of backpropagation

Yet, for a domain expert, these are just mere distractions.

Following the introduction, the author describes the methodology followed by the presentation of results and discussion. I appreciate that the author keeps explaining new concepts. Still, both sections could better guide a reader. Yet, sometimes, I am missing a more in-depth comment or explanation. For example:

- On page 49 the author chooses to use alpha-carbon for finding close residues. Yet, on page 42 a rare absence of alpha-carbon is given as a reason to not use it as a residue center.

- In section 3.5.2 the author decides to merge train and test splits from Yu dataset. This makes the results of baselines incomparable to other results.
- I would expect a comment on over-fitting in sections around figures 4.3 and 4.5.
- Why not use p2rank datasets?

There are visual distractions like the formatting of lists on pages 18, and 51.

The software aspect of the thesis is mostly focused on data preparation, experiment execution, and evaluation. From this perspective, the main difficulties are the integration of the right libraries and execution in specific environments. Duo to the objectives of the thesis the code is more of a support rather than the main output. This is reflected in the code quality, selected issues are:

- Hardcoded values and paths
- Commented code
- Code structure

Especially the first issue renders it quite tedious to replicate the experiments.

Práci doporučuji k obhajobě.
Práci nenavrhuji na zvláštní ocenění.

V Praze dne 31.5.2024

Podpis: