Posudek oponenta diplomové práce

Jméno posuzovatele: David Hoksza

Autor práce: Bc. Gesa-Maret Neitzert

Název práce: Enzyme optimization using sequence homology and machine learning

The thesis aims to the evaluation of different computational approaches to enzyme optimization. The thesis is split into four parts – introduction/motivation, review of existing approaches for enzyme optimization, experimental part, and discussion.

The **introduction** is quite brief, and I would appreciate more background on enzyme optimization, also from the biological perspective. Considering the thesis has only 30 pages, there is surely enough space for that. On the other hand, I think all necessary facts and motivation are given. What I am missing, however, is some formal explanation of what are the goals of the thesis. What were the inputs and what are the expected outputs.

The **second section** aims at describing existing approaches, but to me, it seems more like it describes one paper (Xu et al. (2020)) although there seems to be quite a lot of literature on this topic. This seems also to be the reason why the experimental part is actually not about conversion rates or stability improvement (which tends to be the goal of enzyme optimization as also stated by the author), but about "shift in peak absorption, increased enantioselectivity, or increased fluorescence" (citation from the thesis). In section 2.1.1., if I understood correctly, the author presents the results of (Xu et al. (2020)) (this was not stated directly, so I am not sure about it). The section is missing an explanation of the task, i.e. what are the inputs, and what was about to be achieved. It starts with "Input sequences were aligned without gaps" so only then, the reader finds out there are some input sequences and can start forming some mental image about the task being solved. Only later, when mutated positions are mentioned, one is starting to understand that the MSA are mutants of a single sequence (probably). At the end of section 2.1.1. there is the sentence "While even simple descriptors perform relatively well, they fail to discover new mutations" – the lack of description of the task hinders my understanding of this sentence because I thought that the mutated sequences were provided on the input.

The second part of the second section (2.2) is then dedicated to the description of different computational methods. However, I simply do not understand the choice of methods. Some of them are then used in the experimental part, but some do not (if I did not miss anything, which is completely possible given my problems with comprehension of the experimental section). For example, evo-velocity or Envision (sections 2.2.5 and 2.2.6) are not used later. I understand that they might be used somehow in enzyme optimization but i) it is not explained how exactly, and ii) there are surely many other such methods, so why this particular choice? Section 2.2. is also not very well structured as it mixes different types of methods, where each is given a separate subsection (2.2.x). For example, SeqVec (2.2.1) and ProtTrans (2.2.3) are embeddings, bio_embeddings (2.2.2) is a software library, Transformers (2.2.4) is a concept (used by ProtTrans), EnzymeMiner is a pipeline. The shallow structure of 2.2.x also causes that much information there is redundant. For example, the fact that SeqVec "produces continuous vectors,

so-called embeddings, which represent the proteins' biophysical properties" (section 2.2.1) holds also true for 2.2.2 and 2.2.3). Similarly "When viewing protein sequences as a form of language of their own, NLPs can be used in an attempt to teach a computer to understand the "grammar of the language of life" (section 2.2.3) holds also for SeqVec (2.2.1); or section 2.2.4 states that "When used upon protein sequences, amino acids become the characters for the language," which is again true for section 2.2.1 and 2.2.3.

The **third section** describes the experimental part of the thesis. Here, several datasets and four approaches are described.

The datasets part introduces four datasets which "were partially used within the analysis done by Xu et al. (2020).". **How** exactly and what "partially" means is missing. Table 3. Summarizing the dataset is missing information on what "size" means. It is clear later from the context but would be nice to have this information in the table description. What I am missing is some statistics that actually might help in the interpretation of the results. Namely regarding the number of mutations. Printscreens of the tables (btw using print screens of tables, especially in low resolution, does not feel very professional) with the dataset do not say much about the number of mutations, but based on the few values, it seems that "EMB_T50" has more mutations, while "EMB_absorption" maybe has only a single mutation? Hard to say and the data were not provided so that one could check that directly (discussed below). When describing each of the datasets, the text mentions "train/test split descriptors" and I am not sure what that means. Maybe train/test labels? Also, the text mentions that column names were normalized, but does not say **how** (again, the source codes were not provided to check). The description of the ENA dataset states that "Train/test splits for this dataset were done according to number of mutations" – again, **how**? I am also missing any mention of where the data were obtained from and what was their format.

As for the approaches, each approach seems to be doing something else. The first two approaches (baseline method and mutagenesis pipeline) give information about probabilities of residues at positions while the second two (Random Forests [RF] on features from ProtParam and RF on embeddings) are predictions. Actually, it's not very clear why the predictions are presented as separate approaches if the only thing they differ in seems to be the set of features they use to represent individual residues. Anyway, the introduction to the experimental section states that it "attempts to find a correlation between residue frequencies and their effect on the respective target value for each dataset as well as predicting target scores based on a training set of sequences" – it would be nice to say which approach does what and not to let the reader guess. This could also help to better structure the thesis if the two aims were clearly separated. It also states that "If successful, a method will be able to single out mutations that improve the target value for the respective dataset" – I think this is important and I am missing a section that would explain exactly this – **how** the mutations would be picked and **how** the results would be used in practice. To be honest, the first "how" is partially addressed by two sentences at the beginning of the results section; specifically "Both the baseline and the bio_embeddings mutagenesis pipeline suggest substitutions in the parent sequence based on the analysis of sets of sequences. The baseline does so by picking the mutations with the highest frequencies from homologous sequences, while the mutagenesis pipeline suggests the most probable amino acid at each position as learned from a much larger and less closely related sequence database, namely BFD (Jumper et al. 2021).". This also suggests that there is, indeed a structure in the methods (2 methods for suggestions based on sequence, 2 based on ML).

The baseline method is a relatively straightforward application of sequence search where the results are limited to 100 sequences irrespective of the resulting e-value. The obvious reservation would be that among those 100 sequences there can be sequences that are not evolutionary related so wouldn't it be better to threshold based on a specific e-value? Also, what would be the effect of selecting different thresholds (fewer or more sequences)? What would happen if the sequence under evaluation was a rare sequence? None of these questions seems to be tested or even discussed in the discussion or anywhere else in the thesis (unless I missed it). The results show a "relationship between the mean residue frequency in natural homologs compared to the experimentally measured target value for each dataset". The sentence is missing an explanation of **what** the mean is computed. The result of a pipeline is a vector of frequencies for a sequence (at least I think so), so I have vectors of vector. So is the mean value mean of means? Or do I take the mutated sequence and for mutated positions I consider the mutated residue and pick its frequency computed for the parental sequence and take mean over all the mutated positions? That seems more likely, but the reader should not be forced to guess.

The "mutagenesis pipeline" uses the ProtTrans embeddings. Here, I do not follow why sequences containing mutations suggested by the consensus sequence and the rest were selected and their target values compared. What is the motivation? I could also choose sequences that had mutations that were highly probable, but not the most probable. Seems reasonable as well.
In the mutagenesis section, the thesis also mentions technical details, such as using YAML config files. I don't think it is a good idea to mix method descriptions with implementation details. Especially considering no implementation was provided; so picking one technical detail has zero added value.
Similarly to the baseline, I am not sure how the averaging was done.
Finally, for the mutagenesis pipeline, the pairwise correlation of residue probabilities was compared to BLOSSUM62. It was not discussed why BLOSSUM62 was used. Considering that the parent sequence and mutated sequences are evolutionary very close, it seems more reasonable to use a lower value than 62. Anyway, what was the point of this comparison? How does it help in the enzyme optimization process or what relevant information does it bring with respect to enzyme optimization?
As for the prediction methods, no correlation between predicted and true values was observed. Now, it seems that in both cases, RFs were used on a vector describing the whole protein (in the first case, on 7-dim feature vector, in the second on flattened residue-level feature vectors). Why did the author not try some sort of windowing and aggregation, so that only parts of the protein (maybe the ones around the mutation) would be considered? It seems highly unlikely that any protein-level annotation (especially such, which does not consider structure) would be able to pick up a signal based on one or just a few mutations (information on exactly how many mutations there were on average in the sequences was not provided). Also, it was not specified how much dimensional the vector describing a protein in the second case was, but I guess a lot. Considering the size of the data set and dimensionality of the vector it is no wonder the model was not able to learn the mapping. This should have been discussed in the thesis, but I did not find any such analysis. I am also missing any information on the RF models, i.e. how was it trained, how deep the trees were, how many trees were used and so on.

What I did not find in the whole thesis is any discussion **how** the developed methods and results could be used in actual enzyme optimization.

The final part of the thesis is the discussion. Apart from the things described above, I am missing here **how** the results could be applied in an actual enzyme optimization scenario and whether they are generalizable. Considering the methods are tested on 7 sequences in total, it should be discussed if and why the results could be generalized beyond these few sequences.

Follow comments which are **not specific to any particular section**.

In way too many cases, the description of different processes and concepts is very brief, not clear, or rigorous enough. Most of the examples below (I excluded those which are already mentioned in the text above) are not really mistakes per se, but call for a more detailed explanation; which would not be a problem considering how short the thesis is. Also note the bold "hows" above.

- Section 2.1.1. "Input sequences were aligned without gaps (possible because datasets contain substitutions rather than indels) and turned into feature vectors, where columns represent positions" – **how** is a sequence "turned" into a feature vector? Also talking about "column" requires a table, but there was none introduced (I can understand what the author means, but this an example of the "not rigorous enough" category).
- Section 2.1.1: "for other machine learning methods the input was reduced to the mutated columns" – first, the input was not defined, thus, what it means to "reduce it to mutated columns" is not entirely clear either (I think it means that only mutated columns were taken into account in the further analysis, but that is just my guess).
- Section 2.1.1: 44 combinations of descriptors and ML methods were taken, but from Fig.1 it follows there are 48 methods, so it would make sense to tell which ones were considered
- Section 2.2.4: "The protein landscape modelled for evo-velocity connects the k closest sequences in an embedding space with edges describing the likelihood…" - Landscape is typically defined by a fitness function. **How** is landscape represented by a graph?
- Section 2.2.1: "Probability distributions are learned on unlabeled data overcoming the problem of lacking annotations" – there is no explanation on **how** probability distributions are obtained from SeqVec and thus the reader can't know how they could be utilized
- Section 3.3.1: "The frequencies served on one hand to determine the likelihood of finding sequences" – **How** exactly were the likelihoods computed?
- Section 3.4.1.: "The mutagenesis pipeline can also compute a heatmap of residue probabilities" – **How** exactly? I guess every position was taken, this position was represented then as a vector of probabilities and correlation was computed. However, there are multiple positions with the same residue and each of them has different probability vector due to different context. So how exactly is the correlation between multiple vectors computed? I know there are ways how to compute this, but as I reader I should not have to guess.

Related to the previous point is that the thesis mentions quite a few concepts without explanation, for example, "evolutionary predictability" (section 2.2.5) or "mutagenesis" (first mention in section 2.2.6).

In some cases, there are statements that are not completely true or simply false:

- Section 2.2.4 states that transformers are "deep learning model similar to recurrent neural networks, with an advantage in speed stemming from the parallelization that is made possible by processing the entire input at once". This is not true, because transformers do not take entire input at once.

- Section 2.2.1: SeqVec… “Its goal is to outperform common protein function prediction methods”. This is not true as the goal of SeqVec is to learn contextualized representations which can be used for various tasks, one of which can be function prediction.
- Section 2.2.1: “SeqVec produces continuous vectors, so-called embeddings, which represent the proteins’ biophysical properties” – this is questionable. Which properties exactly? Is there some work showing correlation with a given set of biophysical properties?
- Section 2.2.2: “The information contained within embeddings has been shown to reflect biophysical and biochemical properties of proteins and can be further used with help of transfer-learning to train common machine learning models for whichever predictive task at hand” – embeddings can be used directly, without any transfer-learning; for example, I can use them directly as feature vectors for random forests.

Regarding formatting of the thesis I have only two comments:

- What strikes me most are low-resolution images causing pixelated figures. Surprisingly, this happens not only when the images are taken from a publication (Fig.2), but also when they are directly generated by the author. Sometimes an image is not centered (e.g. Fig. 1); sometimes it even contains underscore lines from language correction software (Fig. 5).
- Variables or file names should be written in italics, e.g. “k” in “k closest sequences” (section 2.2.5), NaN (btw. this is not explained and also is missing in the list of abbreviations), config.yml, …

Language problems and typos:

- “Results were once again downloaded as a FASTA file and serve as the input“ -> served
- “any analysis originating only with the parental sequence was done threefold, once for each” -> three times
- “to determine 10 high scoring mutations” -> highest
- “Additional to the comparison of overall residue frequencies” -> Additionally
- Missing commas
    - “Within biology evolution is…” -> “Within biology, evolution is…”
    - “Within drug discovery these processes” -> Within drug discovery, these processes
    - “Within this thesis different methods” -> Within this thesis, different methods

Regarding citations, I noticed only one missing citation: “transformer models outperform LSTM models in most tasks such as remote homology detection or secondary structure prediction”.

In conclusion, the thesis itself could be, in my opinion, much better written. It could have better structure and more detailed explanation of how things were done. The review part of the thesis seems to only present methods taken from a single publication (Xu et al.) and a few other cherry-picked methods (or at least it seems so – an explanation of why those particular methods were selected is missing). As for the software part, the methods were not explored fully (hyperparameter optimization, more thought-through ML techniques). I can’t comment on the implementation as there is no data or software attached and no software repository referred to. For that reason, it is difficult to say how complicated the work actually was, how much software work has been done and how replicable the

work is. The thesis appeals to me more as a bachelor thesis rather than a mature diploma thesis. For all those reasons I am slightly hesitant to recommend the thesis for defense.