JOHANNES KEPLER
UNIVERSITY LINZ

Univ.-Prof.
Dr. Sepp Hochreiter
Head of Institute for
Machine Learning

P +43 732 2468 4521
F +43 732 2468 4539
hochreit@ml.jku.at

Office:
Birgit Hauer
Ext.4520
birgit@ml.jku.at

Linz, March 15, 2021

Supervisor Review for the Bachelor Thesis: "Understanding Protein Function Prediction using Deep Learning"

The Bachelor Thesis of Fathy Shalaby „Understanding Protein Function Prediction using Deep Learning“ investigates the influence of multi-task settings on the predictive performance of deep neural networks for characterization of proteins based on their amino acid sequences. The basis of this thesis are the publications "DeepLoc: prediction of protein subcellular localization using deep learning" (Almagro Armenteros, J. J., 2017), "DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier" (Kulmanov, M., 2018), and "DeepTox: Toxicity prediction using deep learning" (Mayr, 2016). In the first two works, successful applications of deep learning methods for protein characterization are shown, without closer investigation of the influence of multi-task settings. In the last work, the advantage of multi-task settings in deep learning is demonstrated for toxicity prediction.

Mr. Shalaby has trained a long short-term memory (LSTM) recurrent network on publicly available amino acid sequence data to obtain classification models for proteins. He trained these models in single-task settings and multi-task settings, where a model should predict a single main task and additional auxiliary tasks. Finally, he analyzed the differences in performance between the models trained in single- and multi-task settings. The thereby created code base could be used for further application and investigation of LSTM models on amino acid sequence data.

Mr. Shalaby has worked relatively independently and needed little input for his practical work. He was able to extract relevant information either directly from scientific publications or also from conversation with his supervisor. From the methodological point of view, there are a few open question that were not explored sufficiently enough but would have required more computational resources. E.g. a larger hyper-parameter search for the LSTM architectures, more combinations of auxiliary tasks, and other output designs, such as hierarchical targets.

In his thesis, Mr. Shalaby showed that he can cite appropriate references, explain his approach, and report and analyze results. **Shortcomings:** The thesis is missing clarity and substantiation at some parts, due to bloomy or ambiguous statements. E.g. using unnecessary convoluted sentences and referring to his own results as "impressive" or "noteworthy" without substantiation. In the used formulas, some (minor) variables remain undefined. The number of reported decimals is inconsistent. Citation is done properly, with one minor exception in the conclusion, where the method names would benefit from repeated citations, and a formatting error in bibliography entry "[HHO]". **Strengths:** Overall, the report and analysis of the results appear thorough. The provided plots are sufficiently discussed in the text and illustrate the model performance and behavior well. Mr. Shalaby also demonstrated his capabilities at re-implementing and adapting algorithms.

We can confidently state that Mr. Shalaby is worthy of obtaining a Bachelor of Science title, given the work performed in this thesis.

Sepp Hochreiter