

---

FROM: DAVID ŠAFRÁNEK, PH.D.  
ASSOCIATE PROFESSOR

E-MAIL: SAFRANEK@FI.MUNI.CZ

Review of PhD Thesis of Ing. Jiří Hon

Herewith I provide the review of PhD thesis “MINING OF SOLUBLE ENZYMES FROM GENOMIC DATABASES” of the doctorate candidate Ing. Jiří Hon (BUT FIT). The review is structured with respect to the individual evaluated aspects that lead to the final assessment presented at the end of the review.

### **Thesis Topic**

The thesis targets an important and relevant interdisciplinary topic of computational methods for design of efficient enzymes. One of the open challenges in bioinformatics is to establish a computer-aided platform that will allow to design enzyme vehicles working efficiently for a given (bio)chemical task in a given environment. The thesis of Jiří Hon contributes to this challenge with new methods for mining and rational selection of novel soluble enzymes. Solubility is an important aspect that allows practical implementation of protein design in given conditions.

There are two main methodological goals solved in the dissertation. The first goal was to develop a working method for database-based mining of enzymes based on the input protein sequence and the description of the essential amino acids with incorporation of structural and environmental annotations. The second goal was to develop a new method for predicting solubility of a protein based on its amino acid sequence. In particular, the objective was to improve the state-of-the-art. Besides the mentioned individual goals, there was also a general challenge of integrating the solubility prediction method within the enzyme mining procedure. The combination of the goals naturally displays the dissertability of the thesis and a challenging research question to be solved.

In general, the contribution of the thesis is original and in-line with current needs and trends in bioinformatics. To the best of my knowledge, there are no robust working solutions addressing the problem effectively while incorporating solubility metrics within mining. The most accepted existing (not optimal and partial) solutions are mentioned in the thesis as the related work and compared against the new results of the candidate. In my opinion, the topic of this dissertation, especially the uniqueness of the developed method, clearly confirms originality of this dissertation.

### **Research Contribution**

The main contribution is provided in the form of two software packages – EnzymeMiner and SoluProt. The tools implement the developed methods and that way bring the new algorithms towards the exact future impact in protein engineering. Every tool is published as a self-contained result together with the relevant algorithmics and a test-bench evaluation showing the practicability.

The contribution targeting EnzymeMiner has been published in a highly impacted journal – Nucleic Acids Research (2020), IF>11. The candidate acts as the first author of the paper with a significant contribution (45%) to the informatics part of the work.

The contribution targeting SoluProt has been published in a top bioinformatics journal – Bioinformatics (2021), IF>5. Again, the candidate acts as the first author with strong significance of his contribution (60%) covering almost all computer scientific and mathematics tasks.

Additionally, the candidate has co-authored four other publications published in high quality journals. The most relevant is ACS Catalysis (2018), IF>12 – that seems to be the initial publication where the candidate has took his initial review work while making 30% contribution to the manuscript.

In conclusion, the research targeting the thesis goals has been published in three high quality journal papers. All papers are of high impact thus displaying the quality of the candidate’s research contribution. The papers have already obtained citations that show the candidate’s contribution is relevant for the research community.

## **Thesis Text**

The text of the thesis is organised in the form of a collection of published papers.

In general, the text is well-organised and gives enough information allowing the reader to systematically understand the main contributions of the research. The content is sound.

What is missing is some comprehensive summary of the bioinformatics innovations and novel technical achievements including an extensive discussion of performance issues. That would definitely help to judge the actual technical contribution of the candidate. Moreover, the general goal of integrating computed solubility metrics within the mining procedures seem not to be discussed at the end.

I suggest the candidate to focus on the above mentioned aspects during the thesis defense.

## **Overall Evaluation**

To sum up, the research presented in the thesis displays the following attributes: a non-trivial problem very relevant for the current needs; a systematic work of the candidate giving an interesting and original solution to the problem; main research outcomes published in highly-impacted journal papers; significant role of the candidate in conducting this research. The thesis meets the standard requirements imposed on a dissertation thesis in the field of bioinformatics. Based on the facts mentioned in this review, I am pleased to suggest to the committee to accept the thesis and nominate the candidate for the title Ph.D.

## **Questions**

1. How the software tools perform (in terms of computation time) in a typical user scenario?
2. Identify the potential bottlenecks in the computational efficiency. Is the technology already prepared for high-performance hardware or what needs to be done to improve scalability if that is an issue?
3. Where do you see places for further improvements in both quality of the results and efficiency of the computation?

Brno, 20th February 2022

David Šafránek, Ph.D.