Review of Doctoral Thesis

Mgr. Radoslav Krivák

Faculty of Mathematics and Physics, Charles University, Prague

Prediction of ligand binding sites from protein structure

In his dissertation thesis, Mgr. Radovan Krivák deals with three topics dealing with ligand binding from protein structure – (i) rescoring of ligand binding sites predictions with PRANK machine learning scoring function, (ii) preparation of own ligand binding site predictor implemented in P2Rank package and PrankWeb web service and its integration to P2Rank-Pept service and PDBe-KB database, and finally (iii) search for Apo-Holo structural pairs differing in ligand binding site occupation and conformation in protein database.

The thesis is written in good English on 135 pages. After the short introduction into the topic of binding site prediction follows an overview of the author's contribution to the topic on 7 manuscripts (3 as a first author) and 3 first-author's conference proceedings. Two major chapters then follows - i) ligand binding site prediction, and ii) Apo-Holo protein search and Part I is concluded. Part II lists individual publications together with author's highlights and bibliography.

Despite few minor typos, I can but only cite the author (page 6): "There is no reason to pretend that the work presented in the thesis was a *liner* process of first setting set of goals and then gradually accomplishing them." I find the presentation of the thesis and the work behind it and its acceptance in community via nice usage of PrankWeb and PDBe-KB implementations to accomplish the major goal of any thesis – its relevance and significance to the field. Hence, this thesis shows that the candidate is ready for its defense. I recommend the acceptance of the submitted thesis as a part of the procedure of awarding the Ph.D. degree.

I have attached my questions for the author on the following page.

Conclusion: **I suggest the acceptance of the thesis**

Olomouc 20th March 2023

doc. RNDr. Karel Berka, Ph.D.
Department of Physical Chemistry
Faculty of Science, Palacký University in Olomouc

Questions to the author:

1. Can you elaborate more on how do you use Bayesian optimization to optimize several arbitrary parameters simultaneously for the development of prediction methods (see page 11)?
2. On page 18, you state that P2Rank has more than 100 documented configurable parameters. Do they have and physical meaning or interpretability?
3. How hard would it be to train P2Rank for specific types of substrates, e.g. cofactors, steroids, peptides, nucleic acids, ions, magnesium, …?
4. Would it be possible to predict ligand binding sites with PrankWeb not only on AlphaFold structures, but also on structures from large language models such as ESM from ESMAtlas?
5. Given that now we can have structural model for almost any protein sequence, would it be possible to train P2Rank to predict ligand binding sites to work directly on sequences?
6. On page 23, you mention that the prediction of P2Rank-pept is not yet practically useful in its current state. What is missing?
7. Can AHoJ analyze how often are AlphaFold models in Apo or Holo state?
8. What do you expect as the new direction forward in the field of ligand binding prediction? What is missing in current benchmarks used in the field?