FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Univerzita Palackého v Olomouci

Přírodovědecká fakulta

October 18, 2025

Reviewer's report of the habilitation thesis *New methodologies of Machine-Learning modeling of complex chemical systems: mixtures, reactions and ligand-protein complexes*

The submitted thesis addresses a wide range of topics in cheminformatics revolving around the representation of molecular systems and their application in a wide range of issues, including the prediction of molecular properties, the interpretation of models, the exploration of the chemical space, all the way to the design of bioactive molecules. As I am familiar with the work of Dr Polishchuk, I noticed that these do not even cover some of his recent research towards molecular docking, such as the development of the EasyDock framework (although this is mentioned in the list of the developed software at the end of the thesis). The thesis is arranged chronologically, which helps to see how approaches developed earlier in Dr. Polishchuk's career fed into the research in the later stages. This shows that the research presented in the thesis does not lead to a dead end, but is later actively used and extended. Of course, this is also reflected by more than 3000 citations of the work (Google Scholar) with an increasing trend in the last few years. The last chapter, where Dr. Polishchuk gives three examples of virtual screening projects, shows that he sees the domain in its full breadth and can integrate the individual facets of the field, resulting in real-world applications. This is further supported by many projects Dr. Polishchuk has been involved in and which are cited in the thesis. This shows that although the presented work is computational, Dr. Polishchuk is able to transfer his knowledge to practical applications.

Department of Software Engineering
Malostranské nám. 2/25, 118 00 Praha 1, Czech Republic
ksi@mff.cuni.cz
+420 95155 4250

As mentioned above, the work of Dr. Polishchuk is widely cited and is published in more than 50 journal articles in top journals in the field, such as Journal of chemical information and modeling, Journal of computer-aided molecular design or Journal of cheminformatics.

As a computer scientist, I also greatly appreciate that the published work is accompanied by publicly available software enabling further spreading of the methods. Especially, the later software developed in the lab of Dr. Polishchuk is published according to standards in the field, available at GitHub, with clear installation instructions, examples of usage, or "dockerized" versions of the applications enabling easy execution. This might seem like a small thing, but it can be a lot of work and, most importantly, helps other researchers in the field and ensures reproducibility of the research.

Furthermore, it is essential to note that the title of associate professor (Doc) not only signifies scientific expertise but also holds academic significance. In this regard, Dr. Polishchuk has also committed himself to the transfer of his knowledge to students (and hopefully a new generation of cheminformaticians), as evidenced by having many PhD students who are part of this lab.

Questions for the habilitation thesis defence

1. At the end of section 1.2, the newly developed SiRMS is compared to COSMO-RS, showing comparable performance. What is then the advantage of SiRMS? The same question then goes to SiRMS-react as at the end of section 1.3. It is again stated that the introduced approach is "only" competitive with the SOTA methods. I understand that SiRMS is a nice set of easily interpretable features and is thus suitable for model interpretation, but that research came only later. So I am wondering if there is another advantage to using SiRMS. Or does it simply show that this is another way of representation that has not been tried before?
2. In section 2, a method for showing the contribution of a fragment to a given property of interest. Wouldn't it be better to use 3D structure here to cluster the compounds instead of using Gaussian Mixture Modeling? Or maybe using it alongside the GMM. I imagine that compounds with a similar mode would also overlap in the conformational space. This actually relates to my other question, which is how often I see multimodal distributions. The example outlined in Fig. 24 states that 118 out of 311. Is this typical?
3. In section 3.2, 3D ligand pharmacophores are used to screen 3D poses of ligands of receptors from PDB. I am wondering what the similarity was between the ligands in the train set and the ones in PDB? Is this something considered in the field? I would

imagine that more dissimilar molecules would be more difficult to detect and thus methods being able to discover such molecules would be more valuable.

4. In section 3.3 pharmacophore modeling is applied on MD simulations of protein-ligand complexes. Did you check whether the MD simulation was able to recapitulate known protein-ligand complexes in the first place? For instance, one could take all available known-holo structures of a ligand from PDB and see if the MD simulation is able to get close to them in terms of pocket and ligand RMSD. For instance, I know that it is actually quite challenging to get from HOLO to APO states so I was wondering whether it would hold for HOLO modelling as well.
5. I missed it – did using MD actually help compared to not using it?
6. When using multi-instance learning, none of the more advanced methods seem to work better than the simple instance wrapper. Wasn't the reason overtraining or underfitting? Considering that the more advanced approaches are more complex in terms of the number of parameters, having enough training data could be a limiting factor. Did the training and validation curve converge?
7. Later in the same section, Bag-AttentionNet is used to identify the biologically relevant conformers. In the process, the detected ligand conformations are compared to those in PDB and statistics are computed. I was wondering if the electron density maps for the ligands were looked at. Because having a 2A match might not tell much if the ligand of interest is not well modelled. I don't know which ligands were tested, but for instance, in structure 4KCQ, one of the targets, most atoms of one of the ligand (1QF 506) are not covered by electron density at all, and so there is no reason to believe in the positions of those atoms.

Conclusion

I strongly believe that the habilitation thesis entitled "New methodologies of Machine-Learning modeling of complex chemical systems: mixtures, reactions and ligand-protein complexes" by Dr. Pavlo Polishchuk fulfills the requirements expected of a habilitation thesis in the field of physical chemistry.

doc. David Hoksza, Ph.D.