

Jury Member Report – Doctor of Philosophy thesis.


Name of Candidate: Sergey Sosnin

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Exploration of chemical space by machine learning

Supervisor: Professor Maxim Fedorov

Name of the Reviewer:

<p>I confirm the absence of any conflict of interest</p> <p>(Alternatively, Reviewer can formulate a possible conflict)</p>	<p>Signature:</p>  <p>Date: DD-MM-YYYY</p> <p>28. 11. 2020</p>
---	---

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The topic of the thesis is rather broad, encompassing a majority of the cheminformatics field, and, accordingly, the solved problems are rather diverse. While somebody may consider this as a weakness, we all know how it works with PhD topics in reality and, actually, I see this as candidate's advantage. The reason is that during his work the candidate successfully mastered a wide variety of computational, data analysis and cheminformatics techniques, as demonstrated by an impressive list of his publications.

The thesis is written with a minimum number of mistakes and typos. The figures are well chosen and they sufficiently illustrate thesis concepts. However, the applicant should have pay a bit more attention to details. Here are some problems I identified:

1. Chapter 2.6 Bioconcentration factor paragraph does not contain the description of data sets used, while the next two paragraphs (Acute animal toxicity and Bioactivity datasets) do.
2. Page 69. The number of records for each endpoint is missing in Table 4.1, in contrary to what is stated.
3. Table 5.1. The horizontal blocks (i.e., kNN, SVM, XGBoost, RF) are not described which makes the understanding more difficult.
4. Page 110-111. The last paragraph on the page 110 does not, apparently, belong here. It looks like the response to the reviewer comment.

Regarding the science, I have several points for a discussion.

BCF modelling

1. I don't fully understand how "the merging of BCF values of different experiments is possible". Doesn't the dissimilarity between species imply differences in the distribution of the organic compound within them? Please, could you elaborate more on why is it that BCF of different species can be compared?
2. Are there any strongly hydrophobic molecules in test data set and did you check the predictions of your models for strongly hydrophobic molecules?

Acute toxicity modelling by multitask learning

3. In this chapter, several claims about the superiority of MT_DNN models are made. In Conclusions part (page 75) you even claim that MT_DNN significantly (it doesn't mean statistical significance, does it?) improve toxicity prediction. However, in my opinion, such bold claims should be supported by statistical model comparison. For example, you state that ST_DNN models are comparable with XGBoost and RF but, based on the visual inspection of the Figure 4-4, I would say that by the same logic these models are also comparable with MT_DNN models. What lead you to such conclusion about MT_DNN performance?
4. In Table 4-4, the Feature Net is compared with ST_DNN and MT_DNN models. Why didn't you include also RF into the comparison? According to the Figure 4-4, RF is the best performing model from all single task models. Please, could you compare the performance of MT_DNN and RF and specify they advantages and disadvantages?

Chemical space visualization

5. If I understand it well, the cost function (the Kullback-Leibler divergence in your case) can be used as the measure of map quality (i.e., the map with lower cost function can be regarded as better). This is useful when you want to compare different maps generated with different algorithm settings. Please, could you comment on how such defined quality relates to the real-use quality of the map? By this I mean if it can, e.g., happen, that maps with lower cost functions may reveal relationships not observed in “better” maps.
6. Can you compare computational requirements (speed and memory demands) of the parametric t-SNE with its non-parametric version, as well as with other embedding methods such as MDS, IsoMap and UMAP? IsoMap and UMAP are also very popular, why were these two not included in the study?

Legogram

7. I wonder if compounds generated by the legogram are synthetically accessible. How is this ensured?
8. One of the most important quantities generated compounds to be optimized for is the biological activity against the selected biological target. How would you bias your generator to yield compounds with higher probability being biologically active?

Summarizing, Sergey has performed a large amount of insightful research and obtained new original results which broaden our understanding of chemical space exploration. The dissertation work has been performed at a high scientific level and the candidate demonstrated his capability of critical thinking and of independent scientific work in this specific field. All important results presented in the thesis are published in peer-review journals. Judging by the thesis, the candidate merits the PhD degree and I clearly recommend its acceptance.

Provisional Recommendation

I recommend that the candidate should defend the thesis by means of a formal thesis defense

I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report

The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense