Skoltech
Skolkovo Institute of Science and Technology

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: Prof. Maxim Fedorov

Name of the Reviewer: Maxim Panov, Assistant Professor, CDISE, Skoltech

I confirm the absence of any conflict of interest.	**Signature:**  **Date: 30-11-2020**

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 considered thesis targets the important area of efficient processing chemical data by machine learning methods. The author considers several important problems:

1. Construction of 3D spatial descriptors for molecules based on convolutional neural networks.
2. Application of multitask learning to toxicity modelling
3. Chemical space visualization by deep learning-based embedding methods
4. Usage of molecular grammars for sampling new compound from chemical space.

All these topics are very relevant for modern chemoinformatics. The author shows fluency in application of state-of-the-art machine learning methods to challenging problems involving molecular data. Moreover, some of the methods require non-trivial modifications and lead to the algorithms which are interesting for machine learning in general.

In my opinion, thesis results are very relevant from the application perspective. Importantly, the author provides all the details on the algorithm construction and training procedures, which allows to directly apply them to drug-discovery problems. I should note that in some application areas having very complex data such as chemistry, it might be complicated to construct robust models which beat competitors by a significant margin. That's clearly the case for the present study. For example, in chapter 4 we see that multitask learning doesn't give a significant improvement for the considered problem. However, I don't think that it decreases the quality and significance of the results.

Finally, the results of the thesis research were published in well-reputed journals including six Q1 journals. Thus, the quality of the publications well supports the overall good scientific quality of Sergey's thesis research.

While I have overall positive opinion about the research contents of the thesis I think that the text deserves serious improvement. I am especially concerned with the chapter 2 where the author tried to give a detailed coverage of the algorithms used in the thesis. However, this attempt was not very successful as many essential details are missing while some of the formulas contain serious errors. There are also multiple misprints and minor language issues appearing throughout the manuscript. I also suggest to change the citation style as in current text there are some issues with differentiating between citation and the text itself. The list of issues with Chapter 2 is given below.

To sum up, I think that the issues found do not decrease the scientific quality of the thesis and Sergey Sosnin deserves to be awarded with Skoltech PhD degree.

The list of issues (in the order of appearance in the text):

1. [Formula 2.2] It is not clear with respect to what expectation is taken. I also can only deduce that it is expectation as this symbol is not explained in the text.
2. [The sentence before formula 2.3] Should be WiEner index.
3. [Formulas 2.3 and 2.4] The same letter is used for degrees and distances, might be confusing.
4. [Explanation of formula 2.5] Should be F_i is a molecular fragment.
5. [Formula 2.6 and everywhere else] If upper index for summation is provided then lower index also should be present, i.e. sum_{i = 1}^N.
6. [Formula 2.8] \theta is not introduced.
7. [Formula 2.9] x_i is not introduced, it is not clear how y_t depends on x_i.
8. [Formula 2.10] It is not clear for what reason we have some constant C here.
9. [Formula 2.12] Again, index i should be somewhere in y_t.

10. [Formula 2.15] Some w_i appear, again not clear who are they.
11. [Formula 2.18] There is problem with indices.
12. [Section 2.3] It appears that square is missing in both starting formulas of the section for the part of the formula to the left of equality sign. Also, the variance term should include only Var[\hat{f}] as Var(y) is exactly sigma^2 (next summand).
13. [Section 2.4] \rho_\gamma is not defined.

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*