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: Professor Maxim Fedorov

Name of the Reviewer:

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** **Date: 20-11-2020**

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 thesis “Exploration of chemical space by machine learning” by Sergey Sosnin dedicated to the development of new methods and application of current machine learning methods to address specific biological and chemoinformatics challenges. Three main topics were selected: bioconcentration factor prediction; toxicity prediction; and visualization of the chemical space by 2D representation and sampling of chemical compounds directly from the desired regions of chemical space. The results are presented in structured way; however, the logic flow does not seem smooth. In particular, the chapter 2 Material and Methods looks like the overview of the current methods in the field and should be considered as a literature overview. The methods used in current study are sparsely introduced in the mentioned chapter, but should be in a separate section/s. The contents page does not include all subchapters and headings presented throughout of the thesis, which makes it difficult to navigate through the thesis. The methods overview for the separate topics is introduced at the beginning of each section but it is not reflected in the heading and subheadings. Many figures are not within the section where are discussed and the numbering is not in the sequence upon their mentioning. The overall quality of the research is high, and the explanation of the obtained results allow readers to understand the problematic aspects of the conducted research.

The navigation in the chemical space is hot and dynamic topic as it is directly applicable for the drug design. In this thesis a 3D convolutional neural networks were used to develop a prediction model using single descriptor for the estimation of the bioaccumulation property of the different chemical compounds. Other models were also tested but where less accurate. This finding demonstrate that the solvent density distribution parameter was successfully chosen to build a satisfactory prediction model.

For acute toxicity modelling multi-task deep neural network was used in direct comparison with single-task DNN and other machine learning models. The visualization of chemical space was achieved by the application of the t-SNE method and clusters of toxic compounds were revealed. A specific issue with the toxicity data values was also addressed by a modification of a loss function, which also demonstrates the author’s efficiency in his approaches.

Applicability of the parametric t-SNE approach was successfully tested and applied for the visualization of the GPCR and nuclear receptor ligands as well as for the TAAR1 receptor agonists. This modification would be useful for the visualization and analysis of compounds with the different scaffolds for others pharmaceutically relevant drug targets.

The final part of the dissertation describing the modeling of new compounds from the chemical space is the most complicated and interested part of the work. It described the alternative representation of the chemical structures and its applicability for de-novo generation of chemical compounds with desired properties. Legogram library was created as a tool to generate molecular structures from local regions of chemical space and placed on GitHub. This application would be definitely interesting to the medicinal chemists and bioinformaticians working in drug discovery.

All the methods used are relevant. The newly developed methods and tools were combined into the platform to be used for drug design and development. The high quality of the developed approaches is supported by the publications in peer-review scientific journals.

The results of the research work are published in Q1 and Q2 journals (three and one publications, respectively). 3 papers relevant to the subject of this thesis are also published in Q1 journals. The quality of publications is high. All papers were peer-reviewed.

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*