

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: Nikolay Brilliantov

<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: 29-11-2020</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

Report on the Thesis of Sergey Sosnin “Exploration of Chemical Space by Machine Learning”

The Thesis of Sergey Sosnin entitled “Exploration of Chemical Space by Machine Learning” is dedicated to the elaboration of methods and tools based on the modern Machine Learning (ML) techniques, for the exploration of chemical space. It comprises the combination of the molecular modelling, machine learning, and deep learning methods. The author demonstrates an application of the general methodology to a couple of practical problems. These include the prediction of bioconcentration factor (BCF) with the use of 3D Convolutional neural networks, application of multitask learning to the animal acute toxicity modelling, elaboration of a method for the visualization of broad chemical space and of a Legogram framework for the construction of molecular structures from "building blocks" which preserve chemical validity.

The dissertation is mainly based on four research papers and some other results related to the dissertation topics are published in three other papers. The structure of the dissertation consists of an abstract, table of contents, introduction, chapter, that addresses the materials and methods and chapters reflecting the content of the published papers and conclusion. Additionally, the author includes the Glossary and Additional Resources, which is definitely very beneficial for a reader. The overall quality of the papers is rather high, which is confirmed by the ranking of the journals – six papers are published in Q1 journals.

In Chapter 1 the author provides the detailed motivation of the study and demonstrates the location of the research area in the scientific landscape of the field. Here he also stresses a practical importance of the addressed problems and discuss the overall structure of the thesis.

In Chapter 2 Sergey presents a description of the exploited methods. He briefly sketches the most essential features of the methods in the context of the thesis. He also explains the logical links between different methods and provides a solid argumentation for the choice of the methodology. The level of the detail in the discussion of the methods is adequate; it helps to prepare a reader for the subsequent discussion of the practical application of the methodology.

In Chapter 3, the author discusses a new method for the bioconcentration factor prediction. The main idea is utilizing 3D Convolutional neural networks. Sergey proposes a novel approach for the molecular representation, based on 3D Reference Interaction Site Model (3D RISM). This approach has been applied for the bioconcentration factor prediction and demonstrated slightly better performance than descriptor-based models. However, this approach also shows that analysing solely the distribution of solvent atom density, one can achieve a high computational performance. This study has a potential in further development for the universal 3D chemical descriptors.

In Chapter 4, the candidate compares multitask learning and single-task learning for acute toxicity modelling. He uses the toxicological data from RTECS database. The applicant defines several toxicological endpoints based on different animal species, administration routes, and types of toxicity. The main conclusion that follows from this study is that for acute toxicity modelling the multitask learning overperforms the single task learning. The author also demonstrates that one can emulate the multitask learning by single-task learning, yielding a comparable performance.

In Chapter 5, a new hybrid method for the visualization of chemical space is described. This method is based on parametric t-SNE, where a feed-forward artificial neural networks (ANN) projects chemical compounds to 2D planes. The author demonstrates that this method can provide a meaningful projection, such that similar chemical structures group together. This research has a substantial practical impact as it opens a door for the new visualization tools.

In Chapter 6, the Legogram library is introduced. This library is designed for the generation of chemical compounds with desired properties. Using this library, one can generate valid molecular structures that possess the desired properties. Interestingly, the author combines the generative models based on Legogram with the chemical mapping approach, described in Chapter 5. He demonstrates the ability of sampling drug-like compounds from the specific regions of chemical space.

The overall quality of the Thesis is rather high. Moreover, a couple of new, interesting and scientifically important results have been obtained. The author demonstrates a variety of skills and knowledge from very diverse areas, ranging from molecular simulations, theoretical chemistry and graph theory to deep neural networks and reinforcement learning. The thesis is written in a clear and logical way.

Although I consider the Thesis as a solid scientific work satisfying all the criteria of a PhD qualification document, I still have comments listed below:

1) Why the author uses only data obtained by the 3D RISM simulations and does not apply available experimental data for the structure of the solvent shell around a solute molecule. This could be e.g. data from the mean residence time of water molecules in the shell, measured by NMR, data on thermal neutron scattering, etc. I can admit that the author strives to develop a regular approach, while such experimental data are very sparse. Nevertheless, the additional cross-check with the experimental data would be very beneficial. At least I expect a short discussion in the text about the possibility of the usage of experimental data.

2) It is not clear, why the author believes that the structure of the solvation shell of a chemical substance is a reliable toxicity indicator. Is this a just hypothesis, confirmed by the subsequent analysis? Or there exist some regular studies which prove the importance of the solvation shell structure for the toxicity? In the latter case, the according references should be provided, and their content briefly discussed.

3) Why the graph convolution model shows notably worse results than the baseline model (US EPA)? Although it is difficult to formulate the reason rigorously, a discussion of possible explanations is very welcome.

4) Is the ability of Logogram to produce “absolutely correct” molecules is determined by the “block” nature of this approach, that is, by the fact that it operates with existing chemical groups, which chemical combination produces correct molecules with much high probability that chemical combination of single atoms?

5) The application of the expression (5.1) in page 81 is not sufficiently justified/explained. The author writes: “...where the parameter β is found by binary search to achieve the predetermined entropy of the distribution”. What is the “predetermined entropy”? What kind of entropy is used? Shannon entropy, Reni entropy or may be Tsallis entropy? At least the structure of the Eqs. (5.2)

and (5.3) in page 82 leads to such a guess. The motivation of the usage of Eq. (5.1) and of its practical application should be better articulated.

6) This comment is related to the previous one: What is the motivation of the application of the loss function L in the form (5.3)? Structurally, it looks as a definition of a Complexity of a 2D Pattern with the cross-probabilities of its elements p_{ij} . Although these quantities are the elements of the distance matrix, does it have any relation to the complexity of the 2D structure which results from the mapping from the high-dimensional descriptor space? If such a connection exists, it would be worth to discuss it.

7) From the nature of the 2D patterns that arise in the process of mapping from the high-dimensional descriptor space, one can expect that the distance matrix will be a sparse one. High-dimensional sparse matrixes demonstrate many important properties. These properties of the sparse matrixes may be additionally used to understand the nature and topology of the chemical space. Did the author consider such a possibility?

8) The next comments refer to the exposition of the material:

- The statement in page 51 “only conformers with mutual RMSD (computed on the heavy atoms) more than 0.5Å have been kept” is very vague and is not well explained: What are “heavy atoms”? How do you define “mutual RMSD”? Please reformulate this piece of the text accordingly
- Page 25: The letters in the diagram are hardly visible, the picture is fuzzy, a better quality figure is required
- Page 25: I believe that there is a discrepancy in the notations – the author should use either p_i or Δp_i in Eq. (2.6) and the text below the equation.
- Page 28: In Eqs. (2.16) and (2.17) are the quantities w_{ij} and the same quantities with a hat the same?
- Page 26: I expect that the upper limits in the summation in Eq. (2.7) should be L and R in the first and second term respectively
- Page 20: the awkward expression: ...”we are use to using”.... please correct
- Page 27: after Eq. (2.8) “We can thing”..... please correct
- Page 31 and 38: Please read carefully to remove typos

These comments do not undermine, however, the overall high quality of the Thesis, which satisfies all the requirements of Skoltech for the PhD defenses. Based on this I recommend the Committee to award Sergey Sosnin the according Skoltech PhD degree.

Faithfully Yours



Nikolay Brilliantov
Full Professor, CDISE
Skoltech, Moscow,
Russia
e-mail: n.brilliantov@skoltech.ru
web: <https://faculty.skoltech.ru/people/nikolaybrilliantov>

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