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: Dr. Peter Ertl, Novartis Institutes for BioMedical Research, Basel, Switzerland

I confirm the absence of any conflict of interest	**Signature:** Pete **Date: 16-11-2020**

Reviewer's Report

The Thesis is focused on application of cheminformatics and machine learning methods to analyse relationship between structure of molecules and their properties as well as to exploration of chemical space. The coverage ranges from applications to the development of novel methodologies. These topics are well chosen, all of them are relevant to the drug discovery and pharmaceutical industry.

The thesis begins with a thorough overview of cheminformatics and machine learning methodology. The candidate has shown here good knowledge of application of various statistical and machine learning methods in chemistry and chemical descriptors, as well as previous literature in the field. The detailed description of the methods and shown examples show familiarity of the candidate with these technologies.

The major part of the Thesis consists of 4 chapters focusing on different cheminformatics topics. The Chapters 3 and 4 are more application oriented - bioconcentration is predicted with use of the combination 3D RISM and convolution NNs and multitask learning is applied to predict toxicity - although also here the novel methodology is applied. In Chapter 5 the chemical space is visualized using deep learning and in Chapter 6 a molecular grammar is described that may be used to generate novel, diverse molecules with desired properties.

All these 4 chapters provide novel and scientifically interesting results. They are clearly written, the problems are well stated and the methodology competently explained. I particularly like the chapters about chemical space and the Legogram methodology. These both are very interesting, enhancing the global cheminformatics toolset. I also value that the Legogram code is available as open source at Github.

Below some points for discussion and follow-up studies:

- In the overview of using SMILES in deep learning, discuss also variant of SMILES designed specifically for deep learning: O'Boyle N, Dalke A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. https://europepmc.org/article/ppr/ppr56249

- Chaper 3 - quite sophisticated method was used to generate conformations, would not much simple approach using standard RDKit conformation generator provide results of the same quality? Is the correlation shown in Figure 3.5 for training or test set?

- Chapter 4 "processing of intervals" this is well discussed topic in QSAR - such values are called normally qualified data or censored data, see for example

  https://www.tandfonline.com/doi/abs/10.1080/15459621003609713
  https://academic.oup.com/biomet/article-abstract/66/3/429/232342

- Chapters 3 and 4, quite sophisticated methods are used to predict data that show quite high experimental variation (bioconcentration and toxicity) sometimes even in ranage of tens of percent. It would be good to look into this, learn about variation and reliability of data and look at how this affects the reliability of predictions; it would be useful to show also some representative molecules for these datasets to see what part of chemical space we are dealing with; how the training and test sets have been selected (random or diversity-based) ?

- Chapter 5 - why the tSNE was used for initial dimensionality reduction, this method does not keep well distances between the molecules, has the other dimensionality reduction methods been considered? Would it be possible to skip the tSNE step completely and train the network to map molecules directly to the 2D space?

- Chapter 6 - it would be interesting to see some simple statistics of properties of generated molecules, for example % of purely aromatic or aliphatic molecules, how many of them contain rings, are macrocycles, how many are pure hydrocarbons etc.

- Some small formal changes: literature: change Robert C. Glem to Robert C. Glen, Figure 6.6 - is not clear, atom labels not visible, redraw more clearly

Summary

In summary, I consider this Thesis very good. It is well and clearly written, used methodology and results are well documented. The applicant has shown good knowledge of relevant scientific literature, cheminformatics, and machine learning methods The results obtained are novel, scientifically interesting, relevant to the field of cheminformatics and useful also from the point of view of medicinal chemistry and pharmaceutical industry.

The applicant clearly deserves to be awarded the degree "Doctor of Philosophy".

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*