Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Bogdan Kirillov

PhD Program: Life Sciences

Title of Thesis: Uncertainty Quantification and Neural Network Interpretation for studying CRISPR mechanics

Supervisor: Assistant Professor Maxim Panov

Name of the Reviewer: Marko Djordjevic

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 13-09-2023**

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

CRISPR/Cas-based technologies, a general topic of this thesis, are often mentioned to have revolutionized science. The thesis is well written, with interesting and relevant results. The main focus of the thesis is to address the sequence editing by Cas proteins, where gRNA in complex with Cas effector should recognize desired targets, but evade cutting other places in the genome (so-called off-targets). gRNA design has become a classic problem in bioinformatics, where the thesis research provides important contributions, particularly by exploring novel white-box deep learning approaches to the problem.

The introduction carefully assesses previous neural net (deep learning) approaches to this problem, including an explanation of the advantages of the new approach, which is the subject of the thesis.

The results were published in two credible journals, one with a high impact factor (Nucleic Acid Research) and the other from Nature Portfolio (Scientific Reports).

The PhD candidate is the first author of both papers, and his contribution to the presented research is clear/substantial. The thesis presents research results clearly and succinctly, with carefully done visualizations. The results might be applicable outside the considered problem, specifically in other bioinformatics problems that involve the recognition of short degenerate motifs in the genome (see my general comment no. 2 below).

I have one minor technical and two more general comments. I leave it to the thesis author whether to address the general comments in the thesis or to discuss it as a question on the thesis defense.

Minor comment: I think that the Figure 2.7 from the thesis (bias-variance trade-off) is from "The elements of statistical learning" by Hastie, Tibshiranie, and Friedman. The thesis quoted a website that probably adapted the figure (or a similar variant) from this classic book, so the authors should check that.

General comments:

1) The thesis discusses only neural-net approaches to this problem. However, other machine learning methods have shown significant success in addressing a number of problems, including those in bioinformatics. In particular, I am thinking about the problems in which ensembles of decision trees, such as Gradient Boosting or Random Forest, are employed. Gradient Boosting is also often a winner in different machine learning competitions. Did the author try these approaches in his work to compare performance with deep learning? In general, have other people tried machine learning approaches (other than neural nets) to approach this problem, and if yes, to what success?
2) This problem seems similar to the problem of regulatory element recognition (transcription factor binding site recognition problem), which is an extensively researched problem in bioinformatics. Recently (and even before), a number of machine-learning approaches have also been employed for this problem. In this context, can the author discuss whether some of these methods can be effectively used in the problem considered in the thesis? The other way around, can some of the new bioinformatics methods developed by the author in this thesis be employed for other problems, such as the recognition of transcription factor binding?

In summary, this is a well-written thesis with excellent research results addressing an important topic. I look forward to the thesis defense!

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*