

## 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:** Assistant Professor Ekaterina Khrameeva

I confirm the absence of any conflict of interest

(Alternatively, Reviewer can formulate a possible conflict)

**Date: 25-09-2023**

*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*

**Reviewer's Report**

- **Brief evaluation of the thesis quality and overall structure of the dissertation.**  
The thesis describes an interesting case of machine learning techniques application for studying a biological mechanism - CRISPR mechanics. In particular, the author studies potential off-target event detection and cleavage efficiency estimation. For this purpose, he applies a set of state-of-the-art methods - Deep Neural Networks, Explainable Machine Learning, and Uncertainty Quantification. The thesis consists of three main parts - one describing an application of Capsule Networks approach (and matching the author's Scientific Reports paper), and other two describing gRNA selection and off-target events topics.
- **The relevance of the topic of dissertation work to its actual content.**  
As the thesis presents applications of neural network methods for studying CRISPR machinery, the topic of the thesis is relevant to its actual content.
- **The relevance of the methods used in the dissertation.**  
Methods used in the thesis are relevant and applied correctly, to my best knowledge. The used methods are well described and presented with enough details.
- **The scientific significance of the results obtained and their compliance with the international level and current state of the art.**  
Application of neural networks for studying biological questions is a hot topic at the international level, as well as CRISPR. Therefore, the thesis has high scientific significance. The author applies state-of-the-art methods for studying CRISPR mechanics.
- **The relevance of the obtained results to applications (if applicable).**
- **The quality of publications.**  
High enough to pass the PhD program requirements.

**The summary of issues to be addressed before/during the thesis defence.**

The thesis is clearly written and presents high-quality research work, which passed the review process in respectful scientific journals. Therefore, I only have two questions regarding the presented research:

- 1) In Chapter 6, the author applies 10-fold CV as a method for benchmarking different methods. However, K-fold CV might be dangerous in terms of overfitting. If the dataset has some intrinsic noise or a technical artifact, which is specific to this particular dataset, models might learn this artifact instead of the real biological signal. And because the models are trained and tested on the same dataset (even though on different chunks of it), the model that learns this artifact best gets the best performance metrics. A safer benchmarking strategy would be to train models on one dataset and test them on another dataset. Two datasets obtained in different labs would fit perfectly for this task. If two such datasets are available. Did the author try this approach?
- 2) The author applies advanced neural network models for solving a classification task in this thesis. But did he try a simple logistic regression? I wonder what performance metrics would it show. If it is applicable at all.

The literature review presents a comprehensive description of existing knowledge on CRISPR-Cas system and biological applications of machine learning methods used in the thesis. It is well written and contains all the details necessary for understanding of the research presented in next chapters. Other chapters are clearly written as well and I have no suggestions on possible improvements of their contents.

**Provisional Recommendation**

*V 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*