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: Oksana Maksimenko

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 12-09-2023**

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

This dissertation is devoted to an investigation of CRISPR-Cas machinery as a powerful tool for both fundamental investigations and biotechnology applications. The author focuses on the careful usage of Deep Neural Networks, Explainable Machine Learning, and Uncertainty Quantification. Both articles associated with this dissertation aim to investigate potential off-target event detection and cleavage efficiency estimation of a gRNA-Cas protein complex. The obtained results have made it possible to create a new prediction instrument with a novel axis for off-target cleavage efficiency analysis, enabling a new approach to selecting the optimal gRNA for gene editing.

This is a very interesting and profound work on a relevant topic. All of my comments are of a recommendatory nature and do not question the obtained results, nor do they diminish their importance. Some of the comments represent questions regarding moments that I did not understand in the work. Perhaps some questions arose for me because I am not an expert in deep learning or bioinformatic algorithms. However, a more detailed or more consistent description of them may make them understandable even to a non-specialist.

The dissertation work is presented according to the traditional scheme and includes the following sections: Abstract, Introduction, Background, Thesis objectives, Materials an Methods, Results, Discussion and conclusions, Bibliography. All sections are presented quite fully. The preparation of the dissertation did not raise any questions. Below I will provide comments and observations on individual sections of the work.

Background.

The review is, in my opinion, very well written and illustrated, reflecting the current state of affairs in the field and citing actual sources. It is interesting and understandable even to a specialist in related topics.

From the comments:

In section 2.4, when describing approaches for analyzing binding efficiency, while the main results achieved in the field are given for rule-based approaches, only algorithm examples are given for other groups of methods. It would be great to indicate what these groups of methods have provided, what findings have been obtained thanks to them. Have new important features of sequences reproducibly associated with binding efficiency been identified? Or has the prediction of binding efficiency improved? It is particularly interesting to know how much the quality of prediction has improved using deep learning compared to classical non-neural network-based machine learning. In addition to the example given in Figure 2-6, it seems to me that it would be interesting to characterize in this section the main features that are usually used as input. Are the sequences themselves, or is feature design performed beforehand? Are there approaches that use information not contained in the sequences, such as chromatin accessibility (when assessing the probability of off-target binding)?

Also I came across a certain number of typos (examples: auxilliary -> auxiliary, posess -> possess, "a lot of exaptation cases was found" -> "a lot of exaptation cases were found", "grnas -> gRNAs"), therefore I would recommend checking the spelling and grammar additionally.

Materials and Methods.

This section appeared quite challenging for me to comprehend. While there is no doubt that a great deal of meticulous work has been done, the description of methods in the dissertation itself (as opposed to articles) seems somewhat incomplete. During my reading, I had to frequently refer to the author's articles to clarify unclear points. Each individual method section was understandable, but they did not provide me with an overall understanding of the picture. I believe the main reason for this is that the work described in the two articles is presented in parallel rather than sequentially. If a sequential approach, as in the Results section, had been chosen, it would have made comprehension easier. Many of the unclear points for me were clarified only after reading the Results section. If I were to give more specific feedback, I would highlight the following:

1) At the beginning of the methods, I missed a brief summary map of the project in the form of text or a table, where all the tasks set, models developed to solve them, and the data used for training and testing each of the models would be indicated.

2) The methods begin with section 4.1 on data description and preprocessing. Without an introduction and project map, it remains unclear how section 4.1.3 differs from the second half of section 4.1.1, where data for predicting off-target binding is also described. Moreover, in 4.1.3, information and descriptions of preprocessing already described in 4.1.1 are duplicated. As I understood after reading the results, apparently these descriptions ended up in different sections because they correspond to data for models from different publications, i.e., research on off-target binding using GuideHOM and the anomaly detection method. However, this is not explained in the Methods.

3) It is also unclear until reading the Results for what purposes the data obtained in section 4.1.2 is used. It seems to me that it would have been more understandable if this section followed the description of GuideHOM. In particular, because what is described in points 3 and 4 on page 41 already, in my opinion, relates not to data preprocessing but to model training. Moreover, the word "model" used here is unclear to which model it refers since no model has yet been described at this stage of the work.

4) It is stated that the NRG sequence (resulting in NGG or NAG) was used as the PAM for Cas9, while in the Background on page 23, NGG was described in the example. It would be useful to provide a brief justification for why this particular PAM was chosen in this case. It is also interesting why only sequences from genes were taken into analysis, while possible off-targets in intergenic regions were ignored.

5) With a summary map of the project, it would become more clear that the non-CRISPR-Cas data is an extension of the anomaly detection model to other research objects. And, it seems to me, in the Results, the anomaly detection section would look more harmonious if the application of the model specifically for CRISPR-Cas was described first. In the current version of the work, where the goals regarding CRISPR-Cas are clearly defined, the description of data and results for unrelated biological objects, such as photographs of skin lesions, is somewhat discouraging without any introduction or explanation.

6) Section 4.2, which describes the tasks, seems to reflect only the part about GuideHOM, and does not include a section about anomaly detection. At least that's the impression created. But the objectives must include all the work done in the dissertation. Maybe it was worth emphasizing in subsection 4.2, related to off-target detection, that this problem was solved by both GuideHOM and the anomaly detection method?

7) In the methods and results there are references to the Supplementary, but it is not always indicated which article is meant.

Results.

The results of the current dissertation are presented in three parts. First one describes potential off-target event detection through investigation of inequalities in coupling coefficients within Capsule Networks. The second part describes Uncertainty-aware and Explainable Machine Learning for gRNA selection. The third part presents the diversity of an off-target cleavage space that was discovered through application of Uncertainty Quantification.

Inequality in capsule networks for detection of potential off-target events.

1) In this section, I did not see any information about the training schemes used. It is unclear how the data was divided into training and testing sets, and whether a validation set was left out for optimizing network parameters. It would also be interesting to see the amount of data in each set and the number

of network parameters used. This information may be available in Supplementary materials, but it seems important enough to mention in the main text.

2) It is not entirely clear what data was used in section 5.4 (described in 4.1.3?).

Uncertainty-aware and Explainable Machine Learning for gRNA selection

1) The phrase " We use the same data for training and testing in cases where the actual training and testing sets are available" sounds like the model was trained and tested on the same data. However, it seems that the author meant that the training and testing sets were borrowed from publications with corresponding data. If this is the case, it would be better to rephrase the sentence. It would also be interesting to include the total amount of data in each case in Table 6.1 (perhaps in parentheses). It is not entirely clear to me what the dashes mean in this table. If the dashes are for the validation set, does that mean these data were only used for training the model but not for testing?

2) It would be helpful to provide explanations for the abbreviations CNN and RNN. Additionally, in section 4.5.1 describing GuideHOM, it would be useful to provide the abbreviations CNN for convolutional preprocessing and RNN for LSTM, if I understood correctly what they correspond to. This would make it clearer where the different results come from. I also did not understand why CNN preprocessing was used for some GuideHOM datasets and RNN for others, according to Table 6-2. Was this choice based on higher quality metrics?

- Figure 6-2. "The numbers at the plot (A) and at the plot (B) denote the same gRNAs." Is there a typo here (C instead of B)? If not, I do not understand what the numbers refer to. It is also unclear to me what is meant by probability density on (C).

- Figure 6-3. I think the upper and lower panels have different messages, and it would be more logical to split the figure into two.

- Figure 6-4. "at the left hand side (E) and (G)" is missing "for Cas12a"?

Uncertainty Quantification highlights the hidden diversity of off-target events

- Pages 74 and 81 compare the model with the Jost dataset. Am I correct in understanding that the same result is described in both chapters (0.625, as compared with 0.617)? I do not understand why this is described in two different chapters, and I am not sure what task was solved for this dataset - assessing binding to off-targets or true targets?

Discussion and conclusions

The author conducts a correct analysis of the data obtained in the work and draws fully justified conclusions based on them. I would have liked to see more specific indications of possible limitations of the study and possible ways to improve the model in the discussion. Specifically, it would be interesting to hear the author's position on such issues:

1) The paper emphasizes that the developed models only take into input sequence information. Could future models incorporate information about chromatin accessibility, or distribution of specific genomic elements, or other useful information not contained in the sequence?

2) If I'm not mistaken, in capsule networks it is possible to investigate which features individual capsules are responsible for. Could this be an additional way to investigate model interpretability?

3) Most of the testing schemes I saw in the paper did not include cross-validation, although I believe this is a fairly typical practice. Is this due to the long training time of the models? It would be interesting to explain this in the discussion.

In general, the work makes an excellent impression in terms of the volume of modern methods used and the mentioned comments do not reduce its value. The author of the work obtained new important results that make it possible to more effectively use Explainable Machine Learning and Neural Network Interpretation to increase the efficiency and specificity of gene editing. The conclusions of the work are well substantiated by the results obtained and do not raise doubts. The results of the work will be useful in scientific laboratories and biotechnology involved in gene editing. The work was published in reputable foreign journals with a fairly high impact-factor. The work meets all the requirements for PhD dissertations, and its author deserves to be awarded the required degree.

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*

Oksana G.Maksimenko, PhD,
Deputy Director, Institute of Gene Biology RAS,
Head of the Center for Precision
Genome Editing and Genetic Technologies
for Biomedicine IGB RAS

12/09/23