Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Bogdan Kirillov

PhD Program: Life Sciences

Title of Thesis: Uncertainty Quantification and Neural Network Interpretation for Studying CRISPR Mechanics

Supervisor: Prof. Maxim Panov

The thesis document includes the following changes in answer to the external review process.

Dr. Nikolai Kulemin

> However, in the text of the work there is no information of the hardware characteristics of the computer equipment used. In such works, they usually form a separate chapter in the "materials and methods" section, where all software and hardware characteristics are described, at least briefly. If there is no description of the characteristics, it can be very hard to reproduce the work published.
I have added the detailed information about hardware and software in "Description of hardware and software used in the studies" – a new section in "Materials and methods".

> Another nuance is that when comparing data obtained by different consortia, there is not only directly measured error, but also summarized error in conclusions, which includes the type of protocol used to generate the final data loaded into open databases. Of course, when solving a specific task, such indirect errors can be neglected, but most likely they would be worth mentioning in the text.
During the study design, I have made substantial effort to evade the problem of comparing two different model-unrelated sources of errors (e.g. preprocessing, data source and so on) instead of comparing two models. To do so, I have tried to reproduce all pipelines that I compare my works to as closely as possible, so the only difference in the pipelines are the machine learning models. That procedure minimizes the indirect component of the error.

Dr. Mikhail Gelfand

Considering that the suggestions provided by Dr. Gelfand during the current stage of thesis review mainly focused on preferences for writing style and substantial editing remarks have already been addressed after the pre-defense, I have made the decision to leave the text as it is.

Dr. Ekaterina Khrameeva

> In Chapter 6, the author applies a 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 ar fact, 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 be perfect for this task. If two such datasets are available. Did the author try this approach?
The 10-fold CV study was done as a response to a reviewer from the journal, so it is not really what I would prefer to do. Cross-dataset testing would work perfectly as a good benchmark strategy, but unfortunately there are no two openly accessible datasets that correspond perfectly in terms of cell lines, preprocessing, sequencing pipeline and other significant bias-introducing factors. The lack of such datasets makes cross-dataset testing, as Dr. Khrameeva suggested, unfeasible in practice.

> 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 it would show. If it is applicable at all.
The baselines (including logistic regression) were covered in the previous studies upon which I base my work. I personally did not do any baselines.

Dr. Marko Djordjevic

> 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.
I have corrected the citation.

> 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?
I have personally not tried classical methods like Gradient Boosting or Random Forest because I have considered it not necessary since there is a large corpus of works that applied these methods to the same data I have used (“Background” chapter, “Evolution of bioinformatical tools for gRNA activity evaluation” section) and these results were surpassed by deep learning methods.

> 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?
This question will be discussed in the presentation.

Dr. Martin Takac

>It appears that some captions (both tables and figures) are not finished with “.”. I recommend always cast them as a sentence and hence finish the sentence to avoid the inconsistency.

>In multiple places, after a bold word follows a new sentence that sounds incomplete. E.g. Page 21. Last paragraph, sentence after "Inference".
>Figure 2-6. This Figure is from [77]. I would suggest clearly making a note of this by explicitly saying that this is not motivated by [77] but is the copy from it. E.g. write (image source [77]).
>Section 2.6. You accidently added an extra line in your LaTeX source that created an intent.
>Figure 2-7 – add (image source [4]) in caption.
>Page 32 – after stating Uncertainty Quantification add (UQ).
>Page 54. I feel you should remove the empty line in your LaTeX source above the equation for $O_i(X)$. Also, I would define Sigmoid, BatchNorm and Linear as operators in LaTeX.
>Page 62. Remove "below" from "is shown on the Fig 5-1 below".
Corrected throughout.

>I believe (4.4) is incorrect.
>Page 50. Are in the definition of $P(Y=C_j)$ missing norms for v_j and v_ks'?
For some reason, latex did not recognize the \norm symbol from mathtools and showed nothing in (4.4) and page 50 equations instead. I have corrected that.

>Page 26, 2nd sentence after the "Rule-based systems" should be reformulated
I have rewritten two sentences after the 'Rule-based systems' as '*Within the rule-based framework, cleavage efficiency of gRNA is inferred using a single rule or a set of rules that are relatively easy to compute and self-explanatory. For example, the first rule-based systems, introduced by Gagnon [70] and Wang [71], consider the GC-content of the gRNA as a primary factor influencing cleavage efficiency.*'

>Page 55. Align equations.
Since the current center alignment of equations is the default in the Skoltech latex style file, I have decided to leave the equations as they are.

>Section 2.7 – rewrite (for example here [115], here [116] and here [117]).
I have rewritten this part as '*Both in various literature, such as Gilpin et al.[115], Ribeiro et al.[116] and Linardatos et al.[117].*'.

>Section 2.7 – rewrite (for example here [115], here [116] and here [117]).
I have rewritten this part as '*Both in various literature, such as Gilpin et al.[115], Ribeiro et al.[116] and Linardatos et al.[117].*'.

>Page 50. To which "Supplementary note" are you referring?.
I have removed the reference to the supplementary note since all hyperparameters are explained in the text and no such note is needed. Most likely during drafting the first version of the text, I planned to do a supplementary note, but then included everything in the main text and forgot to remove the reference.

Dr. Oksana Maksimenko

> Figure 6-4. "at the left hand side (E) and (G)" is missing "for Cas12a"?
> It is not entirely clear what data was used in section 5.4 (described in 4.1.3?),
> In the methods and results there are references to the supplementary, but it is not always indicated which article is meant.
> 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.

> 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 would recommend checking the spelling and grammar additionally.
Corrected throughout.

> Pages 74 and 81 compare the model with the Jost dataset. Am I correctly 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?
The results on page 74 introduces the results for Jost et al dataset, which solves assessment of binding to off-targets (section "GuideHOM solves off-target cleavage regression with acceptable confidence intervals"). Page 81 references this result as a ground for our approach applied to off-target study (*"The consistency and slight but significant superiority of results obtained with our method compared to those in the original study ($r2$ value of 0.625 versus 0.617 for the Jost et al. model, see details above) supports the utility of our approach."*). To clarify this, I have rewritten this sentence as follows: *"The consistency and slight but significant superiority of results obtained with our method compared to those in the original study ($r2$ value of 0.625 versus 0.617 for the Jost et al. model, as described in section 7.1) supports the utility of our approach."*

> 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.
Yes, the upper part explains that 2D visualization of capsule features shows the gradient of cleavage efficiency regardless of visualization method and the lower part describes the performance of the models on Jost et al dataset. Considering that these two parts were joined in the paper for brevity, I have decided not to split the figures and leave them as they are.

>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.
I have added the following to the subsection about rule-based system: '*Therefore, the algorithm that Wang and Gagnon's groups offer, may be summarized as follows: if GC-content is higher than 0.5 and guanine is present in seed region and/or PAM, then the gRNA is active. This rule provides a natural baseline for gRNA classification by efficiency, but it is not enough for experimental design as GC-content does not correlate well with efficiency – there is a significant positive correlation but also a lot of noise (for example, Figure 2C from [72].). Application of machine learning helps address the problem of accuracy by building constructs that correlate with cleavage efficiency better and have narrow confidence intervals. '.*

>1) At the beginning of the methods, I missed a brief summary map of the project in the form of text or 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.
>) 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. 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.

To improve readability and understandability of the methodology chapter, I have implemented the following changes to the text:

1. I have moved the 'Problem setups' section at the beginning of the methodology chapter;
2. I have added the links for the appropriate sections of Material and Methods into 'Problem setups' section:
    a. *'In the current dissertation, the problem of off-target event detection is solved by the anomaly detection pipeline described in sections 4.2.3, 4.4 and 4.5 of "Materials and Methods" and the results are presented in chapter 5. The solutions are based on data that are described in "Description of the dataset and preprocessing routine" (subsection 4.2.2) of "Materials and Methods".'*
    b. *'In the current dissertation, the problem of cleavage efficiency prediction is solved by the GuideHOM pipeline described in sections 4.6, 4.7, 4.8, and 4.9 of "Materials and Methods" and the results are presented in chapters 6 and 7. The solutions are based on data that are described in sections "Description of the dataset and preprocessing routine" (subsection 4.2.1) of "Materials and Methods" chapter.'*
3. I have moved the section about off-target data before the image data.

Those changes convert the 'Problem setups' into the requested brief summary map of the project.

>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 are 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.

I have moved the 4.1.3 into a new section 4.2.2 'Data collection and preprocessing for CRISPR-Cas9 Off-target event detection via anomaly detection' and indicated explicitly that sections 4.2.1 and 4.2.2 belong to different tasks.

>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?

Section 4.1 (former 4.2) reflects both papers. Detection of off-target events is solved only by the anomaly detection pipeline and is not solved by GuideHOM. GuideHOM estimates cleavage efficiency, including cleavage efficiency for off-targets (solves the second task) – it is not the same as classifying potential off-target sites into 'probably an off-target/harmless' which is the event detection done by anomaly pipeline. GuideHOM may be used for the event detection if we supply it with a decision rule e.g. 'predicted cleavage efficiency for a potential off-target is higher than 0.15 -> probably an off-target' but in this dissertation we separate event detection and cleavage efficiency estimation and solve them by different methods.

I have added the following line to 'CRISPR-Cas off-target event detection': '*In the current dissertation, the problem of off-target event detection is solved by the anomaly detection pipeline.*'

>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?

Yes, it is possible to do generally, but one of the approaches introduced in this dissertation makes it rather difficult. There is no obvious way to explain contributions of individual capsules in routing-less network GuideHOM. Nevertheless, it is possible for generic capsule networks used in the anomaly detection study. I will elaborate this in the presentation.

> 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.
Added the explanations to the “Abbreviation” section.

> 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 publication with corresponding data. If this is the case, it would be better to rephrase the sentence.
I have rewritten this sentence as ‘*We use the same data for training and testing as the original works in cases where the actual training and testing sets are available and the same train-test split ratio in other cases.*’.

> 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?
Dashes indicate that the validation set was not used in the original study – in this work I tried to follow the original training routines as closely as I could. In the paper, the table contained the train-test-val splits from the original studies and some of them did not use train-test-val splits (using CV instead), while in the dissertation it should have contained my train-test-val splits. I have addressed this discrepancy by adding information about splits I have used according to the source code at www.github.com/bakirillov/uace.
I have added the data amounts for each case.

>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?
I have done both kinds of preprocessing for all models (and both types of loss function), the results are included in the Supplementary Table of the NAR paper. For Table 6-2, I used my best models to compare with the original papers, some of the best models had CNN preprocessing, and some had RNN one.

>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.
There is no real need for cross-validation in this task because of the data nature (sequences that may have phylogeny-like structure and thus introduce bias in the CV process if we do 10-fold CV). I have actually performed a 10-fold CV while answering a similar question of the reviewer, the results are given in the Supplementary Table of the NAR paper.

>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 the case.
I have added NRG instead of NGG PAM in the picture.

>It is also interesting why only sequences from genes were taken into analysis, while possible off-targets in intergenic regions were ignored.

I have decided to exclude intergenic regions from the study of off-target effect diversity because the effect of such off-target events if they happen would not be direct – the genes contain most of the functionality and preventing the disruption of the genes would be of most importance during experimental planning. That said, it's important to note that while intergenic regions are excluded from off-target searches in the current work, they are not devoid of function. Some intergenic regions may contain elements involved in gene regulation, such as enhancers and silencers, and others are transcribed into non-coding RNAs with various roles. Studying intergenic off-targets is a promising area of future research – for example, comparison of intergenic off-targets to ones in the genes in terms of uncertainty and activity. Considering that the pipeline introduced in the current dissertation relies heavily on usage of CasOffinder and/or FlashFry for its search for potential off-target sites and the neural networks are used to filter the outputs of CasOffinder or FlashFry, the decision about whether to include intergenic regions is essentially left on the end user of the pipeline – the end user is free to include them or not.

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?
Future models indeed can include non-sequence information, but this work was deliberately focused on using sequential features and considering chromatin accessibility and other information is out of scope of the current study.

>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)?
I have added a short description for features used as input into the Section 2.4 'Evolution of bioinformatical tool for gRNA evaluation' as follows:
'Most deep learning approaches use one-hot encoding of the sequence [14, 76, 77] while classical machine learning approaches may also rely on feature design (e.g. geCRISPR [72] uses thermodynamic properties like melting temperature of RNA, and sgRNA-DNA binding energy, and also structural features like Shannon entropy). However, some deep learning approaches also use sequence embeddings (e.g. [95]) and may incorporate additional non-sequential information like chromatin accessibility (e.g. chromatin-based model of [14]). A thorough study of input features can be found in a review work of [96], in the current work we have used only one-hot encoding for sequence and only sequential features due to the initial focus on studying the sequence feature influence on cleavage efficiency.'

>In this section [Chapter 5], 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.
The experimental design for this part requires no fixed train-test ratio and no validation set. The splits are defined as described in the section 4.2.3 of methodology chapter. I have added the following to Chapter 5: *"We use Adam optimizer [161] with default settings and train-test splits defined according to Section 4.2.3 of Chapter 4."*

>It is also unclear to me what is meant by probability density on C.
The probability density plots on Figure 6-2C denote the distributions of cleavage efficiency of the gRNAs in question. I will explain this in detail in the presentation.