Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Satyarth Mishra Sharma

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Predicting phenotype from genotype with machine learning

Supervisor: Professor Dmitry Yarotsky

Name of the Reviewer:

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 16-12-2024**

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

The candidate has demonstrated the application of machine learning across multiple domains in computational genomics and proteomics, including polygenic risk scoring, ancestry prediction with federated learning, protein stability prediction, and accurate variant calling. While the breadth of work is satisfactory, the chapters feel somewhat disconnected. This lack of cohesion is not uncommon in Computational Biology. However, some effort to integrate these works into a cohesive narrative would enhance the thesis's overall impact and readability. Below, I summarize my observations and provide suggestions for improvement.

The work on complex traits is intriguing, but I noticed that DNA embedding models, which are increasingly relevant, were not explored. The study appears to rely on a limited set of standard machine learning techniques, with deep learning methods conspicuously absent. Clarifying the rationale behind the choice of methods and detailing the cross-validation and hyperparameter tuning settings would add transparency and rigor to the analysis. If standard algorithms are used, it is advisable to leverage libraries like AutoML to ensure comprehensive model evaluation. Additionally, incorporating explainability techniques such as LIME, SHAP, or saliency maps could significantly enhance the interpretability of the results. For traits with low prevalence, metrics such as Positive Predictive Value (PPV) and Negative Predictive Value (NPV) should ideally be reported alongside AUC for a more balanced evaluation.

The chapter on federated learning for ancestry prediction addresses an interesting and timely problem. However, the study appears to have tested only a single, standard federated learning algorithm. Exploring and comparing multiple algorithms would have provided deeper insights and strengthened the conclusions. It would be helpful to understand why this broader exploration was not undertaken.

In the protein solubility work, the UMAP visualization in Figure 4-1 appears unusual and could benefit from further clarification or adjustments to improve its interpretability.

Lastly, the chapter on variant calling presents ideas that are not clearly articulated, making it challenging to follow the details of the approach and results. Greater effort to explain the methodology and findings would enhance this chapter’s clarity and accessibility.

While the candidate has accomplished a significant amount of work, bridging the gaps through improved writing and a discussion of future research directions would add cohesion and elevate the thesis's overall quality.

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*