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	**Date: 08-12-2024**

Reviewer's Report
The thesis covers several topics linked by the theme of predicting structure and function given sequence. At that the structure of the thesis is somewhat uneven: the review in Chapter 1 introduces prediction of phenotype from genotype that is the topic of Chapter 2. Other chapters have their own introductions. The presented results are interesting and novel, as well as the developed algorithms. The main result of Chapter 2 seems to be negative: accounting for non-additive interactions does not add much to the prediction quality; the author's program XGBoost sometimes, but not always outperforms the existing program Snpnet, while comparisons to published results are not very informative due to differences between the testing schedules. On the biological side, these results do not mean that all epistatic interactions are negligible; some gene pairs might still demonstrate strong epistasis (the extreme case being synthetic letals), and, given the enormous amounts of already processed and prepared data, it might be interesting to search for such pairs directly. This is not a criticism per se, as the author's model seems to be very useful in several different settings, rather a suggestion for further analysis. Chapter 3 presents an approach to federated learning. It is an important issue, as major data holders are increasingly concerned with the safety of personal data. The phonotype prediction algorithm is different from the one developed in Chapter 2; the candidate might wish to include the discussion of perspectives of federating his XGBoost. Introduction to Chapter 4 on protein structure is rather short and surprisingly lacks any mention of AlphaFold. The latter does not seem to perform well on prediction of mutation consequences, but this should

have been mentioned as a possible motivation to continue working in this area. Aside of that, the chapter is interesting, even if somewhat distant from the main theme, and the reported results quite promising, especially the performance on unseen data (probably due to careful train / test / holdout splitting of data). The author mentions that "the embeddings encode a rich representation of protein structure and thermodynamic properties"; it might be interesting to analyze what features are learned by the network attempting to make a prediction.

Chapter 5 is dedicated to a technical, but important problem of genotype calling, using the XGBoost approach of Chapter 2 and comparing it to logistic regression.

The final Chapter 6 presents the author's thoughts on social, economic, ethical problems surrounding the precision medicine. It makes an instructive reading.

What I feel missing, is a common discussion bringing under one umbrella all diverse problems, generalizing the lessons gained, and providing a concise list of conclusions.

The results are published in two papers (of which ref. 1 is a first-co-author one) and two preprints (ref. 3 in "Published papers" is a preprint); impact factors of both journals are greater than 2, hence the publication requirements are met. They were also presented as posters at three international conferences including ISMB in 2020 and 2023, so this requirement is also fulfilled.

The thesis is generally well written, although I would prefer a somewhat more academic style, especially in the Introduction. However, I have some editorial comments (the list is not exhaustive due to the timing / deadline issue).

§ 1.1.1 – "Mendel's law's" – which law?

§ 1.1.1 – there is nothing Lamarckian in Darwin's notion of heredity: he has assumed random variation, not inherited change due to training.

§ 1.1.3, 1st para. – the second ("In this model…") and third ("The model…") sentences are almost the same.

§ 1.1.4 – eye color is not monogenic.

§ 1.1.4, "Most non-communicable disorders with a major … impact … highly polygenic" – I'd appreciate a reference.

§ 1.2.1, SNPs "are the most common type of human genetic variation" – again, this depends on definitions (chromosome rearrangements may involve more genetic material).

§ 1.2.1, SNPs "occur once per 1000 nucleotides" – two individual genomes would have this level of difference, polymorphic positions are more frequent.

§ 1.2.1 – non-coding CNPs do not alter proteins but may influence gene expression and hence also affect the phenotype.

§ 1.2.1, last para. "information … is orders of magnitude larger" – how is the information measured?

Tables 2.2 and 2.3 are difficult to use for the comparison of methods and parameters; it might be better to group the lines by the phenotype.

In many places (first time on page 20) the author writes about "non-linear interactions" while the epistatic interactions are better characterized as non-additive.

The L_1 regularization is written in at least three different ways: *L1*, l_1, ℓ_1; performance is measured by r^2 and R^2. The notation should be uniform.

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*~~