Skoltech

Skolkovo Institute of Science and Technology

# Jury Member Report – Doctor of Philosophy thesis.

**Name of Candidate:** Polovnikov Kirill

**PhD Program:** Physics

**Title of Thesis:** On connection between sparse graphs and hyperbolic geometry

**Supervisor:** Professor Mikhail Gelfand, Skoltech

Professor Sergey Nechaev, Interdisciplinary Scientific Center Poncelet

**Name of the Reviewer:**

| | |
|---|---|
| I confirm the absence of any conflict of interest<br>(Alternatively, Reviewer can formulate a possible conflict) | **Signature:**<br>**Date: DD-MM-YYYY** |

*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 dissertation has a non-standard structure. The text starts with an introduction that provides some preliminary information about random matrices and eigenvalues distribution, In particular, the spectrum of sparse random matrices was analyzed. This part is well written but the references 59-62 are empty.

The next chapters are just a copy of the papers. The chapter "order and stochasticity in the folding of individual Drosophila genomes" represents the submitted manuscript where the figures are given at the end of the text. This form of presentation is inconvenient for the reader.

Here I will focus on the "biological" chapters. In the chapter "Order and stochasticity in the folding of individual Drosophila genomes" was described the results of the single-cell HiC experiment. The experimental data and part of the bioinformatics analysis were performed by other co-authors. The HiC data presents information on contacts of chromosomes and it can be presented as a large sparse matrix. Kirill Polovnikov applies spectral methods to the HiC data analysis. In particular he shows that the experimental sparse Hi-C matrices are not equivalent to random realizations of the configuration model graphs with conserved contact probability. The HiC matrix has a specific block structure. The blocks in this matrix are defined as a topologically associated domains (TAD). To find the TADs algorithmically Kirill Polovnikov proposed the method to annotate TADs in sparse Hi-C matrices, based on the non-backtracking walks. He demonstrated that the found domains are biologically significant, namely, that they exhibit high persistence of boundaries across the ensemble of single cells and the boundaries are enriched with various epigenetic markers. But here I have to give some remarks.

1. One should take into account that the single-cell data presents only a part (about 20%) of the real contacts. In this paper, a subsampling approach to the bulk (many-cells) data was applied and the robustness of the TAD calling method was shown.
2. It is impossible to get simultaneously the epigenetic data and HiC data on the same cell. The data of the epigenetic markers were obtained on different cells and this is a population data.

The last chapter, "Non-backtracking walks reveal compartments in sparse chromatin interaction networks" also devoted to the HiC data analysis. Here Kirill Polovnikov developed the polymer stochastic block model and the non-backtracking flow operator, neutralized to the polymer contact probability. He generalizes the modularity operator to take into account the scaling. He established the connection with the generalized modularity and have proved that partition of a chromatin network into two compartments by means of the leading eigenvector of the proposed operator responds to the maximum entropy principle. He also realized the approach to real sparse data and have demonstrated the biological significance of the annotation by profiling the single-cell domains using the GC content and the leading eigenvector of the population-averaged Hi-C matrix.
I have some questions

1. Does the author compare the compartmentalization for different cells and in the bulk data?
2. Does the author compare the compartmentalization with other biological features, such as gene activity and epigenetic data?

3. In first papers on the HiC data analysis, two compartments were foud, but in some later publications, six compartments were postulated. How the author can comment on this?

Kirill Polovnikov got some significant results. The results are well published in 5 papers and one paper is submitted. The papers are published in international journals with impact factor up to 4.1, and one paper is submitted to Nature Communications

**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*