Nalecz Institute of Biocybernetics and Biomedical Engineering
Polish Academy of Sciences

Dr hab.
Jan Poleszczuk

Warsaw, 23 September 2025

Review of the doctoral dissertation

mgr inż. Ruby Khan entitled *Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems*

Advisor: Dr hab. inż. Krzysztof Puszyński, prof. Pol. Śl.

1. Scientific problems considered in the dissertation

The subject of the dissertation is the proposal, implementation, and preliminary verification of a computational method for the automatic acquisition, organization, and analysis of data, aimed at facilitating the identification of key elements within complex biological networks. For this purpose, the Author employed Python, a programming language widely used in the scientific community and offering an extensive range of libraries for various research applications. The developed solution includes the implementation of methods for retrieving and processing data from different omics repositories, whose integration can subsequently be analysed using established approaches to network complexity analysis. For the preliminary verification of the method, i.e., the correct identification of key regulatory nodes, the Author applied it to two well-known biological networks: the cell cycle signalling pathway and the MAPK signalling pathway.

The problem of biological network analysis, as well as the attempt to develop a novel computational method, is of great importance for understanding the mechanisms underlying the development of diseases such as cancer, which in the long term may find applications in medical practice. For many years, researchers have been seeking new therapeutic targets, the relevance of which is largely determined by their role within a given signalling pathway. Moreover, in the era of combination therapies acting simultaneously on different elements of regulatory networks, a deeper understanding of the functioning of specific biological pathways may contribute to increasing the effectiveness of such treatments through the optimal adjustment of drug dosages. Taking this into consideration, it should be emphasized that the research problem addressed in the dissertation is timely and aligns well with the current directions of scientific work. The subject matter of the dissertation clearly falls within the field of Biomedical Engineering.

The dissertation distinguishes three research hypotheses concerning the possibility of identifying key elements in regulatory networks. In my opinion, however, the way they are

formulated is not fully appropriate, as they reflect more the description of the obtained results than the definition of research directions. In particular, the third hypothesis combines a statement regarding the stability and generalizability of the proposed method with a percentage result for two exemplary regulatory networks under consideration. Unfortunately, because of how the dissertation is written, it is unclear whether the hypotheses were verified—this will be discussed further in the next sections.

2. Structure of the dissertation

The dissertation consists of abstracts in English and Polish, acknowledgements, a table of contents, a list of the Author's publications and conference presentations, an introduction, a chapter describing the research methodology, a chapter presenting the obtained results, a discussion, a bibliography, and supplements containing the scripts used in the research.

The introduction briefly outlines the history of the field of systems biology and presents the motivation for undertaking the topic, along with a description of the research hypotheses. The next two chapters are dedicated to the discussion of the methodology and the presentation of the results. In the discussion, the Author summarizes the obtained findings and briefly addresses their consistency with the work of other research groups.

The overall structure of the dissertation is appropriate, although the materials included in the supplements could be better organized, particularly by clearly separating the scripts from other tabular data. Furthermore, the scripts could be presented more coherently in terms of content, for example, structured according to individual classes and components.

3. Analysis of sources

The bibliography comprises 130 entries, with entries [112] and [113], [98] and [114], [117] and [102] being duplicated (these are only detected repetitions). I estimate that only about half of the cited works are related to the topic of the dissertation. The reader may encounter numerous entries concerning, for example, epidemic modeling, artificial neurons, pharmacodynamics, or predator-prey systems, which are not relevant to the presented research.

Among the interesting but not necessarily related references, entries [92], [92], and [93] can be highlighted, which discuss, respectively, the optimization of plant growth processes, epidemic waves, and climate models. At the same time, the list of publications directly related to the dissertation's topic, in my opinion, does not sufficiently include works introducing key issues central to the study, such as gene expression, transcription factors, or graph theory.

In summary, the selection of the bibliography does not allow me to positively assess the Author's familiarity with the international scientific literature in the area addressed by the dissertation.

4. Originality and strengths of the dissertation

Undoubtedly, a major strength of the dissertation lies in the chosen research topic and the attempt to provide a comprehensive solution to the problem - from data acquisition to the presentation of results based on the proposed analytical methods. The work employs relatively standard methods; however, the way they are applied and the results obtained appear to be original. These results have not yet been published, but with appropriate additions and revisions during the review process, they have the potential to be accepted in international scientific journals.

Unfortunately, the manner in which the results are presented in the dissertation does not currently allow me to confidently determine which of the reported findings can be considered fully correct and particularly valuable from a scientific perspective. A detailed description of the reservations is provided in the subsequent sections of this review.

5. Weaknesses of the dissertation

The weaknesses of the dissertation include:

- The Author clearly states that her goal is to develop a comprehensive computational protocol. However, in the dissertation, this protocol is not presented in a manner that would allow the reader to fully implement it or understand it completely – only fragments of code and general descriptions are provided, which, despite the Author's claims, do not constitute pseudocode. For such a stated objective, it is crucial today to provide a repository with the code and full documentation enabling others to apply and understand the proposed methods.

- The Author repeatedly emphasizes that the protocol is automated. However, according to the information presented in the dissertation, this does not appear to be the case. There are stages of manual modification and supplementation of the analyzed networks, and the code contains hard-coded information regarding specific data sources and formats. The method does not seem readily applicable to the analysis of other data sources, nor easily usable without conducting additional and complex analyses.

- The introduction (Chapter 2, Literature Review) contains very general information and does not provide the reader with the basics of the key issues addressed in the dissertation. While a general overview of the history of systems biology is presented, there is no coverage of topics central to the main subject of the work, such as the definitions of genes, proteins, transcription and translation processes, or the basics of graph-based modeling. The introduction discusses predator-prey systems, Michaelis-Menten kinetics, pharmacodynamics and pharmacokinetics, as well as neuron models – topics not directly related to the dissertation's focus. Furthermore, the dissertation does not provide a

detailed description of the data sources used by the Author, i.e., what each dataset contains, how it has been curated, and which analyses were performed, which is necessary to justify their use.

- It is not clearly justified in the dissertation that, as the Author proposes, elements from different omics sources can be merged into a single large network. These elements are distinct entities, and simply placing them into one table might not be correct – the functional aspects of the considered elements are crucial. This is a key assumption of the dissertation and requires strong supporting argumentation.

- Overall, the dissertation is not well written, with numerous errors, some of which are highlighted in later sections of the review. In its current form, I cannot confidently determine what has been done and how, or whether each step is correct both technically and scientifically. Consequently, it is unclear whether the original research problem has been effectively addressed.

Unfortunately, the originality and strengths of the dissertation do not outweigh its weaknesses.

6. Detailed substantive and editorial comments

In the dissertation, there are errors, omissions, or ambiguities in elements crucial for the substance and overall clarity of the work:

6.1. Page 13: “This scalable computational pipeline” – In what sense is the proposed computational protocol scalable? How was this ensured and verified?

6.2. Page 13: “python based networks” – What exactly are these networks?

6.3. Page 13: Incorrect spelling of Cyoscape (should be Cytoscape).

6.4. Page 14, section 1.4.3: “constracted” → constructed.

6.5. Page 14: Incomplete sentence at the end of the paragraph: “Using iterative modeling…”.

6.6. Introduction: References are often incorrect. For example, on page 16, it reads “The Lotka-Volterra model [7]”, where [7] refers to an extended predator-prey model with additional terms. The same sentence continues with references to the Hodgkin-Huxley model, Michaelis-Menten kinetics, and epidemic models, yet only reference [8] is provided, which addresses only the neuron model. Similarly, on page 17: “…enzymatic reaction rates were understood thanks to the establishment of an enzyme kinetics model [10]”, where [10] is titled “Densely packed matrices as rate-determining features in starch hydrolysis.”

6.7. Page 18: “The extraction of mathematical models […]” – This statement is unclear.

6.8. Section 2.1.2 refers to the period 1980–1990, yet the Author cites and discusses a 2023 publication ([14]).

6.9. Section 2.5.1 is not a summary of the previous text and does not clarify the “Transition to methodology”.

6.10. Page 26: What are high-throughput databases?

6.11. Page 26: “Primary purpose of this project is to develop an integrated network model that can understand the key connections […]” – How can a model “understand” these connections?

6.12. Table 3.1: The Author claims the table contains pseudocode, but this is not accurate; the method cannot be reproduced in any programming language based on the provided information.

6.13. Table 3.1: Why is “Database Design” included under “data processing and integration”? Typically, databases are designed before data processing and integration.

6.14. Table 3.1 and Figure 1: Consistency should be ensured (e.g., Figure 1 lacks Step 10, which is present in Table 3.1).

6.15. Page 31: It is stated that error handling is implemented in the code (“Error-handling”), but no such functionality is visible in any listing. Even basic error handling for connection retries is missing.

6.16. Page 31, “Processing Downloaded Data”: It mentions CSV files, yet some data are downloaded as tar.gz archives.

6.17. Page 31: Table 3.2 seems redundant. Additionally, section 3.2.4 mentions using the KEGG component from Pathways Commons, but this is not reflected in the supplementary code.

6.18. Page 32: The Author writes about reading CSV files in chunks of 1000 rows, but the attached code contradicts this; steps 4 and 5 are missing, and the message “data downloaded” does not match the code.

6.19. Table 3.3: Unnecessary repetition of information from the previous page.

6.20. Page 34, step 4: “dataset underwent meticulous processing. The raw data was systemically organized into a structured format, enhancing coherence and accessibility” – This is not evident in the code. Aren’t the datasets already structured and organized when downloded?

6.21. Separate sections describe data acquisition for each source, which is redundant and inflates the volume of the dissertation unnecessarily, especially as the code is attached. Moreover, code listings and data acquisition descriptions are repeated in subsequent sections – entirely unnecessary.

6.22. Page 36: The Author claims rigorous quality control of samples was performed. First, it is unclear how; second, this process is not visible in the attached code.

6.23. Page 37: “a standardized format for gene and transcription factor identifiers was adopted” – The method for creating these identifiers is not clearly described.

6.24. Page 38, “Database structure and optimization”: The Author discusses normalization, foreign keys, and indexes, but none of these are visible in the database-related code.

6.25. Section 3.3.3: What was the purpose of this analysis?

6.26. Page 45: “the algorithm converged in a reasonable number of iterations” – This is not a precise or appropriate statement for scientific writing.

6.27. Page 46, section 3.5.2: Assuming a weight of 1, as the Author does, it is unclear what the difference between “uniform” and “biased” distributions is.

6.28. Section 3.5.3: The definition of Ei is not precise for the directed graphs used.

6.29. Section 3.5.4 is repeated in 3.5.10 (e.g., formula 3.6).

6.30. The Author frequently mentions assigning weights to edges based on mutations, but no precise engineering definition or justification for this procedure is provided.

6.31. Page 53, Normalization: The formula provided does not match the description.

6.32. Table 4.2: A txt file appears, whereas the Author previously only mentioned CSV formats.

6.33. Tables 4.3, 4.5, 4.6: These do not appear in the code, and it is unclear where or when they are used.

6.34. Page 63: “The script systematically filtered out redundant and low-quality data, including missing values (null or NaN), before importing to database” – This is not visible in the code, and it is not explained how it was implemented; for example, the source of null or NaN values is unclear.

6.35. Section 4.1.5: This is another unnecessary repetition.

6.36. Page 70: Keys, normalization, data cleaning, indexing, partitioning – none of these are visible in the code, and it is unclear how they were performed. Moreover, indexing and partitioning make little sense for tables with only a few thousand records, as these operations are typically applied to much larger datasets.

6.37. Section 4.2.2: The Author mentions 1,492 nodes, whereas Figure 4.7 shows 1,500 nodes. Additionally, Figure 4.7 is completely unreadable and does not match the description.

6.38. Figure 4.10 caption: “see interactive version in Supplementary Materials” – What does it mean that the network in the supplement is interactive?

6.39. Figure 4.11 and subsequent results: The Author repeatedly states that networks are constructed based on omics data (genes, proteins, transcription factors), yet nodes in the figures

are labeled as “apoptosis” or “cell cycle arrest” – cell states, not elements previously described. It is unclear how the proposed computational protocol operates. Furthermore, these states are absorbing, making it difficult to interpret previously discussed random walk simulation results.

6.40. Figures 4.13 and 4.14: No references to them are found in the text.

6.41. Bibliography formatting requires correction – e.g., entry [18] lacks capitalization in the title (miRNAs and DNAa), entry [32] should read Lotka-Volterra. Similar errors appear in multiple references.W pracy pojawił się braki lub niejasności w elementach istotnych dla merytoryki pracy, do których Doktorant powinien odnieść się w trakcie obrony. W szczególności:

7. Final evaluation of the dissertation

In summary, I conclude that the doctoral dissertation presented by M.Sc. Eng. Ruby Khan, entitled “Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems”, in the version submitted for my review, does not meet the requirements for doctoral dissertations as set forth in the Act on Higher Education and Science of 20 July 2018 (Journal of Laws of 2024, item 1571, as amended).

These shortcomings do not concern the choice of the topic or the conceptual design of the conducted research, which I assess positively, but rather the insufficient presentation of the results and numerous editorial deficiencies, as indicated in my review. Considering the above, I must, unfortunately, give a negative assessment of the dissertation submitted for review.

Dr hab. Jan Poleszczuk