dr hab. Zuzanna Szymańska, Assoc. Prof.
Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw
zk.szymanska@uw.edu.pl

Doctoral Thesis Review

Author: Ruby Khan
Title: "Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems"
Supervisor: Dr hab. inż. (D.Sc. Eng.) Krzysztof Puszyński, Assoc. Prof., Silesian University of Technology

The review of the doctoral thesis was prepared at the request of the Chair of the Discipline Council of Biomedical Engineering at the Silesian University of Technology, Prof. Dr. hab. inż. Robert Michnik, as expressed in letter RDIB.0211.71.2025 dated July 10, 2025.

Overview of the Thesis

The doctoral thesis entitled "Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems" was prepared by Ruby Khan at the Department of Systems Engineering and Biology, Silesian University of Technology. Its primary objective is the design of a computational pipeline for the automated extraction and analysis of multiomics data, aimed at identifying critical variables within complex biological networks.

In support of her thesis, the PhD candidate refers to three publications, with two already published or accepted and one presently under editorial review.

1. Khan, R., Pari, B., & Puszynski, K. (2024). Comprehensive Bioinformatic Investigation of TP53 Dysregulation in Diverse Cancer Landscapes. Genes, 15(5), 577.

2. Khan, R., Khan, S., Pari, B., & Puszynski, K. (2025). Optimizing Machine Learning for Network Inference through Comparative Analysis of Model Performance in Synthetic and Real-World Networks. Accepted in Scientific Reports, Springer.

3. Khan, R., Khan, S., Pari, B., Almohaimeed, H. M., & Puszynski, K. (2025). Vancomycin-Resistant and Multidrug-Resistant Bacteria in Raw Milk Pose Critical Public Health Risks. Under review in Scientific Reports, Springer.

The thesis consists of five chapters, containing, respectively: an introduction, a literature review, a description of the methodology, research results, and discussion, with a more detailed description provided below.

Chapter 1 – Introduction, defines the central research problem, namely the need for systematic, automated approaches to construct and analyse biological networks based on diverse data sources. The candidate outlines the research objectives, which focus on developing computational pipelines for the integration of data from multiple databases into a biological network, its refinement, and analysis aimed at identifying critical variables, particularly those that may serve as potential therapeutic targets. The introduction also highlights the broader biomedical motivation, especially the identification of therapeutic targets.

Chapter 2 – Literature Review situates the candidate's work within the broader development of systems biology, tracing the evolution of the field from the early mathematical models of the 1950s, through the incorporation of control theory and engineering principles, to its contemporary reliance on high-throughput omics data. The author emphasises the persistent challenge of bridging abstract mathematical frameworks with concrete biomedical applications, thereby motivating the computational strategies developed in this thesis.

Chapter 3 – Methodology presents a multi-step computational framework for inferring the properties of components of large biological networks. It outlines procedures for large-scale data extraction and integration from multiple repositories, leading to the construction of structured biological interaction networks. The chapter introduces several analytical strategies, including signal-flow modelling to capture network dynamics, the PageRank and Random Walk algorithms to assess node and edge importance, and Boolean modelling to simulate system behaviour. It also covers visualisation techniques, performance considerations, and systematic validation across various network configurations.

Chapter 4 – Results demonstrates the effectiveness of the proposed approaches, showing how integrated databases can be transformed into detailed biological networks, which are then analysed using topology, flow, and centrality measures. The application to ovarian cancer highlights potential drug targets such as NF-κB and MdM2, while additional case studies on the cell cycle and MAPK signalling pathways reveal further key regulators. Boolean modelling identifies stable states and pivotal control nodes, confirming the utility of the computational pipeline. Comparative analyses underscore the robustness of the results, while algorithmic outputs are contextualised with existing biological literature.

Chapter 5 – Discussion situates the findings within the context of current knowledge in systems biology. The candidate identifies consistencies with prior research, underscores the novelty of her methodological contributions, and highlights their broader implications for therapeutic discovery. The thesis concludes with reflections on the potential of computational systems biology to guide experimental validation and clinical applications, as well as directions for future research.

The work is supported by a detailed **Bibliography** and extensive **Appendices**, which include supplementary data, database integration scripts, and code developed for network construction, analysis, and simulation.

Content Evaluation

Chapter 1 – Introduction would benefit from a clearer and more thorough description of the candidate's original contribution. As written, the novelty of integrating data from multiple sources into a comprehensive model of biological networks is unclear. The text suggests that the contribution may be limited to coding and combining existing solutions.

Chapter 2 – Literature Review would benefit from a clearer focus and more concise presentation. It currently includes some general statements and digressions that could be streamlined to strengthen the main narrative. The selection of references could better reflect the candidate's grasp of the broader theoretical background within the discipline. More detailed critical observations are provided in later sections of this review. For future scientific writing, the candidate is encouraged to aim for greater precision and brevity, emphasising key information and minimising redundancy. If artificial intelligence language models are used, they should primarily assist with proofreading rather than text generation.

Chapter 3 – Methodology begins with a description of the study's purpose, followed by an overview of data collection. The main criticism of this section is that it focuses too heavily on successive procedural steps, which are relatively minor and make the text read like an installation manual. In contrast, key elements such as the formats of input and output data and the database structure are omitted. Moreover, there is no clear information about the exact contents of each database, and the candidate's original conceptual contribution remains unclear. A positive aspect of the proposed solution is the inclusion of information on carcinogenic mutations in the constructed network database. The examination of coherence between the constructed network and pathways reported in the literature is also commendable, as it offers an opportunity to supplement and consolidate knowledge on well-established models of signalling pathways. Although not articulated very precisely, if it indeed represents the candidate's original contribution, the consideration of time-varying interaction strengths between components would constitute an interesting extension of the model. In summary, the candidate has clearly invested substantial effort in developing a coherent biological interaction network database, which is a valuable research achievement. However, the presentation of the results could be improved for clarity and emphasis.

Chapter 4 – Results begins with a description of data extraction from the various databases used in the study. The impression is that the division of the thesis into separate Methods and Results chapters is not entirely consistent with their actual content. In essence, the chapter explores the dynamics of signal flow in biological systems. The importance of individual components within large biological networks was examined using several advanced network methodologies, including the Random Walk model, the PageRank algorithm, Regularized Collaborative Co-Occurrence Networks (RCCN), and the Boolean model. Inferring the roles and relative importance of specific variables within the network from its topology appears

to be a promising research direction. The aspects discussed in subsection 4.12.8 Model Limitations and Areas for Improvement could have been a particularly valuable component of the thesis but are treated too briefly. In contrast, the findings concerning the Cell Cycle Network and the MAPK Signalling Network are noteworthy and of clear scientific interest.

Chapter 5 – Discussion provides a summary of the results obtained in the study. However, as in the previous sections of the thesis, the candidate has not entirely avoided a degree of verbosity and a lack of focus on specific issues.

Detailed Comments and Suggested Corrections

Chapter 2 — Literature Review:

1. Page 16: Incomplete sentence: ”Using iterative modeling” ...
2. Page 17: Typo - ”An more nuanced...”
3. Page 21: The sentence ”The Michaelis–Menten framework, which characterizes the rate of enzymatic reactions as a function of substrate concentration, has historically been used to model enzyme kinetics, a fundamental area of biochemical research [36].” is problematic because reference [36] does not contain information on Michaelis–Menten kinetics. Please replace [36] with the correct source or provide an appropriate reference.
4. Page 23: Typo - ”Mathematicsians design...”
5. Page 24: Typo - ”Collaborative platforms and open data efforts aelerate research...”

Chapter 3 — Methodology:

1. Page 26: I suggest rephrasing or restructuring ”Table 3.1: Consolidated Pseudocode for Methodology”, as in its current form it provides little additional value for understanding the content. The ”Description” column often fails to add significant information beyond what is already conveyed in the ”Steps” column. For example, the explanation ”Design a structured database schema to store and efficiently retrieve integrated biological data” for the step ”Database Design” is trivial. Similarly, the explanation of the step ”Data Preprocessing” — ”Clean and preprocess collected data to remove inconsistencies and missing values” — is not particularly informative.
2. Page 26: The acronym ”KEGG” is used but has not yet been defined. Please provide its full form at this stage. Repeated definitions of acronyms occur in several other parts of the text as well, for example, ”TCGA” on page 36. I recommend providing the definition at the first occurrence of each acronym and adding a glossary of acronyms at the end of the thesis.

3. Page 30: Network Modelling, point 1. The description does not clearly specify what type of data structure is used to store the draft network or how the network is constructed from the Pathway Commons data.

4. Page 30: The text states: "The first step involves accessing the database by using the GDC database API, accessible at https://api.gdc.cancer.gov." Please verify the URL and update it to a valid, functioning link.

5. Page 31: Section 3.2.3 ends with the statement: "In summary, this methodology effectively facilitated the extraction of necessary high-throughput data related to ovarian cancer from the GDC database. Table 3.2 is given below, providing all the details and ensuring a streamlined process for its integration into subsequent analyses." First, a period appears to be missing after "database". More importantly, it is not clear what the data format is and how the data are translated into a network.

6. Page 31: The acronym KEGG (Kyoto Encyclopedia of Genes and Genomes) appears earlier in the text and should be defined at its first occurrence.

7. Page 32: Table 3.2: Data Extraction Steps from Genomics Data Commons (GDC) is not explained clearly enough. For example, the statement "Query File IDs; Request specific file IDs corresponding to genomic data of interest from the GDC repository" does not clarify how to interact with the database in order to obtain the appropriate "genomic data ID."

8. Page 32: Again, there is a repeated definition of the acronym KEGG (Kyoto Encyclopedia of Genes and Genomes) that should be removed to avoid redundancy.

9. Page 36: Point 3 of the list Methodology for Combining TF and Signaling Networks in Humans contains the description of the Network Construction, which is too brief. It contains two points: i) TFs were mapped to their corresponding signalling pathways in humans based on their target genes and regulatory roles. This mapping was achieved by overlapping TF target genes with genes involved in specific signalling pathways in human cells, and ii) The resulting network was represented as a graph, where nodes represent biological entities (e.g., TFs, genes, proteins) and edges represent interactions (e.g., regulatory relationships, protein–protein interactions) specific to human systems. However, it is not clear what exactly is meant by the statement "TFs were mapped to their corresponding signalling pathways" or what the graph illustrating the constructed signalling network actually looks like. There is not even any information on whether the graph is directed. The content would be easier to understand if the candidate had provided appropriate examples.

10. Page 36: Point 4 of the list Methodology for Combining TF and Signaling Networks in Humans contains the subpoint "Whole-genome sequencing data provided insights into genomic alterations (e.g., mutations, copy number variations) in human samples that could influence TF activity or signaling pathway

dynamics," which is not sufficiently informative and does not clearly explain what exactly occurs at this stage of the procedure.

11. Page 36: Point 5 of the list Methodology for Combining TF and Signaling Networks in Humans contains information on Functional Analysis; however, very little is explained about what is actually done at this stage.

12. Page 37: Table 3.6 Data Processing and Integration Methods does not provide any meaningful explanation. For example, the first row, which concerns Preprocessing, lists the action "Check for anomalies, biases, and technical issues" and the corresponding details "Analyze whole-genome and RNA sequencing data obtained from GDC," which do not offer a clear or substantive explanation.

13. Page 40: The text states: "The integration focused on adding mutation information as node attributes, enabling the identification of genes with somatic mutations that may influence signalling pathways and regulatory mechanisms." However, it is not clear whether retaining mutation information merely as an attribute is sufficient. For instance, what if a mutation causes a gene not only to lose its original functionality but also to gain a new, potentially harmful one? The candidate should provide a more detailed justification for adopting this methodological approach.

14. Page 41: The purpose and benefits of employing "Signal Flow Modeling Methods" are not sufficiently clear and should be elaborated upon further.

15. Page 44: Typo - a period is missing at the end of the sentence: "This implementation can be adapted to analyze biological networks, such as protein–protein interaction networks or signaling pathways".

16. Page 46: The definition of a biased initialisation of the Random Walk Algorithm effectively reduces to an unbiased one, which suggests an error in the description.

17. Page 52: Typo - unnecessary comma at the beginning of the fourth line before "making it suitable."

18. Page 52: The paragraph is poorly structured because, immediately after the sentence "The process of assigning edge weights is as follows:", the subsection "3.5.13 Initial Weight Assignment and Biological Relevance" begins.

Chapter 4 — Results:

1. Page 60: The description of "Table 4.3: Specifications of the combined genetic data set derived from GDC, integrating gene and clinical information" does not clearly explain how the data set is constructed, particularly how expression level data and clinical information are stored.

2. Page 61: It is not clear what exactly is included in ”Table 4.4: Example of ligand–receptor interaction pairs extracted from CellTalkDB, illustrating key aspects of intercellular communication”. In particular, the column ”Interaction Characteristics” is not clearly explained.

3. Page 61: Similarly, it is not entirely clear what is included in ”Table 4.6: Pathway and molecular interaction details from Pathway Commons have been integrated, comprising crucial information on pathway components and interactions.” In particular, the meaning of ”Interaction Type” in the context of a biological pathway should be clarified.

4. Page 64: Figures 4.2 and 4.3 seem not to be referenced in the main text. They appear to be subfigures associated with Figure 4.3, which is a bit confusing.

5. Page 70: Typo – an extra ”T” appears at the beginning of ”TTo preserve.”

6. Page 72: Figure 4.7 is difficult to understand and should be improved for better clarity.

7. Page 74: Figure 4.9 is difficult to understand and should be improved for better clarity.

8. Page 76: At this point, the issue of methodology reappears. Although the inclusion of mutations is a positive aspect of the thesis, the way it has been implemented is unclear, as the term ”mutation-based weights” is not sufficiently informative.

9. Page 77: Typo - ”For all the procedure the main operating environment was Python utilize ...” - probably ”utilized”.

10. Page 77: Typo - an extra ”of” in ”The generated network highlights the interconnected connections communication of of nodes...”.

11. Page 77: There appears to be an unfinished sentence: ”4.3 my complete network’s of nodes, edges.”

12. Page 78: It is not clear what is contained in subsections ”4.4.4 Detailed Network Metric Visualizations” and ”4.4.5 Whole Network Visualization”.

13. Pages 79–81: Figures 4.12, 4.13, and 4.14 do not appear to be discussed or referenced in the main text.

14. Page 85: The section title ”4.5 General Overview” lacks specificity and does not clearly indicate the content of the section.

15. Pages 86–87: Figures 4.25 through 4.29 should have been larger to improve readability.

16. Page 93: It appears that two versions of the same fragment were included by mistake, making the text redundant: ”My network analysis using Python shows a scale-free topology, with many nodes with few connections and a few highly linked clusters. This structural characteristic suggests that while the network is resistant to sporadic failures, it is susceptible to deliberate attacks on strategic hubs.” and ”The findings show that the biological network has a scale-free topology, which is defined by a large number of nodes with few connections and a few hubs with numerous connections. This structural characteristic demonstrates that although the network may tolerate sporadic failures, it is susceptible to intentional disruptions of critical hubs.”

17. Page 93: Typo: ”... for the node-to-node interactions .” — an extra space remains before the period.

18. Page 94: Typo: ”significant therapeutic advantages. .” — an extra space and period.

19. Page 94: Typo: ”activity rather than individual molecules. . — an extra space and period.

20. Page 103: Incomplete sentence: ”How the nodes”.

21. Page 104: What is the colour interpretation in Figure 4.38?

22. Page 108: What is the colour interpretation in Figure 4.40?

23. Page 110: Figure 4.41 is poorly formatted, as part of it is obscured by the legend.

24. Page 112: The underlying ideas of motif analysis were not introduced sufficiently, so it is not fully clear what this analysis contributes.

Chapter 5 — Discussion:

1. Page 118: The Louvain algorithm is mentioned in the text, but it is not explained.

2. Page 120: References to figures are missing — the text reads ”Illustrative figures such as ?? and ?? capture...”.

3. Page 121: The acronym ATM (Ataxia-Telangiectasia Mutated) should be explained upon its first use.

Bibliography:

1. Page 129: Reference 106 is incorrectly formatted and should be revised in accordance with the citation style guidelines.

Summary

In summary, the candidate has undertaken an important research topic with potentially significant implications for the advancement of science. The thesis is interdisciplinary, and although it contains several noteworthy shortcomings, I conclude that it meets the requirements set out for doctoral thesis under the Act of 20 July 2018 – Law on Higher Education and Science (consolidated text: Journal of Laws of 2023, item 742, as amended). I recommend that the Council of the Biomedical Engineering Discipline at the Silesian University of Technology approve the thesis of Ruby Khan and permit it to advance to the subsequent stages of the doctoral process.

dr hab. Zuzanna Szymańska

Warsaw, 23 September 2025