

dr hab. Zuzanna Szymańska
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego,
Uniwersytet Warszawski,
zk.szymanska@uw.edu.pl

Recenzja Pracy Doktorskiej

Autorka: Ruby Khan

Tytuł Pracy: „Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems”

Promotor: dr hab. inż. Krzysztof Puszyński, prof. Politechniki Śląskiej

Recenzję rozprawy doktorskiej sporządzono na prośbę Przewodniczącej Rady Dyscypliny Inżynierii Biomedycznej Politechniki Śląskiej Prof. dr hab. inż. Roberta Michnika, wyrażoną w piśmie RDIB.0211.71.2025 z dnia 10 lipca 2025 roku.

Charakterystyka Rozprawy

Rozprawa doktorska pt. „Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems” została przygotowana przez Ruby Khan w Katedrze Inżynierii Systemów i Biologii Politechniki Śląskiej. Głównym celem pracy było opracowanie narzędzia umożliwiającego zautomatyzowane pozyskiwanie i analizę danych multiomicznych, ukierunkowanego na identyfikację kluczowych zmiennych w złożonych sieciach biologicznych.

Na poparcie swoich osiągnięć autorka rozprawy doktorskiej powołuje się na trzy publikacje, z których jedna została już opublikowana, jedna przyjęta do druku, a jedna znajduje się obecnie w trakcie recenzji redakcyjnej.

1. Khan, R., Pari, B., & Puszyński, K. (2024). Comprehensive Bioinformatic Investigation of TP53 Dysregulation in Diverse Cancer Landscapes. *Genes*, 15(5), 577.
2. Khan, R., Khan, S., Pari, B., & Puszyński, K. (2025). Optimizing Machine Learning for Network Inference through Comparative Analysis of Model Performance in Synthetic and Real-World Networks. *Przyjęta do druku Scientific Reports*, Springer.
3. Khan, R., Khan, S., Pari, B., Almohaimeed, H. M., & Puszyński, K. (2025). Vancomycin-Resistant and Multidrug-Resistant Bacteria in Raw Milk Pose Critical Public Health Risks. W recenzji *Scientific Reports*, Springer.

Rozprawa składa się z pięciu rozdziałów, obejmujących kolejno: wprowadzenie, przegląd literatury, opis zastosowanej metodologii, wyniki badań oraz dyskusję, których bardziej szczegółowy opis przedstawiono poniżej.

W rozdziale 1 – **Introduction** autorka definiuje główny problem badawczy, jakim jest opracowanie zautomatyzowanych procedur umożliwiających integrację informacji pochodzących z wielu baz danych w jedną sieć biologiczną, a następnie jej udoskonalanie i analizę ukierunkowaną na identyfikację kluczowych zmiennych, w szczególności tych, które mogą stanowić potencjalne cele terapeutyczne. We wprowadzeniu podkreślono również szerszy kontekst biomedyczny, w szczególności znaczenie podejmowania badań o charakterze zbliżonym do przedstawionych w niniejszej rozprawie.

Rozdział 2 – **Literature Review** umiejscawia badania doktorantki w szerszym kontekście rozwoju biologii systemów, śledząc ewolucję tej dziedziny, od wczesnych modeli matematycznych z lat 50., poprzez integrację zasad teorii sterowania i inżynierii, aż po jej współczesny etap, obejmujący włączenie wysokoprzepustowych danych omicznych. Autorka podkreśla wyzwanie, jakim jest połączenie abstrakcyjnych modeli matematycznych z konkretnymi zastosowaniami biomedycznymi, co stanowi uzasadnienie dla opracowania strategii obliczeniowych przedstawionych w niniejszej rozprawie.

Rozdział 3 – **Methodology** przedstawia procedurę służącą do wnioskowania o właściwościach składowych złożonych sieci biologicznych. Zawiera on opis procesu pozyskiwania i integracji dużych zbiorów danych pochodzących z wielu repozytoriów, a następnie konstrukcji sieci interakcji biologicznych. W rozdziale zaprezentowano szereg strategii analitycznych, w tym modelowanie przepływu sygnału w celu opisu dynamiki sieci, zastosowanie algorytmów PageRank i Random Walk do oceny znaczenia węzłów i krawędzi, a także modelowanie boolowskie wykorzystywane do symulacji zachowania systemu. Omówiono również techniki wizualizacji, aspekty wydajnościowe oraz systematyczną walidację w różnych konfiguracjach sieciowych.

Rozdział 4 – **Results** przedstawia sposób działania opracowanego narzędzia programistycznego, w szczególności proces integrowania danych pochodzących z różnych baz w rozbudowane sieci biologiczne. Następnie opisuje analizę tych sieci z wykorzystaniem narzędzi grafowych oraz omawia wnioski płynące z uzyskanych wyników. W dalszej części rozdziału doktorantka demonstruje zastosowanie proponowanego narzędzia do analizy sieci regulatorowych zaangażowanych w patogenezę raka jajnika, wskazując potencjalne cele terapeutyczne, takie jak szlaki NF- κ B i Mdm2. Kolejno omawia dodatkowe przykłady regulacji cyklu komórkowego oraz szlaku sygnałowego MAPK, ujawniając ich kluczowe elementy regulatorowe. Modelowanie boolowskie umożliwia identyfikację stanów stabilnych oraz węzłów kontrolnych o istotnym znaczeniu, potwierdzając przydatność opracowanego schematu obliczeniowego. Analizy porównawcze podkreślają spójność uzyskanych wyników z obecnym stanem wiedzy, a rezultaty algorytmiczne zostały osadzone w kontekście istniejącej literatury biologicznej.

Rozdział 5 – **Discussion** umiejscawia uzyskane wyniki w kontekście aktualnego stanu wiedzy w dziedzinie biologii systemów. Autorka wskazuje zgodność otrzymanych rezultatów z wcześniejszymi badaniami, jednocześnie podkreślając nowatorski charakter przedstawionych rozwiązań metodologicznych oraz zwracając uwagę na

ich potencjalne zastosowanie w projektowaniu strategii terapeutycznych. Rozprawa kończy się refleksją nad znaczeniem podejścia biologii systemów w ukierunkowywaniu badań eksperymentalnych i zastosowań klinicznych, a także nad możliwymi kierunkami dalszych badań.

Rozprawa jest uzupełniona szczegółową bibliografią (rozdział **Bibliography**) oraz obszernymi apendyksami (rozdziały **Appendices**), które zawierają dane uzupełniające, w tym skrypty integracji baz danych oraz kod opracowany na potrzeby konstrukcji, analizy i symulacji sieci.

Ocena merytoryczna

Rozdział 1 – **Introduction** nie zawiera przejrzystego ani dostatecznie szczegółowego opisu oryginalnego wkładu doktorantki. W obecnej formie nie zostało jasno określone, na czym polega nowatorski charakter proponowanego podejścia, które zakłada integrację danych z wielu źródeł w celu budowy kompleksowego modelu sieci biologicznych. Z treści rozdziału można odnieść wrażenie, że wkład autorki ogranicza się głównie do implementacji kodu oraz integracji istniejących rozwiązań. Rozdział wymaga doprecyzowania i wyraźniejszego podkreślenia elementów autorских.

Rozdział 2 – **Literature Review** jest nadmiernie rozbudowany i pozbawiony wyraźnego wątku przewodniego. Zawiera liczne ogólne stwierdzenia oraz dygresje, które należałoby ograniczyć w celu zwiększenia klarowności i spójności wyводу. Dobór literatury sprawia wrażenie przypadkowego i nie w pełni dostosowanego do kontekstu badań, co osłabia merytoryczną wartość tej części rozprawy. W przyszłych opracowaniach naukowych doktorantka powinna dążyć do większej precyzji oraz zwięzłości wypowiedzi, koncentrując się na prezentacji treści istotnych dla podjętego problemu badawczego i unikając powtórzeń. W przypadku korzystania z modeli językowych opartych na sztucznej inteligencji należy wykorzystywać je tylko do korekty językowej i redakcyjnej tekstu, a nie do jego generowania. Bardziej szczegółowe uwagi krytyczne dotyczące tej części pracy zostały przedstawione w dalszych fragmentach niniejszej recenzji.

Rozdział 3 – **Methodology** rozpoczyna się od przedstawienia celu badania, a następnie omawia metody pozyskiwania informacji z baz danych. Głównym zarzutem wobec tej części pracy jest nadmierne skupienie na szczegółowym opisie kolejnych kroków proceduralnych, mających drugorzędne znaczenie merytoryczne, co sprawia, że fragment ten przypomina raczej instrukcję obsługi niż opis badawczy. Niestety, pominięto istotne elementy, takie jak formaty danych czy struktura wykorzystanych baz. Brakuje również jednoznacznego przedstawienia zawartości poszczególnych baz danych, a oryginalny wkład koncepcyjny doktorantki pozostaje nie w pełni klarowny. Pozytywnym aspektem proponowanego rozwiązania jest uwzględnienie informacji o mutacjach nowotworowych w konstruowanej sieci oddziaływań. Na uwagę zasługuje także analiza spójności między opracowaną siecią a danymi dostępnymi w literaturze, gdyż pozwala ona na uzupełnienie i utrwalenie wiedzy

dotyczącej dobrze znanych modeli szlaków sygnalizacyjnych. Choć nie zostało to wyrażone w sposób w pełni precyzyjny, uwzględnienie zmienności w czasie siły interakcji między komponentami mogłoby stanowić interesujące i oryginalne rozszerzenie modelu, o ile rzeczywiście jest to autorski wkład doktorantki. Podsumowując, doktorantka włożyła znaczący wysiłek w opracowanie spójnej bazy danych sieci interakcji biologicznych, co należy uznać za osiągnięcie badawcze. Jednocześnie prezentacja wyników mogłaby zostać ulepszona w celu zwiększenia przejrzystości oraz lepszego wyeksponowania kluczowych informacji.

Rozdział 4 – **Results** rozpoczyna się od opisu procesu pozyskiwania informacji z różnych baz danych. Początkowa część rozdziału ma jednak charakter bardziej metodologiczny niż prezentacyjny, co sprawia, że podział rozprawy na rozdziały Metodologia i Wyniki nie wydaje się w pełni spójny z ich rzeczywistą treścią. W dalszej części rozdział koncentruje się na analizie dynamiki przepływu sygnałów w systemach biologicznych. Znaczenie poszczególnych zmiennych w rozbudowanych sieciach biologicznych zostało ocenione z wykorzystaniem kilku zaawansowanych metod analizy sieci, w tym modelu Random Walk, algorytmu PageRank, modelu Regularized Collaborative Co-Occurrence Networks (RCCN) oraz modelu boolowskiego. Wnioskowanie o roli i znaczeniu poszczególnych zmiennych w sieci na podstawie jej topologii stanowi obiecujący i uzasadniony kierunek badań. Zagadnienia przedstawione w podrozdziale 4.12.8 – Model Limitations and Areas for Improvement mogłyby stanowić szczególnie wartościowy element rozprawy, jednak zostały omówione zbyt ogólnikowo. Natomiast wyniki dotyczące Cell Cycle Network oraz MAPK Signalling Network są potencjalnie interesujące, gdyż posiadają potencjał aplikacyjny.

Rozdział 5 – **Discussion** zawiera podsumowanie wyników uzyskanych w ramach badań. Jednak, podobnie jak w poprzednich częściach pracy, autorce nie udało się uniknąć rozwlekłości oraz braku skupienia na kluczowych zagadnieniach.

Uwagi i sugerowane poprawki

Rozdział 2 - Literature Review:

1. Strona 16: Niepełne zdanie – fragment „Using iterative modeling” wymaga dokończenia lub poprawy, aby był pełnym zdaniem.
2. Strona 17: Literówka – „An more nuanced...” należy poprawić na „A more nuanced”.
3. Strona 21: Zdanie „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].” jest problematyczne, ponieważ źródło [36] nie zawiera informacji dotyczących kinetyki Michaelisa–Menten. Należy zastąpić [36] właściwym odniesieniem.

4. Strona 23: Literówka – „Mathematicians” należy poprawić na „Mathematicians”.
5. Strona 24: Literówka – „Collaborative platforms and open data efforts aelerate research...” należy poprawić na „Collaborative platforms and open data efforts accelerate research...”.

Rozdział 3 - Methodology:

1. Strona 26: Sugeruję przeformułowanie tabeli „Table 3.1: Consolidated Pseudocode for Methodology”, ponieważ w obecnej formie nie pomaga ona w zrozumieniu treści pracy. Kolumna „Description” często nie dostarcza dodatkowych informacji w stosunku do tych już zawartych w kolumnie „Steps”. Na przykład objaśnienie „Design a structured database schema to store and efficiently retrieve integrated biological data” dla kroku „Database Design” jest dość oczywiste. Podobnie opis kroku „Data Preprocessing” jako „Clean and preprocess collected data to remove inconsistencies and missing values” nie jest szczególnie informatywny.
2. Strona 26: Skrót „KEGG” został użyty, zanim został formalnie zdefiniowany. Proszę podać jego pełne rozwinięcie w tym miejscu. Natomiast powtórzone definicje skrótów pojawiają się w innych częściach tekstu, na przykład „TCGA” na stronie 36. Zalecam podanie pełnej definicji przy pierwszym wystąpieniu każdego skrótu oraz dodanie słownika skrótów na końcu rozprawy.
3. Strona 30: Network Modelling, punkt 1. Opis nie precyzuje jasno, jakiego typu struktura danych jest używana do przechowywania sieci ani w jaki sposób sieć jest konstruowana na podstawie danych z Pathway Commons.
4. Strona 30: W tekście napisano: „The first step involves accessing the database by using the GDC database API, accessible at <https://api.gdc.cancer.gov>.”. Proszę zweryfikować podany adres URL i zaktualizować go do prawidłowego, działającego linku.
5. Strona 31: Sekcja 3.2.3 kończy się stwierdzeniem: „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.” Po pierwsze, wydaje się, że po słowie „database” brakuje kropki. Co ważniejsze, nie jest jasne, w jakim formacie znajdują się dane ani w jaki sposób są przekształcane w sieć.
6. Strona 31: Skrót „KEGG” (Kyoto Encyclopedia of Genes and Genomes) pojawia się wcześniej w tekście i powinien być zdefiniowany przy pierwszym wystąpieniu.
7. Strona 32: Tabela 3.2: Data Extraction Steps from Genomics Data Commons (GDC) nie jest wystarczająco informatywna. Na przykład stwierdzenie „Query File IDs; Request specific file IDs corresponding to genomic data of interest

- from the GDC repository” nie wyjaśnia, w jaki sposób należy wchodzić w interakcję z bazą danych, aby uzyskać odpowiedni „genomic data ID”.
8. Strona 32: Ponownie pojawia się powtórzona definicja skrótu „KEGG” (Kyoto Encyclopedia of Genes and Genomes).
 9. Strona 36: Punkt 3 listy „Methodology for Combining TF and Signaling Networks in Humans” zawiera opis etapu „Network Construction”, który jest zbyt skrótowy. Obejmuje on dwa punkty: 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, oraz 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. Jednak nie jest jasne, co dokładnie oznacza stwierdzenie „TFs were mapped to their corresponding signalling pathways” ani jak wygląda graf ilustrujący skonstruowaną sieć. Brakuje również informacji o tym, czy graf jest skierowany. Treść byłaby łatwiejsza do zrozumienia, gdyby doktorantka zamieściła odpowiednie przykłady.
 10. Strona 36: Punkt 4 listy „Methodology for Combining TF and Signaling Networks in Humans” zawiera podpunkt „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”, który nie jest wystarczająco informatywny i nie wyjaśnia, co dokładnie dzieje się na tym etapie procedury.
 11. Strona 36: Punkt 5 listy „Methodology for Combining TF and Signaling Networks in Humans” nie zawiera wystarczającego wyjaśnienia, co dzieje się na tym etapie procedury, w szczególności nie jest wyjaśnione pojęcie Functional Analysis.
 12. Strona 37: Tabela 3.6 „Data Processing and Integration Methods” nie dostarcza żadnego znaczącego wyjaśnienia. Na przykład pierwszy wiersz dotyczący Preprocessing wymienia działanie „Check for anomalies, biases, and technical issues” oraz odpowiadający mu opis szczegółowy „Analyze whole-genome and RNA sequencing data obtained from GDC”, który nie daje wystarczającego wyjaśnienia.
 13. Strona 40: W tekście znajduje się stwierdzenie: „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.” Jednak nie jest jasne, czy zachowanie informacji o mutacjach jedynie jako atrybutu węzła jest wystarczające. Na przykład, co w przypadku, gdy mutacja powoduje, że gen nie tylko traci swoją pierwotną funkcję, lecz także uzyskuje nową, potencjalnie szkodliwą? Doktorantka powinna przedstawić bardziej szczegółowe uzasadnienie przyjętego rozwiązania.

14. Strona 41: Cel i korzyści wynikające ze stosowania „Signal Flow Modeling Methods” nie są wystarczająco jasne i powinny zostać dokładniej omówione.
15. Strona 44: Literówka – brakuje kropki na końcu zdania: „This implementation can be adapted to analyze biological networks, such as protein–protein interaction networks or signaling pathways”.
16. Strona 46: Definicja inicjalizacji z preferencją (biased initialisation) dla algorytmu Random Walk w praktyce sprowadza się do inicjalizacji bez preferencji (unbiased), co sugeruje błąd w opisie.
17. Strona 52: Literówka – zbędny przecinek na początku czwartej linii przed „making it suitable”.
18. Strona 52: Fragment pracy jest niezręcznie ułożony, ponieważ bezpośrednio po zdaniu „The process of assigning edge weights is as follows:” rozpoczyna się podsekcja „3.5.13 Initial Weight Assignment and Biological Relevance”.

Rozdział 4 - Results:

1. Strona 60: Opis „Table 4.3: Specifications of the combined genetic data set derived from GDC, integrating gene and clinical information” nie wyjaśnia w sposób dostateczny, w jaki sposób zbiór danych jest skonstruowany, w szczególności jak przechowywane są dane dotyczące poziomów ekspresji genów oraz informacje kliniczne.
2. Strona 61: Nie jest jasne, co dokładnie zawiera „Table 4.4: Example of ligand–receptor interaction pairs extracted from CellTalkDB, illustrating key aspects of intercellular communication”. W szczególności kolumna „Interaction Characteristics” nie jest jasno wyjaśniona.
3. Strona 61: Nie jest jasne, co zawiera „Table 4.6: Pathway and molecular interaction details from Pathway Commons have been integrated, comprising crucial information on pathway components and interactions.” W szczególności należy wyjaśnić znaczenie kolumny „Interaction Type” w kontekście ścieżki biologicznej.
4. Strona 64: Rysunki 4.2 i 4.3 chyba nie są cytowane w tekście głównym. Wydaje się, że są rysunkami powiązanymi z Rysunkiem 4.3, co może wprowadzać pewne zamieszanie.
5. Strona 70: Literówka – na początku zdania „TTo preserve” pojawia się dodatkowa litera „T”.
6. Strona 72: Rysunek 4.7 jest trudny do zrozumienia i powinien zostać poprawiony w celu zwiększenia czytelności.
7. Strona 74: Rysunek 4.9 jest trudny do zrozumienia i powinien zostać poprawiony w celu zwiększenia czytelności.

8. Strona 76: W tym miejscu ponownie pojawia się kwestia metodologiczna. Choć uwzględnienie mutacji stanowi pozytywny element rozprawy, sposób ich implementacji nie jest jasny, ponieważ określenie „mutation-based weights” nie jest wystarczająco informatywne.
9. Strona 77: Literówka – w zdaniu „For all the procedure the main operating environment was Python utilize ...” prawdopodobnie powinno być „utilized”.
10. Strona 77: Literówka – dodatkowe „of” w zdaniu „The generated network highlights the interconnected connections communication of of nodes...”.
11. Strona 77: Zdanie: „4.3 my complete network’s of nodes, edges.” jest niedokończone.
12. Strona 78: Nie jest jasne, co zawierają podrozdziały „4.4.4 Detailed Network Metric Visualizations” oraz „4.4.5 Whole Network Visualization”.
13. Strony 79–81: Rysunki 4.12, 4.13 i 4.14 nie są omawiane ani cytowane w tekście głównym.
14. Strona 85: Tytuł podrozdziału „4.5 General Overview” jest mało precyzyjny i nie wskazuje wyraźnie na jego zawartość.
15. Strony 86–87: Rysunki 4.25–4.29 powinny zostać powiększone, ponieważ w obecnej formie nie są czytelne.
16. Strona 93: Zamieszczono dwie wersje tego samego fragmentu, co powoduje zbędne powtórzenie tekstu: „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.” oraz „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. Strona 93: Literówka: „... for the node-to-node interactions .” - przed kropką znajduje się dodatkowa spacja.
18. Strona 94: Literówka: „significant therapeutic advantages. .” - dodatkowa spacja i kropka.
19. Strona 94: Literówka: „activity rather than individual molecules. .” - dodatkowa spacja i kropka.
20. Strona 103: Niepełne zdanie: „How the nodes”.
21. Strona 104: Należy wyjaśnić interpretację kolorów na Rysunku 4.38.
22. Strona 108: Należy wyjaśnić interpretację kolorów na Rysunku 4.40.

23. Strona 110: Rysunek 4.41 jest źle sformatowany, część jego treści jest zasłonięta przez legendę.
24. Strona 112: Podstawowe założenia analizy motywów (ang. motif analysis) nie zostały wystarczająco wyjaśnione, w związku z czym nie jest w pełni jasne, co wnosi ta analiza.

Rozdział 5 - Discussion:

1. Strona 118: W tekście wspomniano algorytm Louvain, ale nie jest wyjaśnione do czego on służy ani jak działa.
2. Strona 120: W tekście znajduje się fragment „Illustrative figures such as ?? and ?? capture...”, w którym brakuje odwołań do właściwych rysunków.
3. Strona 121: Skrót „ATM” (Ataxia-Telangiectasia Mutated) powinien być wyjaśniony przy pierwszym użyciu.

Bibliografia:

1. Strona 129: Odniesienie 106 jest nieprawidłowo sformatowane i powinno zostać poprawione zgodnie z przyjętym stylem cytowania.

Posumowanie

Podsumowując, Doktorantka podjęła ważny temat badawczy, który o potencjalnie istotnym znaczeniu dla rozwoju nauki. Rozprawa ma charakter interdyscyplinarny i stwierdzam, że mimo faktu, że zawiera wiele zauważalnych niedociągnięć, spełnia wymagania określone dla rozpraw doktorskich przez Ustawę z dnia 20 lipca 2018 roku - Prawo o szkolnictwie wyższym i nauce (j.t. Dz. U. z 2023 r. poz. 742, z późn. zm.). Biorąc pod uwagę wyniki i dorobek naukowy Ruby Khan oraz obowiązujące przepisy dotyczące stopni i tytułów naukowych, rekomenduję Radzie Dyscypliny Inżynierii Biomedycznej Politechniki Śląskiej dopuszczenie pracy do kolejnych etapów postępowania doktorskiego.

dr hab. Zuzanna Szymańska

Warszawa, 23 września 2025