



Instytut Biocybernetyki i Inżynierii Biomedycznej
im. Macieja Nałęcz
Polskiej Akademii Nauk

Dr hab.
Jan Poleszczuk

Warszawa, 23 września 2025

Recenzja rozprawy doktorskiej

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

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

1. Zagadnienia naukowe rozpatrywane w pracy

Przedmiotem rozprawy jest zaproponowanie, implementacja oraz wstępna weryfikacja metody obliczeniowej służącej do automatycznego pozyskiwania, porządkowania i analizy danych, której celem jest ułatwienie identyfikacji kluczowych elementów złożonych sieci biologicznych. Autorka wykorzystała w tym celu szeroko stosowany w środowisku naukowym język programowania Python, dysponujący rozbudowanym zasobem bibliotek dedykowanych różnym zastosowaniom badawczym. Opracowane rozwiązanie obejmuje implementację metod do pobierania i przetwarzania danych z różnych repozytoriów danych omicznych, których integracja może być następnie analizowana przy użyciu znanych metod badania złożoności sieci. W ramach wstępnej weryfikacji metody, tj. poprawnego wyłaniania kluczowych węzłów regulacyjnych, Autorka odwołała się do dwóch dobrze znanych sieci biologicznych: szlaku sygnałowego cyklu komórkowego oraz szlaku sygnałowego MAPK.

Podjęty problem analizy sieci biologicznych, a także próba opracowania nowej metody obliczeniowej, mają istotne znaczenie dla zrozumienia mechanizmów rozwoju chorób, takich jak nowotwory, co w dłuższej perspektywie może znaleźć zastosowanie w praktyce medycznej. Od wielu lat badacze poszukują nowych celów terapeutycznych, przy czym ich znaczenie w danym szlaku sygnałowym stanowi podstawę do oceny potencjału terapeutycznego. Co więcej, w dobie terapii kombinowanych, oddziałujących równocześnie na różne elementy sieci regulatorowych, pogłębione zrozumienie działania poszczególnych szlaków biologicznych może przyczynić się do zwiększenia skuteczności leczenia poprzez optymalny dobór dawek. Mając to na uwadze, należy stwierdzić, że problem badawczy podjęty w rozprawie jest aktualny i dobrze wpisuje się w nurt współczesnych badań naukowych. Tematyka rozprawy niewątpliwie mieści się w obszarze dyscypliny Inżynieria Biomedyczna.

W rozprawie wyróżniono trzy hipotezy badawcze dotyczące możliwości identyfikacji kluczowych elementów w sieciach regulatorowych. W mojej ocenie ich sformułowanie nie jest jednak w pełni poprawne i bardziej odzwierciedla opis uzyskanych wyników niż wyznacza kierunki prowadzonych badań. W szczególności trzecia hipoteza stanowi połączenie stwierdzenia dotyczącego stabilności i możliwości uogólnienia proponowanej metody z podaniem wyników procentowych dla dwóch przykładowych analizowanych sieci regulatorowych. Z uwagi na sposób przedstawienia rozprawy, pozostaje dyskusyjne, czy hipotezy zostały w niej faktycznie zweryfikowane - na co wskazują również uwagi zawarte w dalszej części recenzji.

2. Struktura pracy

Rozprawa składa się ze streszczenia w języku angielskim i polskim, podziękowań, spisu treści, listy publikacji oraz wystąpień konferencyjnych Autorki, wprowadzenia, rozdziału opisującego metodologię badań, rozdziału prezentującego uzyskane wyniki, dyskusji, bibliografii oraz suplementów zawierających skrypty wykorzystane w pracy badawczej.

Wprowadzenie przedstawia krótko do tematyki biologii systemów oraz motywację podjęcia tematu wraz z opisem badanych hipotez. Kolejne dwa rozdziały poświęcono omówieniu metodologii oraz prezentacji wyników. W dyskusji Autorka podsumowała uzyskane rezultaty i odniosła się w skrócie do zgodności swoich wyników z pracami innych zespołów badawczych.

Układ pracy oceniam jako prawidłowy, choć możliwe byłoby lepsze uporządkowanie materiałów zawartych w suplementach, w szczególności wyraźne oddzielenie skryptów od innych danych tabelarycznych. Dodatkowo skrypty mogłyby być bardziej przejrzyste pod względem merytorycznym i przedstawione w formie poszczególnych klas oraz komponentów.

3. Analiza źródeł

Spis literatury obejmuje 130 pozycji, przy czym pozycje [112] i [113], [98] i [114], [117] i [102] są powtórzone (nie wykluczam, że powtórzeń może być więcej). Szacuję, że zaledwie około połowa cytowanych prac odnosi się do tematyki rozprawy. Czytelnik może znaleźć wiele pozycji dotyczących np. modelowania epidemii, sztucznych neuronów, farmakodynamiki czy układów drapieżnik-ofiara, które nie mają związku z rozważanymi w pracy zagadnieniami.

Spośród ciekawych, lecz niekoniecznie powiązanych z tematyką rozprawy pozycji można wyróżnić [92], [92] i [93], gdzie omawiane są odpowiednio optymalizacja procesu wzrostu roślin, fale zachorowań w epidemii oraz modele klimatyczne. Jednocześnie lista publikacji związanych z tematyką pracy, moim zdaniem, nie obejmuje wystarczająco pozycji wprowadzających w kluczowe zagadnienia dla omawianego tematu, takich jak ekspresja genów, czynniki transkrypcyjne czy teoria grafów.

Podsumowując, dobór bibliografii nie pozwala mi ocenić pozytywnie rozeznania Autorki w światowej literaturze naukowej w obszarze, którym się zajmuje.

4. Oryginalność i silne strony rozprawy

Niewątpliwie silną stroną rozprawy jest podjęta tematyka badawcza oraz próba kompleksowego rozwiązania problemu - od pozyskania danych po przedstawienie wyników w oparciu o zaproponowane metody analityczne. W pracy wykorzystano stosunkowo standardowe metody, jednak sposób ich zastosowania oraz uzyskane wyniki wydają się być oryginalne. Wyniki te nie zostały jeszcze opublikowane, lecz po uzupełnieniach i poprawkach w procesie recenzyjnym mają szansę znaleźć miejsce w międzynarodowych czasopismach naukowych.

Niestety sposób przedstawienia wyników w pracy nie pozwala mi obecnie z pełnym przekonaniem stwierdzić, które z prezentowanych rezultatów można uznać za w pełni prawidłowe i szczególnie wartościowe z naukowego punktu widzenia. Szczegółowy opis zastrzeżeń przedstawiony jest w dalszej części niniejszej recenzji.

5. Słabsze strony rozprawy

Do słabszych stron pracy należą:

- Autorka jasno wskazuje, że jej celem jest opracowanie kompleksowego protokołu obliczeniowego. Jednak w rozprawie protokół ten nie został przedstawiony w sposób umożliwiający czytelnikowi jego pełną implementację ani pełne zrozumienie – można znaleźć jedynie fragmenty kodu i ogólne opisy, które, mimo zapewnień Autorki, nie stanowią pseudokodu. W przypadku tak sformułowanego celu kluczowe jest dziś przygotowanie repozytorium z kodami i pełną dokumentacją, umożliwiającą innym zrozumienie i wykorzystanie proponowanych metod.
- Autorka wielokrotnie podkreśla, że protokół jest zautomatyzowany. Jednak zgodnie z informacjami zawartymi w rozprawie wydaje się, że tak nie jest. Występują etapy ręcznej modyfikacji i uzupełniania analizowanych sieci, a w kodach wpisane są na stałe informacje dotyczące specyficznego źródła i formatu danych. Metody nie wydają się być łatwe do zastosowania do analizy innych źródeł danych ani użyteczne bez przeprowadzenia dodatkowych, skomplikowanych analiz.
- Wprowadzenie do rozprawy (Rozdział 2, Literature Review) zawiera bardzo ogólne informacje i nie pozwala czytelnikowi zapoznać się z podstawami kluczowych zagadnień poruszanych w pracy. Przedstawiono ogólny zarys historii rozwoju biologii systemów, ale brak w nim omówienia zagadnień kluczowych dla tematyki pracy, takich jak definicje genu, białka, procesy transkrypcji i translacji, czy też podstawy modelowania grafowego. W wprowadzeniu omawiane są układy drapieżnik-ofiara, kinetyka Michaelisa-Menten,

farmakodynamika i farmakokinetyka, a także model neuronu – tematy niezwiązane bezpośrednio z problematyką rozprawy. Co więcej, w pracy nie przedstawiono dokładnego opisu źródeł danych wykorzystywanych przez Autorkę, tj. co zawiera dany zbiór, jak został uzupełniony oraz jakie analizy zostały na nim przeprowadzone, co jest niezbędne do uzasadnienia jego wykorzystania.

- W pracy nie jest jasno uzasadnione, że można, tak jak proponuje Autorka, scalać elementy z różnych źródeł omicznych w jedną dużą sieć. Elementy te są różnymi bytami i nie wystarczy po prostu umieścić je w jednej tabeli – kluczowe są funkcjonalne aspekty rozważanych elementów. Jest to istotne założenie przyjęte w rozprawie i wymaga odpowiednio silnej argumentacji.
- Rozprawa jest ogólnie bardzo niestarannie napisana, z licznymi błędami, część z nich wskazana jest w dalszej części recenzji. W obecnej formie nie jestem w stanie z pełnym przekonaniem stwierdzić, co i w jaki sposób zostało wykonane ani czy każdy przyjęty krok jest poprawny zarówno implementacyjnie, jak i merytorycznie. W konsekwencji nie jest jasne czy postawiony problem badawczy został rzeczywiście rozwiązany.

Oryginalność i mocne strony rozprawy niestety nie przeważają nad jej słabościami.

6. Szczegółowe uwagi merytoryczne i redakcyjne

W pracy pojawił się błędy, braki lub niejasności w elementach istotnych dla merytoryki i odbioru pracy:

- 6.1. Strona 13: This scalable computational pipeline -> w jakim sensie proponowany protokół obliczeniowy jest/ma być skalowalny? Jak o to zadbano i jak sprawdzono, że spełnia założenia?
- 6.2. Strona 13: "python based networks" – czym są takie sieci?
- 6.3. Strona 13: nieprawidłowa nazwa Cyoscape (powinno być Cytoscape).
- 6.4. Strona 14, sekcja 1.4.3: „constracted” -> constructed
- 6.5. Strona 14: niedokończone zdanie „Using iterative modeling” na końcu paragrafu.
- 6.6. Wprowadzenie: Odwołania do pozycji literaturowych są często niepoprawne. Np. na stronie 16 czytamy „The Lotka-Volterra model [7]” gdzie [7] to pozycja literaturowa na temat rozbudowanego modelu drapieżnik-ofiara z dodatkowymi członami. W tym samym zdaniu czytamy dalej o modelu Hodgkina-Huxleya, modelu Michaelis-Menten i o modelach epidemicznych, przy czym na końcu jest tylko odwołanie do [8], które traktuje tylko o modelu neuronu. Podobnie na stronie 17 czytamy „[...] enzymatic reactions rates were understood thanks to the establishment of an enzyme kinetics model [10]” gdzie [10] jest zatytułowana „Densely packed matrices as rate determine features in starch hydrolysis.”.
- 6.7. Strona 18 „The extraction of mathematical models [...]” - jest to dla mnie niezrozumiałe stwierdzenie.

- 6.8. Sekcja 2.1.2 odnosi się do okresu 1980 do 1990 a Autorka cytuje i opisuje pracę z 2023 roku (pozycja [14]).
- 6.9. Sekcja 2.5.1 nie jest żadnym podsumowaniem wcześniejszego tekstu, a już zdecydowanie nie pozwala zrozumieć „Transition to methodology”.
- 6.10. Strona 26: czym są high-throughput databases?
- 6.11. Strona 26: „Primary purpose of this project is to develop an integrated network model that can understand the key connections [...]” – w jaki sposób model może zrozumieć te połączenia?
- 6.12. Tabela 3.1.: Autorka twierdzi, że w tabeli zawarty jest pseudokod, ale nie można się z tym zgodzić, bo na podstawie tego co jest napisane nie można w żaden sposób odtworzyć metody w dowolnym języku programowania.
- 6.13. Tabela 3.1.: Czemu „Database Design” jest zawarte w kroku „data processing and integration”? Zwykle bazy projektuje się zanim dojdzie do przetwarzania danych i ich integracji.
- 6.14. Tabela 3.1. i Rysunek 1: Należałoby zmienić jeden albo drugi element, żeby był zachowana spójność (np. na Rysunku 1 nie ma Step 10, który jest w Tabeli 3.1.).
- 6.15. Strona 31: Jest informacja, że w kodach jest zaimplementowana obsługa błędów „Error-handling”, ale w tego nie widać w żadnym listingu. Nie ma nawet najprostszej obsługi błędu połączenia i podjęcia kolejnej próby pobrania danych za jakiś czas.
- 6.16. Strona 31: Punkt „Processing Downloaded Data”: jest informacja o tym, że są pobierane pliki CSV, ale w kodach widać, że niektóre dane ze źródeł są ściągane jako archiwum tar.gz.
- 6.17. Strona 31: Tabela 3.2 wydaje się zbędnym powtórzeniem. Ponadto na początku sekcji 3.2.4 jest informacja o wykorzystaniu komponentu KEGG z Pathways Commons, ale tego nie widać w kodach znajdujących się w suplemencie.
- 6.18. Strona 32: Autorka pisze o czytaniu pliku CSV po 1000 wierszy na raz, ale załączone kody temu stwierdzeniu przeczą; kroków 4. i 5. nie widać w załączonych kodach; wiadomość „data downloaded” nie jest zgodna z tym co jest w kodach.
- 6.19. Tabela 3.3. Niepotrzebne powtórzenie informacji ze strony wcześniej.
- 6.20. Strona 34, krok 4.: “dataset underwent meticulous processing. The raw data was systemically organized into a structured format, enhancing coherence and accessibility” – tego nie widać nigdzie w kodach. Czy bazy danych, z których Autorka korzysta nie są już przypadkiem ustrukturyzowane i przystępnie zorganizowane?
- 6.21. Generalnie dla każdego źródła danych przedstawiono niezależną sekcję opisującą jak pobrano dane. Jest to zupełnie niepotrzebne powtórzenie, które nic nie dodaje oprócz zwiększenia sztucznego wolumenu pracy (szczególnie, że są załączone kody). W dodatku listingi i opisy pobierania danych pojawiają się w następnych częściach pracy – to zupełnie niepotrzebne.
- 6.22. Strona 36: Autorka pisze, że przeprowadzono rygorystyczną kontrolę jakości próbek. Po pierwsze nie jest jasne jak to zostało zrobione, po drugie znowu nie widać tego procesu w załączonych kodach.

- 6.23. Strona 37: „a standardized format for gene and transcription factor identifiers was adopted” – nigdzie nie widzę jasnego przedstawienia jak te identyfikatory były tworzone.
- 6.24. Strona 38: „Database structure and optimization”: Autorka napisała o formie normalnej w bazach danych, kluczach obcych w tabelach, czy też indeksach. W kodach odpowiadających bazie danych nic z tych elementów nie widać.
- 6.25. Sekcja 3.3.3: Jaki był cel przeprowadzenia tego typu analizy?
- 6.26. Strona 45: „the algorithm converged in a reasonable number of iterations” – to nie jest prawidłowe i ścisłe sformułowanie w tekstach naukowych.
- 6.27. Strona 46, sekcja 3.5.2: zakładając wagę równą 1 jak to czyni autorka nie rozumiem jaka jest różnica między rozkładami „uniform” i „biased”.
- 6.28. Sekcja 3.5.3: definicja E_i nie jest precyzyjna w przypadku grafów skierowanych, na których operuję Autorka.
- 6.29. Sekcja 3.5.4 jest ponownie powtórzona w 3.5.10 (np. wzór 3.6).
- 6.30. Autorka wielokrotnie pisze o przypisywaniu wag do krawędzi na podstawie mutacji, ale nigdzie w pracy nie znalazłem jasnej i ścisłej inżynierskiej definicji oraz uzasadnienia przyjętych założeń takiego postępowania.
- 6.31. Strona 53, Normalization: Przedstawiony wzór nie odpowiada opisowi.
- 6.32. Tabela 4.2: Pojawia się plik w formacie txt, a Autorka pisze wcześniej tylko o formatach csv.
- 6.33. Tabela 4.3., 4.5. i 4.6.: Te rzeczy nie pojawiają się nigdzie w kodach i nie jest jasne, gdzie i kiedy są wykorzystywane.
- 6.34. Strona 63. „The script systematically filtered out redundant and low-quality data, including missing values (null or NaN), before importing to database”: nie widać tego w załączonych kodach i nie jest opisane jak to dokładnie zostało zaimplementowane. Np. nie jest jasne skąd się biorą wartości null albo NaN.
- 6.35. Sekcja 4.1.5: To kolejne powtórzenie tego co już było wcześniej, które wydaje się być kompletnie zbędne.
- 6.36. Strona 70: klucze, normalizacja, czyszczenie danych, indeksowanie, partycjonowanie: nic z tych elementów nie widnieje w kodach załączonych przez Autorkę i nie jest kompletnie wiadome jak zostało zrobione. W dodatku jaki sens ma indeksowanie i partycjonowanie w przypadku tabel z kilkoma tysiącami rekordów? Takie rzeczy się robi przy dużo większych zbiorach.
- 6.37. Sekcja 4.2.2: Autorka pisze o 1492 węzłach, natomiast na Rys. 4.7 mamy 1500 węzłów. Co więcej, Rys. 4.7 jest kompletnie nieczytelny i nie przystaje do niego opis.
- 6.38. Rysunek 4.10, podpis: „see interactive version in Supplementary Materials”: co to znaczy, że wersja sieci w suplemencie jest interaktywna?
- 6.39. Rysunek 4.11 i późniejsze wyniki: Autorka wielokrotnie pisze, że buduje sieci na podstawie danych z baz omicznych (z genów, białek, czy też czynników transkrypcyjnych). Natomiast na rysunki widać nody takie jak „apoptosis”, czy też „cell cycle arrest”. To są stany komórek, nie elementy, o których wcześniej pisała Autorka. Nie rozumiem zatem jak działa proponowany

protokół obliczeniowy. Co więcej, stany te są pochłaniające i nie rozumiem, jak interpretować omawiane wcześniej wyniki symulacji typu błędzenia losowego.

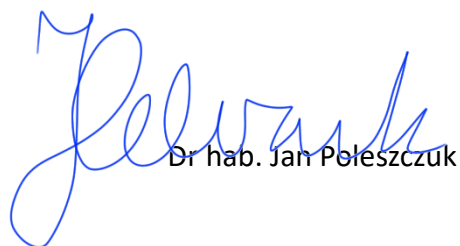
6.40. Rysunki 4.13 i 4.14: nie widzę odwołań do nich w tekście.

6.41. Formatowanie bibliografii do poprawy – np. poz. [18] nie ma dużych liter w tytule (powinno być miRNAs i DNAa,) czy też poz. [32] gdzie powinno być Lotka-Volterra. Analogiczne błędy są w wielu pozycjach literaturowych.

7. Ocena końcowa rozprawy

Podsumowując, stwierdzam, że przedstawiona przez mgr inż. Ruby Khan rozprawa doktorska pt. "Development of Methods for Identifying Key Variables in Complex Mathematical Models of Biological Systems", w stanie przedstawionym mi do recenzji, nie spełnia wymagań stawianych pracom doktorskimi przez ustawę Prawo o szkolnictwie wyższym i nauce z dnia 20 lipca 2018 r. (Dz.U. z 2024 r., poz. 1571, z późn. zm.).

Niedociągnięcia te nie dotyczą wyboru tematyki ani koncepcji wykonanych prac badawczych, które oceniam pozytywnie, lecz dotyczą niewystarczającej prezentacji wyników oraz licznych niedoskonałości redakcyjnych, które wskazałem w recenzji. Biorąc pod uwagę powyższe, moja ocena przedłożonej mi do recenzji rozprawy jest niestety negatywna.



Dr hab. Jan Połeszczyk