

Recenzja rozprawy doktorskiej “Analiza topologiczna biomolekuł”

Autor: Bartosz Ambroży Greń

Promotor: dr hab. Joanna Sułkowska, prof. UW

Recenzent: dr hab. Paweł Dłotko, prof. IMPAN

Uniwersytet Warszawski, Szkoła Doktorska Nauk Ścisłych i Przyrodniczych

Dziedzina: nauki ścisłe i przyrodnicze; dyscyplina: nauki fizyczne

1. Uwagi ogólne i charakter pracy

Rozprawa dotyczy analizy komputerowych rekonstrukcji struktur białek oraz RNA, ukierunkowanej na wykrywanie i ilościowy opis ich stopnia zawężenia. Przedstawiono metody obliczeniowe umożliwiające automatyczną identyfikację węzłów, struktur typu “kajdanki” oraz pokrewnych motywów topologicznych zarówno w strukturach eksperymentalnych, jak i w modelach przewidywanych przez algorytmy sztucznej inteligencji, w tym przez AlphaFold. Praca ma wyraźnie interdyscyplinarny charakter: łączy elementy matematyki (teoria węzłów), biofizyki i biologii strukturalnej (relacja sekwencja–struktura–funkcja), a także inżynierii oprogramowania i analizy dużych zbiorów danych.

W obszarze, którego dotyczy rozprawa, w ostatnich latach dokonała się rewolucja w metodach predykcji struktur (AlphaFold i inne podejścia oparte na AI), otwierając możliwość badań struktur białkowych na niespotykaną dotąd skalę. Autor konsekwentnie wykorzystuje dane powstałe w wyniku tej zmiany: z jednej strony rozwija podstawy teoretyczne oraz algorytmy, a z drugiej przekuwa je w praktyczne narzędzia i bazy danych dostępne dla szerokiej społeczności naukowców. Szczególnie istotne jest to, że rozwiązania te zostały przygotowane także z myślą o użytkownikach nieposiadających kompetencji programistycznych, tak aby mogli oni efektywnie wyszukiwać motywy topologiczne w analizowanych przez siebie strukturach — zarówno eksperymentalnych, jak i uzyskanych metodami symulacyjnymi czy generatywnymi.

Rozprawa ma charakter publikacyjny, co jest podejściem powszechnie stosowanym w krajach zachodnich i, w mojej ocenie, szczególnie adekwatnym dla tej tematyki. Autor opiera doktorat na bogatym, spójnym tematycznie i licznie cytowanym zbiorze publikacji. Część opisowa zawiera omówienie i syntetyczne podsumowanie wyników uzyskanych w trakcie doktoratu, streszczając poszczególne publikacje, a następnie przedstawia zestaw włączonych artykułów (w tym publikacje narzędziowe dotyczące baz danych oraz prace metodologiczne). Całość uzupełnia preprint o wyraźnie matematycznym charakterze, stanowiący istotny wkład teoretyczny w rozwój narzędzi matematycznej teorii węzłów. Taki format dobrze eksponuje mierzalne osiągnięcia autora i w pełni odpowiada standardom współczesnych badań interdyscyplinarnych.

2. Najważniejsze osiągnięcia rozprawy

Poniżej streszczam kluczowe wątki i rezultaty, zwracając uwagę na elementy szczególnie istotne w recenzji doktorskiej (oryginalność, dojrzałość warsztatu, wkład autora, znaczenie dla dziedziny).

2.1. Mocny dorobek publikacyjny i oddziaływanie (track record)

Autor prezentuje dojrzały, spójny tematycznie dorobek, obejmujący prace metodologiczne, aplikacyjne oraz narzędziowe (bazy danych). Z perspektywy recenzji doktoratu istotne są dwie rzeczy: (i) obecność publikacji w renomowanych czasopismach z obszaru biologii obliczeniowej i biologii strukturalnej (m.in. *Nucleic Acids Research*, *PLOS Computational Biology*, *Journal of Molecular Biology*), (ii) relatywnie duża liczba cytowań jak na etap kariery, wskazująca na realne wykorzystanie wyników przez społeczność.

Warto podkreślić, że wpływ dorobku wynika nie tylko z pojedynczych odkryć, ale również z wytworzenia infrastruktury badawczej (oprogramowanie, serwery, bazy), która z definicji generuje

użyteczność i cytowalność.

2.2. Połączenie teorii (tangle enumeration od Conwaya) z obliczeniami i analizą danych (AlphaFold)

Jednym z najmocniejszych elementów rozprawy jest rzadko spotykana spójność osi: *teoria* → *algorytmy* → *implementacja* → *analiza dużych danych biomolekularnych*. Wkład teoretyczny (klasyfikacja algebraicznych splątków/tangles) wyrasta wprost z klasycznych idei Conwaya i odpowiada na realną potrzebę: efektywnej parametryzacji i rozpoznawania bardziej złożonych motywów topologicznych, w tym motywów grafowych (handcuff, θ -curve), które pojawiają się w biologii strukturalnej i w związku z tym wymagają języka formalnego do ich klasyfikacji.

Jako matematyk stosowany bardzo cieszy mnie, że przedstawiona rozprawa nie jest “matematyką dla matematyki”: enumeracja i klasyfikacja splątków stanowi fundament do systematycznego katalogowania złożonych topologii, co następnie jest wykorzystywane w analizie struktur białek przewidywanych przez AlphaFold. Autor pokazuje tym samym dojrzałość naukową: potrafi zaproponować narzędzia formalne, a następnie włączyć je do dużych procedur obliczeniowych i wykorzystać na tzw. danych omicznych.

2.3. Pierwszy przykład naturalnie występującego zawężłonego łańcucha RNA

Znalezienie pierwszego przykładu naturalnie występującego węzła w RNA (trefoil knot) ma duże znaczenie porządkujące dla dziedziny. Po pierwsze, zamyka wieloletnią lukę poznawczą (odpowiada na pytanie: dlaczego węzły występują w białkach i DNA, a nie były dotąd obserwowane w RNA?), po drugie, dostarcza nowego punktu odniesienia dla biologii RNA i dla metod predykcji struktur RNA.

W rozprawie ważnie wybrzmiewa także wniosek metodyczny: obecne metody AI, które świetnie działają dla białek, nie przenoszą się automatycznie na łańcuchy RNA — najpewniej potrzebna jest dodatkowa praca, dotrenowanie metod, być może z uwzględnieniem topologicznych składników funkcji kosztu. Innymi słowy, przedstawiony wynik ma charakter uogólniający: pokazuje potrzebę uwzględniania kryteriów teorii węzłów w ocenie i treningu modeli AI (np. poprzez kary/topologiczne cechy, dodatkowe filtry bądź dotrenowanie na lepszych zbiorach referencyjnych). To jest istotny wkład na styku nauk podstawowych i rozwoju narzędzi AI dla biologii.

2.4. Implementacje, bazy danych i interfejsy dla użytkowników nieprogramujących

Na podkreślenie zasługuje fakt, że Autor nie kończy pracy na prototypach, lecz dostarcza kompletne narzędzia badawcze: serwery (np. do detekcji węzłów i lasso), bazy danych z wstępnie obliczonymi wynikami dla dużych kolekcji struktur oraz pakiet programistyczny integrujący metody obliczeniowe teorii węzłów. Interfejsy — z tego, co widzę — zostały zaprojektowane pod użytkowników, którzy nie muszą programować (choćby dokumentacja i przykłady mogłyby być bogatsze): to realnie poszerza grono odbiorców metod topologicznych (biolodzy, chemicy).

Uważam, że poniższe aspekty pracy są istotne:

- **transfer do praktyki:** udostępnienie serwerów WWW i API, nie tylko kodu,
- **skalowalność:** przetwarzanie bardzo dużych zbiorów danych (predykcje AlphaFold liczone w dziesiątkach tysięcy),
- **reprodukowalność:** pipeline’y, wersjonowanie, zautomatyzowane obliczenia,
- **walory edukacyjne i popularyzacyjne:** obniżenie bariery wejścia do metod topologicznych.

2.5. Krótkie omówienie pozostałych prac włączonych do rozprawy

Uzupełniając, poniżej zamieszczam zwięzłe streszczenie pozostałych artykułów (poza omówionymi wyżej wątkami):

- **Artykuł 1 (Lasso proteins):** praca porządkuje i systematyzuje zjawisko tzw. *lasso proteins* oraz analizuje częstość ich występowania w danych PDB. Autor bada redundancję/konserwację motywów (klastry homologii, domeny PFAM) i formułuje wniosek, że topologia lasso bywa funkcjonalnie istotna.
- **Artykuł 3 (Slipknotted topology in transmembrane transporters):** przedstawiono przegląd (slip-)węzłów w transporterach błonowych oraz pokazano, że złożona topologia koreluje z mechaniką zmian konformacyjnych. Wnioski wspierają tezę, że motywy topologiczne mogą wspomagać koordynację ruchów elementów strukturalnych i potencjalnie stanowić cel farmakologiczny przy projektowaniu nowych leków.
- **Artykuł 5 (Knots in the human AlphaFold proteome):** wykonano systematyczny screening ludzkiego proteomu przewidzianego przez AlphaFold pod kątem węzłów oraz zaproponowano procedury odróżniania przypadków wiarygodnych od artefaktów (w tym analizę pewności lokalnej oraz kontrolę fragmentaryzacji modeli). Wyniki pokazują, że surowa detekcja węzłów w predykcjach AI generuje liczne fałszywe trafienia, ale pozostawia też pulę kandydatów potencjalnie nowych, biologicznie istotnych węzłów.
- **Artykuł 7 (Knotted artifacts in predicted 3D RNA structures):** praca analizuje jakość predykcji struktur RNA (m.in. w ramach CASP15) z perspektywy topologicznej i wykazuje, że metody oparte na uczeniu maszynowym są istotnie bardziej podatne na generowanie niepoprawnych topologii (węzłów i lokalnych splątań) niż metody klasyczne. Wniosek ma charakter metodologiczny: kryteria topologiczne powinny stać się elementem oceny jakości i/lub funkcji kosztu w predykcji RNA.

3. Dodatkowe aspekty

1. **Oryginalność i nowość:** Autor wyraźnie zaznacza, które elementy są jednoznacznie nowe (np. kanoniczna reprezentacja tangles i skala klasyfikacji; identyfikacja RNA-knot; nowe klasy motywów w danych AlphaFold). W mojej ocenie są to wyniki, które znacznie wykraczają poza prace o charakterze inkrementacyjnym.
2. **Samodzielność i wkład autora:** Autor jasno wskazuje, gdzie był liderem koncepcji/metody, a gdzie współautorem.
3. **Poprawność metodyczna:** (i) dobór niezmienników i procedur upraszczania/zamykania otwartego łańcucha, by uzyskać matematyczny węzeł, (ii) analiza wrażliwości na parametry i zaszumienie danych wejściowych (np. brakujące nukleotydy/fragmenty, jakość predykcji pLDDT itp.), (iii) kryteria rozróżniania rzeczywistych motywów od artefaktów.
4. **Znaczenie dla społeczności:** przedstawione w rozprawie narzędzia mają potencjał stać się standardem (łatwość użycia, dostępność, możliwość integracji z innymi pipeline'ami, API, dokumentacja) w dość dużej dyscyplinie naukowej, jaką jest biologia molekularna.
5. **Perspektywy dalszych badań:** bardzo mocno wybrzmiewa tu naturalna agenda na kolejne lata: (i) rozszerzenia klasyfikacji grafów przestrzennych i ich zastosowania w biologii, (ii) rozwój metryk/topologicznych składników funkcji celu w predykcji RNA przez modele generatywne, (iii) współpraca z eksperymentem w weryfikacji najciekawszych kandydatów z predykcji AI. Mam nadzieję, że Autor będzie kontynuował pracę naukową w tych (i innych) kierunkach.

6. **Strona formalna rozprawy:** należy zwrócić uwagę na klarowność narracji, jakość wprowadzenia, spójność terminologii i kompletność bibliografii. Językowo rozprawa mogłaby być lepsza, ale drobne potknięcia nie wpływają negatywnie na zrozumienie treści.

4. Mocne strony pracy

W świetle przedstawionych wyników, za najmocniejsze strony rozprawy uważam:

- **Rzadkie połączenie głębokiej matematyki z praktyką obliczeniową i biologią strukturalną** (pełny łańcuch: teoria–algorytm–narzędzie–hipoteza biologiczna).
- **Silny komponent narzędziowy:** serwery i bazy danych są wkładem o wysokiej użyteczności i długim czasie życia. Jako osoba zajmująca się również rozwojem oprogramowania naukowego bardzo doceniam ten komponent i zdaję sobie sprawę z jego czasochłonności.
- **Wyniki przełomowe/porządkujące:** pierwszy na świecie przykład węzła w RNA oraz identyfikacja i katalogowanie nowych motywów topologicznych w epoce AlphaFold. To bardzo ważne wyniki, które będą przywoływane w licznych pracach w przyszłości.
- **Dojrzałość w obchodzeniu się z artefaktami danych AI:** świadomość, że predykcje masowe wymagają weryfikacji jakości, walidacji i ostrożnej interpretacji.

5. Nieliczne uwagi krytyczne i sugestie

Poniżej przedstawiam kilka standardowych, konstruktywnych uwag krytycznych (zwyczajowo oczekiwanych w recenzji doktoratu):

- **Walidacja wyników:** część wniosków (zwłaszcza opartych na predykcjach AI oraz interpretacjach funkcjonalnych) wymaga potwierdzenia eksperymentalnego. Warto doprecyzować, jakie obserwowalne sygnały (np. w danych z mikroskopii elektronowej) mogłyby weryfikować bądź falsyfikować kluczowe hipotezy oraz jaką procedurę kontroli jakości zaproponować eksperymentalistom.
- **Artefakty predykcji:** mimo stosowania filtrów jakości, w zbiorach takich jak AlphaFold nieuniknione są błędy modelu prowadzące do struktur topologicznie podejrzanych, ale niewystępujących w rzeczywistości. Pożądane byłoby wyraźniejsze wskazanie kryteriów odróżniania przypadków wiarygodnych od potencjalnych artefaktów oraz rekomendacji dalszej weryfikacji.
- **Spójność narracji:** miejscami można rozważyć lepsze zbalansowanie warstwy matematycznej i biologicznej; tj. bardziej intuicyjne omówienie konsekwencji klasyfikacji splątków dla biologii oraz (tam, gdzie to potrzebne) bardziej formalne domknięcie argumentów teoretycznych.

6. Wniosek końcowy

Rozprawa stanowi spójny, oryginalny i wartościowy wkład w rozwój metod topologicznych w analizie białek i łańcuchów RNA oraz w ich praktyczne zastosowania w biologii strukturalnej w erze masowych predykcji AI. Autor wykazuje dojrzałość naukową, umiejętność prowadzenia badań interdyscyplinarnych, a także rzadką kompetencję w doprowadzaniu wyników do postaci narzędzi użytecznych dla społeczności.

Na tej podstawie oceniam rozprawę bardzo wysoko i uważam, że spełnia ona wymagania stawiane rozprawom doktorskim. Wnoszę o dopuszczenie Autora do dalszych etapów przewodu doktorskiego i do publicznej obrony rozprawy. Wnioskuje o wyróżnienie rozprawy.

Paweł Dłotko
Centrum Dioscuri w Topologicznej Analizie Danych, IMPAN



Signed by /
Podpisano przez:
Paweł Dłotko
Date / Data:
2026-02-02 15:28