

Gdańsk 24.08.2026

dr hab. Krzesimir Ciura
Katedra i Zakład Chemii Fizycznej
Gdański Uniwersytet Medyczny
Wydział Farmaceutyczny
Al. Gen. J. Hallera 107
80-416 Gdańsk

**Recenzja rozprawy doktorskiej Karola Barana
pt. „Few-shot machine learning for multimolecular systems modeling”**

Przedstawiona do recenzji rozprawa doktorska mgr. inż. Karola Barana powstała w Katedrze Chemii Fizycznej Politechniki Gdańskiej a jej promotorem był prof. dr hab. inż. Adam Kloskowski. W ramach prowadzonych badań doktorant podjął się opracowania i zastosowania metod uczenia maszynowego, ze szczególnym uwzględnieniem technik uczenia z niewielkiej liczby przykładów (ang. *few-shot learning*), w celu modelowania i przewidywania właściwości cieczy jonowych.

Podjętą tematykę uważam za wartościową i aktualną w dwóch niezależnych płaszczyznach. Pierwszą wyznacza sam przedmiot badań. Ciecze jonowe (ang. *ionic liquids*, ILs) są obiecującą grupą związków chemicznych ze względu na znikomą prężność par, szeroki zakres ciekłości, stabilność termiczną oraz dwuskładnikową, jonową naturę, pozwalającą na celowe projektowanie właściwości poprzez dobór pary kation–anion. Ta sama elastyczność strukturalna stanowi jednak zasadniczą przeszkodę w racjonalnym doborze cieczy jonowej do konkretnego zastosowania. Liczba możliwych par jonowych znacznie przekracza możliwości badań eksperymentalnych a zależności struktura–właściwość w układach wieloskładnikowych, modyfikowanych dodatkowo przez temperaturę i ciśnienie, nie poddają się prostym opisom addytywnym. Wiarygodne modele predykcyjne są więc warunkiem racjonalnego przeszukiwania tej przestrzeni chemicznej.

Druga płaszczyzna ma charakter metodyczny i wykracza poza chemię cieczy jonowych. Doktorant zmierzył się z problemem modelowania danych o małej liczności, czyli sytuacją w której liczba dostępnych obserwacji jest o rzędy wielkości mniejsza niż wymagana przez typowe architektury głębokiego uczenia. Jest to trudność powszechna wszędzie tam, gdzie dane powstają w wyniku kosztownego i czasochłonnego eksperymentu, np. chemii medycznej, ocenie bezpieczeństwa chemicznego czy inżynierii materiałowej.

Magister inżynier Karol Baran podjął się próby rozwiązania trzech następujących problemów badawczych: (i) ocena ograniczeń klasycznych metod uczenia maszynowego i uczenia głębokiego w przewidywaniu właściwości cieczy jonowych; (ii) określenie rzeczywistej przydatności oraz ograniczeń technik uczenia z niewielkiej liczby przykładów (*few-shot learning*) w modelowaniu układów multimolekularnych; (iii) opracowanie modyfikacji istniejących algorytmów i protokołów treningu, poprawiających dokładność modeli w wybranych zastosowaniach.

Dysertacja została przygotowana w formie cyklu pięciu powiązanych tematycznie publikacji, z obszernym komentarzem do uzyskanych wyników. Całość ma formę spójnego opracowania liczącego 172 strony, napisanego w języku angielskim. W teoretycznym wstępie omówione zostały zagadnienia dotyczące sztucznej inteligencji i uczenia maszynowego, podstaw chemoinformatyki oraz sposobów reprezentacji struktury cząsteczek na potrzeby modelowania. Następnie przedstawiono przegląd literatury, cele i pytania badawcze a także wykorzystaną metodykę, zbiory danych, protokoły walidacji, reprezentacje molekularne oraz zastosowane algorytmy. Kolejny rozdział stanowi syntezę wyników uzyskanych w poszczególnych publikacjach cyklu. Rozprawę uzupełnia wykaz 175 pozycji literaturowych oraz załącznik zawierający pełne teksty publikacji tworzących cykl wraz z oświadczeniami autorów.

Cykl tworzą prace opublikowane w renomowanych czasopismach: *The Journal of Physical Chemistry B*, *Science of the Total Environment*, *ACS Sustainable Chemistry & Engineering*, *Chemical Research in Toxicology* oraz *Journal of Chemical Information and Modeling*. Wiodąca rola magistra inżyniera Karola Barana jest od razu widoczna, ponieważ w każdej z prac jest on pierwszym autorem oraz autorem korespondencyjnym, trzy z pięciu prac są dwuautorskie. Potwierdzają to również oświadczenia o wkładzie autorskim, zgodnie z którymi Doktorant odpowiadał za koncepcję, metodykę, oprogramowanie, analizę danych, wizualizację oraz przygotowanie pierwotnej wersji manuskryptu.

W mojej ocenie na podkreślenie zasługuje samodzielność Doktoranta w pozyskiwaniu finansowania. Magister inżynier Karol Baran jest kierownikiem projektu badawczego Narodowego Centrum Nauki w ramach konkursu PRELUDIUM 22 (nr UMO-2023/49/N/ST5/01043, lata 2024–2027), a także kierownikiem dwóch grantów obliczeniowych (nr PLG/2024/017245 i nr PT00996). Uzyskanie własnego finansowania na etapie studiów doktorskich świadczy o dojrzałości badawczej i zdolności do samodzielnego prowadzenia projektu naukowego.

Ocena merytoryczna:

Pierwszym badanym przez Doktoranta zagadnieniem była ocena przydatności klasycznych metod uczenia maszynowego do modelowania układów z udziałem cieczy jonowych przy bardzo ograniczonej liczbie danych. Analizę przeprowadzono na dwóch zbiorach o dużym znaczeniu praktycznym: rozpuszczalności ditlenku węgla wyrażonej stałą Henry'ego (76 rekordów) oraz wydajności izolacji ligniny z biomasy (ok. 110 rekordów). W pierwszym przypadku wykazano, że włączenie ułamkowej objętości swobodnej (FFV) jako deskryptora molekularnego istotnie poprawia jakość modeli nieliniowych, dla sieci MLP współczynnik determinacji na zbiorze testowym wzrósł z 0,738 do 0,833. W przypadku ligniny Doktorant połączył w jednym modelu skład biomasy, parametry procesu i strukturę cieczy jonowej, uzyskując R^2 w zakresie 0,63–0,77. Modele poddano interpretacji z użyciem wartości Shapleya, metody LIME oraz reguł asocjacyjnych CN2, co pozwoliło wskazać czynniki decydujące o przebiegu obu procesów.

Drugim podjętym problemem badawczym było ustalenie sposobu reprezentacji układów dwucząsteczkowych w grafowych sieciach neuronowych. Doktorant porównał trzy warianty konstrukcji grafu oraz pięć modeli ładunków cząstkowych, wykazując że najlepsze i najbardziej powtarzalne wyniki daje osobna optymalizacja jonów połączona w jedną strukturę grafową ($R^2 = 0,79 \pm 0,05$) a sposób optymalizacji geometrii ma większy wpływ na jakość modelu niż wybór metody wyznaczania ładunków. Uzyskano także wynik nieoczywisty: dla gęstości i napięcia powierzchniowego modele trenowane na danych surowych przewyższały modele oparte na danych oczyszczonych przez doktoranta, co Autor tłumaczy szerszym pokryciem przestrzeni chemicznej dla zbioru surowego. Ta część badań ujawniła zarazem podstawowe ograniczenie uczenia głębokiego, jakim jest wymagana liczność zbioru treningowego, co uzasadniło sięgnięcie po metody meta-uczenia.

Trzecim zagadnieniem była ocena przydatności uczenia z niewielkiej liczby przykładów oraz opracowanie modyfikacji algorytmu MAML. Doktorant wykazał, że przy małej puli zadań meta-inicjalizacja nie różni się w praktyce od losowej i zaproponował dwuetapowy protokół treningu z wykorzystaniem danych symulowanych metodą COSMO-RS, podnoszący R^2 z 0,518 do 0,643. Właściwą przewagą meta-uczenia wykazano w modelowaniu toksyczności cieczy jonowych, gdzie MAML zachowywał zdolność predykcyjną przy zbiorach adaptacyjnych liczących 4–16 związków a więc w zakresie, w którym metody drzewiaste zawodzą. Na tej podstawie zaproponowano autorską modyfikację Att-MAML, w której w pętli wewnętrznej aktualizowana jest wyłącznie warstwa skalująca istotność cech. Modyfikacja ta wyraźnie ograniczyła odchylenia standardowe metryk, zmniejszając zależność wyniku od losowego doboru zbioru adaptacyjnego, co przy kilku dostępnych pomiarach ma zasadnicze znaczenie praktyczne.

Ostatnim wyzwaniem, które podjął Doktorant, było przeniesienie wniosków dotyczących podobieństwa zadań na grunt modelowania sorpcji. Na zbiorze współczynników aktywności w rozcieńczeniu nieskończenie wielkim wykazano, że skuteczność modelu zależy istotnie statystycznie (test Kruskala-Wallisa, $p = 0,04$) od podobieństwa zadania testowego do zadań meta-treningowych, przybliżanego współczynnikiem Tanimoto między cząsteczkami solutów. W odpowiedzi zaproponowano drugi autorski algorytm, TSA-Reptile, w którym funkcja straty jest skalowana odwrotnie do podobieństwa zadania, dzięki czemu zadania nietypowe uzyskują większą wagę podczas meta-treningu. Metoda okazała się skuteczniejsza od MAML dla pięciu z sześciu zadań odległych, co prowadzi do praktycznego wniosku o doborze algorytmu w zależności od położenia zadania względem danych treningowych.

Wysoko oceniam uzyskane wyniki, które w całości zostały już poddane weryfikacji przez niezależnych recenzentów w renomowanych czasopismach naukowych. Za szczególnie wartościowe uważam dwie autorskie modyfikacje algorytmów meta-uczenia oraz systematyczne rozstrzygnięcie kwestii reprezentacji grafowej układów dwucząsteczkowych. Warto podkreślić, że Doktorant udostępnił opracowane kody źródłowe w publicznym repozytorium GitHub, zgodnie z zasadami otwartej nauki, co umożliwia weryfikację i powtórzenie przeprowadzonych analiz. Język angielski, którym posługuje się Doktorant, jest na wysokim poziomie a przygotowane grafiki są czytelne i przejrzyste.

Lektura rozprawy nasunęła mi dwa pytania, które przedstawiam poniżej.

1. W rozdziale 5 Autor zwraca uwagę na znaczną zmienność wyników predykcji w reżimie *few-shot*, wyrażoną wysokimi odchyleniami standardowymi metryk przy zbiorach adaptacyjnych liczących 2–8 związków i próbuje ją ograniczyć poprzez autorską modyfikację Att-MAML. Jednocześnie z przedstawionych wyników wynika, że przewaga poszczególnych metod zależy od liczności zbioru adaptacyjnego — meta-uczenie dominuje poniżej pewnego progu, powyżej którego skuteczniejszy okazuje się Gradient Boosting. W związku z tym nasuwa się pytanie, czy Autor rozważał zastosowanie modelowania konsensusowego, łączącego predykcje wypracowanych algorytmów meta-uczenia z klasycznymi modelami zespołowymi, takimi jak Gradient Boosting? Uśrednienie lub ważenie predykcji modeli o odmiennym charakterze błędów mogłoby stanowić alternatywną, a przy tym tańszą obliczeniowo drogę do redukcji wariancji niż modyfikacja samego algorytmu meta-uczenia.
2. W rozdziale 5.2 wykazano, że modele grafowych sieci neuronowych trenowane na danych surowych przewyższają modele oparte na danych poddanych eksperckiemu czyszczeniu, co wyjaśniono szerszym pokryciem przestrzeni chemicznej przez zbiór surowy. Alternatywnym wyjaśnieniem tej obserwacji pozostaje jednak sama większa

liczność tego zbioru. Czy przeprowadzono porównanie przy wyrównanej liczbie rekordów, pozwalające rozdzielić wpływ różnorodności strukturalnej od wpływu wielkości zbioru treningowego?

Podsumowując, bardzo pozytywnie oceniam rozprawę doktorską Magistra inżyniera Karola Barana. Podjęta przez niego tematyka jest wymagająca, sytuując się na styku chemii cieczy jonowych i zaawansowanych metod uczenia maszynowego a uzyskane wyniki świadczą, że bardzo dobrze opanował on wyżej wymienione zagadnienia. Całość dorobku Doktoranta, pięć publikacji w renomowanych czasopismach, we wszystkich jako pierwszy autor oraz samodzielnie pozyskane finansowanie badań ze środków Narodowego Centrum Nauki wskazuje, że bardzo dobrze opanował on warsztat badawczo-naukowy i jest przygotowany do samodzielnego prowadzenia pracy naukowej.

Jednoznacznie stwierdzam, że przedstawiona mi do oceny rozprawa doktorska Magistra inżyniera Karola Barana spełnia wymogi stawiane rozprawom doktorskim zgodnie z art. 187 ustawy z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2023 poz. 742) i zwracam się o dopuszczenie Pana mgr. inż. Karola Barana do dalszych etapów postępowania w sprawie nadania stopnia doktora. Jednocześnie ze względu na wysoki poziom prowadzonych badań wnioskuję o wyróżnienie niniejszej rozprawy doktorskiej.