Prof. dr hab. inż. Krzysztof Krawiec
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 11.08.2025

Recenzja rozprawy doktorskiej mgr Łukasza Maziarki p.t. "Adapting Deep Learning Architectures for Drug Discovery"

1. Tematyka rozprawy

Przedmiotem rozprawy doktorskiej mgr Łukasza Maziarki są modele uczenia głębokiego dla zadań typowych dla zastosowań biochemicznych, w szczególności w zakresie predykcyjnym oraz generatywnym. Tematyka ta plasuje się w uczeniu maszynowym (*machine learning*), co pozwala mi stwierdzić że praca jednoznacznie wpisuje się w obszar dziedziny Informatyka Techniczna i Telekomunikacja.

2. Ocena treści rozprawy i wkładu oryginalnego

2.1 Treść rozprawy

Rozprawa podsumowuje przyczynki i powiązane z nimi opublikowane prace mgr Maziarki w dwóch grupach: metod predykcyjnych (Część I, rozdział 2) i generatywnych (Część II, Rozdział 3).

W części I, w sekcji 2.1 (i powiązanej z nią publikacji P1) Doktorant przedstawił komponent Set Aggregation Network (SAN), który można interpretować jako uogólnienie operacji agregacji (pooling). Autor wykorzystał SAN jako komponent splotowej sieci grafowej (Graph Convolutional Network, GCN) zaprojektowanej dla zadania oceny toksyczności związków chemicznych, i porównał jego przydatność z tradycyjną agregacją przez sumę, wykazując zauważalną przewagę sieci GCN wyposażonej w SAN.

W sekcji 2.2, Autor zaproponował model Spatial Graph Convolution Network (SGCN, powiązany z artykułem P2), który w mojej ocenie stanowi pewną nowość w zestawieniu z GCN, dzięki bezpośredniemu uwzględnieniu współrzędnych węzłów oraz uczalnemu odwzorowaniu wektora różnicowego na czynnik ważący stany sąsiednich węzłów (Eq. 2.7), co uważam za zasadny zabieg,

gdyż znacząco wzbogaca on reprezentację danych wejściowych. Autor wykazał teoretycznie że model ten generalizuje klasyczne sploty na kratach/rastrach (Tw. 2, s. 27) i zilustrował przydatność tego modelu w prostym zadaniu rozpoznawania ręcznie pisanych znaków (benchmark MNIST) oraz predykcji właściwości związków chemicznych (benchmark MoleculeNet).

Sekcja 2.3 prezentuje drugi z kluczowych przyczynków rozprawy (powiązany z artykułem P3), Molecule-augmented Attention Transformer (MAT), model bazujący na architekturze typu transformer, dedykowany do zastosowań w analizie związków chemicznych. Jak w przypadku innych zastosowań wzorca Transformer, kluczową wartością dodaną jest mechanizm samo-uwagi (self-attention), który pozwala modelowi przywiązywać różną wagę do poszczególnych komponentów danych wejściowych (tu: atomów wchodzących w skład analizowanych cząsteczek) oraz uczyć się modelować to ważenie. Doktorant zaproponował rozbudowanie tego mechanizmu o informację topologiczną (pochodzącą z macierzy incydencji grafu reprezentującego molekułę) oraz geometryczną (skonstruowaną na bazie macierzy odległości pomiędzy atomami wchodzącymi w skład molekuły). Empiryczne porównanie MAT (w połączeniu z automatycznym strojeniem hiperparametrów) wykazało jego przewagę nad pięcioma referencyjnymi metodami (oraz jedną ablacją MATgraph) na 6 z 8 rozważanych benchmarków (a wariant MAT z powiększonym budżetem strojenia hiperparametrów osiągnął najlepsze wyniki na wszystkich benchmarkach). Rozdział prezentuje także ciekawą analizę działania nauczonego mechanizmu atencji na wybranej molekule i benchmarku, oraz ilustrację przydatności uczenia wstępnego, które pozwoliło m.in. na znaczącą redukcję nakładów obliczeniowych zainwestowanych w strojenie hiperparametrów (Rys. 2.11).

Sekcja 2.4 (i artykuł P4) proponuje uogólnienie modelu MAT do modelu Relative MAT (R-MAT), którego głównym przyczynkiem jest zastąpienie ‘zwykłego’ mechanizmu samo-uwagi mechanizmem względnym opartym na relatywny kodowaniu pozycyjnym, angażującym trzy specyficzne dla dziedziny źródła informacji: grafową odległość pomiędzy atomami, informację o typie wiązania chemicznego, oraz odległość pomiędzy atomami kodowaną pozycyjnie przy pomocy kilku funkcji o symetrii kołowej (radial basis functions). W tej samej pracy/rozdziale autor proponuje także dwa nowe mechanizmy samo-nadzorowanego uczenia wstępnego, przez maskowanie całych podstruktur (local substructure masking) i predykcję innych zmiennych zależnych (global graph-level prediction, s. 46). Eksperymentalna ocena metody wykazała jej przewagę nad MAT (a czasami równą skuteczność) oraz kilkoma metodami referencyjnymi (m.in. Tabele 2.9 i 2.10).

Sekcja 2.5 (i artykuł P5) prezentują bibliotekę programistyczną i repozytorium modeli Hugging Molecules, w której mgr Maziarka zebrał autorskie i inne modele, udostępniając je społeczności naukowców i praktyków uczenia maszynowego na zasadach wolnego oprogramowania. Biblioteka pozwala na łatwe korzystanie z istniejących architektur i wstępnie nauczonych modeli oraz konstruowanie nowych architektur. Rozdział podsumowuje także wyniki pokaźnego eksperymentu obliczeniowego porównującego modele oferowane w bibliotece.

Sekcja 2.6 i praca P6 mają charakter stricte eksperymentalny i opisują wyniki empirycznego porównania użyteczności różnych atrybutów atomów i cech cząsteczek chemicznych w zadaniach

predykcji właściwości związków chemicznych postawionych w kilku znanych benchmarkach (QM9, ESOL, HUMAN, RAT). Eksperyment zaowocował wartościowymi konkluzjami odnośnie szczególnej istotności niektórych atrybutów (m.in. obecność ciężkich sąsiadów i atomów wodoru) a także mniej oczywistymi (okazyjnie szkodliwy wpływ uwzględniania informacji o ładunku i obecności pierścieni aromatycznych).

Sekcja 2.7 jest mniej powiązana z głównym nurtem pracy i prezentuje propozycję modelu językowego EmBERT, bazującego na modelu BERT. Głównym przyczynkiem rozdziału jest wprowadzenie zanurzeń zadań, umożliwiających wykorzystanie modelu w scenariuszach uczenia wielozadaniowego (multi-task learning). Prezentowane w rozdziale wyniki eksperymentu przemawiają na korzyść proponowanej metody, w zestawieniu z analogicznymi podejściami znanymi z literatury i wariantami metody poddanymi ablacjom.

W części II, autor rozprawy prezentuje osiągnięcia bazujące na zaproponowaniu architektur generatywnych dla zastosowań biochemicznych. W rozdziale 3.1, prezentuje architekturę Mol-CycleGAN, która jest interesującą hybrydą specyficznego wariantu autoenkodera wariacyjnego (Junction Tree Variational Encoder, JT-VAE) oraz znanej koncepcji cyklicznej architektury generatywnej CycleGAN. Docelowym scenariuszem użycia jest poprawa (optymalizacja) właściwości projektowanych molekuł (na bazie wyjściowych molekuł, tzw. leads). Doktorant zastosował proponowany model do benchmarku ZINC-250k, gdzie zadaniem stawianym przed modelem jest strukturalna transformacja molekuł wejściowych (np. dodawanie/usuwanie specyficznych grup funkcyjnych, np. pierścieni aromatycznych). Inne eksperymenty prezentowane w tej sekcji dotyczą ograniczonej i nieograniczonej optymalizacji cząsteczek oraz optymalizacji konkretnych właściwości cząsteczki (dokładniej: binding affinity). Rezultaty potwierdzają zalety metody i jej przewagę nad metodami referencyjnymi (m.in. JT-VAE i GCPN).

Przedmiotem rozdziału 3.2 jest propozycja metody PluGeN, która umożliwia ‘rozplecenie’ (disentanglement) wstępnie nauczonego, zamrożonego modelu bez jego modyfikacji, a następnie warunkowe generowanie przez odpytywanie modelu, dając w ten sposób bogate możliwości kontrolowania charakterystyki produkowanych artefaktów. Proponowana metoda pozwala na kontrolowanie procesu generowania zarówno zmiennymi ciągłymi jak i dyskretnymi, i może być wykorzystywana do generowania zupełnie nowych próbek lub modyfikowanie danych próbek. Przeprowadzony eksperyment wykazał skuteczność PluGeN w procesie generowania cząsteczek ‘de novo’ (Fig. 3.13) oraz w procesie optymalizacji wybranych właściwości cząsteczek (lead optimization, Rys. 3.14).

Rozdział 4 prezentuje profil Doktoranta a rozdział 5 podsumowuje rozprawę, dyskutuje jej przyczynki i oczekiwany impakt, oraz wyznacza kierunki dalszych prac.

2.2 Ocena wkładu oryginalnego i prezentacji pracy

Praca jest relatywnie zwarta i skupiona na docelowych zastosowaniach biochemicznych, mimo dużej liczby zróżnicowanych przyczynków (w sumie 7 sekcji proponujących nowe metody/komponenty). Opisy metod i eksperymentów są zwięzłe i rzeczowe, a dla bardziej detalicznych wyników (np. dowodów twierdzeń) zapewnione są odnośniki do źródłowych artykułów Doktoranta.

Autor adekwatnie bazuje na metodologii uczenia maszynowego, wstępnie aplikując metody do prostych problemów (np. MNIST dla SGCN czy predykcja relacji dwóch podstruktur dla MAT, s. 35) czy stosując m.in. uczenie wstępne modeli (pre-training, np. samo-nadzorowane uczenie wstępne przez predykcję wymaskowanych atomów w MAT, s. 35). Wszystkie elementy metod wydają się dobrze przemyślane i uzasadnione (np. zasilanie metod dodatkowymi danymi wejściowymi, sposoby ich kodowania, dummy node w MAT i R-MAT), a niezbędność znacznej części z nich została empirycznie wykazana w eksperymentach ablacyjnych. Część eksperymentów było powtarzanych celem uzyskania większej wiarygodności wyników, w szczególności porównania poszczególnych metod. Autor przeprowadzał także automatyczną optymalizację hiperparametrów (co prawda bez wykorzystania dedykowanych do tego metod, np. SMAC). Proponowane algorytmy zestawiane były z szeroką gamą metod referencyjnych (np. 9 metod w Tabeli 2.3). Autor stosował także adekwatne metody augmentacji danych, zarówno generycznych (np. obroty molekuł) jak i specyficznych dla dziedziny zastosowania (różne konformacje molekuł, Rys. 2.6) i wykazał ich skuteczność, przejawiająca się wzrostem skuteczności predykcyjnej modelu.

Uważam że wiele (jeśli nie większość) elementów proponowanych metod ma charakter nowatorski, najczęściej w sensie rozbudowania znanej wcześniej metody/architektury (np. wsparcie mechanizmów uwagi dodatkowymi informacjami) a w niektórych przypadkach w sensie bardziej fundamentalnym, np. metoda PluGeN umożliwiająca 'wpięcie' kontroli generowania przykładów do już nauczonego, zamkniętego modelu. Szczególnie interesujące i przekonujące były dla mnie wykorzystanie dostępnej wiedzy dziedzinowej w projektowaniu metod, w tym informacji o geometrii i topologii cząsteczek (głównie w metodach MAT i R-MAT) czy bardziej wyrafinowanych cech, np. wyodrębnienie kategorii cięższych atomów (Tab. 2.4), czy kodowanie informacji o typie wiązania w Tabeli 2.8.

W pracy dopatrzyłem się jedynie pojedynczych uchybień i niejasnych fragmentów. W sekcji 2.1.4 pojawia się sugestia że model SAN wykształca projekcje które są *optymalne*, co w ogólności nie jest prawdą, biorąc pod uwagę heurystyczną naturę gradientowych algorytmów uczenia. Pod równaniem 2.7 na s. 26 pojawia się wyjaśnienie parametru d, podczas gdy we wzorach pojawia się D. Drugi akapit na stronie 77 definiuje ścieżkę (path) rozpiętą na dwóch punktach (x i G(x)), co uważam za nieco dezorientujące (tj. poszukiwałem w pobliżu definicji takiej ścieżki w postaci *listy* punktów), bo ta para punktów wyznacza jedynie *odcinek*, po którym (jak rozumiem) iteruje następnie proces odpytywania dekodera.

Moja jedyna poważniejsza wątpliwość związaną z rozprawą dotyczy nowatorskiego charakteru metody agregacji (poolingu) SAN prezentowanej w sekcji 2.1. Jak wynika z formuły 2.2 i towarzyszącego jej opisu, SAN to złożenie (i) funkcji aplikowanej niezależnie do każdego elementu (wektora) zbioru (realizowanej jako uczalna podsieć neuronowa), (ii) sumy, oraz (iii) uczalnej funkcji (podsieci) mapującej otrzymaną sumę na wektor wyjściowy. Doktorant podkreśla że kluczową 'wartością dodaną' SAN jest parametryczny charakter (i) i (ii). Jednak nawet prosta agregacja przez sumowanie (sum pooling, odpowiednik (ii)) zazwyczaj umieszczana jest w sieciach głębokich pomiędzy uczalnymi podsieciami, co wydaje się prowadzić do funkcjonalnie równoważnego złożenia komponentów (i), (ii) i (iii). Jeżeli moja diagnoza jest poprawna, poprawy skuteczności uzyskiwane przez SAN względem sum pooling (Tabela 2) byłyby prawdopodobnie jedynie efektem zwiększonej liczby parametrów dostępnych w modelu wyposażonym w SAN.

Powyższe wątpliwości nie mają jednak krytycznego charakteru i nie wpływają na moją ogólnie pozytywną ocenę rozprawy.

Prace stanowiące podstawę rozprawy ukazały się w materiałach poważanych konferencji i organizowanych przy nich workshopach (m.in. ICML, ICLR, NeurIPS, AAAI). Udział mgr Maziarki w tych pracach wydaje się istotny, z drobnymi wyjątkami (np. Doktorant szacuje swój przyczynek do artykułu P2 (Sekcja 2.2) na zaledwie 10% i nie pada tam jednoznaczne stwierdzenie o autorstwie kluczowej idei SGCN; niemniej zakładam że udział ten był esencjonalny, biorąc pod uwagę poczesne ostatnie miejsce Doktoranta na liście autorów, a także kontekst pozostałych rozdziałów). Szacowany przez Doktoranta wkład w prace stanowiące podstawę większości sekcji utrzymuje się w okolicach 50% i w tym sensie nie budzi już wątpliwości. Opublikowane przez mgr Maziarkę prace doczekały się razem już niemal tysiąca cytowań na moment przygotowywania recenzji (wg Google Scholar), co jest imponującym wynikiem jak na ten etap kariery naukowej.

3. Konkluzja końcowa

Rozprawa doktorska mgr Łukasza Maziarki zawiera długą listę interesujących i wartościowych przyczynków o nowatorskim charakterze koncepcyjnym i znacznym potencjale aplikacyjnym, potwierdzonym już badaniami empirycznymi, podpartą wartościowymi publikacjami o zasięgu międzynarodowym. Uważam zatem że **spełnia ona z solidną nawiązką warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony, o co wnoszę do Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Uniwersytetu Jagiellońskiego.**

Ponadto, biorąc pod uwagę nowatorski charakter przyczynków, skuteczne wykazanie konkurencyjności proponowanych metod względem istniejących algorytmów, znaczne przełożenie praktyczne wyników, publikacje autora w materiałach czołowych konferencji i czasopismach oraz ich liczne cytowania, **wnioskuję o wyróżnienie rozprawy**.