Wrocław 3.07.2023

dr hab. inż. Henryk Maciejewski
Katedra Informatyki Technicznej
Politechniki Wrocławskiej

**Recenzja pracy doktorskiej pani mgr Magdaleny Wiercioch**
**pt. „Development of universal data representations with application in chemistry"**

## 1. Tematyka, cel i zakres rozprawy

Recenzowana praca doktorska została napisana pod opieką prof. Jacka Tabora na Wydziale Fizyki, Astronomii i Informatyki Stosowanej Uniwersytetu Jagiellońskiego. Rozprawa dotyczy metod tworzenia reprezentacji cząsteczek związków chemicznych, które umożliwią badanie własności tych związków za pomocą metod uczenia maszynowego, przy czym własności, na których skupia się Autorka dotyczą przydatności związków w zadaniach projektowania leków. Problematyka rozważana w pracy sytuuje się w obszarze informatyki określanym jako chemioinformatyka (albo informatyka chemiczna), która rozwija metody analizy i modelowania danych chemicznych. Metody te znajdują zastosowanie w procesie opracowywania nowych leków – szczególnie we wczesnych etapach tego procesu – ułatwiają bowiem znajdowanie molekuł, które są obiecującymi kandydatami na nowo opracowywany lek poprzez komputerowe przeszukiwanie dużych baz związków chemicznych. Rozwiązanie takiego zadania jest możliwe dzięki metodom obliczeniowym, które pozwalają przewidywać, na podstawie analizy struktury molekuły, własności związku, które powinien wykazywać kandydat na lek, takie jak spodziewana aktywność molekuły wobec zidentyfikowanego celu biologicznego dla projektowanego leku (np. białka, transkryptu czy genu), czy też bezpieczeństwo związane z poziomem toksyczności związku. Badania przedstawione w dysertacji dotyczą rozwoju tego typu metod obliczeniowych. Głównym celem doktoratu jest rozwój i zbadanie metod generowania reprezentacji molekuł, na podstawie m.in. ich struktury grafowej oraz opisu struktury w tekstowym formacie SMILES, a także na podstawie własności chemicznych. Reprezentacje tworzone za pomocą proponowanych metod umożliwiają badanie aktywności i bezpieczeństwa związków chemicznych – kandydatów na nowe leki metodami

obliczeniowymi (badania *in silico*), przed wykonaniem chemicznych bądź biologicznych badań laboratoryjnych (*in vitro*, *in vivo*). Przydatność zaproponowanych metod uczenia reprezentacji badana była przez Autorkę w trzech przykładowych zastosowaniach: w zadaniu przewidywania aktywności biologicznej molekuł (zadanie klasyfikacji), w zadaniu oceny toksyczności związku (klasyfikacja i regresja) oraz w zadaniu oceny czy związek chemiczny i cel biologiczny (w pracy jako cel rozważano białko) wejdą w interakcję (zadanie określane jako DTI – *drug-target interaction*).

W mojej ocenie problematyka rozważana w pracy jest ważna i dobrze uzasadniona, zaś metody rozwijane przez Autorkę w dotyczą aktualnych i istotnych wątków badawczych o dużym znaczeniu naukowym i aplikacyjnym. Tematyka doktoratu dobrze mieści się w obszarze zainteresowania dziedziny Informatyka Techniczna i Telekomunikacja.

## 2. Zawartość rozprawy

Praca – napisana w języku angielskim – składa się z 5 rozdziałów, podsumowania oraz dodatku. Pierwsza część pracy obejmuje dwa rozdziały: Wstęp zawierający sformułowanie problemu, cel, zadania wykonane w pracy i główne osiągnięcia oraz rozdział drugi zawierający wprowadzenie do zagadnień chemioinformatyki, głównie w kontekście projektowania leków. W rozdziale tym zawarte też jest zwięzłe wprowadzenie do metod reprezentowania / zapisywania struktury grafowej molekuł, oraz przedstawienie aktualnego stanu badań dotyczących generowania reprezentacji molekuł dla potrzeb modelowania ich własności za pomocą uczenia maszynowego.

Główne wyniki przedstawione są w drugiej części pracy obejmującej rozdziały od trzeciego do piątego. W rozdziale trzecim przedstawiona została metoda reprezentacji molekuł HybNN, która łączy wektor cech generowany na podstawie zapisu struktury cząsteczki w tekstowym formacie SMILES z wektorem cech zbudowanym na podstawie grafowej jej reprezentacji. Graf cząsteczki analizowany jest z wykorzystaniem grafowej sieci konwolucyjnej. Finalny wektor cech dla molekuły wykorzystywany jest do klasyfikacji molekuł jako aktywne/nieaktywne biologicznie w stosunku do zadanych celów biologicznych. W rozdziale tym przedstawiono wyniki predykcji aktywności biologicznej dla 14 zbiorów związków chemicznych, gdzie każdy ze zbiorów dotyczył innego celu biologicznego. Wyniki wskazują, że metoda HybNN osiąga dla większości benchmarków lepszą skuteczność predykcji (raportowaną jako AUC-ROC) niż metody konkurencyjne.

W rozdziale czwartym przedstawiona została metoda Subgraph Encoded Neural Network (SENN). Algorytm ten został opracowany dla potrzeb ilościowej i jakościowej oceny toksyczności molekuł. SENN tworzy cechy na podstawie podgrafów zawartych w grafie opisującym strukturę molekuły; cechy te są dalej łączone z globalnymi cechami związku (takimi jak np. masa cząsteczkowa). Metoda została zbadana w zadaniu klasyfikacji związków pod względem toksyczności. Wyniki tych badań wskazują, że metoda ma lepszą skuteczność (mierzoną jako AUC-ROC) od metod konkurencyjnych (lepszą skuteczność pokazano dla 4 z 12 zbiorów testowych). Metoda SENN została również zbadana w zadaniu ilościowej oceny toksyczności – w tym przykładzie pokazano lepszą dokładność regresji na dwóch z czterech zbiorów testowych. Należy dodatkowo zwrócić uwagę na interpretowalność metody SENN – do tej ważnej własności metody odniosę się bardziej szczegółowo w części 3 recenzji.

W rozdziale piątym Autorka przedstawiła model Triplet Encoded Neural Network (TENN), którego celem jest predykcja, czy związek wykaże aktywność wobec zadanego celu biologicznego (w badaniach celem tym są białka) – zadanie znane jako DTI (*Drug-Target Interaction*). Metoda zakłada (i) generację wektora cech dla związku chemicznego korzystając z jego reprezentacji w formacie SMILES, (ii) generację wektora cech na podstawie struktury grafowej uzyskanej z reprezentacji SMILES, (iii) generację reprezentacji białka na podstawie sekwencji aminokwasów korzystając z algorytmu Word2vec. W rozdziale przedstawiono też wyniki predykcji DTI dla czterech zbiorów testowych, w każdym z tych przypadków wyniki uzyskane przy pomocy proponowanej metody są lepsze niż wyniki metod literaturowych (w trzech z czterech zbiorów poprawa jest statystycznie istotna).

Rozdział szósty zawiera krótkie podsumowanie pracy i zarysowuje otwarte problemy, które mogą być przedmiotem dalszych badań.

## 3. Ocena rozprawy

Główne osiągnięcia Autorki polegają na zaproponowaniu trzech metod generowania reprezentacji cząstek chemicznych – metody HybNN, SENN i TENN. Istota tych podejść polega na łączeniu reprezentacji generowanych na podstawie opisu molekuł w tekstowym formacie SMILES z reprezentacją generowaną na podstawie struktury grafowej molekuł, oraz dodatkowo na uwzględnianiu w tworzonym wektorze cech również globalnych atrybutów związków, takich jak cechy fizykochemiczne. Wyniki zaprezentowane przez Autorkę wskazują, że tak generowane wektory cech umożliwiają budowanie modeli klasyfikujących

i/lub regresyjnych pozwalających na predykcję własności molekuł, takich jak aktywność biologiczna, toksyczność czy interakcja z zadanym białkiem (celem leku). Dodatkowo ciekawym elementem procedur generacji reprezentacji ze struktur grafowych jest propozycja generowania reprezentacji dla podgrafów zawartych w grafie molekuły (takie podejście zastosowano w metodzie SENN i TENN). Otwiera to nowe możliwości – w stosunku do metody generacji wektora cech z całej molekuły (jak w HybNN) – w zakresie badania interpretowalności otrzymywanych reprezentacji.

Poniżej ocenię najciekawsze i najbardziej oryginalne wyniki w ramach poszczególnych metod.

Metoda HybNN (rozdział 3)

Autorka zaproponowała metodę tworzenia reprezentacji związku chemicznego poprzez połączenie informacji wydobytych z opisu struktury związku w tekstowym formacie SMILES oraz informacji z grafowej postaci związku. Najciekawszymi osiągnięciami Autorki w tej części pracy są: (i) propozycja oryginalnych metod uczenia reprezentacji cząsteczek na podstawie każdego z tych źródeł; (ii) empiryczna weryfikacja wykazująca, że metoda jest praktycznie stosowalna - można przy jej pomocy wygenerować reprezentacje związków na podstawie publicznie dostępnych danych dotyczących przewidywania aktywności biologicznej związków, i reprezentacje te w wielu benchmarkach prowadzą do lepszej skuteczności predykcji aktywności niż metody literaturowe; (iii) analiza własności tworzonych reprezentacji związków, w tym wykazanie, że większą wartość informacyjną ma komponent oparty na grafie niż na zapisie SMILES.

Tworzenie reprezentacji związku na podstawie zapisu struktury w formacie SMILES wykorzystuje metody przetwarzania języka naturalnego: (i) reprezentowanie podciągów znaków w ciągu SMILES (traktowanych jak ‘słowa’) przez wektory osadzeń uzyskane z pretrenowanego modelu Word2vec, (ii) wyznaczenie reprezentacji SMILES jako sekwencji ‘słów’ za pomocą rekurencyjnej sieci neuronowego BiGRU. Natomiast tworzenie reprezentacji związku na podstawie jego struktury grafowej wykorzystuje algorytm grafowej sieci konwolucyjnej. Węzły grafu, reprezentujące atomy związku, inicjowane są przez wektory fizykochemicznych cech atomów; dalej algorytm GCN agreguje te wektory z cechami atomów w sąsiedztwie. W ten sposób zaproponowana przez Autorkę metoda uwzględnia zarówno cechy fizykochemiczne atomów wchodzących w skład związku, jak i ich strukturę połączeń. Oceniam obie te metody jako ciekawe i oryginalne osiągnięcia Autorki. Metody te przestawione zostały

jasno, ewaluacja nie budzi zastrzeżeń dotyczących stosowalności i skuteczności tych metod w przedstawionym zadaniu.

Metoda SENN (rozdział 4)

Metoda SENN polega na agregacji reprezentacji tworzonych na podstawie grafowej struktury związku oraz wektora fizykochemicznych cech związku. Występują jednak dwie istotne różnice w stosunku do metody tworzenia reprezentacji struktur grafowych proponowanej w HybNN: (i) reprezentacja oparta na grafowych sieciach konwolucyjnych (GCN) tworzona jest nie na podstawie grafu całego związku (jak w HybNN), ale dla podgrafów zawartych w grafie związku, rozpiętych jako sąsiedztwa wszystkich węzłów grafu (czyli atomów związku), przy czym reprezentacje podgrafów są w końcowej fazie agregowane do reprezentacji związku; (ii) reprezentacje węzłów podgrafów, aktualizowane przez algorytm GCN, inicjowane są losowo (w metodzie HybNN inicjalizacja wykonywana jest na podstawie cech fizykochemicznych atomów). Komentarz – uwagę dyskusyjną dotyczącą tego wątku formułuję w części 4 recenzji.

Za najciekawsze osiągnięcia Autorki w tej części rozprawy uważam: (i) propozycję oryginalnej metody reprezentacji struktury grafowej poprzez agregację reprezentacji podgrafów; (ii) empiryczną ewaluację proponowanej metody i wykazanie jej skuteczności w zadaniu predykcji toksyczności związków chemicznych (ocena toksyczności wykonywana jako zadanie klasyfikacji lub regresji); (iii) pokazanie, że budowanie reprezentacji grafu jako agregacji reprezentacji podgrafów umożliwia badanie interpretowalności reprezentacji – ideę tego podejścia pokazano w rozdz. 4.4.7.

Ten trzeci wynik – dotyczący metod budowania reprezentacji grafowych tak, żeby możliwe (łatwe?) było uzyskanie interpretowalności reprezentacji uważam ze jedno z najciekawszych i najbardziej inspirujących osiągnięć rozprawy.

Opis metody SENN jest generalnie czytelny, ewaluacja jest przekonująca, z jednym zastrzeżeniem dotyczącym raportowania wyników regresji – patrz punkt 4 recenzji.

Metoda TENN (rozdział 5)

Zaproponowana w tej części rozprawy metoda tworzenia reprezentacji służy do badania interakcji pomiędzy związkiem chemicznym – potencjalnym lekiem a białkiem – celem leku (zadanie określane jako *drug-target interaction*, DTI). Główne oryginalne osiągnięcia Autorki w tej części to: (i) propozycja uczenia reprezentacji dla zadania DTI poprzez agregację informacji wydobytych z zapisu SMILES związku, jego struktury grafowej oraz struktury

pierwszorzędowej białka (sekwencji aminokwasów); (ii) opracowanie generatorów reprezentacji dla związku (na podstawie zapisu SMILES) oraz dla białka (na podstawie sekwencji aminokwasów) - wymagane tu było m.in. nauczenie modeli Word2vec do generowania wektorów osadzeń dla podciągów SMILES oraz dla aminokwasów; (iii) opracowanie metody reprezentacji grafu związku na podstawie reprezentacji podgrafów i sieci GAT (Graph Attention Network); (iv) empiryczna weryfikacja stosowalności i skuteczności metody w zadaniu przewidywania interakcji związek – białko. Metoda opisana jest w sposób jasny i kompetentny, weryfikacja empiryczna – zrobiona w przekonujący sposób. Drobne uwagi dotyczące prezentacji szczegółów technicznych metody– p. punkt 4.

Analiza źródeł

Spis literatury zawiera 176 pozycji. W większości są to artykuły w renomowanych czasopismach z obszaru chemioinformatyki, bioinformatyki czy uczenia maszynowego. Autorka korzystała też z prac publikowanych na renomowanych konferencjach międzynarodowych, a także z preprintów w serwisie arXiv. Pozycje literaturowe są dobrze dobrane i aktualne. Dobór literatury świadczy o dużej erudycji Autorki w zakresie metod informatycznych wykorzystywanych w modelowaniu związków chemicznych, metod bioinformatyki oraz metod uczenia maszynowego, w tym uczenia głębokiego. Obszerny zestaw bibliografii świadczy o wnikliwości Autorki i dużej pracy włożonej w analizę aktualnego stanu wiedzy w obszarze, w którym prowadziła badania.

## 4. Uwagi krytyczne i dyskusyjne

(i) Tabela 4.5, str. 74: błędny nagłówek, w tabeli przedstawiono wartości $R^2$, a nie RMSE, a ponieważ dla $R^2$ „higher is better" – wyróżnione powinny być wyniki dla testów IGC50 i LD50.

(ii) Generalna uwaga dotycząca prezentacji wyników dla algorytmów regresji – rozdz. 4.4.3: Autorka ograniczyła się do prezentacji miar RMSE i MAE dla swojej metody i metod konkurencyjnych, co oczywiście pozwala na porównanie skuteczności modeli. Jednak nie pozwala na ocenę na ile model regresyjny jest dokładny/skuteczny/pasuje do danych, stąd dla oceny modeli warto podawać RMSE/(średnia wartość zmiennej przewidywanej) – czyli CV, a także $R^2$. W rozdz. 4.4.6 Autorka podaje $R^2$, ale tylko dla wybranych analiz (możemy się przy okazji zorientować jaki jest zakres wartości zmiennej target – Fig. 4.7, i do tego zakresu

zmienności odnosimy raportowane wcześniej RMSE i MAE, żeby zweryfikować jakość tych modeli).

(iii) Str. 87: niejasny opis połączenia części Unit A i Unit B w generatorze reprezentacji białka: na wyjściu Unit A dostajemy reprezentację białka – czyli „concatenation of amino acid vectors", natomiast w podpisie do Fig. 5.4 czytamy: „an embedding vector with a fixed size from Unit A is used to describe a protein". Dla różnych białek nie możemy oczekiwać, że wyjście w Unit A będzie miało „fixed size".

(iv) Uwaga dotycząca inicjalizacji wektorów cech przypisanych do węzłów grafu, aktualizowanych następnie przez algorytm GCN – metody SENN i HybNN wykonują tą inicjalizację w różny sposób (HybNN uwzględnia cechy atomów, SENN – losowo). Dlaczego Autorka nie wykorzystała pomysłu z HybNN do inicjalizacji wag w SENN? W cechach generowanych przez SENN mamy zawartą informację o strukturze grafu/związku; czy uwzględnienie dodatkowo informacji o cechach atomów mogłoby poprawić skuteczność generowanych cech? Może warto metodę SENN rozwinąć w tym kierunku?

(v) W metodzie TENN (podobnie jak w SENN) korzystamy z podgrafów zawartych w grafie związku. Ciekawy jestem komentarza Autorki czy i w jaki sposób można by badać interpretowalność reprezentacji generowanych przez Algorytm 1 (rozdz. 5.3.3) – czyli analizować, które części związku wpływają na powodzenie interakcji DTI, podobnie jak to zrobiono w rozdz. 4.4.7.

(vi) Str. 89, linia 9 Algorytmu 1: sumujemy wartości $v_a$'', sumowanie po a \in V (a nie w V'') – pomyłki w oznaczeniach.

Powyższe uwagi krytyczne nie zmniejszą mojej generalnie bardzo dobrej opinii o czytelności i jasności przekazu poszczególnych wątków dysertacji, i o starannej edycji rozprawy. Natomiast uwagi (iv) i (v) nie mają charakteru krytycznego - mają na celu zainspirowanie Autorki do dyskusji, a może również rozwoju swoich metod.

## 5. Podsumowanie i wnioski końcowe

W swojej rozprawie p. mgr Magdalena Wiercioch podejmuje ciekawy, aktualny i ważny problem badawczy, który mieści się w obszarze zainteresowania dyscypliny Informatyka

Techniczna i Telekomunikacja. Autorka dysertacji ma ciekawy dorobek publikacyjny, obejmujący artykuły w czasopismach i na renomowanych konferencjach. To wskazuje, że mamy do czynienia z dojrzałą kandydatką do stopnia doktora.

W swojej dysertacji Doktorantka zaprezentowała szereg wyników, które wnoszą ciekawy, oryginalny i wartościowy wkład do dyscypliny naukowej.

W konkluzji stwierdzam, że rozprawa doktorska p. mgr Magdaleny Wiercioch pt. „Development of universal data representations with application in chemistry" spełnia wymagania ustawowe i wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony.