Prof. dr hab. inż. Jacek Rumiński,
Katedra Inżynierii Biomedycznej
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

Gdańsk, 21.04.2023 r.

# Recenzja

poprawionej rozprawy doktorskiej mgr Magdaleny Wiercioch

pt. „Development of universal data representations with application in chemistry".

Promotor: prof. dr hab. Jacek Tabor

Niniejsza recenzja, poprawionej rozprawy doktorskiej, została przygotowana w odpowiedzi na pismo prof. dr. hab. inż. Macieja Ogorzałka, Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Informatyka na Uniwersytecie Jagiellońskim w Krakowie, z dnia 16.03.2023 roku, informującego o powołaniu mnie dnia 19 maja 2022 roku na recenzenta rozprawy doktorskiej Pani mgr Magdaleny Wiercioch. W piśmie stwierdzono, że Doktorantka przedstawiła poprawioną wersję swojej pracy, którą otrzymałem w wersji drukowanej.

Niniejsza recenzja jest oceną aktualnej, tj. poprawionej wersji pracy, niemniej ze względu na niezmienność znacznej części rozprawy pierwotnej opinia ta zawiera liczne powtórzenia z pierwszej recenzji z dnia 22.07.2022 roku wraz z ustosunkowaniem się do wprowadzonych zmian. Wraz z poprawioną wersją rozprawy otrzymałem napisany przez Doktorantkę wykaz odpowiedzi do wcześniejszych uwag.

## 1. Tematyka, cele i tezy rozprawy

Tematyka pracy dotyczy zastosowania metod uczenia głębokiego do uzyskiwania nowych reprezentacji danych przydatnych w różnych problemach uczenia maszynowego związanych z zastosowaniami w chemii. Autorka rozprawy zaproponowała trzy algorytmy pozwalające na uzyskiwanie reprezentacji molekuł związanych z aspektami projektowania leków, w szczególności: klasyfikacji uzyskiwanych związków na aktywne i nieaktywne biologicznie, klasyfikacji i regresji toksyczności związków i oceny występowania interakcji lek – cel biologiczny.

Rozprawa jest napisana w języku angielskim z jasno postawionym pytaniem badawczym sformułowanym na stronie 7: „Does learning representation have an influence on drug discovery and development proces?". Swoiste hipotezy badawcze stanowią zaproponowane w kolejnych rozdziałach trzy metody, a ich weryfikacje przeprowadzono na drodze eksperymentalnej wykorzystując do tego znane, publiczne zbiory danych. Wyniki porównano do rezultatów innych zespołów badawczych.

## 2. Zawartość rozprawy

Rozprawa składa się ze streszczenia w języku angielskim, streszczenia w języku polskim oraz sześciu rozdziałów i jednego załącznika. Załącznik przedstawia ogólne wprowadzenie do uczenia głębokiego prezentując krótko podstawowe zagadnienia związane z metodologią prac badawczych.

W rozdziale pierwszym Doktorantka opisuje zadania badawcze, podkreśla własny wkład w obszarze badawczym zarówno bezpośrednio związany z rozprawą jak i innych projektów oraz przedstawia układ treści rozprawy. Pani mgr Magdalena Wiercioch zamieszcza również link do repozytorium z kodami zaimplementowanych metod.

Drugi rozdział przedstawia ogólne wprowadzenie do obszaru chemoinformatyki, klasycznych metod reprezentacji danych o molekułach oraz bardzo krótką informację o grafowych reprezentacjach danych.

Kolejne rozdziały, tj. 3, 4 i 5 stanowią opisy trzech metod zaproponowanych przez Panią mgr Magdalenę Wiercioch, stanowiących próbę definicji hipotezy/celu szczegółowego, metodologii oraz uzyskanych wyników i ich dyskusji.

W rozdziale trzecim, Doktorantka wprowadza nowy model o nazwie HybNN zaprojektowany do nauki reprezentacji danych w problemie klasyfikacji aktywności biologicznej związków chemicznych. Autorka przedstawia ogólny schemat metod, opisuje podstawowe bloki (np. BiGRU) i ich rolę. Prezentuje i omawia wyniki eksperymentów dla 14 zadań (podzbiorów) identyfikacji molekuł.

Rozdział czwarty to opis proponowanego modelu SENN i jego wykorzystania w problemach klasyfikacji i regresji w ocenie toksyczności związków chemicznych. Podobnie jak w rozdziale poprzednim Doktorantka przedstawia ogólny schemat metod, opisuje metodę konstrukcji podgrafów oraz możliwości interpretacyjne opracowanego modelu. Prezentuje i omawia wyniki eksperymentów dla 12 zadań (podzbiorów) oceny toksyczności ze zbioru Tox21 oraz dla czterech zbiorów w zadaniu regresji.

W rozdziale piątym Pani mgr Magdalena Wiercioch opisuje zaproponowany model TENN dla problemu przewidywania interakcji pomiędzy lekiem a celem biologicznym. Dlatego w przedstawionym modelu wyróżnia i opisuje trzy komponenty związane z reprezentacją sekwencji białek, grafową reprezentacją związku chemicznego oraz tekstową reprezentacją molekuły. Prezentuje sposób przygotowania pięciu zbiorów danych na podstawie informacji z repozytorium BindingDB oraz DrugBank. Przedstawia i opisuje wyniki eksperymentów.

W każdym z rozdziałów wprowadzającym nowy model w chemoinformatyce porównuje uzyskane wyniki do rezultatów innych prac badawczych podkreślając, że uzyskane miary są co najmniej na poziomie wartości metryk otrzymanych przez inne zespoły badawcze.

Rozdział szósty to bardzo krótkie podsumowanie rozprawy z opisem perspektyw dalszego rozwoju.

Autorka cytuje 176 pozycji źródłowych powiązanych z tematyka rozprawy i dobrze wkomponowanych w treść pracy.

Rozprawa jest właściwie uporządkowana strukturalnie, a prezentowane treści są w większości przedstawione w sposób czytelny i łatwy do lektury. W poprawionej wersji rozprawy doktorskiej Doktorantka skorygowała wskazane wcześniej błędy edytorskie. Forma prezentacji niektórych rysunków wymagałaby jednak korekty w celu poprawy kontrastu tekstu względem tła (np. niebieski tekst na zielonym tle, rys. 2.2; podobnie rys. 3.1, 3.2, 3.3).

## 3. Oryginalne osiągnięcia i ocena merytoryczna rozprawy

Do głównych osiągnięć rozprawy zaliczam opracowanie modeli uczenia reprezentacji dla trzech grup problemów dotyczących aspektów chemoinformatyki, tj. klasyfikacji uzyskiwanych związków na aktywne i nieaktywne biologicznie, klasyfikacji i regresji toksyczności związków i oceny występowania interakcji lek – cel biologiczny, w szczególności:

O1: zaproponowanie architektury modelu integrującego sieć rekurencyjną (BiGRU) z podejściem grafowym bazującego na dwóch różnych formach reprezentacji danych wejściowych i wykazanie, że model taki skutecznie uczy się reprezentacji prowadząc do wyników klasyfikacji aktywności związków na poziomie innych modeli znanych z literatury,

O2: zaproponowanie metody i modelu SENN wykorzystującej podgrafy i sieci neuronowe, łączącej dane przestrzenne z właściwościami związku chemicznego i wykazanie, że metoda taka skutecznie uczy się reprezentacji prowadząc do wyników klasyfikacji i regresji w zakresie toksyczności związków na poziomie innych modeli znanych z literatury,

O3: zaproponowanie metody i modelu TENN integrującego trzy komponenty w zakresie reprezentacji sekwencji białek, reprezentacji grafowej związku chemicznego i tekstowej reprezentacji i wykazanie, że wyodrębnienie globalnej informacji dotyczącej związku chemicznego i białka jako celu biologicznego poprawia wyniki klasyfikacji na poziomie innych modeli znanych z literatury.

W każdym przypadku Doktorantka dokonała wnikliwej analizy na różnych zbiorach danych z porównaniem do wyników wybranych metod znanych z literatury.

Zaproponowana metodologia jest zasadniczo poprawna, weryfikacja hipotez w większości właściwa, a rezultaty ilościowe wskazują wysokie wartości miar jakości uzyskiwanych wyników. Wyniki dotyczące osiągnięcia O1 opublikowała w ramach materiałów ELLIS Machine Learning for Molecule Discovery Workshop, 2021. Pani mgr Magdalena Wiercioch opublikowała również szereg powiązanych tematycznie prac.

W zakresie metodologii prowadzonych prac na podkreślenie zasługują również ciekawe analizy w zakresie zależności uzyskiwanego błędu od podobieństwa danych, analiza aspektów wyjaśniania modelu w odniesieniu do O2, analiza wpływu kluczowych komponentów metod na końcowe wyniki (ang. ablation studies).

Oceniam, że zaproponowane metody wraz z uzyskanymi wynikami stanowią ważne osiągniecie w tematyce badawczej związanej z dyscyplinami informatyki techniczna i telekomunikacja oraz inżynieria biomedyczna.

Niemniej, oceniając rozprawę dostrzegam pewne braki czy niedoskonałości omówione w kolejnym punkcie.

## 4. Uwagi krytyczne i pytania do Doktorantki

Autorka w poprawionej pracy wprowadziła korekty wskazywane przez mnie w recenzji wcześniejszej wersji rozprawy. Uważam, że Doktorantka właściwie wprowadziła zmiany w większości moich wcześniejszych uwag czy pytań. Jednak część wyjaśnień nie jest w mojej ocenie kompletna, co wskazuję poniżej.

W recenzji rozprawy pierwotnej wskazałem, że nie jest dla mnie jasne, dlaczego zastosowała w ostatnim kroku modelu HybNN jest funkcję softmax, w szczególności stosując logarytm z funkcji softmax. W poprawionej wersji pracy postać logarytmiczna pozostała (wzór 3.10). Doktorantka w przesłanych odpowiedziach wskazała, że wystarczyłaby sigmoidlana funkcja aktywacji a zastosowana funkcja logarytmiczna została wprowadzona „do zrzutowania wyjść z ostatniej warstwy modelu na rozkład prawdopodobieństwa w skali logarytmicznej". Uzyskując wartości ujemne (logarytm z wartości z zakresu <0,1>) trudno mówić rozkładzie prawdopodobieństwa. Co więcej, taka postać funkcji aktywacji wprowadza więcej problemów (np. log(0), itd.). Uważam, że Autorka powinna przedstawić argumenty, dlaczego nie zastosowała klasycznej wersji sigmoid/softmax, np. wspierając się dowodami z wyników eksperymentów.

Inna uwaga dotyczyła kodów źródłowych, dostępnych we wskazanym przez Doktorantkę repozytorium (link podany na stronie 10 (rozdział 1): https://bitbucket.org/mgdlnwrch.). Nie jest możliwe odtworzenie eksperymentów bezpośrednio z opublikowanego kodu ze względu na problemy z kodem (np. definicje zmiennych, itp.), brak opisu rzeczywistych wymagań (np. brakuje pakietów w requirements.txt), brak niektórych zbiorów danych lub szczegółowych informacji jak je wygenerować (np. „smile_n_gram.in"), itd. Podobne uwagi dotyczą pozostałych projektów, dla modeli SENN czy TENN. Przypuszczam, że repozytorium zawiera wersje eksperymentalne projektów, niemniej brak szczegółów dotyczących modeli (dla których uzyskano konkretne wyniki) głównie w treści rozprawy (ewentualnie we wskazywanych przez autorkę kodach) stanowi niedopatrzenie. Analizując kody źródłowe nie jest dla mnie jasne w jaki sposób Doktorantka wyznaczała miary jakości modeli. Przykładowo, w implementacji modelu HybNN zawarto wyznaczenie wartości RMSE (brak nawet ujętego w komentarz kodu do obliczania AUC-ROC). W rozprawie, w rozdziale 4 (tabela 4.3) wskazuje wyniki AUC-ROC. W modelu SENN do obliczenia AUC (przypuszczenie na podstawie nazw zmiennych „auc_dev", „auc_test") użyto funkcji *accuracy_score* z pamiektu *sklearn.metrics*. W przesłanych odpowiedziach autorkach wskazuje, że „Autorka udostępniła kody źródłowe służące do przygotowania pracy", ponadto wskazała, że „pełną/rozszerzoną wersję implementacji" planowała opublikować po otrzymaniu recenzji. Nie jest dla mnie jasne, dlaczego dopiero po recenzjach. Ponadto Doktorantka otrzymała już komplet recenzji pracy pierwotnej, a nie spowodowało to udostępnienia rozszerzonej wersji

kodów źródłowych. Jeśli autorka nie planowała tego zrobić (co oczywiście może być zrozumiałe z różnych powodów) należało to skomentować przy podaniu linku do kodów źródłowych w treści rozprawy.

Doktorantka dokonała korekty w zakresie większości uwag dotyczących sposobu prezentacji wyników badań (m.in., prezentacji wartości miar oceny wyników). Zmodyfikowała również część opisu w zakresie odniesienia się do innych prac. Niemniej, w dyskusji wyników odwołania do rezultatów innych badań są wciąż ubogie. Przykładowo Doktorantka korzysta ze zbioru Tox21, który związany jest z konkursem ogłoszonym wiele lat temu. W pracy sprzed prawie 6 lat: Mayr A, Klambauer G, Unterthiner T and Hochreiter S (2016) DeepTox: Toxicity Prediction using Deep Learning. Front. Environ. Sci. 3:80. doi: 10.3389/fenvs.2015.00080, autorzy podają bogate zestawienie wyników dla 18 zaproponowanych modeli (w tym ich własny). Przykładowo dla zestawu danych AR wynik AUC-ROC uzyskany przez model *dmlab* wynosi 0,828, a więc jest lepszy od wyniku dla SENN, który Doktorantka wskazuje jako najlepszy w odniesieniu do innych zbadanych metod oraz jawnie wskazuje jako osiągnięcie na stronie 55 (w pierwszej wersji rozprawy na stronie 53): „For instance, on the AR toxicity dataset, our SENN obtains significantly better AUC-ROC score ($0.802 \pm 0.006$) than that by the well-known prediction approaches including GCNN, SVM, and GIN". Oczywiście w przytoczonej przeze mnie pracy (czy cytowanych innych pracach) mogły być inne warunki eksperymentów, niemniej Doktorantka powinna to w mojej ocenie przytoczyć i poddać dyskusji. Autorka niestety nie odniosła się do wyników uzyskiwanych dla tego samego zbioru przez m.in. wskazany wyżej model grupy *dmlab* lub innych.

Wskazałem również, nieprecyzyjną prezentację wyników uzyskanych przez metodę SENN dla zbioru AR. We wskazanym wyżej cytacie ze strony 55 Autorka podaje wynik AUC-ROC dla zbioru AR jako $0,802 \pm 0,006$. Podczas gdy, w zestawieniu w tabeli 4.3 wynik ten podano jako $0,816 \pm 0,007$. Wynik 0,802 uzyskano porównując model SENN w wersji z i bez uwzględnienia cech fizykochemicznych w tabeli 4.6. Doktorantka wyjaśniła, że wynik 0.802 wskazany w cytacie uzyskała dla zbioru testowego, natomiast wynik prezentowany w tabeli 4.3 dla zbioru walidacyjnego. Wskazała również, że w tabeli 4.6 są wyniki dla zbioru walidacyjnego (co dodatkowo powtórzyła w przesłanych odpowiedziach), a tam pojawia się wartość 0,802. Uważam, że jest to niejasne i powinno być precyzyjnie wyjaśnione. Prezentacja wyników własnych w odniesieniu do rezultatów innych prac powinna być spójna, stosując w miarę możliwości te same warunki.

Podsumowując, stwierdzam, że zaprezentowane w rozprawie opisy modeli i wyników uważam ogólnie za wiarygodne, niemniej dokumentacja modeli i wyników powinna być wykonana z większą starannością z uwzględnieniem szerszej dyskusji rezultatów w odniesieniu do stanu wiedzy.

Pytania do autorki rozprawy do przedstawienia w czasie dyskusji nad rozprawą:

- P1: Proszę o przedstawienie szczegółów dotyczących architektury i parametrów modelu HybNN, dla których uzyskano wyniki prezentowane w tabeli 3.4.

- P2: Proszę o przedstawienie i uzasadnienie w zakresie funkcji aktywacji jaką zastosowano na wyjściu modelu HybNN (w kodzie źródłowym: nn.Linear(n_chan + 363, 1), w rozprawie log(softmax(z)).

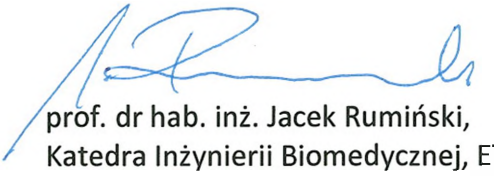
- P3: Proszę o wyjaśnienie dotyczące rozbieżności wyników dla modelu SENN, zgodnie z uwagami opisanymi wyżej w recenzji.

- P4: Posługując się przykładami z własnego kodu źródłowego proszę o wskazanie jak obliczano wartości AUC-ROC w poszczególnych eksperymentach z uwzględnieniem stosowania określonych funkcji aktywacji na wyjściu modeli.

## 5. Konkluzja recenzji

Podsumowując, stwierdzam, że Autorka odpowiedziała na postawione w poprawionej rozprawie pytanie badawcze. Właściwie dobrała i zastosowała metody badawcze, a eksperymenty należycie przeprowadziła. Przedstawiona do oceny praca zawiera pewne niejasności w zakresie dokumentacji modeli i wyników badań, niemniej uważam, że pomimo tych wątpliwości stanowi oryginalne rozwiązanie problemu naukowego. Wskazuje również na ogólną wiedzę teoretyczną i praktyczną Doktorantki z zakresu uczenia maszynowego i chemoinformatyki. Analizując rozprawę doktorską, wskazywane kody źródłowe oraz publikacje stwierdzam, że Autorka rozprawy wykazuje umiejętność samodzielnego prowadzenia pracy naukowej.

Biorąc pod uwagę przedstawione wyżej wnioski z recenzji, uważam, że rozprawa Pani mgr Magdaleny Wiercioch spełnia warunki określone w Ustawie – Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2020 r. poz. 85 z późniejszymi zmianami) i wnioskuję do Rady Dyscypliny Informatyki Technicznej i Telekomunikacji na Uniwersytecie Jagiellońskim w Krakowie, o dopuszczenie Doktorantki do dalszych etapów postępowania w sprawie nadania stopnia naukowego doktora.



prof. dr hab. inż. Jacek Rumiński,
Katedra Inżynierii Biomedycznej, ETI, PG