Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych

dr hab. inż.
Prof. Politechniki Śląskiej
Aleksandra Gruca

Gliwice, 30.11.2023

Recenzja rozprawy doktorskiej mgr-a Oleksandra Myronova

In-silico modeling of antigen recognition during immune response by analyzing the sequential and structural peptide-HLA-TCR data using machine learning

ukończonej na Wydziale Matematyki i Nauk Informacyjnych
Politechniki Warszawskiej

pod opieką promotora Profesora Dariusza Plewczyńskiego

Tematyka i cel pracy, problem badawczy i jego znaczenie

Przedstawiona mi do recenzji rozprawa doktorska powstała na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej pod kierunkiem Profesora Dariusza Plewczyńskiego. Praca ta powstała w formule tak zwanego „doktoratu wdrożeniowego" i realizowana była przez Doktoranta zatrudnionego w firmie Ardigen.

Tematyka rozprawy obejmuje zastosowanie metod sztucznej inteligencji, a w szczególności głębokich sieci neuronowych do modelowania procesu rozpoznawania antygenów przez układ immunologiczny. Rozpoznawanie antygenów to złożony mechanizm, na który składają się elementy takie jak problem doboru właściwego peptydu, problem prezentacji peptydu przez białka układu HLA, rozpoznanie peptydu przez odpowiedni receptor limfocytu T (ang. *T-cell receptor, TCR*) oraz wystąpienie odpowiedzi immunologicznej. W szczególności w pracy Doktorant koncentruje się na trzech wybranych problemach z dziedziny immunologii obliczeniowej: przewidywanie prezentacji peptyd-HLA, przewidywanie immunogenności peptydu zaprezentowanego przez białka układu HLA oraz przewidywanie wiązania peptyd-TCR.

Problem, który podjął rozwiązać się Doktorant jest istotny i aktualny. W ostatnich latach, dzięki rosnącej ilości danych pochodzących z wyskoprzepustowych technik laboratoryjnych, a w szczególności dzięki możliwości połączenia wyników eksperymentów głębokiego sekwencjonowania z zaawansowanymi technikami znakowania i sortowania komórek, możliwe było stworzenie zbiorów zawierających eksperymentalne wyniki dotyczące mechanizmów immunologicznych. Rosnąca liczba dostępnych zbiorów danych zawierających informacje odnośnie możliwych interakcji peptyd – HLA – TCR umożliwiła opracowanie modeli predykcyjnych

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Katedra Sieci i Systemów Komputerowych
ul. Akademicka 16, pok. 414, 44-100 Gliwice
+48 32 237 21 51 / +48 237 27 33
Aleksandra.gruca@polsl.pl

NIP 631 020 07 36
ING Bank Śląski S.A. o/Gliwice 60 1050 1230 1000 0002 0211 3056

dla mechanizmów biologicznych związanych z odpowiedzią immunologiczną, a w szczególności wykorzystanie do tego celu głębokich sieci neuronowych, które z założenia wymagają dostarczenia ogromnej ilości przykładów w procesie trenowania.

Uzyskane w ramach rozprawy wyniki mają niewątpliwie duże potencjalne zastosowanie praktyczne. Stworzenie narzędzi predykcyjnych dla interakcji peptyd – HLA – TCR może zastąpić kosztowne eksperymenty laboratoryjne. Możliwość przewidywania prezentacji peptydu przez białka układu HLA oraz immunogenności peptydu może przyspieszyć proces projektowania szczepionek nowej generacji. Natomiast przewidywanie wiązania peptydu do receptorów limfocytów T może znaleźć zastosowanie podczas projektowania terapii antynowotworowych w wykorzystaniem limfocytów T pacjenta.

Charakterystyka rozprawy

Praca doktorska została napisana w języku angielskim i składa się z czterech rozdziałów.

Rozdział pierwszy rozpoczyna się wprowadzeniem do tematyki pracy, przedstawieniem motywacji badań i zdefiniowaniem celu pracy. Następnie umieszczono w nim opis najważniejszych mechanizmów immunologicznych, a w szczególności elementy składające się na interakcje peptyd-HLA-TCR istotne z punktu widzenia analiz przedstawionych w rozprawie: wiązanie peptyd-HLA, prezentacja peptydu przez białka układu HLA, oraz rozpoznanie antygenu przez receptor limfocytu T.

Rozdział drugi zatytułowany jest „*Materials and Methods*" i zawiera wprowadzenie do wybranych zagadnień głębokiego uczenia takie jak charakterystyka wybranych funkcji straty i metody ich optymalizacji oraz elementy optymalizacji procesu uczenia głębokich, w szczególności regularyzację, różne metody normalizacji oraz funkcje aktywacji. W dalszej części rozdziału Autor omawia architektury wybranych typów sieci głębokich takie jak sieci głębokie sieci konwolucyjne (ang. *convolutional networks*), głębokie sieci rekurencyjne (ang. *recurrent neural networks*) wraz z sieciami typu LSTM (ang. *Long Short-Term Memory*), warstwę uwagi (ang. *attention*), a także architekturę transformera w szczególności w kontekście modeli językowych. W ostatniej cześć Autor przedstawia zbiory danych wykorzystane w badaniach zaprezentowanych w rozprawie.

Ostatni, największy objętościowo rozdział rozprawy zatytułowany jest „*Results*" i zawiera wyniki modelowania z wykorzystaniem głębokich sieci. Rozdział ten w sumie obejmuje 80 stron i jego zawartość jest połączeniem informacji teoretycznych, opisu metod *state-of-the-art* dla analizowanych problemów, charakterystyki zbiorów danych oraz wyników badań dla wybranych przez doktoranta problemów z dziedziny immunologii obliczeniowej. Pierwsza część tego rozdziału dotyczy metod modelowania interakcji peptyd-HLA. W szczególności przedstawiono model prezentacji peptyd-HLA i dwa modele immunogenności peptyd-HLA: model immunogenności wirusowej na bazie konwolucyjnych sieci neuronowych oraz model immunogenności nowotworowej na bazie sieci rekurencyjnych.

W każdym z trzech podrozdziałów podano architekturę sieci, procedurę uczenia i oceny oraz wyniki testów porównawczych ze znanymi metodami z wykorzystaniem benchmarkowych zbiorów danych. Następnie zastosowano wspomniane modele do analizy dwóch rzeczywistych problemów z immunologii obliczeniowej: poszukiwania potencjalnych epitopów dla szczepionki przeciw wirusowi Sars-CoV-2 oraz analiza mechanizmu ucieczki immunologicznej w kontekście szczepionki przeciwnowotworowej. W dalszej części tego rozdziału Doktorant przedstawia autorski model BERTrand, to jest model przewidywania wiązania peptydu do receptorów limfocytów T w oparciu o architekturę transformerów. Z uwagi na fakt, iż w danych eksperymentalnych brakuje negatywnych przykładów, istotnym wkładem autora jest tu również opracowanie nowej metody generowania takich przykładów.

Ostatni rozdział rozprawy zawiera jej krótkie podsumowanie.

Opinia o rozprawie

Należy podkreślić, iż cel pracy jest w jasno sformułowany a problem, który podjął się rozwiązać Doktorant jest niewątpliwie ważny i trudny. Zastosowane metody analizy danych wpisują się w trendy najnowszych badań w dziedzinie maszynowego uczenia i sztucznej inteligencji. Przedstawione rozwiązania mają potencjalnie wysokie zastosowanie praktyczne, na co również wskazuje fakt umieszczenia modelu prezentacji oraz modeli immunogenności w komercyjnym produkcie *ArdImmune Vax* firmy Ardigen.

Cześć teoretyczna pracy i przegląd stosowanych metod, choć rozproszone pomiędzy rozdziałami pracy, pokazują wystarczająco głęboką i aktualną wiedzę Doktoranta dotyczącą tematyki podjętej w rozprawie doktorskiej. Z uwagi na fakt, iż praca jest interdyscyplinarna i łączy ze sobą dwie dziedziny, Autor musiał zmierzyć się z niełatwym zdaniem wyjaśnienia zarówno mechanizmów biologicznych jak i problemów z dziedziny analizy danych. Mimo moich pewnych uwag odnośnie struktury pracy, należy uznać, iż autor poradził sobie z tym zdaniem w sposób wystarczający.

Bibliografia zawiera 98 pozycji i w znaczącej większości, poza pracami seminalnymi dla dziedziny, składa się z artkułów wydanych na przestrzeni ostatnich dziesięciu lat. Świadczy to o tym, iż Doktorant orientuje się w aktualnych badaniach prowadzonych zarówno w dziedzinie immunologii jak i głębokiego uczenia. Jest to również dowód na to, że problem, który podjął się rozwiązać Doktorant oraz wykorzystane w tym celu metody wpisują się w najnowsze trendy badań w obydwu dziedzinach.

Pozytywnie też pozytywnie oceniam fakt, iż kod źródłowy modelu, który nie stanowi części komercyjnego systemu został udostępniony środowisku naukowemu na plaformie GitHub.

Uważam, iż zastosowane metody badawcze są odpowiednie do rozwiązywanego problemu badawczego i wskazują na dobrą znajomość przez Autora rozprawy nowoczesnych i efektywnych metod oraz technologii stosowanych dziedzinie informatyki, a w szczególności w obszarze maszynowego uczenia oraz immunologii obliczeniowej. Zaprezentowane w pracy wyniki pozwalają stwierdzić, że przedstawione w rozprawie cele:

- *wytrenowanie oraz ocena jakości modelu głębokiego uczenia w celu przywidywania rozpoznawania antygenów na podstawie ich sekwencji białkowych. Na rozwiązanie to składają się dwa oddzielne problemy, które należy rozwiązać: przewidywanie interakcji peptyd-HLA oraz przewidywanie interakcji peptyd-TCR*

- *praktyczne zastosowanie modeli przewidywania rozpoznawania antygenów do rozwiązania dwóch problemów z dziedziny immunologii obliczeniowej: (1) priorytetyzacja peptydów w celu stworzenia potencjalnej szczepiony przeciwko COVID-19; (2) analiza relacji między immunogennością, a mechanizmem ucieczki immunologicznej w chorobach nowotworowych.*

zostały zrealizowane.

Uwagi krytyczne i dyskusyjne

Na początku tej części chciałbym podkreślić, że nie znalazłam w przedstawionych wynikach żadnych zasadniczych błędów merytorycznych czy niewłaściwych rozumowań. Wszystkie poniższe uwagi wynikają z chęci podjęcia dyskusji i dialogu na temat niektórych aspektów pracy. Uwagi te nie obniżają mojej pozytywnej oceny pracy.

- Zaprezentowana struktura pracy nie jest w mojej ocenie najlepsza i powinna zostać bardziej przemyślana przez Doktoranta w trakcie jej pisania. W szczególności:

    - Rozdział pierwszy powinien zostać rozdzielony na dwa odrębne rozdziały: wstęp oraz część teoretyczną związana z mechanizmami autoimmunologicznymi

    - Rozdział trzeci pracy jest zbyt obszerny i nie posiada przemyślanej i spójnej struktury. Powinien być on podzielony na kilka części. Czytający rozprawę ma wrażenie, że Autor próbował umieścić tu wszystko co udało mu się osiągnąć w tematyce rozprawy bez próby usystematyzowania wiedzy i wyników. Przykładowo znajdziemy tu charakterystykę zbiorów danych, która powinna zostać umieszczona we wcześniejszej części Datasets w rozdziale 2, albo opis mechanizmów ucieczki immunologicznej, które powinny znaleźć się w części teoretycznej pracy opisującej mechanizmy immunologiczne w rozdziale 1. Rozdział trzeci zawiera również opis skryptów składających się na analizy z wykorzystaniem modelu BERTrand oraz sposoby ich wywołania, które w mojej opinii są zbyt techniczną informacją i jako takie powinny się znaleźć w suplemencie. Niektóre z podsekcji znajdują się na piątym poziomie wyliczenia (!!). Takie przeładowanie informacjami jednego rozdziału powoduje, że pracę ciężko się ją czyta, a także utrudnia powiązanie ze sobą kolejnych elementów wykonanych w ramach badań.

- Czy możliwe byłoby dołączenie informacji strukturalnej sekwencji peptydów do danych wejściowych modeli? Czy miałoby to szansę poprawić wyniki predykcji?

- Czy możliwe jest zastosowanie podjeść XAI w celu interpretacji decyzji podejmowanych przez zaproponowane modele. Czy i w jaki sposób taka informacja mogłaby zostać wykorzystana na dalszym etapie analizy wyników?

- kolejne pytanie odnosi się do wyników analizy przeżyciowej zaprezentowanych w sekcji 3.8.3.5. Czy są jakieś konkretne cechy nowotworu LUSC, które sprawiają, że w przypadku tego zbioru danych *immunogenic load* oraz *IEM status* sprawdzają się jako biomarker dla analizy przeżycia w porównaniu do pozostałych zbiorów danych, gdzie nie zaobserwowano istotności statystycznej między grupami?

- jakie parametry narzędzia BLASTp zostały wykorzystane przy filtrowania toksycznych peptydów. Czy zastosowano optymalizację dla krótkich sekwencji?

Praca napisana jest zrozumiałym językiem jednak zauważyłam w niej nieliczne błędy redakcyjne, które wymagałaby poprawek:

- w sekcji 1.3 powinno być “however there are 5 major steps involved”, a nie “4 major steps”

- rysunek 37 zawiera frazę TODO

- rysunek 53 zawiera w legendzie opisu „FALSE” i „TRUE”. Nigdzie nie podano co one oznaczają, czytelnik musi się tego domyślać.

- w elektronicznej wersji pracy, którą dostałam spis treści jest niepełny (nie ma tego problemu w wersji papierowej), a linki do odnośników literaturowych przenoszą do bazy paperpile, do której nie mam dostępu.

Dla większości opisów rysunków jest krótka, jednozdaniowa i brak w nich objaśnienia co znajduje się na rysunku. Umieszczenie rozszerzonych opisów pod rysunków umożliwiającymi zrozumienie ich bez konieczności szukania opisu w tekście, znacznie zwiększyłoby czytelność pracy.

Dobrze by też było, gdyby w części podsumowującej Doktorant umieścił listę publikacji związanych z Doktoratem wraz z opisem jakich elementów rozprawy dotyczą te publikacje.

Podsumowanie

Pan mgr Oleksandr Myronow przedstawił rozprawę doktorską rozwiązującą aktualny problem naukowy, która przyczyni się do rozwoju reprezentowanej dyscypliny naukowej. Rozprawa zawiera oryginalne rozwiązanie problemu naukowego, a kandydat wykazał, że zarówno posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja oraz umiejętność prowadzenia pracy naukowej.

Część wyników zaprezentowanych w pracy została wdrożona jako część komercyjnego produktu *ArdImmune Rank* będącego jednym z modułów produktu *ArdImmune Vax* firmy Ardigen. Druga część wyników, model BERTrand, nie będąca częścią produktu komercyjnego, została opublikowana w bardzo dobrym czasopiśmie naukowym *Bioinformatics*, a Doktorant jest wiodącym autorem tej publikacji.

Udało mi się znaleźć informację, iż Doktorant jest współautorem dwóch artykułów naukowych. Jak już wspomniałam, w jednym z nich jest autorem wiodącym (jest to czasopismo *Bioinformatics* 200 punktów MNiSW), a w drugim autorem współautorem (czasopismo *Frontiers in Genetics,* 140 punktów MNiSW). Obydwa artykuły są opublikowane w dobrych czasopismach i obudowa związane są z tematyką doktoratu. W doktoratach wdrożeniowych, biorąc pod uwagę charakter prac, które koncertują się bardziej na aspekcie wdrożenia produktu, zwyczajowe wymagania publikacyjne są zwykle niższe niż w przypadku doktoratów tradycyjnych. Dlatego też bardzo wysoko oceniam fakt, iż Doktorant podjął wysiłek publikacyjny zakończony sukcesem w tak dobrych czasopismach.

Biorąc pod uwagę powyższą ocenę, stwierdzam, że przedstawiona do oceny praca doktorska w pełni odpowiada warunkom określonym w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2023 r. poz. 742 z późn. zm.) i na tej podstawie wnoszę do Wysokiej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr Oleksandra Myronowa do dalszych etapów przewodu doktorskiego.

Signed by / Podpisano przez:
Aleksandra Helena Gruca
Politechnika Śląska
Date / Data: 2023-12-03 13:58

dr hab. inż. Aleksandra Gruca