

Czeladź, 28 lutego 2023

Recenzja pracy doktorskiej Pana Krzysztofa Korasa
„*Computational methods for anti-cancer drug sensitivity problem*”

Rozprawa opiera się na pracach

1. K. Koras, Dilafruz Juraeva, Julian Kreis, Johanna Mazur, Eike Staub, Ewa Szczurek, *Feature selection strategies for drug sensitivity prediction*, Scientific Reports 10, 9377 (2020).
2. K. Koras, Ewa Kizling, Dilafruz Juraeva, Eike Staub, Ewa Szczurek *Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines*, Scientific reports 11, 15993 (2021).
3. K. Koras, Marcin Możejko, Paulina Szymczak, Eike Staub, Ewa Szczurek, *A generative recommender system with GMM prior for cancer drug generation and sensitivity prediction*, Machine Learning in Computational Biology, Vol. 200, 21-22, 2022 (PMLR 200:61-73).

Zanim przejdę do dokładniejszego opisu wyników, chciałbym podkreślić, że wszystkie rozważane prace należą do dziedziny cheminformatyki, czyli leżą na styku informatyki i chemii. Rozważana tematyka jest wartościowa zarówno naukowo, jak i z punktu widzenia potencjalnych zastosowań. Niewątpliwie prace te stanowią także spójny zestaw, którego myślą przewodnią jest interpretowalność w tematyce związków chemicznych mogących w przyszłości pomóc w leczeniu raka. Warto także podkreślić, że czasopisma w których prace się ukazały to czasopisma o poważnej renomie naukowej, oraz to, że Pan Krzysztof Koras jest głównym autorem wszystkich wymienionych prac. Pierwsze z dwóch rozważanych prac doczekały się licznych cytowań (18 i 8 na podstawie scholar.google), co oznacza, że zostały zauważone i docenione w środowisku naukowym.

Chciałbym tutaj także zaznaczyć moją drobną wątpliwość formalną odnośnie przedstawionej rozprawy, a mianowicie rozprawa ta według mojej oceny jest rozprawą z dorobku, czyli składa się z przedstawionych prac naukowych. Cytując za Ustawą o szkolnictwie wyższym, Art. 187.3: „Rozprawę doktorską może stanowić praca pisemna, w tym monografia naukowa, zbiór opublikowanych i powiązanych tematycznie artykułów naukowych, praca projektowa, konstrukcyjna, technologiczna, wdrożeniowa lub artystyczna, a także samodzielna i wyodrębniona część pracy zbiorowej”, moim zdaniem przedstawiona praca jest de facto zbiorem powiązanych artykułów (nie jest to moim zdaniem monografia, gdyż stanowiłoby to autoplagiat, jako przekopiowanie wydrukowanych artykułów). Z drugiej strony spisane w rozprawie kluczowe rozdziały w zasadzie prawie dosłownie (z małymi wyjątkami) powielają wymienione przez autora prace naukowe. W konsekwencji w swojej ocenie zająłem się tylko rozważonymi artykułami (pomijam ocenę rozdziałów które im nie odpowiadają).

W pierwszej z rozważanych prac autorzy rozpatrują problem przewidywania wrażliwości na leki dla między innymi komórek nowotworowych. Problem ten jest w sposób oczywisty istotny dla medycyny spersonalizowanej. Głównym zadaniem jest wykorzystanie wiedzy na temat celów leków i sposobu ich działania, w celu wybrania wiarygodnych cechy opisujących linie komórek nowotworowych. Zaprezentowane podejście jest zweryfikowane dla szerokiego spektrum związków przeciwnowotworowych. W szczególności autorzy wskazują zestaw leków, dla których można zidentyfikować łatwe do interpretacji zestawy cech. Autorzy rozważają w sposób naturalny dwie kategorie podziału na cechy: metody selekcji motywowane biologicznie i automatyczne metody selekcji oparte bezpośrednio na danych. Co ciekawe, okazuje się, że różnica w wydajności predykcyjnej modeli motywowanych biologicznie w porównaniu z modelami obejmującymi cały genom jest niewielka, pomimo użycia znacznie mniejszej liczby cech wejściowych. Podsumowując, praca ta to analiza strategii wyboru cech do przewidywania wrażliwości na leki. Z czysto informatycznej perspektywy praca praca nie zawiera zbyt wielkiego wkładu, gdyż użyte modele są dosyć bazowe, a mianowicie to płytkie modele regresyjne używające regresji liniowej (w wersji elastic net) oraz lasów losowych.

Celem drugiej pracy jest stworzenie nowego systemu rekomendacji sieci neuronowej do przewidywania czułości inhibitora kinazy. Problem przewidywania wrażliwości na leki można określić jako problem rekomendacji, w którym linie komórek nowotworowych i leki są analogiczne odpowiednio do użytkowników i przedmiotów. Celem jest zalecenie najlepszego leku dla danej

linii komórkowej. Zaproponowany model, nazwany „Drug Efficacy Estimation Recommender System” (DEERS), wykorzystuje cechy molekularne linii komórek nowotworowych i profile hamowania kinaz przez leki. Jak pokazują autorzy, w analizie porównawczej DEERS przewyższa dwa inne modele systemu rekomendacyjnego opartego na MF i osiąga podobnie dobre wyniki jak najlepiej działający algorytm XGBoost, przy czym jest interpretowalny. DEERS zawiera dwa płytkie¹ autoenkodery do projekcji cech linii komórkowych i leków w 10-wymiarowych ukrytych reprezentacjach oraz sieć neuronową typu feed-forward. Wartościowe jest zaproponowane podejście do interpretacji. W przeciwieństwie do pierwszej pracy która zasadniczo ma aspekt głównie chemiczny (z niewielkim użyciem standardowych modeli płytkich), praca druga zawiera także interesujący wkład informatyczny. Pozwala także na analizę modelu w zależności od architektury (w szczególności wymiarowości przestrzeni ukrytej) rozważanych autoenkoderów.

Ostatnia z przedstawionych prac personalnie była dla mnie najciekawsza, gdyż używała zaawansowanych metod z głębokiego nauczania maszynowego. Zaproponowany model nazwany VADEERS, to system rekomendacji oceny skuteczności leku oparty na wariacyjnym autoenkoderze (VAE). Generowanie związków jest oparte na częściowo nadzorowanym priore w przestrzeni ukrytej zadanym przez mieszaną rozkładów normalnych. Pozwala to uzyskać poklastrowanie przestrzeni ukrytej, powiązane z określonymi właściwościami rozważanych związków chemicznych. Model bazuje na reprezentacji SMILES. Oceniane warianty VADEERS osiągają wysoką korelację ($r = 0,87$) Pearsona między rzeczywistą i przewidywaną wrażliwością na lek. Pracę tę przeczytałem z dużą przyjemnością, gdyż jest to jedna z tych wartościowych prac która efektywnie i twórczo używa zaawansowanych narzędzi z nauczania maszynowego w interpretowalnym generatywnym podejściu do generowania leków. Zastosowany model w swojej podstawowej idei jest podobny do modelu SeGMA [Śmieja, Marek, et al. "SeGMA: Semi-supervised gaussian mixture autoencoder." *IEEE transactions on neural networks and learning systems* 32.9 (2020): 3930-3941] (cytowane w wersji artykułu [56]). Pytaniem które tu mi się tu w związku z tym nasunęło, jest czy możliwe byłoby zastosowanie w modelu VADEERS także transferu stylu, który w SeGMA się dobrze sprawdził? I czy taki transfer miałby tu naturalną interpretację chemiczną?

Pod koniec recenzji chciałbym zaznaczyć, że prace te dobrze świadczą o warsztacie naukowym Pana Krzysztofa Korysa, co więcej w każdej kolejnej

¹Chociaż w pracy nazywane są błędnie głębokimi.

pracy widać coraz szerszy zakres stosowanych przez niego narzędzi informatycznych, co dobrze rokuje pod kątem jego przyszłej kariery naukowej. Podsumowując, rozprawa spełnia wszelkie zwyczajowe i ustawowe kryteria stawiane rozprawom doktorskim, w związku z czym wnioskuję o dopuszczenie doktoranta do dalszych etapów przewodu doktorskiego.

prof. dr hab. Jacek Tabor