

Recenzja pracy doktorskiej mgr Marty Stępniewskiej-Dziubińskiej
„Predicting binding affinity of small molecules in ligand-receptor
systems using neural networks”

Teza przedstawiona przez panią magister Martę Stępniewską-Dziubińską ma formę oprawionego tomu rozmiaru A 4, który zawiera anglojęzyczną rozprawę; jej polskojęzyczny tytuł: „Wykorzystanie sieci neuronowych do przewidywania aktywności związków niskocząsteczkowych w układach ligand-receptor” zamieszczono na stronie tytułowej wraz z informacją o miejscu wykonania badań, którym jest Zakład Bioinformatyki Instytutu Biochemii i Biofizyki Polskiej Akademii Nauk w Warszawie. Promotorem rozprawy jest dr hab. Paweł Siedlecki. Format tezy jest minimalistyczny; autorski wywód pomieszczono na 18 stronach plus wnioski zaprezentowane w jednym bardzo krótkim akapicie. Tak skondensowana forma prezentacji jest pewnym zaskoczeniem dla recenzenta, który nie jest nowicjuszem w dziedzinie opiniowania treści merytorycznych prac naukowych z różnych dyscyplin nauk przyrodniczych, lecz nie ma bezpośredniego doświadczenia z kodyfikacją oraz reprezentacją wyników badań z obszarów informatycznych. Refleksja nad formatem nie ma charakteru zarzutu; zawartość tezy jest bardzo bogata pod względem merytorycznym i zawiera znaczny ładunek inwencji i innowacji, które przybrały formę nowych narzędzi metodycznych o znacznej wartości teoretycznej i praktycznej. Merytoryczne treści które prezentują te wartości są zawarte w oryginalnych publikacjach bioinformatycznych, które przytoczono w dysertacji *in extenso*. Wspomniany już krótki anglojęzyczny komentarz metodyczny poprzedzono dodatkowo jednostronnicowym polskojęzycznym streszczeniem rozprawy, które nawiązuje do relacji metod komputerowego projektowania leków uznanych obecnie za klasyczne (CADD) z nowoczesnymi metodami angażującymi uczenie maszynowe (ML) z wykorzystaniem sieci neuronowych (CNN). Wspomaganie komputerowe metod poszukiwania nowych leków uchodzi słusznie za obszar kluczowych technologii, istotnych dla usprawnienia powszechnie przyjętego lecz mało skutecznego, wieloetapowego procesu weryfikacji nowych cząsteczek chemicznych (ewentualnie związków naturalnych) kandydujących do roli specyfików dopuszczonych przez systemową (np. EMA) rejestrację do obrotu rynkowego jako leki o ściśle określonym wskazaniu terapeutycznym.

Na wstępie wypada nadmienić że tytuły rozprawy nie są jednoznacznie tożsame. Właściwy tytuł, anglojęzyczny, ściśle odpowiada treści rozprawy, odnosząc się do powinowactwa: ligand-białko, podczas gdy polski sugeruje przewidywanie aktywności a te pojęcia są w farmakologii, farmacji i naukach medycznych – rozłączne. Rozprawa składa się z siedmiu części, odpowiadających numerom rozdziałów. W pierwszej autorka wyjaśnia pojęcie uczenia maszynowego (ML) oraz krótko charakteryzuje jego warianty i modele. Omawia przy tym oba rodzaje wspomagania komputerowego, których rozróżnienie polega na wyróżnieniu obiektu molekularnego pierwotnej charakterystyki cheminformacyjnej – jest nim albo receptor białkowy (Structure based design) albo mała cząsteczka liganda (Ligand based design); obie nieco odrębne metodycznie gałęzie mieszczą się w strategii CADD i do obu nawiązują opracowywane przez autorkę nowe narzędzia badawcze. Rozprawa prezentuje cztery odrębne osiągnięcia badawcze o charakterze nowych narzędzi informatycznych uzyskanych przez wykorzystanie zasobów metodycznych dyscypliny do podejmowania nowych zadań z zakresu uczenia maszynowego i konwolucyjnych sieci neuronowych do oceny interakcji pomiędzy znanym białkiem a wybraną małą cząsteczką. Interakcja z białkiem o charakterze receptorowym lub enzymatycznym, zdefiniowana do poziomu wiązania niekowalencyjnego w określonej topologicznie domenie proteinowej z policzalną wartością stałej kompleksowania, jest zasadniczym elementem oceny liganda jako potencjalnego leku (druggability), którego mechanizm działania wyznacza funkcja białka będącego przedmiotem eksperymentów dokowania *in silico*. Pierwsze z prezentowanych opracowań, o nazwie DeCAF, ma charakter modułu języka Python, którego celem jest przetwarzanie konwencjonalnych struktur chemicznych małych cząsteczek organicznych na format farmakofora 2D z rozszerzeniami umożliwiającymi włączenie elementów dodatkowej wiedzy eksperckiej. Program jest narzędziem umożliwiającym różne rodzaje porównań strukturalnych (similarity, fingerprinting) przy zastosowaniu różnorodnych metod selekcji zbiorów (filtering). Wyniki opublikowano w czasopiśmie *Molecules* (MDPI) w 2017 r. Kolejna publikacja doktorantki, w *Bioinformatics* (Oxford University Press) z 2018 roku, opisuje szczegółowo konstrukcję programu uczenia głębokiego, dedykowanego do przewidywania powinowactwa liganda do wybranego białka. Opis architektury programu o nazwie Pafnucy, ma także odpowiednią reprezentację graficzną (warstwy konwolucyjne i gęste), oraz jest uzupełniony szczegółowymi informacjami technicznymi na temat wyboru danych (podstawowych, treningowych i walidacyjnych) których użyto w procesie uczenia i oceny sprawności predykcyjnej. Praca ta uzyskała już wysoki wskaźnik cytowań (!) a nawet ocenę opisową w porównaniu tabelarycznym, w publikacji

konkurencyjnego zespołu badawczego (Y. Mu, *ACS Omega*, 2019). W kolejnej, tym razem cztero-autorskiej publikacji (we wszystkich załączono odpowiednie oświadczenia współautorów) z udziałem doktorantki, powraca motyw reprezentacji molekularnej, w konstrukcji fingerprintu PLEC, który stanowi inne podejście do danych wejściowych dla charakterystyki interakcji białko – ligand. Serię opisu nowych narzędzi ML i DL w zastosowaniu do wykrywania i badania wiązalności białko-ligand zamyka prezentacja kolejnego programu autorskiego na bazie języka Python i architektury U-Net, sieci neuronalnej o nazwie Kalasanty, tym razem w *Scientific Reports* (Nature) w 2020. Do treningu i walidacji modelu zastosowano zbiór ponad 1600 białek pochodzących z PDB. Rezultaty typowania domen wiążących porównano (cross validation) z wynikami programu DeepSite o podobnym przeznaczeniu i odnotowano znacząco lepszy wynik po stronie sieci Kalasanty. Spektakularne wyniki badań polskich bioinformatyków to stały powód do radości i dumy lecz także okazja do ogólniejszej refleksji.

Ponad 30 lat temu, Arthur Kornberg, biochemik z Uniwersytetu Stanforda w Kalifornii, rozpoczął swój słynny artykuł „The Two Cultures: Chemistry and Biology” (*Biochemistry*, 1987) poświęcony zróżnicowaniu metodycznemu dwu zasadniczych kierunków badań przyrodniczych, określanych dziś wspólnym mianem Nauk o Życiu (Life Sciences), zdaniem: „Much of life can be understood in rational terms if expressed in the language of chemistry”. Określenia: „język chemii” a nawet „lingwistyka chemiczna” znalazły już swoje stałe miejsca w dyskusjach przyrodoznawczych oraz oczywiście w naukach informatycznych. Można zauważyć z satysfakcją, że wspomniane przez Kornberga dwie kultury ulegają właśnie daleko idącej amalgamacji, na gruncie informatycznym: podobieństwa baz danych, metod ich przetwarzania, oraz wspólnoty środków komunikacji wyników badań. Cheminformatyka ciągle różni się zarówno wyborem celów badawczych jak i metod od bioinformatyki ale już dziś adept inżynierii biomedycznej i telemedycyny nie zastanawia się raczej nad odrębnością obrazu graficznego, sygnału akustycznego czy wzoru struktury cząsteczki chemicznej a jedynie nad systemem kodowania poszczególnych danych i możliwościami ich komputerowego przetwarzania.

Poszukiwanie nowych leków jest ikonicznie interdyscyplinarną dziedziną, która integruje nie tylko nauki biologiczne i chemiczne ale musi także w swych dążeniach i działaniach uwzględniać determinanty biznesowe, społeczne i prawne. Nie brak opinii radykalnych, twierdzących że przemysł farmaceutyczny (big pharma) realizujący programy poszukiwania nowych leków (DDD) stanowi

przedsięwzięcie skrajnie dysfunkcyjne i nieskuteczne, generując przy tym astronomiczne wprost koszty. Rzeczywiście, trudno zaprzeczyć, że fenomenalne postępy podstawowych badań przyrodniczych i dorównujące im osiągnięcia w dziedzinach technicznych i inżynieryjnych, nie przekładają się na dostępność nowych leków i nowych terapii a suma potrzeb medycznych które nie są zaspokajane niepokojąco rośnie. Związek tej konstatacji z tematyką ocenianej rozprawy jest dla recenzenta oczywisty – to że powstają nowe narzędzia informatyczne, usprawniające proces selekcji związków chemicznych jako kandydatów na nowe leki jest bardzo ważne ale chyba nie zainicjuje reformy systemu kwalifikacji i walidacji (DDD pipeline) nowych leków. Decydującym ogniwem tego systemu są szczególnie zaporowe (zarówno metodycznie jak i finansowo), wieloetapowe badania kliniczne oparte na zasadach klasycznej farmakologii, która najwyraźniej nie zdołała się zreformować do poziomu satysfakcjonującego opisu sytuacji medycznej w kategoriach biologii systemów. Szczególne znaczenie aktualnych osiągnięć bioinformatyki zorientowanych na wspomaganie poszukiwania nowych leków polega na radykalnej redystrybucji podstawowych dla tego kierunku narzędzi badawczych. Programy wytworzone w domenie akademickiej, tak jak te opisane w ocenianej rozprawie, trafiają najczęściej do obiegu zasobów otwartej nauki, co zapewnia skuteczną propagację i akumulację tej wiedzy, w odróżnieniu od firmowych, zamkniętych lub komercyjnych baz danych i programów na usługach przemysłu. Tendencja konsolidacji niezależnych badań nad nowymi lekami przy użyciu nowych metod bioinformatycznych staje się ważnym elementem na drodze do reformy sektora który nie spełnia obecnie oczekiwanej roli w racjonalnej ochronie zdrowia.

W konkluzji stwierdzam, że zaprezentowane przez mgr M. Stępniewską-Dziubińską w ocenianej rozprawie modele i programy uczenia maszynowego przez zastosowanie sieci neuronowych stanowią znaczące osiągnięcia zarówno w znaczeniu naukowym jak i aplikacyjnym. Kandydatka z nadmiarem spełnia wymagania Ustawy o Stopniach i Tytułach Naukowych (Dz. U. 2018, poz. 1668 Art. 185 - 197) wobec czego wnoszę o dopuszczenie Jej do dalszych etapów przewodu doktorskiego.



Warszawa 06-10-2020

Prof. dr hab. Grzegorz Grynkiewicz

Profesor emerytowany Instytutu Farmaceutycznego Sieci Badawczej Łukasiewicz