Katowice, dnia 30.04.2024

prof. dr hab. Michał Daszykowski
Instytut Chemii, Wydział Nauk Ścisłych i Technicznych
Uniwersytet Śląski
ul. Szkolna 9
40-006 Katowice

Tel.: 32 359 15 68
E-mail: michal.daszykowski@us.edu.pl
ORCID: 0000-0003-0275-2751

Recenzja pracy doktorskiej pt. *„Właściwości kwantowe fragmentów cząstek jako deskryptory molekularne"*
autorstwa Pana mgr. inż. Bartłomieja Fliszkiewicza

Pan mgr inż. Bartłomiej Fliszkiewicz przeprowadził badania do pracy doktorskiej na Wydziale Nowych Technologii i Chemii Wojskowej Akademii Technicznej. Promotorem jego pracy doktorskiej był Pan dr hab. inż. Marcin Sajdak z Wydziału Inżynierii Środowiska i Energetyki Politechniki Śląskiej.

Główny cel pracy doktorskiej, który jej autor postanowił zrealizować, obejmuje stworzenie nowych deskryptorów molekularnych, uwzględniających właściwości kwantowe fragmentów cząstek chemicznych. W tym miejscu warto zauważyć, że idea poszukiwania matematycznej zależności pomiędzy wybranymi własnościami molekuł, a ich strukturą chemiczną opisaną poprzez zbiór określonych deskryptorów molekularnych wciąż cieszy się dużą popularnością. W zależności od podejmowanego problemu, mówi się o modelowaniu aktywności związków chemicznych (z ang. *quantitative structure-activity relationship*, QSAR) lub ich własności (z ang. *quantitative structure-property relationship*, QSPR), przy czym oba te zagadnienia mogą mieć charakter ilościowy lub jakościowy, w zależności od przyjętej definicji modelowanej zmiennej zależnej. Podejścia QSAR i QSPR bardzo ułatwiają badania określonych grup związków chemicznych w aspekcie ich szeroko pojętego potencjalnego oddziaływania, np. na człowieka, zwierzęta, rośliny, środowisko naturalne, określony układ, itp. Oba podejścia stosowane są często, a sięga po nie zwłaszcza przemysł farmaceutyczny i chemiczny. Wspierają one proces projektowania nowych związków

o pożądanych właściwościach. Oferują zarazem cenną możliwość redukcji kosztów takich poszukiwań. Do tej pory zaproponowano tysiące deskryptorów, które zostały skrupulatnie skatalogowane przez R. Todeschiniego oraz V. Consoni w znakomitej monografii pt. *„Handbook of Molecular Descriptors"*. Swą popularność zyskały głównie dzięki dużej powszechnej dostępności dzięki oprogramowaniu Dragon, którego następcą jest pakiet alvaDesc[1]. Za jego pomocą użytkownik może bardzo szybko wygenerować do 5666 różnych deskryptorów dla wybranych struktur chemicznych. Deskryptory molekularne uwzględniające własności kwantowe fragmentów cząsteczek mogą stanowić ważne uzupełnienie zbioru już dostępnych deskryptorów oraz wskazują ciekawy kierunek dla projektowania nowych deskryptorów. W efekcie, wzbogacenie bazy o komplementarne sposoby opisu molekuł daje duże nadzieje na uzyskanie lepszych modeli predykcyjnych QSAR/QSPR.

Praca doktorska ma klasyczny układ. Wyróżniamy w niej część teoretyczną, w której autor dokonuje przeglądu istniejącego stanu wiedzy w obszarze QSAR/QSPR, część eksperymentalną (zakończoną podsumowaniem i wnioskami) oraz część aplikacyjną.

W pierwszej części pracy doktorskiej jej autor zwraca uwagę na liczne współistniejące sposoby reprezentacji molekuł, a także sposoby zapisu informacji numerycznej. W szczególności omawia stosowane liniowe notacje molekuł (SMILES, SMARTS, SLN, InChI oraz SELFIES), grafy molekularne, postacie macierzowe, tablice połączeń czy sposoby kodowania struktur chemicznych w innych formach, a także stosowane formaty plików (XYZ, Molfile, SDfile, PDB), które zawierają dane na ich temat.

Następnie, autor pracy doktorskiej przedstawia ogólną taksonomię deskryptorów molekularnych, uprzednio przywołując te wymagania (łącznie czternaście), które powinny one spełniać. Zaproponowany podział uwzględnia źródło ich pochodzenia (eksperymentalne i teoretyczne), typ danych, rozmiar cząsteczki (0D, 1D, 2D, 3D i 4D). Ponadto, dopełnieniem tego już bardzo różnorodnego zbioru mogą być deskryptory kwantowe, tzw. pary atomów, odciski palców czy deskryptory tekstowe. Zwraca również uwagę, że bogactwo i rozmaitość deskryptorów molekularnych wynika przede wszystkim z faktu opisywania przez nie jedynie pewnego wycinka cech struktury danej molekuły. Stąd, dokładny i możliwie pełny opis jej cech wymaga użycia co najmniej kilku ortogonalnych deskryptorów, tzn. takich, które uwzględniają

[1] A. Mauri, alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints, w: K. Roy (Ed.), Ecotoxicological QSARs, Springer US, New York, 2020, pp. 801–820.

zupełnie inne informacje. Ponadto, trudno jest z góry powiedzieć, które spośród tysięcy znanych deskryptorów skutecznie sprawdzi się w przypadku modelowania różnych własności, co niestety dodatkowo komplikuje zagadnienie modelowania QSAR/QSPR. Rozwiązanie tej kwestii wymaga zatem zmierzenia się na etapie budowy modeli z relatywnie dużą liczbą wzajemnie skorelowanych zmiennych objaśniających. W praktyce jest to możliwe dokonując w jakiś sposób wyboru użytecznych deskryptorów albo zmieniając ich reprezentację tworząc nowe zmienne.

W kolejnej sekcji pracy doktorskiej znajdujemy bardzo cenny katalog dostępnych niekomercyjnych i komercyjnych programów wraz z ich krótką charakterystyką, które umożliwiają obliczanie deskryptorów molekularnych. Wśród nich znajdziemy takie pakiety obliczeniowe jak Mordred, PaDEL-descriptor, ChemoPy, PyDPI, alvaDesc 2.0, ADMEWORKS ModelBuilder oraz MOE – Molecular Operating Environment. Wszystkie one różnią się przede wszystkim zakresem możliwości generowania deskryptorów molekularnych, różnymi wymogami co postaci wejściowych struktur oraz dodatkowymi przydatnymi funkcjami i udogodnieniami.

Część eksperymentalną pracy doktorskiej jej autor rozpoczyna od opisu parametrów komputerów używanych do obliczeń. Warto zauważyć, że wszystkie z nich można zaliczyć do klasycznych komputerów o nieznacznie większym zakresie pamięci RAM. W mojej opinii stanowi to podwójną zaletę szczególnie pomocną na etapie rozwijania nowych deskryptorów i stosownego oprogramowania, a także dostępności dla potencjalnych użytkowników. Po pierwsze, możliwość użycia klasycznych komputerów o małej lub umiarkowanej mocy obliczeniowej nie wymaga zakupu kosztownej infrastruktury obliczeniowej i jej późniejszego utrzymania. Stąd, obliczenia w oparciu o już dostępne deskryptory są relatywnie tanie, a dodatkowo ogranicza się wydatki ponoszone na realizację eksperymentów wymagających specjalistycznych laboratoriów, syntez, odczynników, energii, etc. jednocześnie w pełni wspierając realizację założeń zielonej chemii. Po drugie, możliwość relatywnie łatwego generowania wielu deskryptorów bardzo wpływa na popularyzację podejść QSAR/QSPR i rozszerzenie grona ich użytkowników, co w efekcie ograniczy negatywne skutki oddziaływania na środowisko intensywnego poszukiwania przez nich nowych związków. Można zatem, z pełnym przekonaniem, powiedzieć, że metody QSAR/QSPR znacznie zwiększają efektywność projektowania w porównaniu do klasycznych i niestety jeszcze często praktykowanych poszukiwań związków o pożądanych własnościach, które w zasadzie opierają się na metodzie prób i błędów. Mając na względzie obecny stan wiedzy jak i dostępne możliwości, świadomy wybór poszukiwania nowych związków czy optymalizacji niektórych procesów przez

niektóre osoby w oparciu o podejście 'po omacku' stoi w całkowitej sprzeczności do zasad zielonej chemii i jednocześnie kwestionuje ich profesjonalizm zawodowy. Następnie, autor pracy doktorskiej wymienia poszczególne biblioteki oraz moduły języka programowania Phyton, które wykorzystuje tworzone przez niego oprogramowanie. Dokonuje tego w takich środowiskach jak Jupyter Lab, Visual Studio Code oraz PyCharm. Później, ma miejsce wskazanie stosowanych algorytmów obliczeniowych i metod, choć niestety pozbawione merytorycznych wyjaśnień i odnośników literaturowych. Ostatecznie, obliczenia kwantowe przeprowadzono na zaawansowanej maszynie z 32 procesorami i 128 GB podręcznej pamięci RAM. Powstały trzy rodzaje deskryptorów kwantowych, które posłużyły do budowy modeli prognostycznych. Z kolei reprezentacje związków chemicznych obejmowały dwie grupy wykorzystujące: (*i*) właściwości kwantowe określonych fragmentów cząstek oraz (*ii*) kwantowo zmodyfikowane pary atomów. W toku prac badawczych powstały liczne skrytpy i tzw. Jupyter Notebooki, przy czym te ostatnie udostępniono publicznie w jednym z ogólnych i uznanych przez naukowców repozytoriów. Są również uzupełnieniem opublikowanych artykułów. Obliczone fragmentaryczne deskryptory kwantowe i deskryptory kwantowe par atomowych zamieszczono na platformie GitHub wraz ze stosownymi interfejsami graficznymi.

W kolejnej sekcji pracy doktorskiej jej autor szerzej omówił fragmentaryczne deskryptory kwantowe i ich zastosowanie do modelowania aktywności biologicznej lub własności wybranych związków chemicznych.

Z kolei drugim istotnym omawianym wątkiem było przewidywanie własności optycznych dla związków organicznych mogących mieć zastosowanie do budowy diod OLED. Autor wykazał zasadność i konieczność stosowania podejścia projektowania własności molekuł uwzględniającego obliczenia komputerowe, co ułatwia selekcję obiecujących kandydatów do dalszych badań. W badaniach użyto bazę związków QM9, zawierającą aż 134 tysiące związków stanowiące podzbiór liczniejszej bazy GDB17. Następnie fragmentaryczne deskryptory kwantowe (czternaście różnych zestawów) posłużyły do budowy modeli QSPR, w których zmienną zależną była długość fali odpowiadająca maksimum emisji określonych związków. Do budowy modeli wieloparametrowych kalibracyjnych zastosowano regresję wieloraką, lasy losowe i wzmocnienie gradientowe, charakteryzując uzyskane modele takimi wskaźnikami jak średni błąd kwadratowy, błąd maksymalny oraz współczynnik determinacji. Zdolności predykcyjne stworzonych modeli testowano za pomocą dziesięciokrotnej walidacji krzyżowej. Ogólna konkluzja, sformułowana przez autora na 45 stronie pracy doktorskiej, obejmująca analizę wyników

modelowania (uzyskanych dla różnych typów modeli), jest taka: „ (...) *dodanie właściwości kwantowych fragmentów cząsteczek do zmiennych modelu predykcyjnego nie wpływa na otrzymany wynik, niezależnie od tego, jak takie deskryptory zostały zdefiniowane*". Z kolei dalej czytamy, że „*w przypadku regresji wielorakiej lepsze rezultaty otrzymuje się stosując tylko deskryptory kwantowe (...). Należy jednak wyraźnie zaznaczyć, że regresja wieloraka daje gorsze rezultaty niż algorytmy lasu losowego oraz wzmocnienia gradientowego.*" – a dlaczego? Te wnioski zostały poparte wynikami zamieszczonymi w Tabeli 6.1.

Następnie, autor pracy doktorskiej przedyskutował różne kwestie związane z rozszerzeniem bazy QM9. Wykonał odpowiednie obliczenia, które zresztą napotkały znaczne przeszkody.

Ponadto, w kolejnej części pracy doktorskiej zostały przedstawione i omówione wyniki modelowania bioakumulacji związków chemicznych w organizmach żywych w oparciu o fragmentaryczne deskryptory kwantowe. Do ich obliczenia wykorzystano bazę pn. 'QM9-extended plus'. W tym przypadku, uzyskane wyniki wskazują, że połączenie obu grup deskryptorów nie daje znaczącej poprawy budowanych modeli, niezależnie od przyjętej miary oceny zdolności predykcyjnych. Co więcej, modele budowane wyłącznie w oparciu o fragmentaryczne deskryptory kwantowe miały gorsze przewidywania niż modele konstruowane stosując deskryptory RDKit. Autor tłumaczy to suboptymalną zawartością bazy danych eksperymentalnych, która uwzględniała krótsze fragmenty związków niż najdłuższe rozważane fragmenty, a także trudnościami w uzupełnieniu brakujących w danych elementów.

Ostatnią część badań autor pracy doktorskiej poświęcił analizie potencjału jaki może oferować zastosowanie podejścia kwantowo informowanych par atomów. Zbudował wieloparametrowe modele regresji opisujące relację matematyczne pomiędzy deskryptorami, a takie ważnymi parametrami jak logP, pIC50, rozpuszczalność, lipofilowość, energia jonizacji i temperatura topnienia. Ogólnie można powiedzieć, że wprowadzona modyfikacja do standardowej koncepcji par atomów daje lepsze wyniki. Niemniej jednak, w rozważanych przypadkach, deskryptory molekularne zapewniają lepsze własności predykcyjne modeli kalibracyjnych i klasyfikacyjnych.

Podsumowanie pracy doktorskiej zawiera krytyczną dyskusję wniosków, odniesienie się przez jej autora do możliwych źródeł niepowodzeń oraz wskazuje na inne kierunki badań, które mogą być realizowane później. Autor zwraca również

uwagę na zagadnienie otwartych danych i potrzebę udostępniania danych dla rozwijania badań.

Ostatnia część pracy doktorskiej krótko opisuje i przedstawia stworzone oraz udostępnione publicznie aplikacje do przewidywania długości fali przy którym obserwuje się maksimum emisji i możliwe jest obliczenie wybranych fragmentarycznych deskryptorów kwantowych i deskryptorów kwantowych par atomów.

Bibliografia, na którą powołuje się w pracy doktorskiej jej autor, obejmuje łącznie 107 pozycji literatury, z czego przeważająca większość to artykuły, które zostały opublikowane w recenzowanych czasopismach o zasięgu międzynarodowym. Można powiedzieć, że są to aktualne pozycje literatury o dużym znaczeniu dla realizacji badań w obszarze QSAR/QSPR.

Praca doktorska została napisana w języku polskim, w relatywnie dostępny dla czytelnika sposób. Wyróżnia się dużą dbałością o graficzne jej szczegóły, a także bardzo wysokim poziomem estetyki. Wszystkie rysunki autor pracy doktorskiej przygotował z dużą starannością. Niewątpliwie wspierają one dyskusję wyników, ułatwiają analizę porównawczą i formułowanie wniosków. Cel pracy zredagowano poprawnie i klarownie. Niemniej jednak, w niektórych miejscach pracy doktorskiej, w mojej opinii, brakuje zdecydowanie szerszej dyskusji podstaw teoretycznych stosowanych metod obliczeniowych i lepszej argumentacji ich użycia.

Wyniki badań przeprowadzonych w trakcie doktoratu zostały również omówione w dwóch oryginalnych artykułach naukowych, które ukazały się w recenzowanych czasopismach o zasięgu międzynarodowym. Były to takie czasopisma jak *ACS Omega* (70 pkt., IF = 4,1) i *Journal of Molecular Graphics and Modelling* (70 pkt., IF = 4,1). W przypadku obu artykułów doktorant jest pierwszym autorem i autorem do korespondencji, a ponadto zostały one opublikowane w formule otwartego dostępu (licencja CC BY). Można zatem wnioskować, że doktorant miał znaczący udział w zaplanowaniu badań, ich przeprowadzeniu, a także w pisaniu manuskryptu i jego późniejszej recenzji.

Z racji powierzonej mi roli recenzenta chciałem zwrócić uwagę na kilka istotnych zagadnień, które wymagają szerszej dyskusji w trakcie publicznej obrony niniejszej pracy doktorskiej. Poniżej zamieszczam ich listę:

1) Strona 1 – warto zauważyć, że sformułowanie *„modele predykcyjne"* obejmuje modele kalibracyjne i klasyfikacyjne/dyskryminacyjne. W pierwszym przypadku przewidywana jest modelowana zmienna zależna, a w drugim przewiduje się przynależność próbki do określonej grupy lub grup.
2) Strona 33 – autor pracy doktorskiej decyduje się w swoich badaniach używać regresję liniową lub logistyczną. W mojej opinii wymaga to szerszego uzasadnienia i bardziej krytycznego namysłu. W szczególności, mając na względzie wielowymiarowość modelowanych danych z którymi autor ma do czynienia, moją dużą wątpliwość budzi pominięcie w dyskusji takich metod jak regresja głównych składowych (z ang. *principal component regression*, PCR) czy regresja częściowych najmniejszych kwadratów (z ang. *partial least-squares regression*, PLSR). Oczywiście, proste modele liniowe, jedno- i wieloparametrowe można używać w obszarze QSAR/QSPR, ale w przypadku wielu deskryptorów budujących zestaw zmiennych objaśniających efektywność takich modeli jest bardzo ograniczona. Ponadto, bezwzględnie wymaga przeprowadzenia wyboru zmiennych objaśniających.
3) Strona 34 – brak stosownych odnośników do poszczególnych metod obliczeniowych i algorytmów.
4) Strona 34 – *„Algorytmy te charakteryzują się, np. (...) wieloma parametrami, które można zmieniać (...)"* – czy zdaniem doktoranta to faktycznie zaleta?
5) Istnieje wiele możliwych wskaźników walidacyjnych opisujących zdolności predykcyjne modeli. Dlaczego właśnie takie zostały wybrane przez autora pracy doktorskiej wybrane i jakie są ich definicje? Nie ma w pracy doktorskiej starannie dobranych odnośników literaturowych na ten temat.
6) Strona 40 – warto zauważyć, że pojęcie *„optycznie czynny związek"* odnosi się do zjawiska skręcalności płaszczyzny światła spolaryzowanego, które przenika przez roztwór zawierający optycznie czynny związek chemiczny.
7) Strona 45 – co w praktyce oznacza tak sformułowany wniosek *„ (...) dodanie właściwości kwantowych fragmentów cząsteczek do zmiennych modelu predykcyjnego nie wpływa na otrzymany wynik, niezależnie od tego, jak takie deskryptory zostały zdefiniowane"*? Jaki realny wpływ na własności predykcyjne wieloparametrowych modeli (pozytywny czy negatywny) ma włączenie do procesu modelowania dodatkowych deskryptorów? Czy uzyskane błędy przewidywania modeli są satysfakcjonujące z perspektywy poszukiwania lepszych związków do budowy diod OLED?

8) Dlaczego autor pracy doktorskiej nie zdecydował się na stworzenie niezależnego zbioru testowego związków, a proces walidacji oparł wyłącznie o walidację krzyżową? Jakiego była ona typu?
9) Moją poważną obawę budzi użycie modeli regresji wielorakiej dla zestawu zawierającego setki deskryptorów molekularnych. Proszę o szczegółowe wyjaśnienie w jaki sposób było to możliwe, znając ograniczenia tej metody regresji w przypadku modelowania danych ze skorelowanymi zmiennymi. Dlaczego właśnie ta metoda regresji została wybrana? Dlaczego w swych badaniach autor pominął regresję PCR czy PLSR, które przecież są obowiązującymi standardami?
10) Strona 49 – nie jest dla mnie jasne w jaki sposób dokonywana była faktycznie ocena istotności zmiennych. Proszę o szczegółowe wyjaśnienie.
11) Strona 54 – w podsumowaniu autor pracy doktorskiej formułuje wnioski, które wskazują na możliwe źródła błędów. Z perspektywy czasu, czy autor potrafi odnieść się do nich krytycznie?
12) Strona 67 – jakie inne sposoby radzenia sobie z brakującymi elementami występującymi w danych znane są w chemometrii? Czy w aspekcie problemów napotykanych podczas modelowania QSAR/QSPR znajdują one zastosowane?

Przedstawiona mi do recenzji praca doktorska Pana mgr. inż. Bartłomieja Fliszkiewicza, w mojej ocenie, spełnia ustawowe wymogi stawiane pracom doktorskim. Przywołane w treści recenzji uwagi krytyczne nie umniejszają wartości poznawczej niniejszej rozprawy. W toku badań autor pracy doktorskiej zaproponował oryginalne podejścia rozwiązania problemu modelowania wybranych własności wprost wynikających ze struktury chemicznej związków. Zaplanował i wykonał stosowne obliczenia, stworzył modele kalibracyjne i klasyfikacyjne, które następnie poddał walidacji. Sformułowane przez niego wnioski w pełni znajdują odzwierciedlenie w wynikach. Zaproponowanie przez Pana mgr. inż. Bartłomieja Fliszkiewicza nowych deskryptorów, ich obliczenie, opracowanie graficznego interfejsu, a także szerokie udostępnienie deskryptorów dla potencjalnych użytkowników podejść QSAR/QSPR to niewątpliwie bardzo duży wysiłek i wiele pracy, co wymaga docenienia.

Mając na względzie powyższej wymienione przesłanki, wnoszę do Rady Dyscypliny Naukowej Nauk Chemicznych Wojskowej Akademii Technicznej o dopuszczenie Pana mgr. inż. Bartłomieja Fliszkiewicza do dalszych etapów przewodu doktorskiego.

Michał Danikowski