Prof. dr hab. Jarosław Polański
Instytut Chemii Uniwersytetu Śląskiego
ul. Szkolna 9, 40-006 Katowice

Katowice, 28 marca 2022

Recenzja rozprawy doktorskiej magistra Piotra Urbaszka pt. „Przewidywanie sorpcji bromo- i chloropochodnych trwałych zanieczyszczeń organicznych (TZO) na powierzchni fullerenu C_{60}"

Przedstawione w recenzowanej pracy wyniki nawiązują do tematyki badań promotora rozprawy prof. Tomasza Puzyna. Tematyka tych prac od lat związana jest między innymi właśnie z problemami nanotechnologii oraz związanymi z tym zagrożeniami toksykologicznymi.

Najogólniej celem pracy Doktoranta było prognozowanie właściwości chemicznych dużego zbioru (1 840 951) kongenerycznych bromo-, chloro- i bromochloro- pochodnych wielopierścieniowych węglowodorów aromatycznych, które Doktorant określa mianem TZO. Ogólnie praca wpisuje się w kierunek badań określany jako modelowanie wielowymiarowych zależności QSAR (Quantitative Structure-Activity Relationships). Ogólnie QSAR zdefiniowane może być jako odwzorowanie przestrzeni właściwości w przestrzeń deskryptorów. Właściwości są przy tym mierzone w eksperymentach, podczas gdy deskryptory to parametry obliczane dla cząsteczek chemicznych. Istotną różnicę w przypadku pracy Pana Urbaszka stanowi fakt, że modeluje on sorpcję na powierzchni fullerenu C_{60}, którą również oblicza w eksperymencie *in silico*. Z drugiej strony duża liczba analizowanych związków chemicznych uwidacznia obecne trendy badań tzw. wielkich danych (big data). Warto wobec tego w kilku słowach opisać obecne paradygmaty projektowania molekularnego z wykorzystaniem big data. Projektowanie molekularne powinno być automatyczne i autonomiczne, tzn. przebiegać bez udziału człowieka (Schneider, G. Automating Drug Discovery. Nat. Rev. Drug Discov. 2017, 17, 97–113). W metodach *in silico* efektywną metodą analizy wielkich danych zapewniać

mają metody tzw. *deep learning*. Obserujemy ostatnio sukcesy takich metod w szachach czy rozpoznawaniu twarzy. Inspirują one zastosowania podobnych metod właśnie w projektowaniu molekularnym, szczególnie w projektowaniu leków (Polanski, J. Unsupervised Learning in Drug Design from Self-Organization to Deep Chemistry. Int. J. Mol. Sci. 2022, 23, 2797). W nowym leksykonie metody QSAR (QSPR) dzielą się na dwa podstawowe typy tzw. *feature learning* oraz *feature engineering* (Chuang, K.V.; Gunsalus, L.M.; Keiser, M.J. Learning molecular representations for medicinal chemistry: Miniperspective. J. Med. Chem. 2020, 63, 8705–8722). Podczas gdy *feature learning* zapewnia całkowicie autonomiczne rozpoznanie cech istotnych dla prognozowania danej właściwości przez algorytm komputerowy, w metodzie *feature engineering* to człowiek typuje deskryptory istotne w danej procedurze prognozowania. W pewnym uproszczeniu można stwierdzić, że *deep learning* są adaptacją metod QSAR w domenie wielkich danych, gdzie to komputer sam dokonuje operacji feature engineering (inżynierii deskryptorów). Chociaż metody *deep chemistry* wciąż są sprawą przyszłości są one przedmiotem intensywnych badań (Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. 2019, 37, 1038–1040). Prace takie są także przedmiotem krytyki. Inaczej niż w szachach, czy rozpoznawaniu twarzy, gdzie osiąga się spektakularne wyniki predykcji, w chemii wciąż obserwuje się brak wystarczającej populacji danych mierzonych właściwości. Reprezentatywność i zróżnicowanie w zbiorze chemotypów chemicznych bywa kwestionowane.

Ten krótki opis metod QSAR w obecnej nomenklaturze pozwoli nam ławo zrozumieć istotę pracy Doktoranta. Generuje on *in silico* dużą wirtualną bibliotekę tytułowych związków. Następnie metodą *feature engineering* dokonuje wyboru deskryptorów molekularnych, które modelują obliczane przez niego *in silico* wartości sorpcji na powierzchni C_{60} dla niewielkiego podzbioru analizowanej biblioteki. Deskryptory analizowane w pracy Doktorant przedstawił w Tabeli 4 (str. 66). Metodę obliczeń oddziaływania TZO-fuleren$_{60}$ optymalizował, stosując różne algorytmy obliczeniowe (rozdz. Etap 3 (str. 67) a dokładne obliczenia dokonał dla wybranych reprezentatywnych halogenowanych dioksyn PXDD@C60, wykorzystując funcjonał M06-2X w bazie typu Peopla 6-31 (str. 67). Kalibracja metody QSAR wykonana została metodą analizy PLS z

algorytmem genetycznym dla szeregu kongenerów dibenzo-p-dioksyn podstawionych Br lub Cl. Autor zachowuje przy tym w pełni poprawną procedurę metodologii modelowania wielowymiarowych QSAR, np. podział na zbiór modelowy (treningowy) i testowy. Autor szacuje wpływ typu chemotypu (określanej jako struktura bazowa), kształtu, liczby, typu oraz lokalizacji podstawnika na badaną właściwość. Na stronie 93 Autor podaje równanie modelujące energię adsorpcji w funkcji sumarycznej liczby atomów wodoru, całkowitej energii związku oraz składowych momentu dipolowego wzdłuż X oraz Y. Taki zestaw rzeczywistych deskryptorów wybrany został spośród wszystkich analizowanych deskryptorów przez algorytm genetyczny. Autor charakteryzuje otrzymany model wartościami dopasowania statystycznego (R^2, walidowanym krzyżowo metodą LOO Q^2_{cv} oraz Q^2 dla zbioru zewnętrznego). Wartości te są wysokie, co świadczy o tym, że model nie jest przeuczony i dobrze prognozuje wartości zewnętrzne poza domeną treningową. Przy okazji pod równaniem brakuje mi liczby związków uwzględnionych dla kolejnych zbiorów treningowych i testowych. W tabeli 5 możemy odnaleźć te liczby (w sumie 32 związki, w tym 8 związków zbioru testowych). Doktorant sprawdza także domenę, dla której taki model może być stosowany. Poza obszarem domeny znalazło się 84 kongenery badanego zbioru PXDD. Na rysunku 15b Autor przedstawił także rozkład wartości E_{ads} obliczone wg otrzymanego modelu dla wszystkich 1701 kongenerów PXDD. W kolejnym etapie Autor wykorzystuje wybrane przez algorytm genetyczny deskryptory do modelowania równania PLS. Model wykorzystuje do obliczenia E_{ads} dla wszystkich (liczby nie podaje) badanych kongenerów PXDD zbioru predykcyjnego. Kolejnym etapem jest rozszerzenie modelu na wszystkie badane chemotypy. Tabela 6 przedstawia struktury dla których wykonuje obliczenia E_{ads} metodą DFT. Następnie prowadzi analizę uzyskanych wyników tak, aby w wielowymiarowej przestrzeni deskryptorów molekularnych uzyskać spójny model poprzez wykluczenie pewnej liczby związków dla których wynik obliczeń odstają od obserwowanych trendów. Omawia także przyczyny takich odchyleń. Analogicznie jak poprzednio modeluje równanie regresyjne str. 108, podając wartości dopasowania statystycznego (R^2, walidowanym krzyżowo metodą LOO Q^2_{cv} oraz Q2 dla zbioru zewnętrznego). Także tutaj wartości te są wysokie, co świadczy o tym, że model nie jest przeuczony i dobrze prognozuje wartości zewnętrzne poza domeną treningową.

Liczbowe wartości dla tych bardziej zróżnicowanych chemotypów są nieco niższe niż w przypadku kongenerycznego szeregu PXDD. Model taki używa do predykcji wartości E_{ads} i na stronie 111 (rys. 22) przedstawia analizę Insurbia domeny stosowalności modelu.

Doktorant omawia znaczenie poszczególnych deskryptorów modelu oraz ich interpretację dla modelu nano-QSPR oraz wybranych chemotypów i konkretnych związków chemicznych. W sumie pracę interpretować można jako klasyczne modelowanie wielowymiarowego QSAR (QSPR), przy czym właściwość mierzona jest w eksperymencie in silico (obliczenia DFT). Ciekaw jestem czy są znane jakieś wartości E_{ads} mierzone w eksperymentach *in vitro*. Jeżeli tak, to jak one mają się do wartości obliczanych. W skrajnym bowiem przypadku można się upierać, że wartości obliczane są także pewnymi złożonymi deskryptorami molekularnymi charakteryzującymi układ adsorbent-adsorbat (Polanski, J. Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. Current medicinal chemistry, 16(25), 3243-3257), a nie typową właściwością chemiczną. Oczywiście dobrze rozumiem dylemat Autora, który dysponuje ograniczoną liczbą danych literaturowych. Niemniej ciekaw jestem jego polemiki z interpretacją *deskryptor-deskryptor*. Następny problem, który wydaje mi się interesujący w aspekcie obecnych problemów modelowania QSAR to pytanie, czy możliwe byłoby w analizowanym przypadku zastosowanie czystej metody *feature learning*. Jakie deskryptory można by zastosować? Jak doktorant widzi ten problem w aspekcie swojej metodyki badań?

Pracę otrzymałem w postaci typowej dla rozprawy doktorskiej. Manuskrypt liczy 148 stron. Składa się ze wstępu, części: literaturowej, hipotezy i celów badawczych, metodyki badań własnych, wyników badań własnych, podsumowania, dodatku (omawiającego dorobek oraz podającego linki do zewnętrznych zbiorów danych. Autor cytuje 187 pozycji literaturowych.

W części literaturowej Doktorant omawia problemy związane z nanocząstkami węglowymi, trwałymi zanieczyszczeniami organicznymi (TZO), modelowaniem zależności QSAR/QSPR, oddziaływaniami między cząsteczką fulerenu i TZO. To dobrze i ciekawie napisany fragment pracy, który poza częścią naukową obejmuje

wprowadzenie do aktualnych regulacji prawnych konwencji i definicji. Informacje te są moim zdaniem często niedoceniane w chemii. Ta część stanowi bardzo zgrabnie napisane studium literaturowe. Dojrzały język naukowy, poprawne sformułowania są potwierdzeniem wysokiego poziomu. Dodatkowym atutem jest przeglądowa publikacja współautorstwa Doktoranta poświęcona tematowi szacowania ryzyka aplikacji nanomateriałów metodami *in silico* (Gajewicz A, Rasulev B, Dinadayalane TC, Urbaszek P, Puzyn T, Leszczynska D, Leszczynski J. Advancing risk assessment of engineered nanomaterials: application of computational approaches. Advanced drug delivery reviews. 2012 Dec 1;64(15):1663-93. IF czasopisma jest wysoki (15,43), a sama publikacja posiada już 234 cytowania (Google Scholar). Na stronie 29 Autor opisuje ciekawą historię DDT. To słynna cząsteczka. W roku 1948 Paul Mueller uzyskał Nagrodę Nobla za odkrycie jej selektywnego działania przeciw owadom, szczególnie komarom, aktywność antymalaryczna. Z kolei w latach 60' DDT stało się motywem pracy Rachel Carson *Silent Spring*, która piętnuje DDT oskarżając je o ekotosyczność (wiosna bez ptaków). Pełna zgoda z historią opisaną przez Doktoranta. Z drugiej jednak strony w ostatnich latach kwestionuje się fakty podawane przez Garson. Książce *Silent Spring* zarzuca się grę na emocjach. Ciekaw jestem, czy Autor zna anegdotyczną historię prof. J. Gordona Edwardsa z Uniwersytetu San Jose State University (**The Lies of Rachel Carson - 21st Century,** https://21sci-tech.com/articles/summ02/Carson.html). W wielu krajach przywrócono możliwość stosowania DDT lecz w sposób bardziej inteligentny; jaki i dlaczego? Jak poradziły sobie z tym komary? (Mandavilli, A. (2006). Health agency backs use of DDT against malaria. Nature, 443(7109), 250-252; Marcin Rotkiewicz, Zbawienna trucizna, Polityka, 38, 2012). Myślę, że historia DDT dobrze ilustruje dylematy chemii w epoce antropocenu.

Nieco bardziej krytycznie pod względem redakcyjnym oceniam część badania własne oraz podsumowanie. W części badania własne czytelnikowi trudno szybko zorientować się jaki jest, w sumie dosyć złożony schemat modelowania QSAR. Przy równaniach brakuje liczby związków uwzględnionych w zbiorach testowych i modelowych (danych trzeba szukać przez dokładne przeszukiwanie tekstu). Także w tabelach wymieniających konkretne związki chemiczne brak ich kolejnych numerów, co ułatwiłoby czytelnikowi orientację. Zwraca uwagę także nieco słabszy język podsumowania. Na przykład: Autor

pisze: *wpływ ilości podstawników* powinno być *wpływ* ***liczby*** *podstawników*, czy niejasne sformułowania, np. *dla grupy z 1 840 951 kongenerów, mimo wykluczenia na etapie kalibracji modelu kilku kongenerów*. Takie błędy przypisałbym pośpiechowi redakcyjnemu. Z załączonego wykazu publikacji naukowych Doktoranta widać, że wiele z prac pochodzi sprzed kilku lat. Jak sądzę zmiany ustawowe zmusiły Doktoranta do szybkiej redakcji rozprawy oraz obrony.

Podsumowując, naukową treść pracy, stanowi ona klasyczne studium wielowymiarowego QSPR w domenie wielkich danych (big data), przy czym P – *property*, mierzona jest w eksperymencie *in silico* (Lach, D.; et al., Functional and Material Properties in Nanocatalyst Design: A Data Handling and Sharing Problem. Int. J. Mol. Sci. 2021, 22, 5176). Autor posiada biegłą znajomość modelowania takich QSPR oraz prawidłowo stosuje metodykę takich badań. Ze względu na obiekt zainteresowań badania Pana Urbaszka pracę można zaklasyfikować do metod nano-QSPR.

Przedstawiłem powyżej bardzo skrótowo treści recenzowanej pracy. Zakres wykonanych prac budzić musi duży szacunek. Na podkreślenie zasługuje fakt, że Autor zajmował się badaniem złożonych problemów oddziaływania nanocząstek węglowych z TZO. Doktorant twórczo wykorzystuje i modyfikuje metody modelowania wielowymiarowych QSPR (regresja wielowymiarowa, regresja PLS). Praca napisana jest w sposób przejrzysty. Na podkreślenie zasługuje dojrzały sposób opisu wyników. Warto tu dodać, że wyniki pracy pana Urbaszka zostały opublikowane w artykułach naukowych w czasopismach z listy filadelfijskiej oraz rozdziale książki. Doktorant jest współautorem dwunastu publikacji naukowych (w jednej jest pierwszym autorem), czterech publikacji o charakterze książkowym oraz 11 wystąpień konferencyjnych. Jest także laureatem dwóch nagród naukowych (2010, 2011 rok).

Podsumowując, uważam, że przedstawiona mi do recenzji praca doktorska spełnia wymogi stawiane rozprawom doktorskim przez ustawę o stopniach i tytułach naukowych, w związku z czym wnoszę o dopuszczenie pana Piotra Urbaszka do dalszych etapów przewodu doktorskiego.

Jarosław Polański