

Poznań, 04.12. 2019 r.

Prof. dr hab. Zenon Foltynowicz
Instytut Nauk o Jakości
Uniwersytet Ekonomiczny w Poznaniu

RECENZJA

rozprawy doktorskiej Pani mgr Urszuli Kuci pt. „*Analiza QSPR wielkich bibliotek związków chemicznych na przykładzie bazy danych ABAMACHEM*”
przygotowana
na zlecenie Rady Instytutu Chemii Uniwersytetu Śląskiego w Katowicach

Podstawą opracowania recenzji jest pismo Pana Prof. dr. hab. Stanisława Kucharskiego, Dyrektora Instytutu Chemii Uniwersytetu Śląskiego z dnia 30 września 2019 roku o powołaniu mnie przez Radę Instytutu Chemii Uniwersytetu Śląskiego na recenzenta w przewodzie doktorskim mgr Urszuli Kuci. Recenzowana rozprawa pod ww. tytułem jest przygotowana pod promotorstwem Pana Prof. dr. hab. inż. Jarosława Polańskiego.

W tym roku mija 50 lat internetu, który głęboko zmienił sposób naszego komunikowania się i informowania. Technologie informacyjno-komunikacyjne zmieniły naukę dzięki umożliwieniu współpracy na odległość, przetwarzaniu ogromnych ilości danych, dostępowi do odległych danych oraz eksperymentom *in silico*. Oczywiście nie tylko internetu nie było by bez komputerów. Jak stwierdza Doktorantka we Wstępie „... w dzisiejszych czasach komputer stał się powszechnym narzędziem badań chemicznych” i „Miało to wpływ na powstanie dyscypliny naukowej jaką jest chemoinformatyka”. Tematyka recenzowanej rozprawy doktorskiej wpisuje się właśnie w problematykę chemoinformatyki, jest zatem bardzo aktualna. Recenzowana rozprawa doktorska dotyczy badań relacji między strukturą związku chemicznego a właściwościami QSPR, co znalazło odzwierciedlenie w tytule rozprawy. Na uwagę zasługuje zwrócenie przez Doktorantkę uwagi na aspekt ekonomiczny projektowania molekularnego. Szkoda, że QSER nie znalazło się w tytule rozprawy. Doktorantka stawia we Wstępie pytanie, które można by uznać za hipotezę badawczą, mianowicie „czy możliwe jest modelowanie efektów ekonomicznych na rynku związków chemicznych”. Szkoda, że nie znalazło to odzwierciedlenia w celu pracy, który został

sformułowany jako „Celem pracy była analiza dostępności, akwizycji i przetwarzania zbiorów danych tego typu”.

Praca składa się z 21 rozdziałów zawartych na 109 stronach uzupełnionych Bibliografią oraz wykazami tabel i rysunków oraz 35 stronami Załączników. Praca zredagowana jest z podziałem na część literaturową/teoretyczną (54 str., 43% rozprawy) i badawczą (42 str., 33%) oraz osobno Podsumowanie i wnioski (30 str., 24%).

We Wstępie Doktorantka dokonała wprowadzenia w tematykę rozprawy, dobrze uzasadniając podjęty temat badań. W krótkim pierwszym rozdziale części literaturowej zasygnalizowała problemy projektowania leków *in silico* oraz koncepcje obliczeniowe stosowane w tych metodach. Doktorantka jest tak przekonana, że metody *in silico* są tak powszechnie znane jak *in vitro* czy *in vivo*, że nie pokusiła się o ich chociażby krótką charakterystykę. W rozdziale 2. Dokonała prezentacji i krótkiego omówienia siedmiu deskryptorów molekularnych. Sposób przedstawienia świadczy o dobrej znajomości tej problematyki przez Doktorantkę. W rozdziale 3. podjęła problematykę pomiarów efektów fizykochemicznych, biologicznych i ekonomicznych. Bez sprecyzowania we wprowadzeniu do tego rozdziału z czym te efekty są związane przechodzi do konkretnych przykładów na poszczególne rodzaje efektów. O ile w przypadku wpływu efektów elektronowych podstawników na reaktywność cząsteczki (efekt fizykochemiczny) przedstawia równanie Hammetta a dla opisu wpływu lipofilowości na efekty biologiczne proponuje metodę Hanscha z równaniem na $\log P$ to w przypadku efektów ekonomicznych prezentuje prawo Erooma i tzw. „wiek leku” zaproponowany przez Promotora rozprawy. Omówienie wieku leku jest bardzo enigmatyczne i niezorientowany czytelnik może mieć trudność w zorientowaniu się jaka „Analiza wykazuje, że leki się starzeją”.

Pewne zastrzeżenia budzi omówienie prawa Erooma. Przede wszystkim Doktorantka cytuje niewłaściwy odnośnik literaturowy (11) a właściwej, bazowej publikacji Scannella [Scannell J.W., Blanckley A., Boldon H., Warrington B., Diagnosis the decline in pharmaceutical R&D efficiency, *Nature Reviews Drug Discovery*, 11, 191-200, 2012] nie ma w ogóle w Bibliografii. Ponieważ Doktorantka nie przytacza brzmienia tego prawa pozwolę sobie je podać: Prawo Erooma wskazuje, że liczba zatwierdzonych nowych leków na miliard dolarów wydanych na badania i rozwój od 1950 r. zmniejsza się o około połowę co dziewięć lat. Pomimo znacznego postępu naukowo-technicznego w zakresie procesu odkrywania nowych leków a także usprawnień w obszarze zarządzania finansami skuteczność prac R&D przemysłu farmaceutycznego nie wykazuje tendencji wzrostowej. Być może przemysł farmaceutyczny musi przyzwyczaić się do rynków o malejącym zwrocie z inwestycji ROI?

W rozdziale 4. Zatytułowanym „Analizy big data w chemii i projektowaniu leków” przedstawione zostały w p. 4.1 cztery „typowe przykłady generowania dużych bibliotek danych stosowanych w analizach QSAR. Sposób prezentacji świadczy o dobrej znajomości tego zagadnienia przez Doktorantkę. W p. 4.2 Doktorantka rozważa możliwości „Analizy dużych populacji danych w chemii i projektowaniu leków”. Jest to znaczący problem ze względu na ogromne i stale rosnące nagromadzenie danych i informacji. Ekstrakcja istotnych informacji staje się zatem coraz trudniejsza. W związku z tym kluczowe stają się metody analizy danych. Doktorantka nie zwróciła uwagi na metody grupowania, tzw. Clustering algorithms, spośród których prawdopodobnie najbardziej znany jest tzw. K-means. W kontekście komputerowego odkrywania leków, w którym przestrzeń chemiczna szacowana jest na 10^{63} związków, grupowanie możemy wykorzystać do wybrania obiecujących podgrup w dużej bibliotece chemicznej, pozbywając się dużej części zbioru danych bez zainteresowania medycznego. Do zagadnienia tego w pewien sposób odnosi się w rozdziale 5. „Badanie architektury chemii organicznej” w którym zwraca uwagę na metody sieciowe. Na szczególną uwagę zasługuje moim zdaniem podrozdział 5.1 omawiający „Znaczenie ekonomii w syntezie organicznej”. Doktorantka podaje przykłady algorytmów planowania syntez z uwzględnieniem kosztów reagentów, nakładu pracy i czasu reakcji. Zwraca również uwagę na istotny fakt, czy optymalne kosztowo ścieżki syntezy sprawdzą się w rzeczywistym eksperymencie w laboratorium, pozwalając uzyskać odpowiednio wysokie wydajności. Doktorantka zwraca na ten fakt szczególną uwagę w podrozdziale 5.3. Ekonomia atomowa, gdzie stwierdza, że „Pomimo przeprowadzenia wielu badań i wielu prób, nie odnotowano w literaturze przedmiotu informacji o zaprojektowaniu kompletnych ścieżek przez komputer, które następnie mogłyby zostać powtórzone w laboratoriach.”. Po takim stwierdzeniu omawia oprogramowanie Chematica i podaje przykłady jego pozytywnego wykorzystania do zaprogramowania syntez, które sprawdziły się w laboratoriach.

Strategie, które wpływają na decyzje firm farmaceutycznych podczas poszukiwania nowych leków i ich komercjalizacji zostały zaprezentowane w Rozdz. 6. Doktorantka przechodzi od przeglądu procesu wykrywania i rozwoju leku do dominujących strategii optymalizacji podejmowania (właściwych) decyzji w zakresie badań nad nowymi lekami. Na uwagę zasługuje zwrócenie uwagi na rozwój historyczny tych strategii, od *in vivo* do *in silico*. Przedstawiony materiał, aczkolwiek nieco skrótowy, świadczy o dobrej orientacji Doktorantki w tych zagadnieniach. Pozwolę sobie zwrócić uwagę na fakt rozbieżności między odwołaniem się do Rys. 20 na stronie 50 a właściwym rysunkiem 22 na stronie 51.

Część literaturową kończy rozdział 7. poświęcony omówieniu wydajności ligandu jako miary stosowanej w projektowaniu leków. Parametr ten jest szeroko opisywany w literaturze. Doktorantka dokonując jej przeglądu uzyskuje informacje o zależności wydajności ligandów od właściwości fizykochemicznych cząsteczek oraz formułuje wnioski o zastosowaniu miar wydajności w optymalizacji leków. Nie wspomina jednakże o ograniczeniach tego parametru, które opisuje wraz z Promotorem w jednej z publikacji.

Przedstawiony w części literaturowej materiał świadczy moim zdaniem o dobrej znajomości wybranych aspektów problematyki projektowania leków przez Doktorantkę.

Badania własne przedstawiła Doktorantka na 45 stronach pracy. Osobno wydzieliła Część eksperymentalną w której zawarła charakterystyki wykorzystywanych oprogramowań oraz etapy analizy i przetwarzania danych Abachem.

Przeprowadzone badania stanowią przykład analizy Big Data. We współautorskiej publikacji Promotora i Doktorantki napisano „modeling in pharma and chemistry..... requires a complete understanding of the metrics that are used and the chemical effects that determine these metrics”. To stwierdzenie jest dobrym mottem przy ocenie przeprowadzonych badań i uzyskanych wyników. Dlatego skoncentruję się głównie na aspektach ekonomicznych.

Doktorantka jako obiekt badań wybrała katalog bloków budulcowych firmy Abachem zawierający ponad 2 miliony związków chemicznych wraz z ceną każdego oferowanego związku. Zastanawia się jaki parametr przyjęto jako podstawę wyceny tych związków, gdyż twórcy tej bazy nie chcieli tego ujawnić?

W ramach badań własnych Doktorantka zaproponowała binowanie danych jako podstawową metodę modelowania statystyk molekularnych. Jako potencjalne korelaty cen bloków budulcowych z biblioteki Abachem wykorzystywała deskryptory SAS.

Badając zależności między wybranymi deskryptorami molekularnymi a ceną cząsteczki określiła m.in. wpływ masy cząsteczkowej, liczby atomów składu pierwiastkowego na cenę cząsteczki. Jednym z wyników tych badań jest macierz korelacji dla ceny molowej i wagowej oraz wybranych deskryptorów molekularnych dla danych z bazy Abachem, na podstawie której stwierdza, że korelacje pomiędzy danymi ekonomicznymi (ceną) a różnymi deskryptorami molekularnymi są stosunkowo niskie. Nie mniej dochodzi do wniosku, że „cena zależy liniowo od molarności substancji oraz, że średnio im większa jest waga sprzedawanej substancji, tym większa jest cena”. (Pytanie: czy cena jednostkowa czy za daną ilość?). Potwierdza to również wysoka wartość współczynnika korelacji dla średniej ceny molowej a binowaną masą, co wskazuje, że na rynku cząsteczek płaci się za ilość materii. Analizując wpływ składu pierwiastkowego cząsteczki na jej cenę ustaliła obecność których

pierwiastków wpływa na ekonomię atomową syntezy. Jest to ważne stwierdzenie w aspekcie „zielonej chemii”.

Rozdział 8. Zatytułowany jest „Wpływ dostępności syntetycznej na cenę biblioteki Abachem”. Tytuł ten jest pewnym skrótem myślowym, których zresztą nie brakuje w pracy, albowiem sugeruje, że może chodzić o wycenę bazy Abachem! (zresztą taki skrót myślowy pojawia się także w Podsumowaniu, gdzie podano „Niniejsza praca opisuje modelowanie zależności między strukturą a ceną dużej biblioteki związków chemicznych Abachem”). Dopiero dalsza lektura tego rozdziału (8.) pozwala stwierdzić, że „W przeprowadzonych badaniach przeanalizowano wpływ syntetycznej dostępności na cenę cząsteczki dla (z ?, uwaga recenzenta) biblioteki Abachem” i stwierdzono, że wysoki współczynnik korelacji między średnią ceną \$/g a binowaną syntetyczną dostępnością SAS1 świadczy o tym, że „chemia” odgrywa ważną rolę w ustalaniu ceny.

Na podkreślenie zasługuje przedstawiona w rozdziale 9. „Ewaluacja statystyczna uzyskanych modeli struktura-cena” w celu odróżnienia efektów losowych od rzeczywistych.

Rozpoczynając „Interpretację uzyskanych wyników” (Rozdz. 10.) Doktorantka ponawia pytanie, czy w aspekcie molekularnych czynników ekonomii atomowej płacimy za liczbę cząsteczek czy ilość materii. Opisana na 12 stronach interpretacja wyników pozwala Doktorantce na stwierdzenie, że „Obserwowane zależności dobrze ilustrują fakt dominowania efektów ekonomicznych na rynku związków chemicznych, gdzie średnia cena jest funkcją ilości oferowanej *masy*”. A przecież trudno nie zgodzić się z konkluzją podaną na str. 81, „że średnio w każdej grupie związków można spotkać takie, których synteza jest trudna oraz takich, która synteza jest łatwiejsza” (pisownia oryginalna) co powinno znajdować odzwierciedlenie w ich cenie.

Nie zgłaszam zastrzeżeń do poprawności sformułowanych wniosków.

W pracy zacytowano 118 pozycji zwartych i netografii. O brakującej pozycji już wspominałem. W przypadku poz. Nr 34, 43 i 67 nie podano nazwy czasopisma, błąd również w poz. 78.

W podsumowaniu posłużę się analizą SWOT, którą przedstawiam w odwrotnej kolejności.

Zagrożeniem dla szerszego wykorzystania opracowanej metodologii oceny wpływu parametrów fizykochemicznych cząsteczki, złożoności jej syntezy, itp. na cenę oferowanych produktów może być trudność w uzyskaniu dostępu do danych z innych komercyjnych baz czy bibliotek związków chemicznych czy elementów budulcowych. Nasuwa się również pytanie, czy w modelowaniu zależności cena-struktura cząsteczki uwzględniono wszystkie

„economical parameteres”, np. koszty uzyskania ochrony patentowej, które mogą w znaczący sposób wpływać na cenę cząsteczki, zwłaszcza, jeżeli jest syntezowana w oparciu o udzieloną licencję. Można przypuszczać, że oferowane przez Abachem klocki budulcowe syntezowane są na zasadzie open access do własności przemysłowej.

Szanse: wiele firm, również w Polsce, stosuje już metody *in silico* do projektowania cząsteczek mogących być potencjalnymi lekami. Wykorzystanie przedstawionych w pracy zależności „złożoność cząsteczki” a jej cena molekularna czy wagowa może być im przydatne do ustalenia wartości rynkowej swoich produktów.

Do **slabej strony** pracy zaliczam brak wykonania końcowej korekty językowej. Praca napisana jest fachowym językiem, ale Doktorantka nie uniknęła stosowania wielu skrótów myślowych, niektóre już wymieniałem. W pracy, zwłaszcza w części badawczej, Doktorantka stosuje formę osobową, podczas gdy ogólnie stosuje się bezosobową formę pisania rozpraw czy artykułów naukowych.

W wielu przypadkach stwierdzić można zawiłą składnię, np.:

- str. 60: Ponieważ intuicyjnie wydaje się, że im mniej dostępna substancja, tym wyższa powinna być cena deskryptory SAS wykorzystywałam jako potencjalne korelaty cen biblioteki Abachem.

- str. 74: W ramach zbliżonych produktów, które moglibyśmy w chemii określić jako kongeneryczne wydaje się to dobrze oddawać zasady ekonomii.

Str. 81: j.w.

Str. 96: Ponieważ statystyka liczny cząsteczek w jednostce masy koreluje z $1/MW$ efekt taki często przybierać może formę hiperboli.

Zamieszczone w pracy rysunki powinny być opisane w języku polskim. Tabele nie powinny być dzielone i umieszczane na 2 stronach, gdyż utrudnia to zapoznavanie się z nimi, np. Tabela 4./str.71/72.

Wykaz skrótów jest pomocny, szkoda, że nie obejmuje wszystkich użytych w pracy, np. PLS, COMFA, CoMSIA, CoMSA.

Doktorantka używa zamiennie terminy big data (po angielsku Big Data) i wielkie dane. Pozwolę sobie zauważyć, że Zespół ds. Terminologii Informatycznej Rady Języka Polskiego propaguje termin polski „gigadane” (do ew. wykorzystania).

Do **mocnych stron** recenzowanej rozprawy doktorskiej zaliczam przede wszystkim podjęcie ambitnej próby zrozumienia i wyjaśnienia ekonomicznych uwarunkowań projektowania molekularnego. Jak podaje Doktorantka w Podsumowaniu „Nasze modele są pierwszymi modelami typu struktura-ekonomia opisanymi w literaturze”. Na podkreślenie zasługuje fakt,

że uzyskane wyniki stanowią podstawę 3 pionierskich publikacji, których Doktorantka jest współautorką, a które ukazały się w czasopiśmie z IF. Materiał podlegał zatem recenzjom, co upoważnia do stwierdzenia, że proponowane metody zostały zweryfikowane przez świat nauki. Na podkreślenie zasługuje fakt, że wnoszą one znaczący wkład do badań z zakresu QSER „structure-economy relationships”. O ich znaczeniu świadczyć mogą również dotychczasowe cytowania tych prac, np. 8 pracy, która ukazała się w Scientific Reports i 10 z Journal of Chemoinformatics, mimo upływu krótkiego czasu od ich publikacji (2016 i 2017, odpowiednio).

Niezależnie od pewnych niewielkich zastrzeżeń i wątpliwości powstałych w trakcie studiowania rozprawy, mogę wyraźnie stwierdzić, że moja ocena pracy doktorskiej Pani mgr Urszuli Kuci jest pozytywna.

Stwierdzam zatem, że przedłożona mi do recenzji rozprawa Pani mgr Urszuli Kuci pt.

„Analiza QSPR wielkich bibliotek związków chemicznych na przykładzie bazy danych

ABAMACHEM” spełnia wymogi stawiane rozprawom doktorskim zgodnie z zapisami

ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach

i tytule naukowym w zakresie sztuki (Dz. U. Nr 65, poz. 595, z późn. zm.).

Wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony.

Wnioskuje również o wyróżnienie ocenianej rozprawy. Argumenty uzasadniające wyróżnienie zawarłem powyżej w punkcie **mocne strony rozprawy**.

Ławon Foltynowicz