Prof. dr hab. Artur Michalak
Zakład Chemii Teoretycznej
Wydział Chemii
Uniwersytet Jagielloński
ul. Gronostajowa 2, 30-387 Kraków
tel. +48-12-686-2381
fax. +48-12-686-2750
e-mail: michalak@chemia.uj.edu.pl

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

Wydział Chemii

ul. Gronostajowa 2
30-387 Kraków
tel. +48 12 686 26 00
fax +48 12 686 27 50
sekretar@chemia.uj.edu.pl
www.chemia.uj.edu.pl

Kraków, 14 grudnia 2018

Recenzja rozprawy doktorskiej mgr inż. Sary Szymkuć zatytułowanej

„Teaching the computer reactivity rules and strategies of automated retrosynthetic planning”

Rozprawa doktorska mgr inż. Sary Szymkuć przygotowana została w Instytucie Chemii Organicznej Polskiej Akademii Nauk w Warszawie, pod promotorską opieką prof. dr Bartosza A. Grzybowskiego.

Badania Doktorantki przeprowadzone w ramach przewodu doktorskiego **związane są z jednym z największych** - w mojej ocenie - **współczesnych osiągnięć informatyki chemicznej / chemii obliczeniowej, jakim jest opracowanie w grupie prof. Grzybowskiego metodologii oraz kodów komputerowych (platforma obliczeniowa *Chematica*), prowadzących do skutecznego komputerowego przewidywania ścieżek syntezy związków organicznych.** Idea zastosowania komputerów do automatyzacji procesu poszukiwania możliwych szlaków syntezy związków organicznych nie jest nowa – sięga lar 60-tych ubiegłego stulecia. Wiele wcześniejszych prób podejmowanych w różnych okresach kończyło się jednak niepowodzeniami. Dotyczy to także współczesnych prób, w oparciu o kilka programów rozwijanych w kilku innych ośrodkach w ciągu ostatniej dekady. **Na tym tle tematyka i cele doktoratu mgr inż. Sary Szymkuć jawią się jako niezwykle aktualne, ambitne, ale także obarczone wysokim stopniem ryzyka.** Ale tym większe staje się znaczenie sukcesu - **podejście rozwijane w grupie prof. Grzybowskiego, przy znacznym współudziale Doktorantki jest pierwszym przykładem skutecznego przewidywania ścieżek syntezy.**

Rozprawa doktorska mgr inż. Sary Szymkuć została przygotowana w formie spójnego tematycznie zbioru 6 artykułów, opatrzonych komentarzem Autorki. Wszystkie te publikacje są pracami wieloautorskimi (liczba autorów 5-20); załączone zostały oświadczenia Doktorantki, Promotora i pozostałych współautorów dotyczących określenia ich wkładu. Prace stanowiące podstawę rozprawy opublikowane zostały w renomowanych czasopismach ***Chem, Angew. Chem. Int. Ed.*** **(3 prace),** ***Sci. Rep.*** oraz ***Chimia*****.** Wszystkie prace, stanowiące podstawę rozprawy, zostały już opublikowane (w l. 2015-2018), przeszły zatem proces recenzji i spełniły wysokie wymagania stawiane w renomowanych czasopismach międzynarodowych, zarówno co do poziomu merytorycznego, jak i językowego oraz edytorskiego. Poziom wszystkich publikacji jest zatem, oczywiście, bez zarzutu. **Łączna wartość współczynnika oddziaływania tych publikacji, *Impact Factor,* wynosi 55,7. Warto dodatkowo odnotować, że praca P2, komunikat w *Angew. Chem. Int. Ed.,* została opublikowana jako *VIP paper*.**

Doktorantka jest współautorką także 3 innych prac oryginalnych opublikowanych w *Chem., Sci. Rep.* oraz *Arkivoc*, a także artykułu w *Wiadomościach Chemicznych*. Łączna wartość współczynnika IF wszystkich publikacji przekracza 75. Niezwykle rzadko spotyka się tak imponujący dorobek publikacyjny na etapie zamykania doktoratu!

Znaczenie prac stanowiących podstawę doktoratu mgr inż. Sary Szymkuć odzwierciedla ich oddźwięk literaturowy. Praca przeglądowa w *Angew. Chem. Int. Ed.* (P5; opublikowana w 2016, a więc zaledwie dwa lata temu) cytowana była już 62 razy; publikacja w *Chem* (P1; opublikowana w tym roku) – 15 razy, a praca w *Scientific Reports* (P4; opublikowana w roku ubiegłym) – 10 razy. Łączna liczba cytowań Doktorantki wynosi już 90 i można spodziewać się, że będzie bardzo szybko rosnąć.

Publikacje stanowiące podstawę rozprawy ułożone są w odwróconym porządku chronologicznym. Jednak taki układ nie stanowi utrudnienia dla czytelnika. **Pierwsza praca P1** (*"Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory"* ***Chem*****, 2018**, 4, 522-532) **jest ukoronowaniem projektu – przedstawia zaplanowane w oparciu o oprogramowanie *Chematica* ścieżki syntezy ośmiu związków** – siedmiu komercyjnie dostępnych, cennych substancji bioaktywnych oraz jednego związku naturalnego, **a także potwierdzenie doświadczalne przewidywań**, poprzez syntezy przeprowadzone w IChO PAN, Sigma-Aldrich oraz Northwestern University. Obszerne materiały uzupełniające (Supplemental Information) załączone do pracy pozwalają czytelnikowi m.in. dość szczegółowo (sekcje S1-S9, 46 stron) zapoznać się z podejściem zastosowanym w platformie *Chematica*, jej kluczowymi komponentami i niektórymi algorytmami.

Jest rzeczą oczywistą, że praca P1 jest publikacją wieloautorską – lista autorów obejmuje 20 osób. Mgr inż. Sara Szumkuć jest co prawda dopiero na pozycji 5 tej listy, ale 4

pierwsze osoby odpowiedzialne były za przeprowadzenie syntez otrzymanych związków, a zatem Doktorantka jest pierwszą, spośród osób rozwijających metodologię stanowiącą podstawę platformy *Chematica*. Z oświadczenia złożonego przez mgr. Tomasza Klucznika – pierwszego autora - wynika, że oprócz syntez, uczestniczył on także w rozwoju bazy danych, ale treść oświadczenia nie budzi wątpliwości, że wkład ten był mniej istotny od wkładu Doktorantki (cyt. : „*I also helped with the development of Chematica's knowledge base and input ca. 3 000 reactions according to specifications and computer routines developed by Ms Szymkuć.*"). Mgr inż. Sara Szumkuć określa swój wkład w publikację P1 w sposób następujący: „*I contributed to the development of methods described in the paper, especially in formalizing the higher order chemical logic and inputting new reactions (ca 7 000 reactions since 2016 and 15 000 in total). I queried Chematica to produce pathways that were subsequently validated experimentally. I also participated in the writing of the manuscript.*" Oświadczenia pozostałych współautorów, którzy uczestniczyli w rozwoju metodologii/oprogramowania nie ułatwiają oceny wkładu Doktorantki, gdyż są często dość nieprecyzyjne: „*I contributed to the development of methods described in the paper"* (mgr inż. Ewa Gajewska, mgr inż. Karol Molga). Bardziej precyzyjnie określają swój wkład dr Piotr Dittwald („*I was one of the key developers of the search algorithms in Chematica."*), dr Rafał Roszak ("*I was working on the Chematica's filter for electrophilic aromatic substitution reactions."*) oraz dr Michał Startek (*"I developed Synturus' search algorithm.*")

Nieco lepszy obraz wkładu mgr inż. Sary Szumkuć w rozwój metodologii daje lektura komentarza Doktorantki poprzedzającego treść publikacji (a także oświadczeń dotyczących pozostałych publikacji), z której wynika, że: **(i) miała ona udział w wyborze sposobu reprezentacji alfanumerycznej cząsteczek oraz reakcji chemicznych, w szczególności w aspekcie regio- i stereochemii; (ii) była główna autorką sposobu i procedur kodowania reakcji oraz drzew decyzyjnych; (iii) miała znaczący wpływ na ostateczną postać struktury bazy danych; (iv) wniosła znaczący wkład w pracach dotyczących pełnej automatyzacji procesu planowania syntezy; (v) była główną autorką zastosowanych funkcji oceny i tworzących ją zmiennych; (vi) wprowadziła osobiście do bazy danych ok. 8 000 pozycji (spośród 18 000). Wymienione elementy stanowią znaczący wkład w rozwój metodologii i programu *Chematica* i pozwalają uznać, że mgr inż. Sarę Szymkuć jest jednym z głównych autorów zastosowanej metodologii.**

Szczegółowy opis zastosowanej metodologii jest treścią pracy przeglądowej P5 ("*Computer-assisted synthetic planning: The end of the beginning*", **Angew. Chem. Int. Ed. 2016**, 55, 5904-5937.). **W przypadku tej pracy mgr inż. Sara Szymkuć jest pierwszą na liście osób zawierającej 8 nazwisk. Oświadczenie Doktorantki potwierdza jej istotny wkład** (cyt. „*I designed Chematica's knowledge base and input ca 8 000 reactions (out of 15 000 up till 2018). I proposed dual scoring functions and their varińles. I contributed to the development of other modules and methods described in the paper. I also participated in the*

writing of the manuscript."). Nie będę w tym miejscu przytaczał oświadczeń pozostałych Autorów, gdyż wynika z nich podobny obraz jak w przypadku pracy P1.

Skrócony (jednostronicowy) opis metodologii i programu *Chematica* jest treścią artykułu konferencyjnego P3 ("*Synthetic design with the Chematica program – the importance of accurate rules and of higher-order logic*" **CHIMIA 2017**, 71, 512.). Doktorantka jest na drugiej pozycji na liście 5 autorów; pierwszym jest Promotor. Swój wkład doktorantka określa jako udział w rozwoju metodologii.

Publikacje P6 ("*A Priori Estimation of Organic Reaction Yields*", **Angew. Chem. Int. Ed. 2015**, 54, 10797-10801.) oraz P4 ("*Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?* " **Sci. Rep. 2017**, 7, 3582.) dotyczą dwóch różnych podejść mających na celu oszacowanie wydajności reakcji, w pierwszym przypadku (P6) w oparciu o model termodynamiczny, opracowany głównie przez teoretyków z Northwestern University, a w drugim – o proces uczenia maszynowego. Swój udział w tych pracach Doktorantka określa w sposób następujący: " *I performed statistical analysis of organic reactions under thermodynamic vs. kinetic control.*" (P6) oraz "*I provided chemical examples and validated chemical conclusions. I also participated in the writing of the manuscript"*

Ostatnia, nie wymieniona jeszcze przeze mnie publikacja, P2 ("*Discovery and Enumeration of Organic-Chemical and Biomimetic Reaction Cycles within the Network of Chemistry*" **Angew. Chem. Int. Ed. 2018**, 57, 2367- 2371") dotyczy identyfikacji istniejących cykli reakcyjnych; **ze względu na potencjalne wpływ na dalszy rozwój chemii organicznej, praca ta została opublikowana ze statusem *VIP paper*.** Swój udział w tej pracy mgr inż. Sara Szymkuć określa następująco: „*I co-performed analyses of chemical cycles within Cyclorg network. I also helped with preparing the dataset of reactions. I also participated in the writing of the manuscript."*

Przejdę teraz do krótkiego omówienia tekstu **komentarza przedłożonego przez Doktorantkę. Obejmuje on 21 stron tekstu, napisanego w języku angielskim oraz 39 cytowanych pozycji literaturowych. Całość jest napisana poprawnym tekstem, w sposób dojrzały, ciekawy, zwarty i przejrzysty; przedstawiony tekst czyta się z przyjemnością.** Autorka formułuje hipotezę badawczą, następnie na 6 stronach prezentuje wstęp historyczny. Pozostała część stanowi bardzo zwarty opis metodologii, ze szczególnym zwróceniem uwagi na te elementy, w których rozwoju Autorka uczestniczyła. Jak wspomniałem wcześniej, Doktorantka przybliża także na czym polegał jej wkład na poszczególnych etapach. W tym miejscu mam dwie uwagi krytyczne.

(1) Uważam, ze bardzo ogólny sposób sformułowania hipotezy badawczej (cyt. „*The main hypothesis of my doctoral work has been that by combining proper representation of organic-chemical knowledge with the power and methods of modern computing, it would finally be possible – after over five decades of effort by various groups – to design a software system*

capable of autonomous planning of synthetic routes leading to arbitrary target molecules.") może być nieco krzywdzący dla innych współautorów badań leżących u podstaw platformy *Chematica,* bo przecież tak sformułowana hipoteza badawcza odnosi się do całości projektu, w którym uczestniczyło wiele osób i ich wkładu w wielu aspektach, który wydaje się być także kluczowy.

(2) W publikacjach stanowiących podstawę doktoratu okres rozwoju platformy *Chematica* określany jest jako *dekada*. Tymczasem z tekstu komentarza wynika, że Doktorantka włączyła się w prace w r. 2013, czyli ok. pięć lat temu. W tekście komentarza brakuje mi wyraźnego określenia, jaki był stan badań prowadzonych w tym kierunku w zespole kierowanym przez prof. Grzybowskiego, w momencie, w którym Doktorantka dołączyła do zespołu, a zatem co zostało zrobione już wcześniej, bez jej udziału. Przedstawione wprowadzenie historyczne dotyczy badań prowadzonych w innych zespołach, a wcześniejsze prace prof. Grzybowskiego cytowane są na samym końcu komentarza (Ref. 34-39), jedynie w kontekście pracy P2. Chciałbym podkreślić w tym miejscu, że nie jest moim celem podważanie wkładu mgr inż. Sary Szymkuć, który – jak podkreśliłem wcześniej – uważam za znaczący. Chciałbym jednak mieć pełny, jednoznaczny i prawidłowy obraz. Liczę na to, ze to, że na obronie pracy doktorskiej uzyskam pełną odpowiedź.

Podsumowując, **moja ocena badań naukowych przeprowadzonych w ramach przewodu doktorskiego mgr inż. Sary Szymkuć, przedstawionych w sześciu publikacjach stanowiących podstawę rozprawy doktorskiej jest zdecydowanie pozytywna**. **Doktorantka podjęła bardzo aktualną tematykę badawczą, wniosła znaczący wkład w wybitne osiągnięcie naukowe Zespołu kierowanego przez prof. dr Bartosza Grzybowskiego, które - jak wspomniałem na wstępie – uważam za jedno z najbardziej spektakularnych i doniosłych osiągnięć współczesnej informatyki chemicznej / chemii obliczeniowej, o dużym znaczeniu dla nauki, w szczególności dla chemii organicznej i informatyki chemicznej.**

Przedstawiony w publikacjach materiał badawczy spełnia z nawiązką przyjęte wymagania stawiane zwyczajowo pracom doktorskim, jak i wymagania ustawowe określone w art. 13 ustawy z dnia 14 marca 2003 r o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki. Dlatego, wnioskuję o dopuszczenie pani mgr inż. Sary Szymkuć do dalszych etapów przewodu doktorskiego.

Jednocześnie, biorąc pod uwagę wysoki poziom doktoratu, duże znaczenie prezentowanego osiągnięcia naukowego oraz imponujący dorobek publikacyjny mgr inż. Sary Szymkuć, wnioskuję o wyróżnienie jej pracy.

A Michalak

Prof. dr hab. Artur Michalak