Uniwersytet Gdański | Związek Uczelni Fahrenheita

Załącznik 5

AUTOREFERAT

Nowe podejścia chemoinformatyczne do komputerowej oceny zagrożenia chemicznego stwarzanego przez mało liczne lub silnie zróżnicowane strukturalnie zbiory związków chemicznych

dr Agnieszka Gajewicz-Skrętna

Gdańsk 2022

Spis treści

I. IMIĘ I NAZWISKO

Agnieszka Gajewicz-Skrętna (do czerwca 2019 roku Agnieszka Gajewicz)

II. POSIADANE DYPLOMY, STOPNIE NAUKOWE – Z PODANIEM PODMIOTU NADAJĄCEGO STOPIEŃ, ROKU ICH UZYSKANIA ORAZ TYTUŁU ROZPRAWY DOKTORSKIEJ

Stopień/tytuł zawodowy	Uczelnia, Wydział	Data uzyskania
doktor nauk chemicznych	Uniwersytet Gdański, Wydział Chemii	12.06.2013 r.
Temat pracy doktorskiej: „Opracowanie metod *in silico* służących przewidywaniu cytotoksycznego wpływu nanocząstek tlenków nieorganicznych na komórki bakterii *E. coli* oraz ludzkie keratynocyty (HaCaT)”		
Promotor: dr hab. Tomasz Puzyn		
Recenzenci: Prof. dr hab. Roman Kaliszan (Gdański Uniwersytet Medyczny), Prof. dr hab. Piotr Stepnowski (Uniwersytet Gdański)		
magister	Uniwersytet Gdański, Wydział Chemii	30.04.2004 r.
Temat pracy magisterskiej: „Projektowanie szczepionek skojarzonych”		
Promotor**:** Prof. dr hab. Zbigniew Maćkiewicz		
Recenzent: Prof. dr hab. Piotr Rekowski (Uniwersytet Gdański)		

III. INFORMACJA O DOTYCHCZASOWYM ZATRUDNIENIU W JEDNOSTKACH NAUKOWYCH

Okres zatrudnienia	Stanowisko	Miejsce zatrudnienia
01.04.2014 – obecnie	Pracownik naukowo-dydaktyczny (adiunkt)	Uniwersytet Gdański, Wydział Chemii, Katedra Chemii i Radiochemii Środowiska, Pracownia Chemoinformatyki Środowiska
01.10.2016 – 30.09.2017	Pracownik naukowy w ramach rocznego stażu podoktorskiego	National Institute for Environmental Studies (NIES), Center for Health and Environmental Risk Research, Tsukuba, Japonia
15.10.2013 – 31.03.2014	Asystent	Uniwersytet Gdański, Wydział Chemii, Katedra Chemii i Radiochemii Środowiska, Pracownia Chemometrii Środowiska
15.12.2011 – 14.10.2013	Starszy referent techniczny	Uniwersytet Gdański, Wydział Chemii, Instytut Ochrony Środowiska i Zdrowia Człowieka

IV. OMÓWIENIE OSIĄGNIĘĆ, O KTÓRYCH MOWA W ART. 219 UST. 1 PKT. 2 USTAWY Z DNIA 20 LIPCA 2018 R. PRAWO O SZKOLNICTWIE WYŻSZYM I NAUCE (DZ. U. Z 2021 R. POZ. 478 Z PÓŹN. ZM.).

Osiągnięcie naukowe stanowi cykl 9 powiązanych tematycznie publikacji, opublikowanych w latach 2017-2022, traktujących o rozwoju metod uczenia maszynowego i narzędzi wspierających proces komputerowej oceny ryzyka chemicznego dla mało licznych lub silnie zróżnicowanych pod względem struktury chemicznej, liczebności i reprezentatywności grup związków chemicznych. Publikacje wchodzące w skład osiągnięcia naukowego zostały uszeregowane w punkcie 2. zgodnie z rokiem publikacji, natomiast w punkcie 3. opisane tematycznie.

1. TYTUŁ OSIĄGNIĘCIA NAUKOWEGO

Nowe podejścia chemoinformatyczne do komputerowej oceny zagrożenia chemicznego stwarzanego przez mało liczne lub silnie zróżnicowane strukturalnie zbiory związków chemicznych

2. CYKL PUBLIKACJI WCHODZĄCYCH W SKŁAD OSIĄGNIĘCIA NAUKOWEGO

H1. **Gajewicz A.**, Jagiello K., Cronin M., Leszczynski J., Puzyn T.
Addressing a bottle-neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available.
Environmental Science: Nano (2017): 4, 346-358, DOI:10.1039/C6EN00399K.
[IF_{2017}=6,645; IF_{2021}=9,473; $IF_{5\text{-letni}}$=9,350; $MNiSW_{2020}$=140; Licz. cyt=40]

H2. **Gajewicz A.**✉
Development of valuable predictive read-across models based on “real-life” (sparse) nanotoxicity data.
Environmental Science: Nano (2017): 4, 1389-1403, DOI:10.1039/C7EN00102A.
[IF_{2017}=6,645; IF_{2021}=9,473; $IF_{5\text{-letni}}$=9,350; $MNiSW_{2020}$=140; Licz. cyt=17]

H3. **Gajewicz A.**✉
What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps.
Nanoscale (2017): 9, 8435-8448, DOI:10.1039/C7NR02211E.
[IF_{2017}=7,233; IF_{2021}=8,307; $IF_{5\text{-letni}}$=7,891; $MNiSW_{2020}$=140; Licz. cyt=39]

H4. **Gajewicz A.**✉
How to judge whether QSAR/read-across predictions can be trusted? Novel approach for establishing model’s applicability domain.
Environmental Science: Nano (2018): 5, 408-421, DOI:10.1039/C7EN00774D.
[IF_{2018}=6,645; IF_{2021}=9,473; $IF_{5\text{-letni}}$=9,350; $MNiSW_{2020}$=140; Licz. cyt=32]

H5. **Gajewicz A.**, Puzyn T., Odziomek K., Urbaszek P., Haase A., Riebeling C., Luch A., Irfan M.A., Landsiedel R., van der Zande M., Bouwmeester H.
Decision tree models to classify nanomaterials according to the DF4nanoGrouping scheme.
Nanotoxicology (2018): 12, 1-17, DOI:10.1080/17435390.2017.
[IF_{2018}=5,955; IF_{2021}=5,881; $IF_{5\text{-letni}}$=6,319; $MNiSW_{2020}$=140; Licz. cyt=39]

H6. **Gajewicz-Skretna A.**✉, Gromelski M., Wyrzykowska E., Furuhama A., Yamamoto H., Suzuki N.
Aquatic toxicity (Pre)screening strategy for structurally diverse chemicals: global or local classification tree models?
Ecotoxicology and Environmental Safety (2021): 208, 111738, DOI:10.1016/j.ecoenv.2020.111738.
[IF_{2021}=7,129; $IF_{5\text{-letni}}$=7,284; $MNiSW_{2020}$=100; Licz. cyt=3]

H7. **Gajewicz-Skretna A.**✉, Kar S., Piotrowska M., Leszczynski J.
The kernel-weighted local polynomial regression (KwLPR) approach – an efficient, novel tool for development of QSAR/QSAAR toxicity extrapolation models.
Journal of Cheminformatics (2021): 13, 9, DOI:10.1186/s13321-021-00484-5
[IF_{2021}=8,489; $IF_{5\text{-letni}}$=9,187; $MNiSW_{2020}$=100; Licz. cyt=2]

H8. **Gajewicz-Skretna A.**✉, Furuhama A., Yamamoto S., Suzuki N. Generating accurate in silico predictions of acute aquatic toxicity for a range of organic chemicals: Towards similarity-based machine learning methods. *Chemosphere* (2021): 280, 130681, DOI:10.1016/j.chemosphere.2021.

[IF_{2021}=8.943; $IF_{5\text{-letni}}$=8.520; $MNiSW_{2020}$=140; Licz. cyt=6]

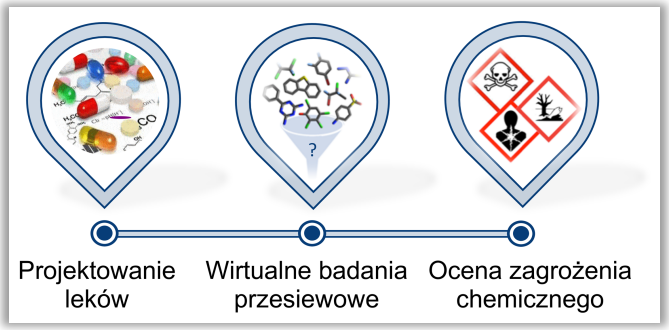
H9. **Gajewicz-Skretna A.**✉, Wyrzykowska E., Gromelski M. Quantitative multi-species toxicity modeling: Does a multi-species, machine learning model provide better performance than a single-species model for the evaluation of acute aquatic toxicity by organic pollutants? *Science of The Total Environment* (2022): https://doi.org/10.1016/j.scitotenv.2022.160590.

[IF_{2021}=10.753; $IF_{5\text{-letni}}$=10.237; $MNiSW_{2020}$=200; Licz. cyt=0]

3. OMÓWIENIE CELU NAUKOWEGO PRAC WCHODZĄCYCH W SKŁAD OSIĄGNIĘCIA NAUKOWEGO I OSIĄGNIĘTYCH WYNIKÓW WRAZ Z OMÓWIENIEM ICH EWENTUALNEGO WYKORZYSTANIA

3.1. Wprowadzenie

Chemoinformatyka jest stosunkowo nową gałęzią nauki, która wykorzystuje metody uczenia maszynowego i sztucznej inteligencji do rozwiązywania problemów chemicznych. Początki współczesnych metod chemoinformatycznych sięgają wczesnych lat 60. ubiegłego stulecia, kiedy to Corwin Hansch i Toshio Fujita opublikowali cykl prac naukowych dowodzących ilościowej zależności pomiędzy strukturą chemiczną związków, reprezentowaną przez parametry steryczne, elektronowe i liofilowe, a aktywnością biologiczną tych związków.[1–3] Współcześnie schemat modelowania oparty na założeniu, że aktywność biologiczna jest matematyczną funkcją struktury chemicznej jest nieodłącznym elementem procesu projektowania leków, wirtualnych badań przesiewowych umożliwiających poszukiwanie struktury o pożądanych właściwościach, czy komputerowej oceny zagrożenia chemicznego (Rysunek 1). Według najnowszych analiz, wartość rynku narzędzi chemoinformatycznych w 2019 roku szacowana była na 1,3881 mld dolarów amerykańskich i przy prognozowanym skumulowanym rocznym 4% wskaźniku wzrostu do 2027 roku wartość ta osiągnie poziom 1,8885 mld US$.[4]

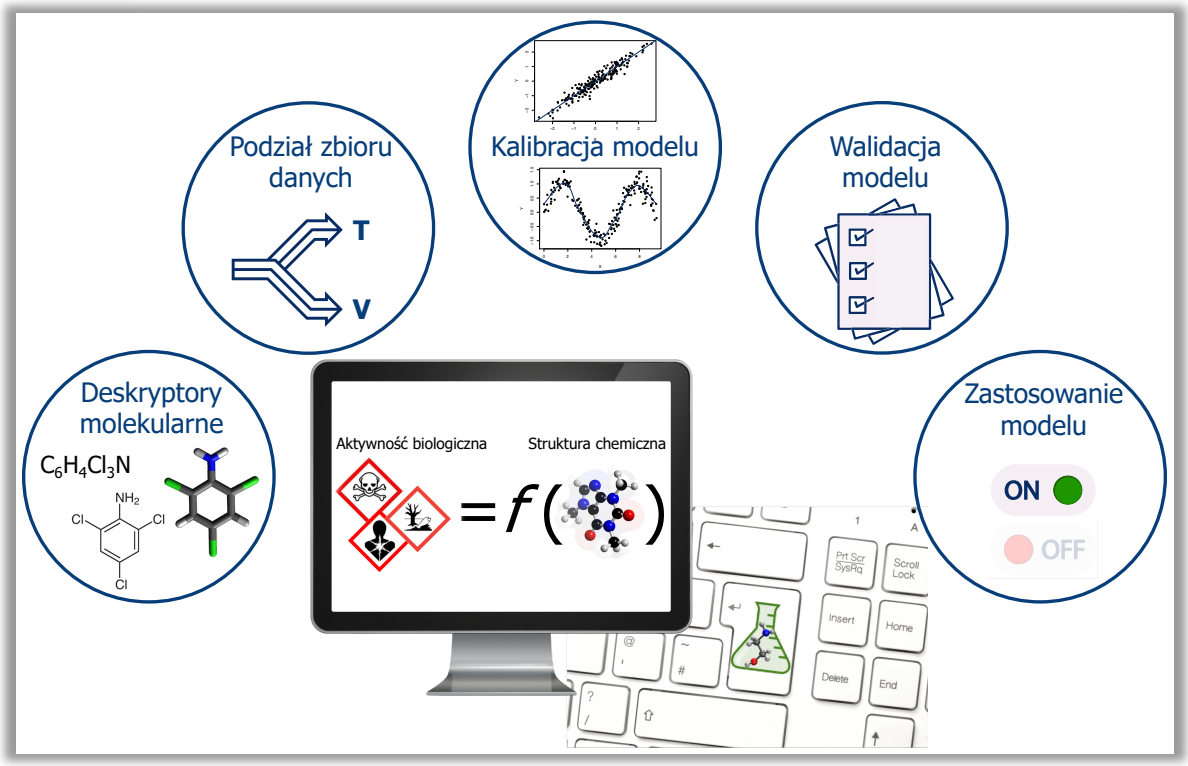


Rysunek 1. Główne obszary zastosowania metod in silico.

W miarę udoskonalania metod i narzędzi chemoinformatycznych, zyskały one również akceptację organów regulacyjnych, czego wyrazem są zapisy w dokumentach legislacyjnych traktujących o bezpieczeństwie chemicznym. Zapisy rekomendujące stosowanie metod komputerowych jako alternatywnych rozwiązań wobec prowadzenia badań na zwierzętach można znaleźć m. in. w europejskim rozporządzeniu REACH[5], amerykańskiej ustawie o kontroli substancji toksycznych[6] czy japońskiej ustawie o kontroli substancji chemicznych.[7] Z opublikowanego w 2020 roku sprawozdania Europejskiej Agencji Chemikaliów wynika, że w przypadku około 70% z 12 tys. substancji podlegających obowiązkowi rejestracji w okresie sprawozdawczym, podmioty rejestrujące zastosowały co najmniej jedno rozwiązania alternatywne, pozwalające uniknąć badań na zwierzętach.[8] Najczęściej wykorzystywanym rozwiązaniem alternatywnym były metody komputerowe, w

tym podejście przekrojowe (ang. *read-across*) oraz modele ilościowej zależności pomiędzy strukturą chemiczną a aktywnością biologiczną (ang. *quantitative structure-activity relationship*, QSAR).

Ogólna zasada działania metod chemoinformatycznych opiera się na powiązaniu za pomocą odpowiedniego modelu matematycznego aktywności biologicznej (**y**) strukturalnie podobnych związków ze zmiennymi kodującymi informację na temat ich budowy chemicznej i właściwości fizykochemicznych, czyli. tzw. deskryptorami molekularnymi (**X**). Model liniowej lub nieliniowej zależności opracowywany jest w oparciu o związki zbioru uczącego (ang. *training set*, T), natomiast jego zdolności do prawidłowego przewidywania oceniane są w oparciu o związki zbioru testowego (ang. *validation set*, V), dla których znana jest eksperymentalna wartość modelowanej odpowiedzi, ale które nie były wykorzystane na żadnym etapie opracowywania tego modelu (Rysunek 2).



Rysunek 2. Idea metod chemoinformatycznych.

Poprawnie opracowany i oceniony model chemoinformatyczny pozwala: (1) zidentyfikować cechy strukturalne analizowanych związków chemicznych determinujące ich właściwości biologiczne i/lub fizykochemiczne; (2) poznać oraz zrozumieć molekularne mechanizmy odpowiedzialne za (nie)pożądane efekty ich działania, jak również (3) wyznaczyć teoretyczne wartości modelowanej aktywności biologicznej lub właściwości fizykochemicznych dla związków, dla których takich danych brakuje (np. nowo projektowanych związków chemicznych).

3.2. Cele i zakres badań

Pomimo iż metody chemoinformatyczne od lat znajdują szerokie zastosowanie praktyczne, to wciąż jednak istnieją pewne ograniczenia metodologiczne. Wśród najważniejszych wyzwań związanych z szerokim wykorzystaniem metod komputerowych w procesie oceny zagrożenia chemicznego i projektowania nowych substancji chemicznych o użytecznych, w ujęciu aplikacyjnym, właściwościach fizycznych, chemicznych i biologicznych wymienić należy te związane z liczebnością i reprezentatywnością zbioru danych wejściowych, a także złożonością relacji struktura - właściwość/aktywność.[9–12] Metody chemoinformatyczne jako metody probabilistyczno-statystyczne, wymagają statystycznie reprezentatywnego zbioru danych o dostatecznie dużej liczebności (N>30) strukturalnie podobnych związków chemicznych o znanej właściwości/aktywności biologicznej. W przypadku niektórych klas związków chemicznych zebranie tak licznej próby jest bardzo trudne, a czasem wręcz niemożliwe. Zbyt mała liczebność próby implikuje także trudności w przeprowadzeniu rzetelnej oceny wiarygodności prognoz uzyskanych w oparciu o modele uczenia maszynowego.

Zdefiniowane powyżej ograniczenia pozwoliły mi wyznaczyć sześć kluczowych obszarów tematycznych stanowiących odzwierciedlenie celów i zakresu badań przeprowadzonych w ramach niniejszej rozprawy habilitacyjnej. Były to:

(i) adaptacja metod interpolacji liniowej, wykorzystujących pojedynczą lub większą liczbę zmiennych objaśniających, do modelowania ilościowej zależności struktura - aktywność mało licznych zbiorów związków chemicznych;

(ii) opracowanie metodyki wykorzystującej estymatory najbliższego sąsiedztwa do ilościowego przewidywania wybranych właściwości biologicznych dla potrzeb oceny zagrożenia chemicznego;

(iii) adaptacja metod bazujących na estymatorach najbliższego sąsiedztwa do międzygatunkowej ekstrapolacji toksyczności;

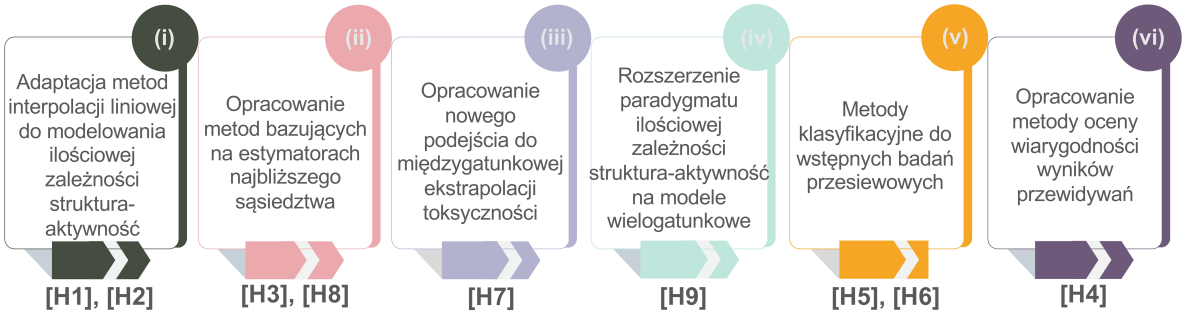
(iv) rozszerzenie paradygmatu ilościowej zależności struktura-aktywność na modele wielogatunkowe;

(v) badania dotyczące wykorzystania metod klasyfikacyjnych do wstępnych, środowiskowych badań przesiewowych mało licznych lub silnie zróżnicowanych pod względem struktury, liczebności i reprezentatywności związków chemicznych, umożliwiających zidentyfikowanie potencjalnie niebezpiecznych substancji wymagających dalszych badań toksykologicznych;

(vi) opracowanie metodyki oceny wiarygodności przewidywań modeli uczenia maszynowego uzyskanych na podstawie mało licznych zbiorów danych wejściowych.

Na szczególne podkreślenie zasługuje fakt, że wyodrębnione obszary tematyczne nie tylko nie są rozłączne, ale wzajemnie się przenikają, a nawet uzupełniają, dzięki czemu oferują możliwość opracowania kompleksowego zestawu narzędzi komputerowych wspierających proces oceny zagrożenia chemicznego oraz zrównoważonego projektowania nowych substancji chemicznych w oparciu o mało liczne lub silnie zróżnicowane strukturalnie zbiory związków chemicznych.

Uzyskane wyniki, które zostały przedstawione w postaci cyklu 9 powiązanych tematycznie artykułów naukowych opublikowanych w latach 2017-2022 są efektem realizacji badań przeprowadzonych podczas rocznego stażu podoktorskiego w *National Institute for Environmental Studies* (Japonia) oraz dwóch międzynarodowych projektów badawczych. Schemat organizacyjny prac stanowiących osiągnięcie naukowe przedstawia Rysunek 3.



Rysunek 3. Schemat prac stanowiących osiągnięcie naukowe.

3.3 Omówienie najważniejszych osiągnięć zawartych w pracach przedstawionych do habilitacji

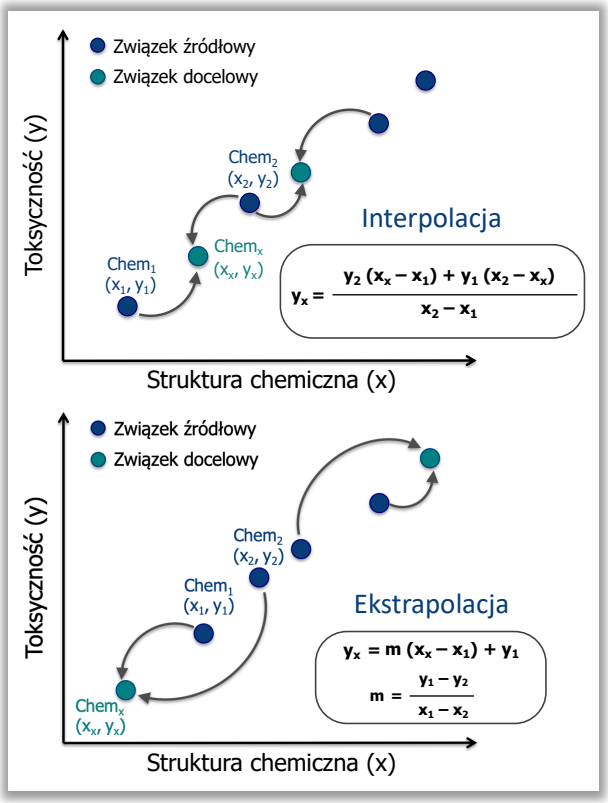
3.3.1 Weryfikacja użyteczności metody interpolacji liniowej do modelowania zależności struktura – aktywność dla mało licznych zbiorów nanocząstek [H1-H2]

Większość opracowanych na przestrzeni ostatnich lat modeli uczenia maszynowego do przewidywania biologicznych, fizykochemicznych lub środowiskowych właściwości różnych klas związków chemicznych bazuje na metodzie regresji liniowej.[13–16] Wyznaczenie równania modelu regresji liniowej w klasycznej metodzie najmniejszych kwadratów sprowadza się do wyznaczenia wektora współczynników regresji w taki sposób, aby zminimalizować w całym analizowanym zbiorze uczącym sumę kwadratów reszt (błędu) modelu. Co ważne, na podstawie centralnego twierdzenia granicznego można przyjąć, że dla odpowiednio dużego zbioru danych (N>30) model regresji liniowej powinien poprawnie przybliżać wartości odpowiedzi także w przypadku danych pochodzących z rozkładu innego niż normalny. Pamiętać jednak należy, że metoda regresji liniowej jest wrażliwa na występowanie obserwacji odstających (ang. *outliers*), które mają duży wpływ na

współczynnik kierunkowy linii regresji. Wpływ wartości odstających jest dodatkowo potęgowany w przypadku mało licznych zbiorów danych. W świetle powyższego, pewne wątpliwości budzić może coraz powszechniejsze wykorzystywanie metod regresji liniowej do opracowywania modeli predykcyjnych w oparciu o mało liczne zbiory danych (N<20). Problem ten dotyczy m. in., nowych grup związków chemicznych, np. nanocząstek, mikro- i nanoplastików, w przypadku których liczebność dostępnych danych empirycznych jest bardzo ograniczona.[17, 18] W pracy **[H1]** postawiłam pytanie: Czy w przypadku mało licznego zbioru danych, w którym występuje liniowa zależność pomiędzy zmienną zależną a zmienną niezależną zasadne jest zastąpienie klasycznej regresji liniowej metodą interpolacji liniowej do prognozowania wartości odpowiedzi? Wybór interpolacji liniowej jako techniki modelowania podyktowany był tym, że jako szczególny przypadek regresji liniowej metoda ta aproksymuje punktowo wyłącznie w bezpośrednim sąsiedztwie badanego punktu, tzw. związku docelowego (ang. *target compound*), wykorzystując wartości leżące bezpośrednio poniżej i powyżej interpolowanej wartości. Dzięki temu, wartość odpowiedzi każdego związku docelowego jest przybliżana na podstawie związków o podobnej budowie strukturalnej/właściwościach fizykochemicznych, a tym samym zakładanej *a priori* podobnej aktywności biologicznej.

W celu zweryfikowania powyższej hipotezy porównałam przybliżone wartości odpowiedzi definiujące cytotoksyczny wpływ nanocząstek tlenków metali (MeOx NPs) na komórki pałeczki okrężnicy (*Escherichia coli*) uzyskane z:

1) modelu, w którym linia regresji dopasowana została do całego 10-elementowego zbioru danych uczących (model Nano-QSAR); oraz
2) modelu, który interpoluje/ ekstrapoluje liniowo wartość odpowiedzi z dwóch sąsiadujących związków źródłowych (ang. *source compounds*) (Rysunek 4).



Rysunek 4. Idea interpolacji i ekstrapolacji liniowej.

Modelowanie przeprowadziłam w oparciu o identyczny podział związków na zbiór uczący/związki źródłowe i zbiór testowy/związki docelowe, wykorzystując jako pojedynczą zmienną niezależną - entalpię tworzenia kationu metalu w fazie gazowej, która opisuje

łatwość uwalniania kationów metali z powierzchni nanocząstki. Analiza porównawcza uzyskanych wyników wykazała, że zdolności przewidywania obu modeli były takie same (współczynnik walidacji zewnętrznej, Q^2_{Ext}=0,80 i średniokwadratowy błąd przewidywania, $RMSE_P$=0,19), jednak niższe wartości reszt modelu interpolacji liniowej dowiodły lepszego dopasowania tego modelu do danych empirycznych (współczynnik determinacji R^2=0,94 i średniokwadratowy błąd kalibracji $RMSE_C$=0,13) w porównaniu z regresyjnym modelem Nano-QSAR (R^2=0,85 i $RMSE_C$=0,20). Ponadto, wyniki porównania z literaturowymi modelami regresyjnymi, których autorzy zwiększali liczebność zbioru uczącego kosztem zbioru testowego i/lub zwiększali kompleksowość modelu poprzez uwzględnienie w równaniu modelu dodatkowych zmiennych niezależnych okazały się jeszcze bardziej optymistyczne dowodząc użyteczności interpolacji liniowej w przypadku mało licznego zbioru danych (H1: Tabela 5).

W praktyce, ze względu na złożoność modelowanych właściwości biologicznych lub fizykochemicznych, często do ich kompleksowego opisu niezbędne jest użycie dwóch lub większej liczby zmiennych niezależnych reprezentujących różne aspekty budowy przestrzennej, właściwości elektronowych, etc. Stanowi to istotne ograniczenie wykorzystania metody interpolacji liniowej do prognozowania wartość oczekiwanej zmiennej zależnej, która umożliwia wykorzystanie tylko jednego deskryptora. Dlatego kolejnymi etapami moich badań było rozszerzenie zastosowania metody interpolacji liniowej na dwie i więcej zmiennych niezależnych.

W tym celu w pracy **[H1]** zaproponowałam przeprowadzenie punktowej interpolacji wartości zmiennej zależnej z położenia na płaszczyźnie przechodzącej przez trzy sąsiadujące ze sobą punkty z wykorzystaniem tzw. metody Sarrusa do obliczenia wyznacznika macierzy trzeciego stopnia (H1: Rysunek 2, Równanie 4). Weryfikację użyteczności tej metody przeprowadziłam opracowując model szacowania przekrojowego (*read-across*) do przewidywania cytotoksycznego wpływu nanocząstek tlenków metali na komórki ludzkich keratynocytów (HaCaT). Odpowiedź modelu została wyrażona jako liniowa kombinacja dwóch deskryptorów kwantowo-mechanicznych obliczonych na poziomie półempirycznej metody PM6: entalpii tworzenia nanoklastra MeOx, reprezentującego fragment powierzchni nanocząstki oraz elektroujemności Mullikena. Ocenę poprawności dopasowania i zdolności predykcyjnych modelu opracowanego w oparciu o zaproponowane podejście przeprowadziłam na podstawie statystyk opisujących zdolności przewidywania dla związków docelowych (zbioru testowego) dodatkowo zestawiając je z wynikami pięciu typów aproksymacji powszechnie stosowanymi w ilościowym podejściu przekrojowym i literalnie wymienionymi w poradniku OECD dotyczącym grupowania chemikaliów.[19] Różnice w wartościach statystyk w zależności od zastosowanej metody aproksymacji (H1: Tabela 6)

dowiodły, że interpolacja z równania płaszczyzny w przestrzeni prowadzi do uzyskanie znacznie bardziej dokładnych prognoz w porównaniu z modelami przybliżenia bazującymi na wartości (i) średniej arytmetycznej, (ii) zachowawczej (najgorszy możliwy scenariusz), (iii) mediany oraz (iv-v) dolnego i górnego kwartyla spośród związków źródłowych (zbioru uczącego).

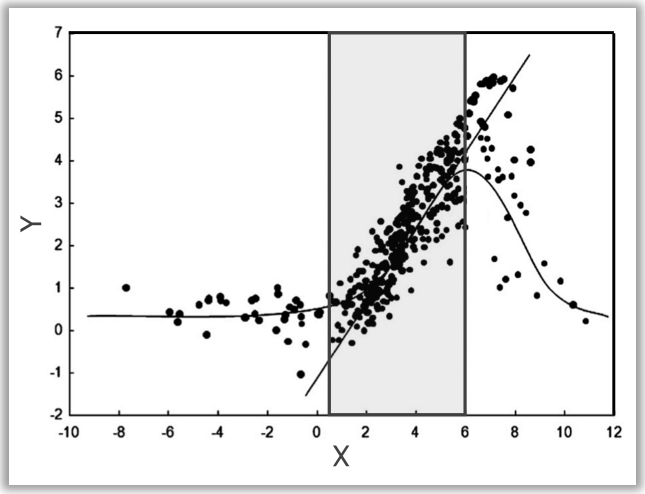
Z kolei, w pracy **[H2]** zaproponowałam modyfikację metody interpolacji liniowej, zastępując pojedynczą zmienną niezależną pierwszą główną składową (PC_1) wyznaczoną w analizie głównych składowych (ang. *principal component analysis*, PCA). Istotą modyfikacji jest zatem wykorzystanie w metodzie interpolacji liniowej głównej składowej, stanowiącej liniową kombinację dowolnej liczby zmiennych pierwotnych. Ocenę użyteczności zaproponowanej modyfikacji przeprowadziłam w oparciu o trzy mało liczne zbiory danych, składające się z 16, 17 i 18 nanocząstek tlenków metali, dla których dostępne były wartości toksyczności wyznaczone eksperymentalnie z użyciem bakterii *E. coli* (w różnych warunkach eksperymentu) oraz w teście z użyciem linii komórkowej HaCaT. Pomimo istotnego zmniejszenia liczby związków źródłowych (zbioru uczącego) i zwiększenia liczby związków docelowych (zbioru testowego) w porównaniu z oryginalnymi modelami Nano-QSAR, wartości statystyk modeli interpolacji liniowej z pierwszą główną składową jako zmienną niezależną były bardzo zbliżone do statystyk klasycznych modeli regresyjnych z dwoma zmiennymi niezależnymi (H2: Tabela 5). Uzyskane wyniki wydają się szczególnie imponujące w odniesieniu do trzeciego studium przypadku traktującego o toksyczności nanocząstek tlenków metali wobec bakterii *E. coli* w warunkach stałego ograniczenia dostępu światła. W oryginalnym badaniu autorstwa Pathakoti i współautorów[20] odpowiedź toksyczna wyrażona została jako funkcja dwóch zmiennych niezależnych: (i) absolutnej wartości elektroujemności atomu metalu (QMELECT) oraz (ii) absolutnej wartości elektroujemności tlenku metalu (LZELEHHO). Model Nano-QSAR opracowany przy użyciu metody regresji wielokrotnej na podstawie 13-elementowego zbioru uczącego charakteryzował się dobrym dopasowaniem (R_2=0,87 i $RMSE_C$=0,47). Z uwagi na zbyt małą liczbę związków w zbiorze testowym (4) autorzy nie przeprowadzili jednak ilościowej oceny jego zdolności przewidywania. Natomiast statystyki Q^2_{Ext}=−0,20 oraz $RMSE_P$=0,53 obliczone, na podstawie przewidzianych przez autorów wartości odpowiedzi dla związków zbioru testowego, wykazały brak zdolności generalizowania modelu na nowe przypadki (związki spoza zbioru uczącego). Dzięki zastosowaniu metody interpolacji liniowej z PC_1, będącej liniową kombinacją QMELECT i LZELEHHO, jako zmienną niezależną uzyskałam model o minimalnie lepszym dopasowaniu (R_2=0,88 i $RMSE_C$=0,44). Na szczególną uwagę zasługuje fakt, że w celu zwiększenia rzetelności oceny zdolności przewidywania modelu interpolacji liniowej, 16-elementowy zbiór nanocząstek MeOx został podzielony na zbiór uczący i testowy

w proporcjach 52,9% ÷ 47,1%, co stanowi odpowiednio 9 związków źródłowych i 7 związków docelowych. Walidacja zewnętrzna przeprowadzona z wykorzystaniem 7-elementowego zbioru testowego dowiodła wysokich zdolności przewidywania opracowanego modelu (Q^2_{Ext}=0,91 i $RMSE_P$=0,33) i tym samym jego użyteczność do przewidywania wartości odpowiedzi dla innych nanocząstek MeOx znajdujących się w obrębie jego dziedziny.

Na podstawie szczegółowej analizy i dyskusji uzyskanych wyników badań przedstawionych w pracach **[H1-H2]** sformułowałam ogólną rekomendację dotyczącą stosowania metod punktowej interpolacji liniowej do mało licznych zbiorów w sytuacji, gdy występuje liniowa zależność między zmienną zależną a zmiennymi objaśniającymi. Oba zaproponowane przeze mnie podejścia zostały ujęte zarówno w raporcie ze spotkania ekspertów OECD poświęconemu metodom grupowania i podejścia przekrojowego do oceny zagrożenia stwarzanego przez nanomateriały (raport *OECD ENV/JM/MONO(2016)59*), jak również w strategicznym dokumencie traktującym o współczesnych wyzwaniach w zakresie nanoinformatyki (*EU US Roadmap Nanoinformatics 2030*).

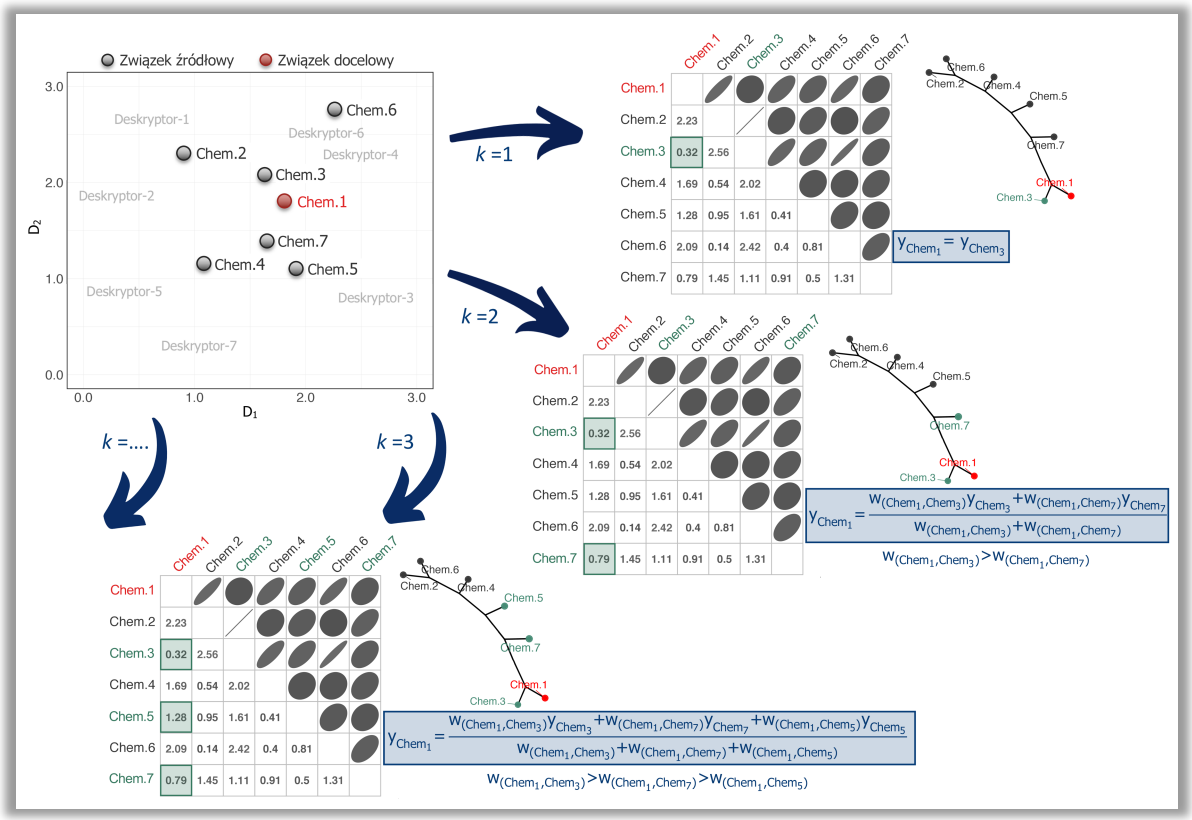
3.3.2 Metody uczenia maszynowego oparte na podobieństwie strukturalnym do efektywnego przewidywania wybranych właściwości biologicznych dla potrzeb komputerowej oceny zagrożenia chemicznego *[H3, H8]*

Pomimo wykazanej użyteczności metod interpolacji liniowej w modelowaniu mało licznych zbiorów danych, podejście to z definicji ograniczone jest wyłącznie do opisu liniowej zależności między zmienną zależną a pojedynczą lub większą liczbą zmiennych niezależnych. Warto w tym miejscu podkreślić, że istotnym wyzwaniem w modelowaniu właściwości biologicznych lub fizykochemicznych związków chemicznych jest nie tylko wielowymiarowość modelowanej zmiennej zależnej, która do poprawnego opisu wymaga użycia kilku zmiennych objaśniających, ale również występowanie złożonej, w tym niemonotonicznej, nieliniowej zależności między zmiennymi **X** i **y** (Rysunek 5). Problem niejednorodności analizowanego zbioru danych, zaszumienia danych i/lub obecności nawet pojedynczych wartości odstających staje się szczególnie istotny w przypadku mało licznych zbiorów danych, wpływając na jakość



Rysunek 5. Przykład złożonej zależności pomiędzy zmienną zależną i zmienną niezależną.

uzyskiwanych wyników modelowania w znacznie większym, stopniu niż w przypadku dużych zbiorów danych. Biorąc pod uwagę powyższe wyzwania, nieparametryczne metody uczenia maszynowego oparte na podobieństwie strukturalnym (ang. *similarity-based machine learning methods*) wydają się najbardziej intuicyjnym wyborem do poprawnego odwzorowania złożonej zależności pomiędzy **X** i **y**. Metody te nie zakładają z góry istnienia zależności pomiędzy zmiennymi ani danego rozkładu danych, dzięki czemu znajdują zastosowania zarówno w rozwiązywaniu problemów liniowych jak i nieliniowych. Ponadto są odpowiednie dla mało licznych zbiorów danych. U podstaw tej grupy metody leży założenie, że związki chemiczne o podobnej budowie strukturalnej wykazują podobną aktywność biologiczną, dlatego algorytm ich działania opiera się na wykorzystaniu estymatorów najbliższego sąsiedztwa. Najczęściej stosowaną nieparametryczną metodą wykorzystującą analizę najbliższego sąsiedztwa jest metoda *k*-najbliższych sąsiadów (ang. *k-nearest neighbors*, *k*-NN). Szczególnie ciekawym, choć stosunkowo rzadko wykorzystywanym w ocenie zagrożenia chemicznego wariantem algorytmu *k*-NN jest podejście ważonych odległości *k*-najbliższych sąsiadów (ang. *distance weighted k-nearest neighbors*). Zasada działania algorytmu ważonych odległości *k*-najbliższych sąsiadów, schematycznie przedstawiona na Rysunku 6, polega na iteracyjnym przeszukiwaniu zbioru związków źródłowych i znalezieniu takich, które są najbardziej podobne do zadanego związku docelowego. Miarą podobieństwa jest miara odległości (np. odległość euklidesowa), która



Rysunek 6. Schemat działania algorytmu ważonych k-najbliższych sąsiadów.

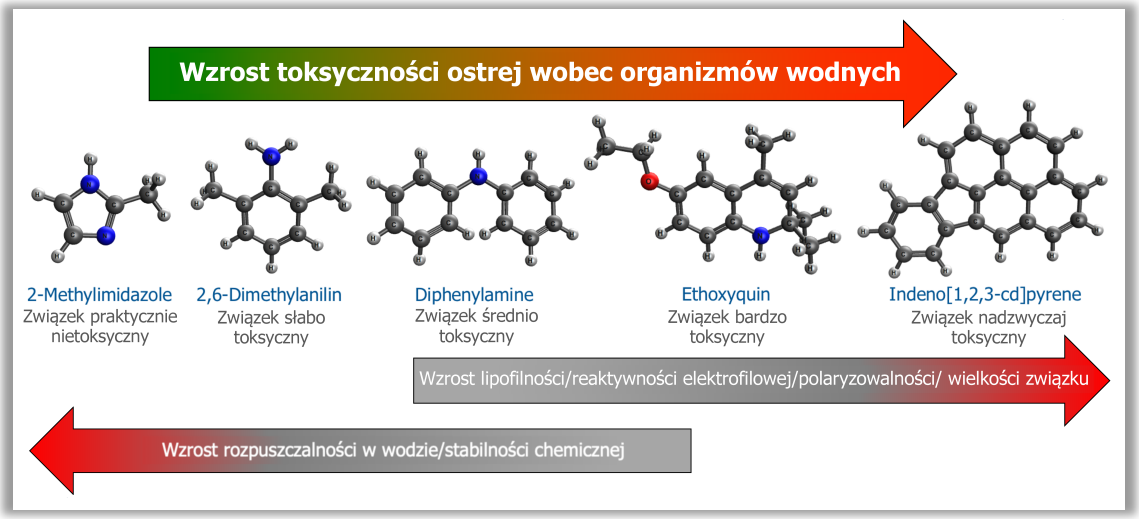
mówi, że związki podobne znajdują się blisko siebie w wielowymiarowej przestrzeni wybranych zmiennych objaśniających. Oszacowanie nieznanej wartości aktywności biologicznej związku docelowego następuje poprzez przyjęcie wartości aktywności najbliższego sąsiada (dla k=1), lub wyznaczenie ważonej aktywności k-najbliższych sąsiadów, gdzie waga (w_i) każdego sąsiada jest odwrotnie proporcjonalna do jego odległości do związku docelowego. Oznacza to, że związki źródłowe znajdujące się w bezpośrednim sąsiedztwie związku docelowego mają dużo większy wpływ na prognozowaną wartość jego aktywności biologicznej, natomiast wraz ze wzrostem odległości od związku docelowego ten wpływ maleje.

W celu oceny potencjału metody ważonych k-najbliższych sąsiadów do modelowania właściwości biologicznych mało licznych zbiorów danych, w pracy **[H3]** przeprowadziłam badania z wykorzystaniem dwóch zbiorów danych dotyczących toksyczności 17 i 18 nanocząstek tlenków metali wobec bakterii *E. coli* oraz ludzkiej linii komórkowej HaCaT. Równolegle, w przedmiotowym badaniu zwróciłam uwagę na ważny problem ograniczający możliwość efektywnego wykorzystania niektórych z opublikowanych w literaturze modeli Nano-QSAR, wynikający z braku oceny zdolności modelu do generalizacji na nowe obserwacje (niespełnienie wymogu walidacji zewnętrznej) lub przeprowadzenie oceny zdolności przewidywania modelu na podstawie kilku, np. 3 związków. Ocena jakości i wiarygodności modelu przeprowadzona w procesie walidacji zewnętrznej z wykorzystaniem tak niewielkiego zbioru testowego budzi wątpliwości ze względu na możliwą korelację losową (ang. *chance correlation*). Dlatego w celu opracowania modeli uczenia maszynowego oferujących bardziej rygorystyczną ocenę rzeczywistych zdolności przewidywania, oba wykorzystane w pracy zbiory danych podzieliłam na zbiór uczący i testowy w taki sposób, aby do walidacji zewnętrznej użyta było ponad połowa wszystkich związków, dla których dostępne były dane eksperymentalne (tj. 58,82% i 55,56%). Analiza uzyskanych wyników wykazała, że pomimo dwukrotnie mniejszego zbioru uczącego, w porównaniu z dostępnymi w literaturze modelami Nano-QSAR, oba opracowane modele ważonych k-najbliższych sąsiadów charakteryzowały się dobrym dopasowaniem (R^2>0,8) i poprawnymi zdolnościami przewidywania (Q^2>0,7) (H3: Rysunek 8). Na szczególną uwagę zasługuje fakt, że w przypadku obu modeli ocena zdolności modelu do generalizacji na nowe obserwacje, które nie były wykorzystane we wcześniejszych etapach modelowania, została przeprowadzona przy użyciu 10 związków, a więc nawet trzykrotnie większej niż w przypadku modeli Nano-QSAR np. autorstwa Sizochenko i współautorów[21] lub Singh i Gupta[22]. Uzyskane wyniki jednoznacznie potwierdziły celność przyjętych założeń i tym samym użyteczność metody ważonych k-najbliższych sąsiadów bazującej na estymatorach najbliższego sąsiedztwa do przewidywania toksyczności dla mało licznych zbiorów nanocząstek.

Oprócz liczebności zbioru danych wejściowych, czynnikiem mającym istotny wpływ na dokładność i wiarygodność modeli uczenia maszynowego jest reprezentatywność danych. Podobnie jak w przypadku mało licznych zbiorów danych, brak reprezentatywności zbioru danych względem populacji może skutkować niedostateczną zdolnością modelu do generalizacji na nowe obserwacje. Wśród przyczyn niedostatecznej reprezentatywności zbioru danych, najczęściej wymieniane są specyfika analizowanej właściwości biologicznej lub fizykochemicznej, ograniczona liczebność danej grupy związków chemicznych, oraz strukturalne i/lub funkcjonalne zróżnicowanie analizowanych związków chemicznych. W praktyce, problem niedostatecznej reprezentatywności danych dotyczy większości rzeczywistych zbiorów danych wykorzystywanych w procesie komputerowej oceny zagrożenia chemicznego. Przykładem takiego zbioru jest baza danych dotycząca toksyczności krótkookresowej (tzw. toksyczności ostrej) wobec organizmów wodnych, wykorzystywana przez japońskie Ministerstwo Środowiska w procesie oceny ryzyka chemicznego wymaganego przy rejestracji nowych substancji chemicznych.[23] W przedmiotowej bazie, dla dwóch podstawowych organizmów wskaźnikowych, tj. rozwielitki wielkiej (*Daphnia magna*) oraz ryżanki japońskiej (*Oryzias latipes*) dostępne są dane eksperymentalne odpowiednio dla 495 i 384 związków organicznych należących do różnych klas chemicznych, w tym np. alkoholi, węglowodorów alifatycznych i aromatycznych, nitrobenzenów, fenoli, estrów, eterów, aldehydów, ketonów i innych. Analiza wykresu radarowego (H8: Rysunek 2) przedstawiającego liczebność reprezentantów poszczególnych klas chemicznych dowiodła, że ponad połowa wszystkich klas, na które można podzielić cały zbiór danych, zawiera mniej niż 15 związków chemicznych. Tak silne zróżnicowanie strukturalne w połączeniu z niedostateczną liczebnością niektórych grup związków chemicznych stanowi ogromne wyzwanie w kontekście opracowania wiarygodnych modeli uczenia maszynowego wspierających proces oceny ryzyka chemicznego. W świetle powyższego nasuwa się pytanie, czy w przypadku silnie zróżnicowanego pod względem struktury chemicznej, liczebności i reprezentatywności zbioru danych możliwe jest opracowanie modelu uczenia maszynowego o dobrych zdolności predykcyjnych?

W celu znalezienia odpowiedzi na powyższe pytanie, w pracy **[H8]** przeprowadziłam modelowanie QSAR do przewidywania toksyczności ostrej związków organicznych o zróżnicowanej budowie chemicznej względem obu wymienionych powyżej organizmów wodnych (*D. magna* i *O. latipes*). Z wykorzystaniem analizy PCA, dla każdego organizmu wybrałam zestaw sześciu zmiennych objaśniających, różnicujących substancje wykazujące wysoką toksyczność ostrą, wyrażoną jako ($L(E)C_{50}$), od substancji, które nie wykazują niepożądanego działania (H8: Rysunek 5). Interpretacja wybranych deskryptorów molekularnych umożliwiła powiązanie toksyczności ostrej wobec rozwielitki i ryby z

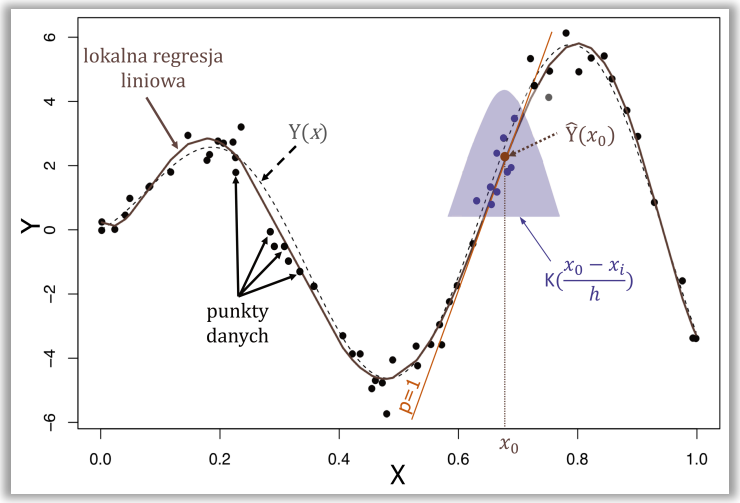
lipofilowością, reaktywnością elektrofilową, polaryzowalnością oraz wielkością cząsteczki związku chemicznego. Najważniejszym parametrem determinującym różnice w toksyczności ostrej wobec organizmów wodnych jest powinowactwo związku do fazy lipofilowej i fazy hydrofilowej, ilościowo wyrażane za pomocą współczynnika podziału *n*-oktanol/woda (LogP). Toksyczność ostra wrasta wraz ze wzrostem wartości LogP. Istotne znaczenie odgrywają również deskryptory związane z stabilnością termodynamiczną i reaktywnością chemiczną związku. Im związek chemiczny staje się bardziej elektrofilowy, tym wykazuje wyższą reaktywność chemiczną, a w konsekwencji wyższą toksyczność ostrą. Toksyczność ostra zwiększa się także wraz z rosnącą polaryzowalnością i wielkością cząsteczki związku chemicznego (Rysunek 7).



Rysunek 7. Zależność między wybranymi właściwościami strukturalnymi i fizykochemicznymi związków organicznych a ich toksycznością ostrą wobec organizmów wodnych.

W kolejnym etapie badań, w oparciu o wybrane zmienne objaśniające porównałam zdolności przewidywania modeli QSAR opracowanych z wykorzystaniem różnych algorytmów uczenia maszynowego. Spośród rozważanych metod uczenia maszynowego, modelami o najgorszych zdolnościach przewidywania były modele uzyskane przy użyciu metody regresji czynników głównych (ang. *principal component regression*, PCR) oraz metody częściowych najmniejszych kwadratów (ang. *partial least squares*, PLS). Żadna z dwóch metod regresji liniowej oparta na zmiennych ukrytych nie poradziła sobie z poprawnym odwzorowaniem złożonej zależności między danymi wejściowymi a odpowiedzią modelu (toksycznością ostrą wobec rozwielitki lub ryby), R^2 i Q^2_{Ext}<<0,5. Szczególnie interesujących obserwacji dostarczyły natomiast wyniki modelowania z wykorzystaniem nieparametrycznych metod uczenia maszynowego opartych na podobieństwie strukturalnym. W pracy **[H8]** oprócz metody ważonych *k*-najbliższych sąsiadów wykorzystałam ważoną lokalnie jądrową regresję liniową (ang. *locally weighted kernel linear regression*, KwLPR). Koncepcyjnie, ważona lokalnie jądrowa regresja liniowa opiera się na

założeniu, że w każdym punkcie zbioru danych można poprawnie przybliżyć dowolną funkcję za pomocą wielomianu niskiego stopnia, biorąc pod uwagę tylko obserwacje z najbliższego sąsiedztwa i wykorzystując jądro jako funkcję wagi. Oznacza to, że metoda ta łącząc w sobie prostotę regresji liniowej z elastycznością regresji nieliniowej, nie definiuje jednej globalnej funkcji, tylko w każdym punkcie zbioru danych dopasowuje ważony model liniowy, przypisując większą wagę punktom leżącym w bezpośrednim sąsiedztwie punktu, którego odpowiedź jest szacowana, a mniejszą wagę punktom bardziej oddalonym. O kształcie i zasięgu lokalnego sąsiedztwa decydują funkcja jądra (ang. *kernel, K*) oraz tzw. parametr wygładzania (ang. *bandwidth, h*) (Rysunek 8).



Rysunek 8. Idea ważonej lokalnie jądrowej regresji liniowej.

Biorąc pod uwagę ogólną dobroć dopasowania modeli do danych uczących oraz poprawność przewidywania ocenioną w procesie walidacji zewnętrznej, z wykorzystaniem 98- i 76-elementowego zbioru testowego, odpowiednio dla *D. magna* i *O. latipes*, najlepszymi wynikami charakteryzowały się modele QSAR opracowane przy użyciu metody KwLPR (H8: Tabela 1). Przewaga ważonej lokalnie jądrowej regresji liniowej nad klasycznymi metodami regresji wynika z faktu, że w odróżnieniu od metod regresyjnych, w których do wyznaczenia współczynników równania regresji najlepiej dopasowujących funkcję liniową do danych empirycznych wykorzystywane są wszystkie związki zbioru uczącego, w metodzie KwLPR wykorzystywana jest tylko niewielka część związków uczących z bezpośredniego sąsiedztwa punktu, którego odpowiedź jest szacowana. Jest to szczególnie istotne w przypadku zaszumionych i niejednorodnych danych. Zaskakująco, wyniki uzyskane w oparciu o metodę ważonych *k*-najbliższych sąsiadów charakteryzowały się zarówno gorszym dopasowaniem modelu jak i nieznacznie słabszymi zdolnościami modelu do prognozowania nowych danych w porównaniu z wynikami uzyskanymi metodą ważonej lokalnie jądrowej regresji liniowej. Możliwym wytłumaczeniem tej różnicy jest większa elastyczność przypisywania wag, które mają decydujący wpływ na oszacowania parametrów modelu lokalnego w metodzie KwLPR w porównaniu z metodą ważonych *k*-NN.

Przeprowadzone badania dowiodły zatem skuteczności ważonej lokalnie jądrowej regresji liniowej w modelowaniu toksyczności ostrej silnie zróżnicowanego pod względem

struktury chemicznej i reprezentatywności zbioru związków organicznych wobec organizmów wodnych. Aktualnie trwają prace zmierzające do wdrożenia opracowanych przeze mnie modeli uczenia maszynowego bazujących na algorytmie KwLPR do japońskiego systemu wstępnej komputerowej oceny ryzyka (*KAshinhou Tool for Ecotoxicity*).

3.3.3 Nieparametryczne modele międzygatunkowej ekstrapolacji toksyczności oparte na estymatorach najbliższego sąsiedztwa [H7]

W ostatnich latach szczególnie dużo uwagi poświęcono zagadnieniom międzygatunkowej ekstrapolacji toksyczności. Ten schemat modelowania umożliwia ekstrapolację (1) wyników badań toksykologicznych między organizmami różnych gatunków (niekiedy występującymi na różnych poziomach troficznych), jak również (2) wyników badań *in vitro* osiąganych w warunkach laboratoryjnych na warunki *in vivo*. Koncepcyjne podstawy modelowania ilościowej zależności aktywność-aktywność (ang. *quantitative activity–activity relationship*, QAAR) lub struktura-aktywność-aktywność (ang. *quantitative structure–activity–activity relationship*, QSAAR) są tożsame z modelowaniem QSAR. Zasadnicza różnica metodologiczna polega na użyciu w modelu QAAR/QSAAR do przewidywania aktywności biologicznej wobec danego gatunku (np. toksyczności ostrej dla ryby) tej samej aktywności biologicznej wyznaczonej w teście dla innego gatunku (np. toksyczności ostrej dla rozwielitki) jako zmiennej objaśniającej. Możliwość opracowania wiarygodnego modelu QAAR/QSAAR o dobrych zdolnościach predykcyjnych pozwoliłoby ograniczyć liczbę badań eksperymentalnych, w tym testów przeprowadzanych z wykorzystaniem organizmów żywych. Jest to szczególnie ważne w kontekście stale rosnącej liczby nowo syntezowanych, a następnie identyfikowanych w środowisku związków chemicznych.

Istotnym problemem ograniczającym jednak rozwój oraz efektywne stosowanie metod międzygatunkowej ekstrapolacji toksyczności są różnice we wrażliwości na działanie substancji toksycznych pomiędzy różnymi gatunkami. Wspomniana wrażliwość gatunkowa, w połączeniu z ograniczoną niekiedy liczebnością i/lub reprezentatywnością danych wejściowych do modelowania, może być niedostatecznie odwzorowywane za pomocą klasycznych technik regresji liniowej. Jednak jak dowodzi przegląd doniesień naukowych opublikowanych na przestrzeni ostatnich kilku lat, większość modeli QAAR/QSAAR do międzygatunkowej ekstrapolacji toksyczności opracowanych zostało z wykorzystaniem metody regresji liniowej.[24, 25] W świetle wykazanej w pracy **[H8]** dokładności przybliżeń uzyskanych za pomocą ważonej lokalnie jądrowej regresji liniowej oraz elastyczności algorytmu modelowania, w kolejnym etapie badań zweryfikowałam użyteczność tego podejścia do międzygatunkowej ekstrapolacji toksyczności.

Istotą badań przedstawionych w pracy **[H7]** była ocena wpływu wykorzystanej metody uczenia maszynowego na poprawność międzygatunkowej ekstrapolacji toksyczności. Na potrzeby badań wybrałam cztery literaturowe modele QSAAR spełniające wszystkie pięć kryteriów jakości OECD. W celu zapewnienia rzetelności i wiarygodności analizy porównawczej, proces modelowania przeprowadziłam z wykorzystaniem takich samych zbiorów uczących i testowych oraz takiej samej kombinacji zmiennych objaśniających jak w oryginalnych modelach QSAAR. Jedynym z rozważanych modeli był model do przewidywania toksyczności ostrej 294 pestycydów wobec ryby z gatunku bass pręgowany (*Lepomis macrochirus*). Autorzy oryginalnego modelu, Basant i współautorzy[25] przy użyciu metody regresji wielokrotnej ilościowo powiązali toksyczność ostrą dla *L. macrochirus* z toksycznością ostrą wobec rozwielitki (*D. magna*) oraz współczynnikiem podziału *n*-oktanol/woda. Powtórzenie modelowania z wykorzystaniem ważonej lokalnie jądrowej regresji liniowej poprawiło zdolność generalizacji modelu z 83 do 91%, w porównaniu z oryginalnym modelem. Również w przypadku pozostałych trzech modeli QSAAR istotnie różniących się liczebnością zbioru danych wejściowych (tj. w zakresie od 41 do 318 związków chemicznych) oraz reprezentatywnością, zastosowanie algorytmu KwLPR opartego na estymatorach najbliższego sąsiedztwa pozwoliło uzyskać modele o znacznie lepszych zdolnościach przewidywania (H7: Rysunek 10).

Wyniki badań przedstawione w pracy **[H7]** potwierdziły funkcjonalność i użyteczność ważonej lokalnie jądrowej regresji liniowej do międzygatunkowej ekstrapolacji toksyczności. W tym samym badaniu **[H7]**, ocena efektywności ważonej lokalnie jądrowej regresji liniowej została dodatkowo rozszerzona o analizę porównawczą ogólnej dobroci dopasowania i zdolności prognostycznych globalnego modelu QSAR do przewidywania toksyczności ostrej związków organicznych wobec ryby z gatunku strzebla grubogłowa (*Pimephales promelas*). Jako główną trudność opracowania wiarygodnego modelu QSAR dla przedmiotowego zbioru danych, autorzy oryginalnej pracy wymienili szeroką dziedzinę modelu obejmującą 908 związków organicznych o znacznym zróżnicowaniu strukturalnym i wykazujących różne mechanizmy toksycznego działania. Do opracowania globalnego modelu QSAR, Cassotti i współautorzy[26] wykorzystali klasyczną technikę *k*-NN oraz sześć deskryptorów ilościowo wyrażających lipofilowość oraz budowę strukturalną i właściwości fizykochemiczne analizowanych związków determinowane obecnością i liczbą heteroatomów w cząsteczce. Powtórzenie kalibracji i walidacji modelu przy użyciu algorytmu KwLPR w oparciu o ten sam podział związków na zbiór uczący i testowy i ten sam zbiór sześciu deskryptorów umożliwiło uzyskanie znacznie lepiej dopasowanego modelu (poprawa R^2 z 0,62 do 0,85) charakteryzującego się dodatkowo nieznacznie lepszymi zdolnościami przewidywania (poprawa Q^2 z 0,61 do 0,68).

W celu ułatwienia korzystania z zaproponowanego algorytmu KwLPR opartego na estymatorach najbliższego sąsiedztwa, wspólnie z mgr Magdaleną Piotrowską napisałyśmy skrypt w darmowym języku programowania R, który umożliwia przeprowadzenie procesu modelowania dla dowolnego zbioru danych. Skrypt został udostępniony w materiałach dodatkowych do publikacji i jest dostępny na stronie internetowej wydawcy https://doi.org/10.1186/s13321-021-00484-5.

3.3.4 Rozszerzenie paradygmatu ilościowej zależności struktura-aktywność na modele wielogatunkowe [H9]

Zarówno klasyczne modele ilościowej zależności struktura-aktywność, jak i modele ilościowej zależności aktywność-aktywność umożliwiają modelowanie, w danym czasie, wyłącznie pojedynczej zmiennej zależnej (np. toksyczności ostrej wobec ryby z gatunku *Cyprinus carpio*). I chociaż podejścia te stanowią ogromne wsparcie w procesie komputerowej oceny ryzyka chemicznego, w ostatnich latach wyraźnie wzrasta potrzeba opracowania bardziej wydajnych narzędzi, które umożliwiałyby jednoczesną ocenę potencjalnego zagrożenia chemicznego indukowanego obecnością związków chemicznych wobec większej liczby organizmów. Możliwość jednoczesnego wyznaczenia odpowiedzi modelu wobec dwóch lub większej liczby organizmów/gatunków/linii komórkowych oferowałaby uzyskanie szerszego i bardziej kompleksowego wglądu w mechanizmy molekularne odpowiedzialne za (nie)pożądane efekty działania analizowanych związków chemicznych. Wydaje się to szczególnie istotne w świetle stale rosnącej liczby środowiskowych zanieczyszczeń chemicznych. Poza oszczędnością czasu i kosztów badań, takie wielogatunkowe modele umożliwiałyby również uzyskanie wiedzy na temat ewentualnej wrażliwości gatunkowej w badaniach toksykologicznych. Dlatego kolejnym etapem moich badań było zweryfikowanie użyteczności paradygmatu ilościowego modelowania toksyczności wielogatunkowej (ang. *quantitative multi-species toxicity modeling*, qMTM) w odniesieniu do strukturalnie zróżnicowanego zbioru związków chemicznych.

Istotą badań przedstawionych w pracy **[H9]** była ocena skuteczności metody analizy korelacji kanonicznej (ang. *canonical correlation analysis*, CCA) do modelowania toksyczności wielogatunkowej. Wybór analizy korelacji kanonicznej jako techniki modelowania był podyktowany faktem, że CCA umożliwia badanie związku między dwoma wielowymiarowymi zestawami danych w taki sposób, aby najlepiej wyjaśniać ogólną zmienność zarówno w obrębie obu zestawów danych (tj. zestawu zmiennych zależnych i zestawu zmiennych objaśniających) jak i między nimi. Skuteczność zaproponowanego podejścia qMTM wykorzystującego algorytm CCA została zweryfikowania w oparciu o opracowany *de novo* model *in silico* do przewidywania toksyczności krótkookresowej 119

silnie zróżnicowanych pod względem struktury chemicznej i reprezentatywności związków organicznych wobec trzech organizmów wodnych z różnych poziomów troficznych: glonów (*Pseudokirchneriella subcapitata*), bezkręgowców (*D. magna*) i ryb (*O. latipes*). Wybór toksyczności ostrej indukowanej obecnością związków organicznych wobec trzech wymienionych powyżej organizmów wodnych jako studium przypadku nie był przypadkowy. Po pierwsze, wybrane organizmy jako przedstawiciele „*pierwotnych producentów*" (tj. algi); „*pierwotnych konsumentów i wtórnych producentów*" (tj. rozwielitki); i „*wtórnych konsumentów*" (tj. ryby) reprezentują organizmy wskaźnikowe najczęściej wykorzystywane w testach toksykologicznych. Po drugie, choć gatunki te należą do różnych poziomów troficznych i grup taksonomicznych to współdzielą mechanizm toksycznego działania, na który łącznie składają się dwa procesy, tj. wchłanianie substancji chemicznej z wody, a następnie jej interakcji z jednym lub kilkoma miejscami działania (np. receptorem komórkowym, docelową tkanką, narządem). Ma to szczególne znaczenie zwłaszcza w kontekście oczekiwanych wysokich zdolności prognostycznych modeli wielogatunkowych. Uznaje się, że podejście oparte na modelowaniu wielogatunkowym może zapewnić dokładne i wiarygodne prognozy, zwłaszcza gdy modelowane zmienne zależne są ze sobą ściśle powiązane mechanistycznie.

Do opracowania wielogatunkowego modelu toksyczności ostrej wobec *P. subcapitata*, *D. magna* i *O. latipes* wybrałam cztery zmienne objaśniające opisujące łatwość wchłaniania zanieczyszczeń organicznych z wody (tj. LogP); reaktywność chemiczną (tj. potencjał chemiczny, μ); oraz wielkość i kształt cząsteczki związku chemicznego (tj. masa cząsteczkowa, MW i średnia atomowa objętość van der Waalsa przeskalowana na atomie węgla, Mv). Analiza parametrów statystycznych opisujących jakość dopasowania (R^2=0,81; $RMSE_C$=0,41) oraz zdolności prognostyczne (Q^2_{F1}=0,80; Q^2_{F2}=0,80; Q^2_{F3}=0,81; $RMSE_P$=0,38) opracowanego modelu qMTM dowiodła jego ogólnej dobroci. Na szczególne podkreślenie zasługuje fakt, że uzyskane parametry były porównywane z wartościami statystyk klasycznych modeli QSAR opracowanych dla poszczególnych organizmów wodnych indywidualnie (H9: Tabela 3). Wyniki przedstawione w pracy **[H9]** dowodzą słuszności przyjętej tezy i wskazują na zasadność stosowania analizy korelacji kanonicznej do modelowania efektu toksycznego wywołanego obecnością substancji chemicznych u wielu badanych gatunków jednocześnie. Najbardziej znaczącą korzyścią zaproponowanego podejścia qMTM, która wykracza poza to, co oferują klasyczne modele QSAR/QAAR, jest uzyskanie kompleksowej wiedzy na temat wpływu substancji chemicznej na wiele organizmów jednocześnie. Zapewnia to głębsze zrozumienie podstawowych mechanizmów molekularnych odpowiedzialnych za niepożądane efekty toksyczne oraz usprawnia potencjalne działania w zakresie zarządzania ryzykiem chemicznym.

Z myślą o ułatwieniu użytkownikom korzystania z algorytmu qMTM, wraz z wynikami badań przedstawionymi w pracy **[H9]** udostępniony został skrypt napisany w języku programowania *Python*, umożliwiający automatyczne przeprowadzenie ilościowego modelowania toksyczności wielogatunkowej. Kod qMTM dostępny w trybie *open source* opublikowany został na stronie internetowej wydawcy: https://doi.org/10.1016/j.scitotenv.2022.160590.

3.3.5 Wstępne, przesiewowe badania środowiskowe mało licznych lub silnie zróżnicowanych pod względem struktury chemicznej i reprezentatywności grup związków chemicznych w oparciu o metody klasyfikacyjne [H5, H6]

Szczególną grupę metod, znajdujących szerokie zastosowanie zwłaszcza na wstępnych etapach komputerowej oceny ryzyka chemicznego, stanowią metody jakościowej zależności struktura-aktywność (ang. *structure-activity relationship*, SAR). Idea metod SAR polega na powiązaniu struktury związków chemicznych z obecnością lub brakiem ich określonych właściwości biologicznych lub fizykochemicznych. W praktyce oznacza to, że na podstawie związków zbioru uczącego, należących do *a priori* znanych klas o zdefiniowanych etykietach wartości zmiennej zależnej, możliwe jest opracowanie modelu klasyfikacyjnego pozwalającego przypisać nowy związek do jednej spośród dostępnych klas. Metody SAR, w tym wspomniane modele klasyfikacyjne, stanowią istotne wsparcie m. in. w środowiskowych badaniach przesiewowych umożliwiając wytypowanie potencjalnie niebezpiecznych związków chemicznych do dalszych badań toksykologicznych. Podobnie jak w przypadku metod ilościowej zależności struktura-aktywność, czynnikami wpływającymi na wiarygodność końcowych wyników przewidywań modeli SAR są: (1) ograniczona liczebność danych eksperymentalnych, (2) niedostateczna reprezentatywność analizowanej próby, rozumiana w tym przypadku jako niezbalansowana dystrybucja klas oraz (3) brak liniowej separowalności danych. W świetle powyższych ograniczeń, interesującym pytaniem jest: Jaka jest wiarygodność i użyteczność klasycznych metod klasyfikacyjnych wykorzystywanych w procesie komputerowej oceny ryzyka chemicznego? Odpowiedź na to pytanie, poprzedzone analizą wyników badań własnych i ich szczegółową dyskusją, przedstawiłam w pracach **[H5 i H6]**.

W obu pracach **[H5 i H6]**, analizę jakościową przeprowadziłam z wykorzystaniem nieparametrycznej metody drzew klasyfikacji i regresji (ang. *classification and regression trees*, CART). Metodę CART, poza prostotą i przejrzystością działania algorytmu, charakteryzuje intuicyjna interpretacja zastosowanych reguł klasyfikacji. W pracy **[H5]** zbudowałam trzy modele drzew klasyfikacyjnych dla 19 nanocząstek (H5: Tabela 2), co umożliwiło przypisanie poszczególnych nanocząstek do jednej z dwóch klas

'aktywna'/'nieaktywna' w teście: (1) zdolności surowicy do redukcji żelaza (III) (ang. *ferric reducing ability of serum*, FRAS); (2) karbonylacji białek oraz (3) krótkoterminowych badań inhalacyjnych na szczurach. Poszczególne modele różniły się liczebnością dostępnych danych eksperymentalnych. Modelem o najmniejszej liczebności danych eksperymentalnych był model do przewidywania wewnętrznego potencjału oksydacyjnego nanocząstek w teście FRAS. W tym przypadku dane dostępne były dla dziewięciu nanocząstek w podziale: pięć aktywnych (μU FRAS/m^{2}h $\geq$ 0,01921) i cztery nieaktywne (μU FRAS/m^{2}h < 0,01921). Natomiast modelem o największej liczebności danych eksperymentalnych był model do przewidywania największego stężenia bez obserwowanego działania szkodliwego (ang. *no observable adverse efect concentration*, NOAEC) w krótkoterminowych badaniach inhalacyjnych na szczurach. W tym przypadku dane dostępne były dla 17 nanocząstek w podziale: 7 aktywnych (NOAEC < 10mg/m^3) i 10 nieaktywnych (NOAEC $\geq$ 10mg/m^3). Aby zminimalizować wpływ mało licznych i nie w pełni zbalansowanych zbiorów danych wejściowych, w procesie modelowania wykorzystane zostało rozmyte podejście typu konsensus, w którym zbiory uczące i testowe wybrane zostały w wyniku kilkukrotnego podziału całego zbioru danych w taki sposób, aby każdy związek przynajmniej raz wykorzystany był do budowy modelu klasyfikacyjnego (zbiór uczący) i przynajmniej raz do oceny poprawności klasyfikacji (zbiór testowy). Jest to szczególnie istotne, w kontekście tzw. „paradoksu dokładności" (ang. *accuracy paradox*) wyrażającego się stronniczością modelu klasyfikacyjnego w kierunku klasy większościowej i nadmiernie optymistycznym oszacowaniem dokładności predykcji modelu. Do budowy modeli klasyfikacyjnych wybrane zostały: (1) pojedyncza zmienna objaśniająca - wielkość cząstek pierwotnych - dla modelu opisującego wewnętrzny potencjał oksydacyjny nanocząstek w teście FRAS; (2) dwie zmienne - wielkość cząstek pierwotnych oraz wielkość powierzchni właściwej - dla modelu karbonylacji białek oraz (3) trzy zmienne - wielkość powierzchni właściwej, obecność otoczki lub jej brak oraz energia najniżej nieobsadzonego orbitala molekularnego (energia LUMO) ilościowo wyrażająca powierzchniową aktywność redoks nanocząstek - dla modelu NOAEC w krótkoterminowych badaniach inhalacyjnych na szczurach. Ocena poprawność klasyfikacji przeprowadzona w procesie walidacji zewnętrznej wynosiła dla tych modeli odpowiednio: 100%, 80% i 97,9%.

Wyniki przedstawione w prezentowanej pracy dowodzą, że nawet w przypadku mało licznych zbiorów danych można poprawnie przewidzieć przynależność nowych związków do określonych klas. Należy przy tym zauważyć, że wynik klasyfikacji nanocząstek, dla których nie były dostępne dane eksperymentalne, nie był zależny od podziału związków na zbiory uczący i testowy (H5: Tabela 6, 9, 12). Zdefiniowane reguły logiczne opisujące przynależność związków do konkretnych klas wraz z wartościami progowymi zmiennych objaśniających

mają praktyczne znaczenie dla projektowania nowych nanocząstek o pożądanej aktywności biologicznej.

W podobny sposób przeprowadziłam ocenę użyteczności metody drzew decyzyjnych CART w zakresie poprawności klasyfikacji strukturalnie i funkcjonalnie zróżnicowanych związków organicznych do jednej z dwóch klas toksyczności: 'wysoka toksyczność' lub 'niska toksyczność/brak toksyczności'. W pracy **[H6]** zbudowałam dwa modele klasyfikacyjne do przewidywania toksyczności ostrej związków organicznych wobec dwóch organizmów wodnych: *D. magna* i *O. latipes*. Do tego celu wykorzystałam opisany powyżej zbiór danych stosowany przez japońskie Ministerstwo Środowiska w procesie oceny ryzyka chemicznego wymaganego przy rejestracji nowych substancji chemicznych. Dodatkowo w ramach tego badania przeprowadziłam analizę porównawczą dwóch schematów modelowania: (1) modeli lokalnych, które z definicji ograniczone są do mniejszej liczby, powiązanych strukturalnie i funkcjonalnie związków chemicznych oraz (2) modeli globalnych, które mają znacznie szerszą dziedzinę zastosowania modelu. Celem analizy porównawczej było uzyskanie odpowiedzi na pytanie: Czy lokalne modele drzew klasyfikacyjnych różnią się jakością przewidywania w porównaniu z modelami globalnymi?

Dziedzinę modeli lokalnych zdefiniowałam z wykorzystaniem analizy podobieństwa z odległością euklidesową jako miarą podobieństwa i hierarchiczną aglomeracyjną metodą Warda. W wyniku analizy (H6: Rysunek 3 i S9), w każdym z dwóch zbiorów danych dotyczących toksyczności ostrej odpowiednio wobec *D. magna* i *O. latipes*, wyróżniłam cztery klasy związków chemicznych o podobnej budowie strukturalnej i mechanizmie działania, tj.: (i) substancje chemicznie obojętne (np. węglowodory alifatyczne, alkohole); (ii) polarne związki o działaniu narkotycznym (np. aminy alifatyczne i aromatyczne, podstawione fenole i (di)nitrobenzeny); (iii) substancje chemiczne o różnych mechanizmach toksycznego działania (np. węglowodory aromatyczne, epoksydy, amidy, imidy) oraz (iv) związki chemiczne wykazujące niespecyficzne mechanizmy toksycznego działania (np. kwasy, estry alifatyczne, aromatyczne, fosforanowe). Do zdefiniowania reguł klasyfikacyjnych umożliwiających rozróżnienie silnie toksycznych związków chemicznych od związków o niskiej toksyczności lub nietoksycznych wykorzystałam deskryptory molekularne związane z lipofilowością, rozpuszczalnością w wodzie, reaktywnością elektrofilową, polaryzowalnością oraz wielkością cząsteczki. Interpretacja fizyczna wybranych zmiennych objaśniających dowiodła, że rozpuszczalne w tłuszczach związki organiczne o dużej masie cząsteczkowej i dużej powierzchni, wykazujące charakter miękkich elektrofili, charakteryzują się wysoką toksycznością ostrą wobec analizowanych organizmów wodnych. Analiza porównawcza parametrów statystycznych opisujących dokładność, czułość, precyzję oraz swoistość modeli lokalnych i modelu globalnego dowiodła lepszej jakości modeli lokalnych, zarówno w

przypadku modeli klasyfikacyjnych dla *D. magna* jak i *O. latipes* (H6: Tabela 1 i 2). Przeprowadzona analiza z zastosowaniem nieparametrycznego testu zgodności McNemara nie wykazała jednak istotnych statystycznie różnic pomiędzy przewidywanymi a rzeczywistymi etykietami toksyczności uzyskanymi zarówno w oparciu o modele lokalne jak i model globalny.

Uzyskane wyniki umożliwiły mi sformułować wnioski dotyczące użyteczności obu schematów modelowania oraz rekomendacje dotyczące wskazań stosowania lokalnych i globalnych modeli klasyfikacyjnych dla strukturalnie i funkcjonalnie zróżnicowanych związków organicznych. Biorąc pod uwagę fakt, że modele lokalne charakteryzują się wyższą zdolnością poprawnego przewidywania przynależności danego związku do określonej klasy toksyczności niż modele globalne, podejście lokalne jest szczególnie preferowane w modelowaniu mechanistycznym umożliwiającym odkrywanie związków przyczynowych między zmienną zależną a zmiennymi niezależnymi. Natomiast modele globalne, spełniające kryteria jakościowe OECD, pomimo nieznacznie gorszej jakości klasyfikacji stanowią istotne wsparcie w procesie zarządzania ryzykiem chemicznym wykorzystywanym m.in. w celach regulacyjnych.

3.3.6 Jak ocenić, czy prognozy modelu in silico są wiarygodne? [H4]

Żaden model uczenia maszynowego nie jest na tyle uniwersalny, aby w sposób wiarygodny przewidzieć wartości modelowanej zmiennej zależnej dla wszystkich istniejących związków chemicznych. Oznacza to, że z każdym modelem związane są pewne ograniczenia jego stosowania. Dlatego aby wykluczyć skrajnie niepewne i mało wiarygodne przewidywania dla związków wykazujących zróżnicowanie strukturalne względem zbioru uczącego, konieczne jest zdefiniowanie granic optymalnej przestrzeni predykcyjnej modelu, w której przewidywania są wiarygodne, tzw. dziedziny zastosowania modelu (ang. *applicability domain*, AD). W literaturze opisanych zostało wiele metod służących do określania granic dziedziny modelu. Metody te można podzielić na pięć głównych kategorii, metody wyznaczania dziedziny modelu w oparciu o (1) zakres zmiennej odpowiedzi lub (2) zakres zmiennych niezależnych; (3) metody geometryczne; (4) metody oparte na odległości oraz (5) metody oparte na rozkładzie gęstości prawdopodobieństwa. Każda z tych metod ma swoje wady i zalety, których krytyczny przegląd Czytelnik znajdzie w pracach.[27–30] W praktyce najczęściej wykorzystywanym podejściem wyznaczania dziedziny modelu jest podejście współczynników dźwigni (ang. *leverages*) sprzężone z wykresem Williamsa.

U podstaw podejścia współczynników dźwigni (h_i), podobnie jak w przypadku innych metod definiowania dziedziny modelu opartych na odległości, leży tzw. zasada „odległości od centroidu”. Zgodnie z tą zasadą współczynnik dźwigni definiuje podobieństwo strukturalne

związku do arytmetycznego środka ciężkości zbioru uczącego wyznaczonego w wielowymiarowej przestrzeni deskryptorów molekularnych (zmiennych niezależnych), co przedstawia poniższy wzór:

$$h_i = \boldsymbol{x}_i^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i$$

gdzie: $\mathbf{x}_i$ – wektor zmiennych niezależnych (deskryptorów) dla *i*-tego związku; **X** – macierz zmiennych niezależnych (deskryptorów) dla wszystkich związków ze zbioru uczącego.

Granicę podobieństwa, tj. krawędź domeny stosowalności modelu w odniesieniu do chemicznej przestrzeni strukturalnej wyznacza wartość krytyczna współczynnika dźwigni *h** obliczana w oparciu o zależność:

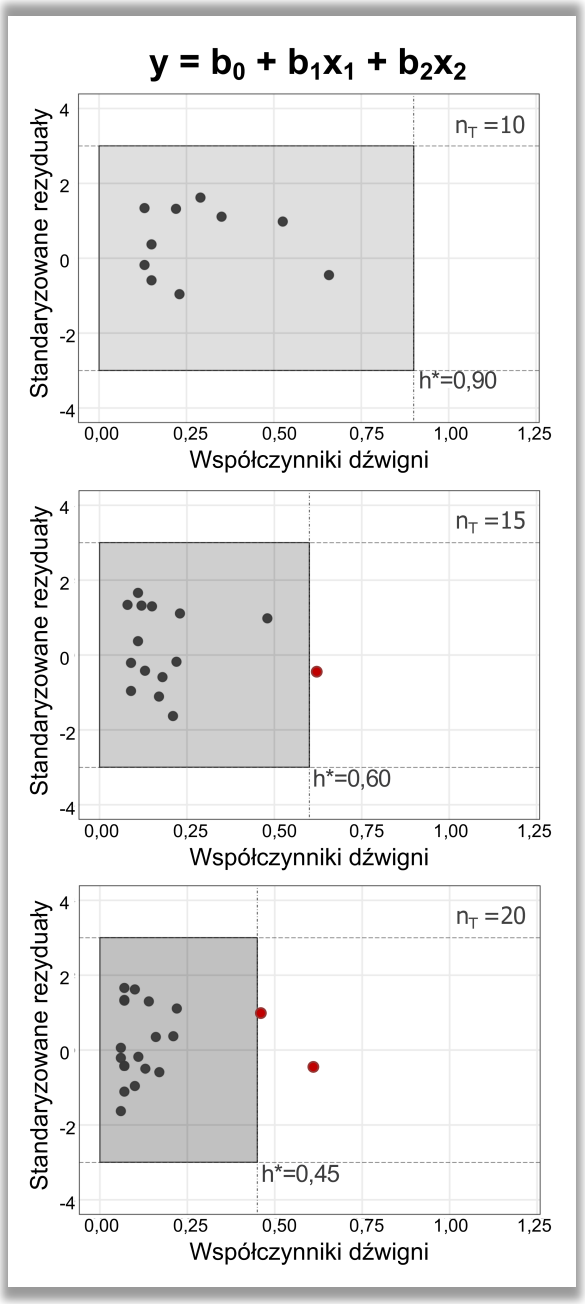
$$h^* = \frac{3p}{n}$$

gdzie: *p* – liczba parametrów modelu (w przypadku modeli zawierających wyraz wolny *p* definiowane jest jako liczba zmiennych niezależnych plus jeden); *n* – liczba związków w zbiorze uczącym.

W ujęciu graficznym, podejście współczynników dźwigni przedstawiane jest za pomocą wykresu Williamsa, poprzez odłożenie na osi odciętych wartości współczynników dźwigni, zaś na osi rzędnych wartości standaryzowanych rezyduałów dla poszczególnych związków ze zbioru uczącego i testowego. W praktyce, wiarygodne przewidywania można uzyskać wyłącznie dla związków leżących w granicach dziedziny modelu, wyznaczonej przez krytyczną wartość współczynnika dźwigni (h*), z jednej strony oraz wartości standaryzowanych rezyduałów równe trzem odchyleniom standardowym ($\pm 3\sigma$), z drugiej. Natomiast przewidywania dla związków, które z uwagi na silne zróżnicowanie w budowie strukturalnej ($h_i > h^*$) i/lub odmienny mechanizm działania (standaryzowane rezyduały > $\pm 3\sigma$) leżą poza granicami optymalnej przestrzeni predykcyjnej modelu, należy interpretować bardzo krytycznie jako potencjalnie niewiarygodne. W procesie oceny wiarygodności prognoz kluczową rolę odgrywa zatem sposób wyznaczenia granic optymalnej przestrzeni predykcyjnej modelu. Jak wspomniałam powyżej, krytyczna wartość współczynnika dźwigni (*h**) wyznaczająca granicę podobieństwa strukturalnego do zbioru uczącego, jest wprost proporcjonalna do liczby parametrów modelu i odwrotnie proporcjonalna do liczby związków użytych do kalibracji modelu. W przypadku mało licznych zbiorów danych prowadzi to do swoistego paradoksu, bowiem zmniejszenie liczby związków zbioru uczącego prowadzi do poszerzenia granic obszaru interpolacji, a w konsekwencji może skutkować błędnym przeświadczeniem co do wiarygodności przewidywań (Rysunek 9). Natomiast powszechnie wiadomo, że model lokalny, którego dziedzina z definicji ograniczona jest do mniejszej liczby

strukturalnie podobnych związków zwykle reprezentujących jedną klasę związków ma węższą dziedzinę modelu w porównaniu z modelem globalnym o większym i bardziej strukturalnie zróżnicowanym zbiorze uczącym.

W pracy **[H4]** zaproponowałam alternatywną metodę definiowania dziedziny modelu, która łączy aspekty metod opartych na odległości z elementami metod opartych na rozkładzie gęstości prawdopodobieństwa (ang. *probability-oriented distance-based method*, $AD_{ProbDist}$). W tej metodzie, granice optymalnej przestrzeni predykcyjnej modelu wyznaczane są w oparciu o elipsy ufności zdefiniowane w dwuwymiarowej przestrzeni wartości standaryzowanych rezyduałów (na osi rzędnych), oraz średniej wartości odległości euklidesowej do zbioru uczącego obliczonej w przestrzeni zmiennych objaśniających (na osi odciętych). Z uwagi na małą liczebność próby elipsy ufności na poziomie 95% i 99% są wyznaczone z rozkładu *t*-Studenta.



Rysunek 9. Wpływ liczebności zbioru uczącego (n_T) na szerokość dziedziny teoretycznego modelu uczenia maszynowego z dwoma zmiennymi objaśniającymi, wyznaczonej w oparciu o podejście współczynników dźwigni sprzężone z wykresem Williamsa.

Aby uczynić interpretację dziedziny modelu bardziej intuicyjną zastosowałam metaforę ulicznej sygnalizacji świetlnej. Zgodnie z którą, prognozy związków znajdujących się wewnątrz elipsy wyznaczonej na poziomie ufności 95% uznawane są za wiarygodne (strefa zielona). Przewidywania związków, które przekroczyły granicę poziomu ufności 95% ale mieszczą się w granicach 99% elipsy ufności należy traktować jako potencjalnie niewiarygodne (strefa pomarańczowa). Natomiast w przypadku związków leżących poza granicą wyznaczoną

przez elipsę na poziomie ufności równym 99% należy traktować jako niewiarygodne (strefa czerwona).

Użyteczność zaproponowanej metody oceny wiarygodności prognoz zaprezentowałam z wykorzystaniem czterech literaturowych modeli predykcyjnych opublikowanych w recenzowanych czasopismach naukowych. Na podstawie szczegółowej analizy i dyskusji uzyskanych wyników badań przedstawionych w pracy **[H4]** wykazałam, że zaproponowana przeze mnie metoda weryfikacji dziedziny modelu pozwala w części zniwelować wpływ bezpośredniej zależności szerokości dziedziny modelu od liczebności zbioru uczącego. Przeprowadzone porównanie wykazało, że szerokość dziedziny modelu Nano-QSAR opracowanego z wykorzystaniem 13-elementowego zbioru uczącego (studium przypadku nr 3) była znacznie węższa zarówno w porównaniu z dziedziną modelu Nano-QSAR opracowanego dla 16-elementowego zbioru uczącego (studium przypadku nr 2) jak i z dziedziną modelu podejścia przekrojowego (*read-across*) z 7-elementowym zbiorem uczącym (studium przypadku nr 4).

Uzyskane wyniki dowiodły również, że szerokość dziedziny modelu wyznaczana za pomocą metody $AD_{ProbDist}$ w istotnym stopniu zależy od natury związków wykorzystanych do kalibracji modelu oraz złożoności modelu predykcyjnego. Przy czym podejście $AD_{ProbDist}$ faworyzuje modele o optymalnej złożoności, w których zachowany jest kompromis pomiędzy obciążeniem a wariancją (ang. *bias-variance trade-off*). Jak można zauważyć analizując wyniki studium przypadku nr 3, w przypadku modelu cechującego się nadmiernym dopasowaniem do danych (ang. *overfitting*) oraz niską zdolnością generalizacji modelu, obszar interpolacji, w którym należy oczekiwać prawdziwych i wiarygodnych przewidywań jest bardzo wąski. Natomiast w przypadku modeli o optymalnej kompleksowości, w których utrzymana jest równowaga pomiędzy błędem wariancji i błędem obciążenia zapobiegająca nadmiernemu przeuczeniu lub niedopasowaniu modelu, granice optymalnej przestrzeni predykcyjnej modelu są znacznie szersze.

Wartą podkreślenia funkcjonalnością metody $AD_{ProbDist}$ jest też możliwość oceny wiarygodności prognoz dla nowych związków, dla których brakuje danych eksperymentalnych. Jest to szczególnie ważne w kontekście komputerowych badań przesiewowych dużej liczby związków oraz projektowania nowych związków chemicznych (np. o pożądanej aktywności biologicznej). Możliwość poznania wiarygodnych prognoz dotyczących właściwości i aktywności nowo projektowanych związków przed etapem syntezy chemicznej, to ogromna oszczędność czasu i pieniędzy oraz korzyść z ograniczenia liczby badań przeprowadzanych z udziałem żywych organizmów (aspekt etyczny).

Podsumowując, głównym osiągnięciem tego etapu badań, było opracowanie metody wyznaczenia granic optymalnej przestrzeni predykcyjnej modelu umożliwiającej rozróżnienie

wiarygodnych prognoz od skrajnie niepewnych. Aby ułatwić korzystanie z zaproponowanej metody oceny wiarygodności prognoz napisałam skrypt w darmowym języku programowania R, który umożliwia automatyczną analizę dziedzinę modelu i wizualizację wyników. Skrypt został udostępniony w materiałach dodatkowych do publikacji i jest dostępny na stronie internetowej wydawcy: https://doi.org/10.1039/C7EN00774D.

3.4. Podsumowanie - elementy nowości naukowej

Otrzymane i przedstawione w ramach osiągnięcia naukowego wyniki badań własnych pozwoliły mi zrealizować zaplanowane cele i odpowiedzieć na główne pytania badawcze. Do najważniejszych, pod względem poznawczym i praktycznym, osiągnięć naukowych zaliczam:

- Potwierdzenie użyteczności metod interpolacji liniowej do efektywnego przewidywania aktywności biologicznych dla mało licznych grup nanocząstek.
- Zaproponowanie modyfikacji metody interpolacji liniowej, polegającej na zastąpieniu pojedynczej zmiennej niezależnej - pierwszą główną składową, wyrażającą liniową kombinację dwóch lub większej liczby zmiennych objaśniających oraz wykazanie skuteczności tego podejścia w komputerowej ocenie ryzyka chemicznego stwarzanego przez nanocząstki, dla których dostępność danych eksperymentalnych jest bardzo ograniczona.
- Wykazanie, że nieparametryczne metody uczenia maszynowego wykorzystując tylko niewielką część związków uczących z bezpośredniego sąsiedztwa punktu, którego odpowiedź jest wyznaczana prowadzą do bardziej dokładnych oszacowań wartości odpowiedzi modelu w porównaniu z klasycznymi (liniowymi i nieliniowymi) metodami regresji.
- Opracowanie efektywnej metody międzygatunkowej ekstrapolacji toksyczności wykorzystującej estymatory najbliższego sąsiedztwa.
- Zaproponowanie rozszerzenie paradygmatu ilościowej zależności struktura-aktywność na modele wielogatunkowe w oparciu o metodę analizy korelacji kanonicznej.
- Wykazanie efektywności metody drzew klasyfikacyjnych jako narzędzia wspomagającego proces komputerowej oceny ryzyka chemicznego w przesiewowych badaniach (eko)toksyczności różnych grup związków chemicznych o zróżnicowanej dostępności danych eksperymentalnych.

- Opracowanie rekomendacji dotyczących wskazań stosowania lokalnych i globalnych modeli klasyfikacyjnych do przewidywania toksyczności ostrej strukturalnie i funkcjonalnie zróżnicowanych związków organicznych.
- Zidentyfikowanie ograniczenia metodycznego utrudniającego przeprowadzenie wiarygodnej oceny dziedziny modelu dla mało licznego zbioru związków chemicznych w oparciu o najczęściej stosowane podejście współczynników dźwigni sprzężone z wykresem Williamsa oraz opracowanie nowego podejścia wyznaczenia granic optymalnej przestrzeni predykcyjnej modelu, umożliwiającej rozróżnienie wiarygodnych prognoz od skrajnie niepewnych.

3.5. Bibliografia

1. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 194: https://doi.org/10.1038/194178b0
2. Hansch C, Fujita T (1964) ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Am Chem Soc 86: https://doi.org/10.1021/ja01062a035
3. Hansch C, Muir RM, Fujita T, et al (1963) The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. J Am Chem Soc 85: https://doi.org/10.1021/ja00901a033
4. The Insight Partners (2021) Quantitative Structure-Activity Relationship Market Forecast to 2027 - COVID-19 Impact and Global Analysis By Application; Industry, and Geography
5. European Comission (2006) Regulation (EC) 1907/2006 of the European Parliament and of teh Council of 18 December 2006 - REACH. Official Journal of the European Union. https://doi.org/http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:396:0001:0849:EN:PDF
6. EPA (2017) Chemicals under the Toxic Substances Control Act (TSCA). In: Environmental Protection Agency
7. CSCL (1973) Act on the Regulation of Manufacture and Evaluation of Chemical Substances
8. ECHA (2020) The Use of Alternatives to Testing on Animals for the REACH Regulation
9. Cronin MTD (2017) (Q)SARs to predict environmental toxicities: current status and future needs. Environ Sci Process Impacts. https://doi.org/10.1039/c6em00687f
10. Cherkasov A, Muratov EN, Fourches D, et al (2014) QSAR modeling: Where have you been? Where are you going to? J Med Chem
11. Muratov EN, Bajorath J, Sheridan RP, et al (2020) QSAR without borders. Chem Soc Rev. https://doi.org/10.1039/d0cs00098a
12. Cronin MTD, Schultz TW (2003) Pitfalls in QSAR. J Mol Struct 622:39–51
13. Yee LC, Wei YC (2012) Current Modeling Methods Used in QSAR/QSPR. In: Statistical Modelling of Molecular Descriptors in QSAR/QSPR
14. Roy K, Kar S, Das RN (2015) Selected Statistical Methods in QSAR. In: Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment
15. Piir G, Kahn I, García-Sosa AT, et al (2018) Best practices for QSAR model reporting: Physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. Environ Health Perspect 126
16. Wu Z, Zhu M, Kang Y, et al (2021) Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinform 22: https://doi.org/10.1093/bib/bbaa321

17. Bigdeli A, Hormozi-Nezhad MR, Parastar H (2015) Using nano-QSAR to determine the most responsible factor(s) in gold nanoparticle exocytosis. RSC Adv 5: https://doi.org/10.1039/c5ra06198a

18. Jagiello K, Halappanavar S, Rybińska-Fryca A, et al (2021) Transcriptomics-Based and AOP-Informed Structure–Activity Relationships to Predict Pulmonary Pathology Induced by Multiwalled Carbon Nanotubes. Small 17: https://doi.org/10.1002/smll.202003465

19. OECD (2017) Guidance on grouping of chemicals, Second Edition

20. Pathakoti K, Huang MJ, Watts JD, et al (2014) Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. J Photochem Photobiol B 130: https://doi.org/10.1016/j.jphotobiol.2013.11.023

21. Sizochenko N, Rasulev B, Gajewicz A, et al (2014) From basic physics to mechanisms of toxicity: The “liquid drop” approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. Nanoscale 6:. https://doi.org/10.1039/c4nr03487b

22. Singh KP, Gupta S (2014) Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. RSC Adv 4: https://doi.org/10.1039/c4ra01274g

23. Japanese Ministry of Environment (2018) Results of aquatic toxicity tests of chemicals conducted by Ministry of the Environment in Japan (- March 2018)

24. Sangion A, Gramatica P (2016) Ecotoxicity interspecies QAAR models from Daphnia toxicity of pharmaceuticals and personal care products. SAR QSAR Environ Res 27:781–798

25. Basant N, Gupta S, Singh KP (2016) Modeling the toxicity of chemical pesticides in multiple test species using local and global QSTR approaches. Toxicol Res (Camb) 5:340–353

26. Cassotti M, Ballabio D, Todeschini R, Consonni V (2015) A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas). SAR QSAR Environ Res. https://doi.org/10.1080/1062936X.2015.1018938

27. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: A review. ATLA Alternatives to Laboratory Animals

28. Sahigara F, Mansouri K, Ballabio D, et al (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules. https://doi.org/10.3390/molecules17054791

29. Gadaleta D, Mangiatordi GF, Catto M, et al (2016) Applicability Domain for QSAR Models. International Journal of Quantitative Structure-Property Relationships 1: https://doi.org/10.4018/ijqspr.2016010102

30. Roy K, Ambure P, Aher RB (2017) How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? Chemometrics and Intelligent Laboratory Systems 162: https://doi.org/10.1016/j.chemolab.2017.01.010

4. PRZYSZŁE KIERUNKI BADAŃ

W ostatnich latach obserwuje się duże zainteresowanie powierzchniowo modyfikowanymi, wielofunkcyjnymi materiałami, które z uwagi na swoją hybrydową budowę charakteryzują się unikalnymi właściwościami fizycznymi i chemicznymi. Wysoka złożoność struktury materiałów takich jak np. nanokompozyty grafenu/TiO_2 i materiałów o strukturze perowskitu, kropki kwantowe lub nanocząstki tlenków metali domieszkowanych jonami pierwiastków ziem rzadkich, w połączeniu z bardzo ograniczoną dostępnością rzetelnych danych eksperymentalnych stanowią duże wyzwanie metodologiczne dla tradycyjnych metod uczenia maszynowego. W świetle powyższych wyzwań, moje przyszłe kierunki badań koncentrują się na rozwoju metod i narzędzi wspierających proces komputerowej oceny ryzyka chemicznego oraz projektowania nowych wielokompozytowych materiałów o

użytecznych, w ujęciu aplikacyjnym, właściwościach fizycznych, chemicznych i biologicznych. Powyższy kierunek badań jest zbieżny z celami projektu: *Advanced High Aspect Ratio and Multicomponent materials: towards comprehensive intelLigent tEsting and Safe by design Strategies (HARMLESS)* finansowanym ze środków Programu Ramowego UE HORYZONT 2020, którego jestem kierownikiem.

Kolejnym problemem badawczym, którego rozwiązania planuję się podjąć w najbliższej przyszłości jest zweryfikowanie użyteczności metod uczenia maszynowego do przewidywania właściwości fizykochemicznych i/lub środowiskowych mikro- i nanoplastików. Kierunek badawczy wydaje się szczególnie istotny w kontekście obserwowanego w ostatnim czasie gwałtownego wzrostu obecności mikro- i nanoplastików w środowisku. W najnowszej literaturze przedmiotu brakuje jednak doniesień na temat możliwości wykorzystania metod komputerowych do oceny ryzyka stwarzanego przez mikro- i nanoplastiki dla zdrowia ludzkiego oraz różnych komponentów środowiska naturalnego.

V. INFORMACJA O WYKAZYWANIU SIĘ ISTOTNĄ AKTYWNOŚCIĄ NAUKOWĄ REALIZOWANĄ W WIĘCEJ NIŻ JEDNEJ UCZELNI, INSTYTUCJI NAUKOWEJ, W SZCZEGÓLNOŚCI ZAGRANICZNEJ

Tytuł zawodowy magistra uzyskałam w roku 2004, kończąc z wyróżnieniem (I lokata) pięcioletnie studia magisterskie na kierunku ochrona środowiska prowadzone przez Wydział Chemii Uniwersytetu Gdańskiego. Po studiach postanowiłam zdobyć dodatkowe wykształcenie, kończąc specjalność „Kierownik projektu z funduszy europejskich" na Studium Prawa Europejskiego w Warszawie. Następnie podjęłam pracę zawodową w założonej przez siebie firmie doradczej, zajmującej się pozyskiwaniem środków unijnych oraz równolegle na stanowisku Inspektora ds. pozyskiwania funduszy unijnych w jednostce samorządu terytorialnego. Praca zawodowa pozwoliła mi zdobyć doświadczenie biznesowe, niezbędne przy pozyskiwaniu środków finansowych na realizację projektów naukowych oraz wdrażanie wyników badań.

W roku akademickim 2008/2009, realizując swoje marzenie o podjęciu pracy naukowej, rozpoczęłam Studia Doktoranckie Chemii i Biochemii przy Wydziale Chemii UG. W ich trakcie, pod kierunkiem prof. dra hab. Tomasza Puzyna, prowadziłam badania mające na celu weryfikację przydatności metod komputerowych w ocenie potencjalnych zagrożeń związanych ze stosowaniem nanocząstek dla zdrowia człowieka i środowiska naturalnego. W toku realizacji badań opracowałam metodykę obliczania deskryptorów ilościowo opisujących strukturę molekularną oraz powierzchnię i kształt nanocząstek, tj. deskryptorów kwantowo-mechanicznych oraz deskryptorów obrazu na podstawie fotograficznej rejestracji obrazu z wykorzystaniem transmisyjnego mikroskopu elektronowego. We współpracy z

naukowcami z Uniwersytetu Gdańskiego oraz *Interdisciplinary Center for Nanotoxicity* (Stany Zjednoczone) uczestniczyłam w opracowaniu pierwszego na świecie, spełniającego wszystkie pięć kryteriów jakości OECD, modelu Nano-QSAR do oceny toksycznego wpływu nanocząstek MeOx na komórki bakterii *E. coli*. Ponadto, w celu zrozumienia podstawowych różnic w mechanizmach toksyczności indukowanej przez nanocząstki tlenków metali i półmetali w komórkach prokariotycznych (*E. coli*) i eukariotycznych (ludzkich) opracowałam model Nano-QSAR do oceny cytotoksycznego wpływu nanocząstek MeOx na komórki ludzkich keratynocytów. W trakcie trwania studiów doktoranckich zdobyłam swoje pierwsze doświadczenia we współpracy z zagranicznymi jednostkami naukowymi uczestnicząc m. in. w sześciotygodniowej szkole letniej *ICN Summer Institute 2009: Summer School in Quantum Chemistry* (Stany Zjednoczone). Ponadto jako wykonawca trzech międzynarodowych projektów naukowych kierowanych przez prof. dra hab. Tomasza Puzyna, w kolejnych latach studiów doktoranckich (2010-2012) corocznie odbywałam 2-3 miesięczny staż naukowy w *Interdisciplinary Center for Nanotoxicity* (Stany Zjednoczone) w zespole kierowanym przez Prof. Jerzego Leszczyńskiego i/lub *National Institute for Environmental Studies* (Japonia) w zespole kierowanym przez Dr. Noriyuki Suzuki. W maju 2013 roku obroniłam rozprawę doktorską zatytułowaną „*Opracowanie metod in silico służących przewidywaniu cytotoksycznego wpływu nanocząstek tlenków nieorganicznych na komórki bakterii E. coli oraz ludzkie keratynocyty (HaCaT)”,* za którą otrzymałam wyróżnienie w konkursie na najlepszą pracę doktorską obronioną w 2013 roku, organizowanym przez Gdański Oddział Polskiego Towarzystwa Chemicznego. Przedstawione w rozprawie doktorskiej wyniki badań zostały opublikowane m. in. w *Nature Nanotechnology*, *Advanced Drug Delivery Reviews*, *Nanotoxicology*.

Po doktoracie kontynuowałam pracę w zespole naukowym prof. dra hab. Tomasza Puzyna, najpierw (w okresie od 15.10.2013 r. do 31.03.2014 r.) na stanowisku asystenta, a od 01.04.2014 r. na stanowisku adiunkta. W tym okresie, pozostając w nurcie zainteresowań naukowo-badawczych dotyczących wykorzystania metod uczenia maszynowego w komputerowej ocenie zagrożeń stwarzanych przez różne grupy związków chemicznych, aktywnie uczestniczyłam w charakterze wykonawcy w 5 międzynarodowych projektach realizowanych m. in. w ramach Narodowego Centrum Nauki, 7. Programu Ramowego, Szwajcarsko-Polskiego Programu Współpracy oraz Programu Ramowego HORYZONT-2020 we współpracy z wiodącymi ośrodkami z zagranicy (m. in. *Liverpool John Moores University* (Wielka Brytania); *Lawrence Berkeley National Laboratory, Computational Research Division* (Stany Zjednoczone); *Joint Research Centre* (Włochy); *Karolinska Institutet* (Szwecja); *Ethniko Idryma Erevnon, National Hellenic Research Foundation* (Grecja); *A.V. Bogatsky Physico-Chemical Institute of National Academy of Science of*

Ukraine, (Ukraina)). Wiedza i doświadczenie zdobyte przy realizacji projektu doktorskiego oraz udział w projektach naukowych, poświęconych rozwojowi metod komputerowych wspierających proces oceny zagrożenia chemicznego stwarzanego przez istniejące nanocząstki oraz projektowania nowych, bezpiecznych dla zdrowia człowieka i środowiska naturalnego nanomateriałów, pozwoliły mi zidentyfikować poważne ograniczenia istniejących podejść komputerowych. Ograniczenia te wynikały w głównej mierze z braku dostatecznej liczby danych empirycznych niezbędnych do opracowania wiarygodnych modeli komputerowych. Powyższe ograniczenie stało się dla mnie impulsem do rozwoju nowych i/lub dostosowania istniejących metod uczenia maszynowego do modelowania mało licznych zbiorów danych. W latach 2015-2018 wiodącym tematem moich badań było opracowanie zestawu metod i narzędzi do modelowania, które niezależnie od: (i) skali pomiaru zmiennej zależnej (tj. ilościowa/jakościowa); (ii) liczby zmiennych niezależnych niezbędnych do opracowania poprawnego modelu uczenia maszynowego oraz (iii) (nie)liniowej zależności między zmienną zależną, a zmiennymi niezależnymi, będą umożliwiały poprawne przewidywanie wybranych biologicznych i/lub fizykochemicznych właściwości dla mało licznych klas związków.

Kolejnym punktem zwrotnym były dla mnie badania zainicjowane podczas rocznego stażu podoktorskiego w *National Institute for Environmental Studies, Research Center for Environmental Risk* (Tsukuba, Japonia), w grupie kierowanej przez Dr. Hiroshi Yamamoto. Staż podoktorski był elementem projektu KATE (*KAshinho Tool for Ecotoxicity*) finansowanym przez *Japanese Ministry of the Environment*. Celem projektu był rozwój metod uczenia maszynowego wspomagających proces komputerowej oceny ryzyka chemicznego wymaganej przy rejestracji nowych substancji chemicznych oraz zastosowanie tych metod do wyznaczenia brakujących informacji na temat biologicznych i/lub fizykochemicznych właściwości wybranych związków chemicznych. W swoich badaniach skupiłam się na wyzwaniach związanych z niedostateczną liczebnością zbiorów związków chemicznych o silnym zróżnicowaniu strukturalnym i/lub funkcjonalnym. Kontynuacja tej tematyki badań możliwa była dzięki uzyskaniu grantu NCN SONATA nr UMO-2016/23/D/NZ7/03973 "*Opracowanie metod szacowania przekrojowego (read-across) wspierających proces oceny ryzyka chemicznego*", którego byłam kierownikiem. Zaproponowana przeze mnie metodyka modelowania oparta na estymatorach najbliższego sąsiedztwa umożliwiła poprawę jakości dotychczasowych modeli KATE. Obecnie podejmowane są działania mające na celu wdrożenie opracowanych przeze mnie modeli uczenia maszynowego do przewidywania toksyczności krótkookresowej wobec organizmów wodnych indukowanej obecnością przemysłowych związków organicznych.

Równolegle do powyższych badań, na przestrzeni kilku ostatnich lat realizowałam badania dotyczące wykorzystania metod uczenia nienadzorowanego do eksploracyjnej analizy danych oraz metod uczenia nadzorowanego do klasyfikacji i predykcji ze szczególnym ukierunkowaniem na wykorzystanie tych metod w ocenie zagrożenia chemicznego. Krótką charakterystykę wybranych aktywności w ramach kontynuacji dotychczasowych lub podjęcia nowych współprac z krajowymi i zagranicznymi ośrodkami naukowymi przedstawiłam poniżej:

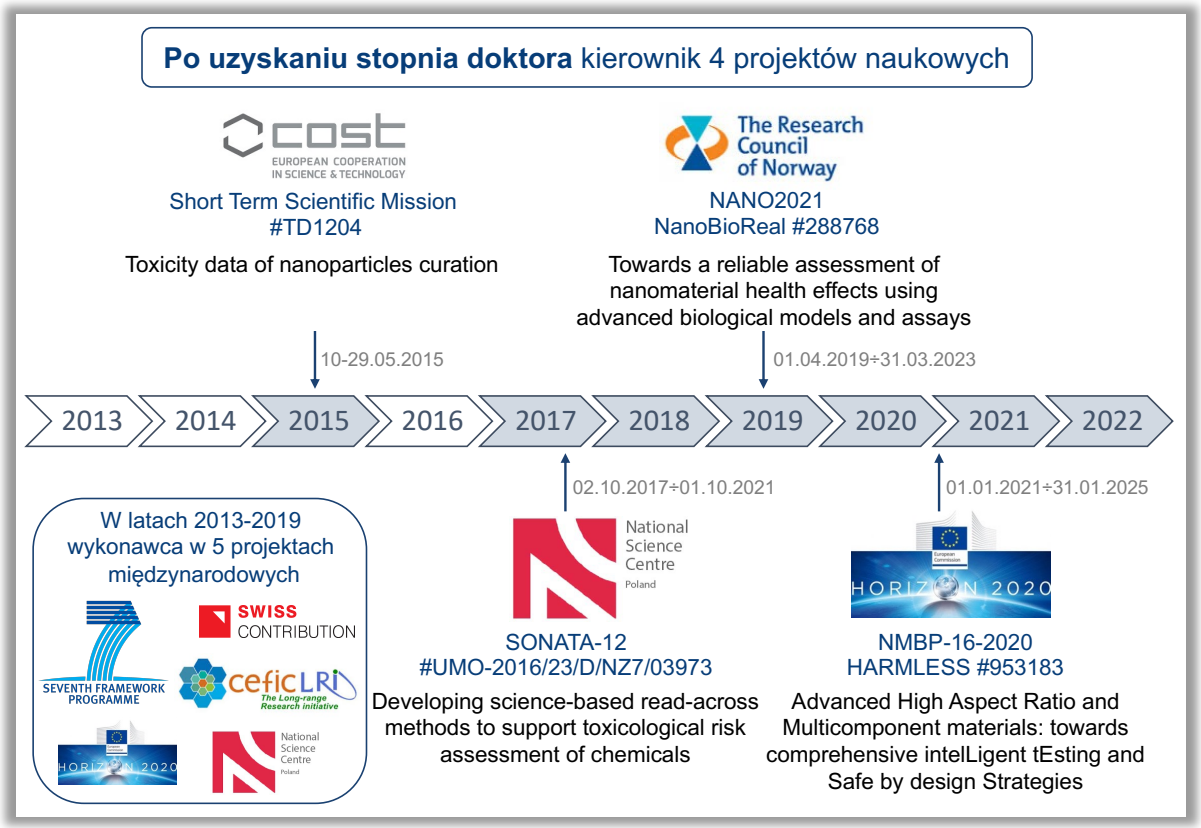
- We współpracy z zespołem dr hab. inż. profesor uczelni Anny Michalskiej-Ciechanowskiej z Katedry Technologii Owoców, Warzyw i Nutraceutyków Roślinnych Uniwersytetu Przyrodniczego we Wrocławiu, w oparciu o chemometryczną analizę i interpretację danych eksperymentalnych, przedstawiliśmy wyniki umożliwiające głębsze zrozumienie indukowanych termicznie zmian zawartości wybranych związków polifenolowych i powstawania hydroksymetylo-*L*-furfuralu w proszkach z soku z aronii (*Aronia melanocarpa L.*). Zastosowanie eksploracyjnej analizy danych pozwoliło nam również odpowiedzieć na pytanie badawcze: Jak różne rodzaje nośników i techniki suszenia wpływają na zmiany właściwości fizykochemicznych proszków z ekstraktów z wytłoków z aronii? Wynikiem tej współpracy są prace opublikowane w *Food chemistry* oraz *Foods* (Załącznik 7 pkt. 4.2 poz. P44 oraz poz. P46).
- We współpracy z zespołem dr hab. inż. profesor uczelni Anny Zielińskiej-Jurek z Katedry Inżynierii Procesowej i Technologii Chemicznej Politechniki Gdańskiej, na podstawie wyników eksploracyjnej analizy danych i nadzorowanego uczenia maszynowego zdefiniowaliśmy wpływ czynników fizykochemicznych, morfologicznych oraz warunków syntezy kompozytów TiO_2/Ti_3C_2 na ich aktywność fotokatalityczną w procesie degradacji acetaminofenu w środowisku wodnym. Wynikiem tej współpracy jest manuskrypt w trakcie recenzji.
- We współpracy z zespołem prof. dr hab. Adriany Zaleskiej-Medynskiej z Katedry Technologii Środowiska na Wydziale Chemii Uniwersytetu Gdańskiego przeprowadziliśmy badania mające na celu integrację metod eksperymentalnych i obliczeniowych umożliwiających komputerowe projektowanie kompozytów TiO_2-$Cu/(Cu_xO_y)$ pokrytych grafenem wykorzystanych w reakcji fotokatalitycznego generowania wodoru. Efektem współpracy jest praca opublikowana w *Catalysts* (Załącznik 7 pkt. 4.2 poz. P45).
- We współpracy z grupą prof. Kunal Roy z *Drug Theoretics and Cheminformatics (DTC) Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata*

(Indie) zaproponowaliśmy nowy zintegrowany schemat blokowy do modelowania mało licznych zbiorów danych. Kontynuując współpracę nad rozwojem narzędzi komputerowych dedykowanych modelowaniu mało licznych zbiorów danych, w kolejnym etapie badań opracowaliśmy nowy algorytm do ilościowego podejścia przekrojowego (*read-across*). Otrzymane wyniki zostały opublikowane w *Journal of Chemical Information and Modeling* oraz *Environmental Science: Nano* (Załącznik 7 pkt. 4.2 poz. P43 oraz poz. P47).

- We współpracy z zespołem prof. Jerzego Leszczyńskiego z *Interdisciplinary Center for Nanotoxicity, Jackson State University* (Stany Zjednoczone) przeprowadziliśmy ocenę efektywności ważonej lokalnie jądrowej regresji liniowej do międzygatunkowej ekstrapolacji toksyczności. Efektem współpracy jest praca opublikowana w *Journal of Cheminformatics* wchodząca w skład osiągnięcia naukowego [H7].
- We współpracy z współpracownikami dr Natalii Fjodorovej z *Department of Chemoinformatics, National Institute of Chemistry, Ljubljana* (Słowenia) przeprowadziliśmy ocenę efektywności algorytmu sztucznej sieci neuronowej i samoorganizujących się map Kohonena do ilościowego i jakościowego modelowania toksyczności nanocząstek tlenków metali. Wynikiem tej współpracy jest praca opublikowana w *Nanotoxicology* (Załącznik 7 pkt. 4.2 poz. P39).

Istotnym elementem dynamizującym, w mojej opinii, podejmowaną przeze mnie aktywność naukową po uzyskaniu stopnia doktora, były zagraniczne staże naukowe. Łącznie po doktoracie odbyłam trzy staże zagraniczne: dwa krótkoterminowe (miesięczny staż w *Interdisciplinary Center for Nanotoxicity, Jackson State University* (Jackson, MS, Stany Zjednoczone) w grupie kierowanej przez prof. Jerzego Leszczyńskiego w 2014 r. i trzytygodniowy staż podoktorski w ramach imiennego grantu *MODENA Cost Short Term Scientific Mission* w *Bundesinstitut für Risikobewertung* (Berlin, Niemcy) w grupie kierowanej przez dr. Andrea Haase w 2015 r.) oraz jeden długoterminowy, roczny staż podoktorski w *National Institute for Environmental Studies, Research Center for Environmental Risk* (Tsukuba, Japonia) w 2016-2017 r. Ponadto, dzięki uczestnictwu w międzynarodowym projekcie ^{Nano}BRIDGES koordynowanym przez prof. dra hab. Tomasza Puzyna, w latach 2014-2015 oprócz wymienionych powyżej staży podoktorskich odbyłam także kilka kilkudniowych wizyt naukowych w zagranicznych instytucjach naukowych, w tym m. in. w *National Center for Nanoscience and Technology of China, Chinese Academy of Science* (Chiny), czy *Jadavpur University in Kolkata* (Indie). Udział w stażach naukowych oraz wyjazdach badawczych poza oczywistymi korzyściami rozwoju warsztatu naukowego umożliwił mi poznanie organizacji pracy naukowej, mechanizmów zarządzania pracą

zespołową oraz współpracą międzynarodową w wiodących zagranicznych ośrodkach naukowych. Co szczególnie ważne z mojej perspektywy, to podczas rocznego stażu podoktorskiego zdefiniowałam przyszłe kierunki badań, zdobyłam większą samodzielność naukową, a także rozwinęłam istniejące oraz nawiązałam nowe współprace z krajowymi i zagranicznymi ośrodkami naukowymi. Wymiernym efektem podjętych współprac naukowych były otrzymane przeze mnie zaproszenia do konsorcjów aplikujących o projekty naukowe, co zaowocowało uzyskaniem dwóch grantów na badania ze środków Programu Ramowego UE HORYZONT-2020 oraz Norweskiej Rady Badań (ang. *Research Council of Norway*). Łącznie po uzyskaniu stopnia doktora nauk chemicznych kierowałam/kieruję czterema projektami badawczymi. Krótką charakterystykę projektów, w których realizację byłam zaangażowana jako kierownik przedstawia poniższy schemat (Rysunek 10, Załącznik 7, pkt. 7.2).



Rysunek 10. Sumaryczna charakterystyka grantów, w których pełnię rolę kierownika projektu.

Istotnym elementem podejmowanej przeze mnie aktywności naukowej w obszarze umiędzynarodowienia badań jest także pełnienie roli opiekuna naukowego w projekcie POLONEZ BIS 2 współfinansowanym przez Komisję Europejską i Narodowe Centrum Nauki w ramach grantu Marie Skłodowska-Curie COFUND. Projekt pt. "*NanoSens: Towards multi-evidence approach for the risk assessment of a diverse group of nanoparticles and advanced materials*" przygotowany przez Dr. Kabiruddin Ikramuddin Khan (Indie) uzyskał

dofinansowanie. Rozpoczęcie realizacji projektu *NanoSens* planowane jest na kwiecień 2023 roku.

W mojej ocenie, równie ważnym elementem mojego rozwoju naukowego jest członkostwo w komitetach redakcyjnych i radach naukowych czasopism. Aktualnie pełnię rolę: (i) zastępcy redaktora naczelnego w czasopiśmie *International Journal of Quantitative Structure-Property Relationships* (IJQSPR); (ii) redaktora recenzji w czasopiśmie *Frontiers in Pharmacology - Predictive Toxicology* oraz (iii) redaktora tematycznego w czasopiśmie *Materials*. Wykonałam ponad 120 recenzji manuskryptów dla wiodących czasopism naukowych podejmujących głównie tematykę komputerowej oceny ryzyka chemicznego, komputerowej (nano)toksykologii oraz rozwoju metod i narzędzi komputerowych wspierających bezpieczne i zrównoważone projektowanie nowych substancji chemicznych (Załącznik 7, pkt. 11). Pełniłam również funkcję współredaktora gościnnego w wydaniu specjalnym czasopisma *International Journal of Quantitative Structure-Property Relationships* (IJQSPR) pt. "*QSPR in Nanotechnology*". Natomiast, w 2020 roku nakładem wydawnictwa *Pan Stanford Publishing* wydana w obiegu międzynarodowym została książka pt. "*Computational Nanotoxicology: Challenges and perspectives*", której jestem redaktorem.

Po uzyskaniu stopnia doktora nauk chemicznych brałam również czynny udział w 24 konferencjach i seminariach naukowych (w tym w 5 krajowych i 19 zagranicznych), prezentując wyniki swoich badań zarówno w formie referatu naukowego (10) jak i prezentacji plakatowej (14). Spośród wygłoszonych wykładów, szczególnie cieszą mnie wykłady na zaproszenie, np. podczas *OECD Expert Meeting on Grouping and read-across for the hazard assessment of manufactured nanomaterials* (Bruksela, 2016) lub *Nanoinformatics: Spanning Scales, Systems and Solutions - Beilstein Nanotechnology Symposium* (Rüdesheim, 2022).

Za swoją dotychczasową działalność naukową otrzymałam liczne wyróżnienia i stypendia naukowe m.in.:

- Laureatka nagrody finałowej Nagród Naukowych POLITYKI w kategorii nauki ścisłe (2019)
- Laureatka ogólnoświatowej nagrody dla wschodzących talentów nauki - *International Rising Talents Awards L'Oréal-UNESCO For Women in Science* (2018)
- Laureatka stypendium habilitacyjnego L'Oreal-UNESCO Dla Kobiet i Nauki (2017)
- Laureatka nagrody za najlepszą prezentację plakatową przedstawioną podczas międzynarodowej konferencji *8th International Nanotoxicology Congress (NanoTox2016)*, Boston, Stany Zjednoczone (2016)
- Laureatka stypendium Ministra Nauki i Szkolnictwa Wyższego dla Wybitnych Młodych Naukowców (2014)

- Laureatka stypendium dla młodych doktorów w ramach projektu „Program rozwoju Uniwersytetu Gdańskiego w obszarach Europa 2020” (2014)
- Wyróżnienie Polskiego Towarzystwa Chemicznego, Oddział Gdańsk za najlepszą pracę doktorską obronioną w 2013 roku.

Ponadto w latach 2015 – 2022 trzykrotnie byłam laureatką zespołowej nagrody Rektora Uniwersytetu Gdańskiego za szczególne osiągnięcia naukowe: nagroda III stopnia (2015 r.), nagroda I stopnia (2019 r.) oraz nagroda II stopnia (2022 r.).

VI. INFORMACJA O OSIĄGNIĘCIACH DYDAKTYCZNYCH, ORGANIZACYJNYCH ORAZ POPULARYZUJĄCYCH NAUKĘ

1. OSIĄGNIĘCIA DYDAKTYCZNE

Działalność dydaktyczna stanowi istotną część podejmowanej przeze mnie aktywności zawodowej. Jako adiunkt w Pracowni Chemoinformatyki Środowiska w Katedrze Chemii i Radiochemii Środowiska prowadziłam/prowadzę zajęcia dydaktyczne: wykłady, laboratoria w pracowni komputerowej oraz seminaria na studiach pierwszego i drugiego stopnia na kierunkach: Chemia, Biznes chemiczny i Ochrona środowiska na Wydziale Chemii oraz na kierunku Bioinformatyka na Wydziale Matematyki, Fizyki i Informatyki, a także na studiach doktoranckich. Wśród wykładanych przeze mnie przedmiotów, oprócz przedmiotów kierunkowych (np. "Statystyka i chemometria w analityce chemicznej", "Analiza danych wielowymiarowych", "Chemometria") znajdują się również autorskie przedmioty, m.in. "Elementy języka R", "QSAR i szacowanie przekrojowe w ocenie ryzyka chemicznego", "Metody regresji i klasyfikacji". Roczną liczbę godzin obciążenia dydaktycznego oraz rodzaj wykładanych przeze mnie przedmiotów sumarycznie przedstawia poniższa tabela (w Tabeli 1. uwzględniłam okres ostatnich pięciu lat).

Ważną częścią mojej aktywności dydaktycznej jest opieka nad projektami licencjackimi i magisterskimi. Łącznie po uzyskaniu stopnia doktora nauk chemicznych byłam opiekunem naukowym 16 prac licencjackich oraz 8 magisterskich. W latach 2016-2017 pełniłam funkcję promotora pomocniczego w przewodzie doktorskim mgr Alicji Mikołajczyk (promotor: prof. dr hab. Tomasz Puzyn).

Ponadto, w latach 2021-2022 aktywnie uczestniczyłam w pracach zespołu programowego dla nowo tworzonej, angielskojęzycznej specjalności studiów II stopnia na kierunku Chemia - „*Digital chemistry*”, w tym w przygotowaniu koncepcji i planu studiów, opracowaniu sylabusów, efektów kształcenia, a także promocji specjalności. W roku akademickim 2022/2023 specjalność „*Digital chemistry*” została uruchomiona z pierwszego naboru.

Tabela 1. Zestawienie prowadzonych zajęć dydaktycznych w latach akademickich 2018/2019 - 2022/2023

Rok akademicki				
2018/2019	**2019/2020**	**2020/2021**	**2021/2022**	**2022/2023**
Liczba godzin pensum dydaktycznego				
120*	120*	120*	210*	165*
* Dzięki uzyskaniu grantu ze środków NCN, w którym pełniłam rolę kierownika projektu, liczba godzin mojego pensum dydaktycznych została zredukowana o 50% (z 240 do 120 godz.)	* Dzięki uzyskaniu grantu ze środków NCN, w którym pełniłam rolę kierownika projektu, liczba godzin mojego pensum dydaktycznych została zredukowana o 50% (z 240 do 120 godz.)	* Dzięki uzyskaniu grantu ze środków NCN, w którym pełniłam rolę kierownika projektu, liczba godzin mojego pensum dydaktycznych została zredukowana o 50% (z 240 do 120 godz.)	* Dzięki uzyskaniu grantu ze środków Horyzont 2020, w którym pełnię rolę kierownika projektu, liczba godzin mojego pensum dydaktycznych została zredukowana o 30 godz. (z 240 do 210 godz.)	* Dzięki uzyskaniu grantu ze środków Horyzont 2020, w którym pełnię rolę kierownika projektu, liczba godzin mojego pensum dydaktycznych została zredukowana o 45 godz. (z 210 do 165 godz.)
Wykładane przedmioty				
▪ Analiza danych wielowymiarowych (ćwiczenia laboratoryjne w pracowni komputerowej, WMFiI, Bioinformatyka) ▪ Chemometria (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Biznes chemiczny) ▪ Statystyka i chemometria w analityce chemicznej (ćwiczenia laboratoryjne w pracowni komputerowe, WCh, Chemia) ▪ Arkusz kalkulacyjny bez tajemnic (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Chemia/Ochrona środowiska)	▪ Analiza danych wielowymiarowych (wykład i ćwiczenia laboratoryjne w pracowni komputerowej, WMFiI, Bioinformatyka) ▪ Chemometria (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Biznes chemiczny) ▪ Arkusz kalkulacyjny bez tajemnic (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Chemia/Ochrona środowiska)	▪ Analiza danych wielowymiarowych (wykład i ćwiczenia laboratoryjne w pracowni komputerowej, WMFiI, Bioinformatyka) ▪ Chemometria (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Biznes chemiczny) ▪ Statystyka i chemometria w analityce chemicznej (ćwiczenia laboratoryjne w pracowni komputerowe, WCh, Chemia) ▪ Arkusz kalkulacyjny bez tajemnic (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Chemia/Ochrona środowiska)	▪ Analiza danych wielowymiarowych (wykład, WMFiI, Bioinformatyka) ▪ Chemometria (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Biznes chemiczny) ▪ Elementy języka R (ćwiczenia laboratoryjne w pracowni komputerowej, WMFiI, Bioinformatyka) ▪ Arkusz kalkulacyjny bez tajemnic (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Chemia/Ochrona środowiska)	▪ Uczenie maszynowe (wykład, WMFiI, Bioinformatyka) ▪ Techniki eksploracji danych wielowymiarowych (wykład, WMFiI, Bioinformatyka) ▪ Statystyka i chemometria w analityce chemicznej (wykład, WCh, Chemia) ▪ Chemometria (ćwiczenia laboratoryjne w pracowni komputerowej, WCh, Biznes chemiczny) ▪ Elementy języka R (ćwiczenia laboratoryjne w pracowni komputerowej, WMFiI, Bioinformatyka)
Promotorstwo prac dyplomowych				
lic. 2		lic. 1	lic. 6	lic. 2
mgr 2		mgr 1	mgr 2	mgr 1
Promotorstwo pomocnicze w przewodzie doktorskim				
			1	

2. DZIAŁALNOŚĆ POPULARYZUJĄCA NAUKĘ

Obok pracy naukowej oraz dydaktycznej podejmowałam również aktywność popularyzującą naukę, w tym m.in.:

- Opublikowanie artykułów w formie wywiadu w prasie o zasięgu ogólnopolskim (m.in. dzienniku Rzeczpospolita, Gazecie Wyborczej, Forum akademickim) oraz w serwisach internetowych (m.in. https://naukawpolsce.pl/, https://innpoland.pl/) przybliżających tematykę komputerowej oceny ryzyka chemicznego i wyzwań z nią związanych np. „Mysz komputerowa zamiast laboratoryjnej", „Niebezpieczne związki okiem chemika", „Ocena ryzyka chemicznego na miarę XXI wieku".
- Wygłoszenie wykładu pt. „Szklanka jest w połowie pusta czy pełna? O bezpieczeństwie chemicznym okiem chemoinformatyka" w ramach otwartych spotkań organizowanych przez Akademia 30+ (luty 2019).
- Od 2011 roku (z przerwami) wraz z innymi wolontariuszami współtworzę Fundację Wspierania Nanonauk i Nanotechnologii NANONET. Misją Fundacji jest promocja i popularyzacja wiedzy oraz wspieranie badań naukowych w obszarze nanotechnologii. Podejmujemy działania mające na celu upowszechnianie wyników badań naukowych i prac rozwojowych oraz wsparcie współpracy pomiędzy sektorem badawczo-rozwojowym a przemysłem.

3. OSIĄGNIĘCIA ORGANIZACYJNE

Mój wkład w działalność organizacyjną obejmuje m.in.:

- Udział w pracach Komisji ds. wdrożenia Europejskiej Karty Naukowca (EKN) i Kodeksu Postępowania przy Rekrutacji Naukowców (C&C) w Uniwersytecie Gdańskim (2016).
- Organizację seminarium naukowego z udziałem dr Ayako Furuhama z *National Institute for Environmental Studies* (Tsukuba, Japonia) w trakcie, którego studenci, doktoranci oraz pracownicy Wydziału Chemii UG mieli okazję wysłuchać wykładu pt. „*Development of chronic aquatic toxicity models based on the quantitative structure–activity–activity relationship (QSAAR) framework*" (2018).
- Udział w pracach zespołu opracowującego strategię rozwoju Wydziału Chemii Uniwersytetu Gdańskiego na lata 2021-2025.

Signed by / Podpisano przez:
Agnieszka Katarzyna Gajewicz-Skrętna
Uniwersytet Gdański
Date / Data: 2022.12.28 06:02.

.....
(podpis wnioskodawcy)