Wojskowa Akademia Techniczna

Wydział Nowych Technogii i Chemii

Wojskowa Akademia Techniczna

Właściwości kwantowe fragmentów cząsteczek jako deskryptory molekularne

Praca doktorska

Bartłomiej Fliszkiewicz

magister inżynier

Promotor

dr hab. inż. Marcin Sajdak
Katedra Ochrony Powietrza
Wydział Inżynierii Środowiska i Energetyki Politechniki Śląskiej

Magdzie i Adasiowi.

„Any alternative viewpoint with a different emphasis leads to an inequivalent description. There is only one reality but there are many viewpoints. It would be very narrowminded to use only one: we have to learn to be able to imagine several."

Hans Primas

w Chemistry, Quantum Mechanics and Reductionism
znaleziono w książce Molecular Descriptors Handbook

Podziękowania

Magdzie za wsparcie nie tylko w pisaniu tej pracy. Profesorowi Witkiewiczowi za pomoc w znalezieniu promotora. Michałowi za polecenie Pythona.

Streszczenie

Właściwości kwantowe fragmentów cząsteczek jako deskryptory molekularne

W niniejszej pracy doktorskiej podjęto się zdefiniowania nowej grupy deskryptorów molekularnych będących pochodną obliczeń metodami chemii kwantowej. Wartości proponowanych deskryptorów wyznaczane są na podstawie właściwości kwantowych wykrytych fragmentów cząsteczek chemicznych. Wyprowadzono trzy definicje nowych deskryptorów molekularnych i zbadano możliwości zbudowania na ich podstawie modeli struktura - aktywność lub właściwość na podstawie w sumie 13 różnych dostępnych baz danych. Ponadto, wykonano obliczenia kwantowe dla 4326 związków z których 4082 zakończyło się sukcesem, co pozwoliło na rozszerzenie zbioru dostępnych fragmentów związków. Przeprowadzone badania wykazały, że modele uczenia maszynowego zbudowane w oparciu o deskryptory molekularne zdefiniowane na podstawie właściwości kwantowych fragmentów związków dają predykcje dokładniejsze niż linie bazowe. W ramach prowadzonych badań przedstawiono jedną definicję, stanowiącą modyfikację znanych w chemoinformatyce par atomów. Zaproponowana definicja pozwoliła na dokładniejsze przewidywanie aktywności lub właściwości niż część tradycyjnych, ugruntowanych już metod chemoinformatycznych.

Abstract

Quantum properties of molecular fragments as molecular descriptors

The aim of the following PhD Thesis was to define a new group of molecular descriptors that are derived from computational quantum chemistry. The values of the proposed descriptors are calculated from quantum properties of detected molecular fragments. Three definitions of the new molecular descriptors are introduced and the possibility to build upon them structure - activity or structure - property predictive models is estimated based on 13 different databases. Furthermore, during the realization of the PhD Thesis additional quantum chemistry calculations were conducted on 4326 molecules of which 4082 succeeded, which allowed to expand the number of possible molecular fragments that are searched for. The conducted research showed that machine learning models built upon the descriptors derived from quantum properties of molecular fragments yield predictions superior than the baseline. As part of the conducted research, one definition was presented, which is a modification of atom pairs known in cheminformatics. The proposed definition allowed more accurate prediction of activity or properties than some of the traditional, well-established chemoinformatics methods.

Spis treści

Spis rysunków

Spis tabel

Spis skrótów

CLI	**C**ommand **L**ine **I**nterface
IUPAC	**I**nternational **U**nion of **P**ure and **A**pplied **C**hemistry
WLN	**W**iswesser **L**ine **N**otation
ROSDAL	**R**epresentation of **O**rganic **S**tructure **D**escription **A**rranged **L**inearly
SMILES	**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification
SLN	**S**YBYL **L**ine **N**otation
InChI	**In**ternational **Ch**emical **I**dentifier
SELFIES	**SELF** referenc**I**ng **E**mbedded **S**trings
SMARTS	**SM**ILES **AR**bitrary **T**arget **S**pecification
MACCS	**M**olecular **ACC**ess **S**ystem
ECFP	**E**xtended **C**onnectivity **F**inger**P**rints
SDfile	**S**tructure-**D**ata file
PDB	**P**rotein **D**ata **B**ank
mmCIF	**m**acro**m**olecule **C**rystallographic **I**nformation **F**ile
QSAR	**Q**antitative **S**tructure-**A**ctivity **R**elationship
QSPR	**Q**antitative **S**tructure-**P**roperty **R**elationship
WHIM	**W**eighted **H**olistic **I**nvariant **M**olecular
CATS	**C**emically **A**dvanced **T**emplate **S**earch
DABE	**D**onor **A**cceptor **B**ulkiness **E**lectropositivity
SAscore	**S**ynthetic **A**ccessibility score
GNN	**G**raph **N**eural **N**etwork
MPNN	**M**essage **P**assing **N**eural **N**etwork
GAN	**G**eneric **A**dversarial **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
t-SNE	**t-**Distributed **S**tochastic **N**eighbour **E**mbedding
MSE	**M**ean **S**quared **E**rror
RMSE	**R**oot **M**ean **S**quared **E**rror
MAE	**M**ean **A**bsolute **E**rror

ROC	**R**eceiver **O**perating **C**haracteristic
OLED	**O**rganic **L**ight **E**mitting **D**iode
HOMO	**H**ighest **O**ccupied **M**olecular **O**rbital
LUMO	**L**owest **U**noccupied **M**olecular **O**rbital
TPSA	**T**otal **P**olar **S**urface **A**rea
VSA	**V**an der Waals **S**urface **A**rea
GBR	**G**radient **B**oosting **R**egressor
API	**A**pplication **P**rogramming **I**nterface
ETKDG	**E**xperimental-**T**orsion basic **K**nowledge **D**istance **G**eometry
B3LYP	**B**ecke **3**-parameter **L**ee-**Y**ang-**P**arr
CCSD	**C**upled **C**luster **S**ingle-**D**ouble
def2-svp	def2-**S**plit **V**alence **P**olarization
RRHO	**R**igid-**R**otor **H**armonic **O**scillator
BCF	**B**ioa**C**cumulation **F**actor
PCA	**P**rincipal **C**omponents **A**nalysis
XGBoost, XGB	e**X**treme **G**radient **Boost**ing
LightGBM, LGBM	**Light G**radient **B**oosting **M**achine
SVC	**S**upport **V**ector **C**lassifier
KNN	**K**-**N**earest **N**eighbours
FQD	**F**ragments **Q**uantum **D**escriptor
PC	**P**rincipal **C**omponent
QAP	**Q**uantum **A**tom **P**airs
BACE-1	**B**eta-site **A**PP **C**leaving **E**nzyme **1**
BBBP	**B**lood-**B**rain **B**arrier **P**enetration
hERG	**h**uman **E**ther-à-go-go-**R**elated **G**ene
IC50	half maximal **I**nhibitory **C**oncentration
AUC	**A**rea **U**nder **C**urve
BRICS	**B**reaking of **R**etrosynthetically **I**nteresting **C**hemical **S**ubstructures
RECAP	**R**etrosynthetic **C**ombinatorial **A**nalysis **P**rocedure
CSV	**C**omma **S**eparated **F**ile
RF	**R**andom **F**orest

Teza i cel pracy

Zastosowanie komputerów w chemii można podzielić na dwie główne grupy: obliczenia kwantowo-chemiczne oraz chemoinformatykę. Pierwsze zastosowanie, pod warunkiem odpowiedniego wykonania charakteryzuje się wysoką dokładnością, jednak wymaga umiejętności, doświadczenia, mocy obliczeniowej i czasu. Alternatywą są zastosowania chemoinformatyczne, które kosztem dokładności umożliwiają szybkie uzyskiwanie wyników, można je zaadaptować w metodach przesiewowych i pozwalają na stworzenie rozwiązania typu „czarna skrzynka" nie wymagających posiadania dużych umiejętności przez użytkownika. Niniejsza praca skupiona jest na poszukiwaniach takiego połączenia rozwiązań chemoinformatycznych z metodami obliczeń kwantowo-chemicznych, aby móc skorzystać z zalet obu grup zastosowań komputerów w chemii. W pracy stawiana jest poniższa teza:

Możliwe jest zbudowanie prostych i dokładnych modeli predykcyjnych lub klasyfikacyjnych, w oparciu o właściwości kwantowe fragmentów cząsteczek chemicznych pochodzących z wcześniej zdefiniowanego zbioru fragmentów opisującego ich właściwości kwantowe.

Pozytywna realizacja tej tezy ma potencjał na stworzenie narzędzi, które wzbogacą istniejący zbiór dostępnych metod chemoinformatycznych.

Nadrzędnym celem pracy jest zdefiniowanie nowych deskryptorów molekularnych opartych o właściwości kwantowe fragmentów cząsteczek chemicznych. Realizacja celu obejmuje zastosowanie dostępnych baz danych zarówno właściwości kwantowych jak i eksperymentalnych w celu budowania modeli QSAR/QSPR. Oprócz definicji samych deskryptorów, praca doktorska obejmuje również identyfikację ograniczeń i potencjalnych błędów mogących wystąpić w takim podejściu. Ponadto sformułowano następujące cele pośrednie:

- dokonanie rozszerzenia bazy danych właściwości kwantowych,
- optymalizacja wykonywania dużej liczby obliczeń kwantowych,
- stworzenie oprogramowania pozwalającego na obliczanie deskryptorów kwantowych opracowanych na potrzeby tej pracy.

Część I

Przegląd stanu wiedzy

Różnorodność sposobów przedstawienia cząsteczki chemicznej

W chemii istnieje wiele sposobów przekazywania informacji dotyczących związków chemicznych. Niektórych nazw zwyczajowych człowiek uczy się jeszcze jako dziecko - np. sól kuchenna. Jednak od czasu kiedy nazwy zwyczajowe były jedynymi nazwami liczba znanych związków chemicznych wzrosła diametralnie. Wzrosła też świadomość dotycząca budowy cząsteczek chemicznych. W celu usystematyzowania nazewnictwa związków chemicznych, z inicjatywy IUPAC, powstały nazwy systematyczne. Jednak nie są one wzajemnie jednoznaczne - jeden związek chemiczny może posiadać więcej niż jedną, poprawną nazwę systematyczną. Ponadto, w zastosowaniach w bazach danych, nazwy systematyczne tylko częściowo pozwalają na przeszukiwanie po fragmentach związków chemicznych. Powstaje zatem potrzeba znalezienia takich sposobów przedstawienia cząsteczki chemicznej, która by uwzględniała skład atomowy związku chemicznego. Najprostszym sposobem jest wzór sumaryczny, jednak każdy chemik wie, że szczególnie w przypadku związków organicznych, wzór sumaryczny często jest niezwystarczający do opisu struktury cząsteczki chemicznej ze względu na swoją niejednoznaczność. Zjawisko, kiedy wzór sumaryczny pasuje do kilku związków chemicznych nazywane jest izomerią.

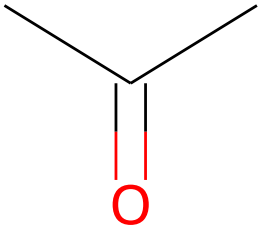
W rozdziale przedstawiono wybrane, historyczne i najczęściej spotykane sposoby reprezentacji, przechowywania i przetwarzania informacji dotyczących struktur chemicznych w systemach informatycznych. Niestety nie ma jednego, standardowego formatu stosowanego w chemoinformatyce, a część rozwiązań i algorytmów nie jest publicznie dostępnych, co sprawia, że można się spotkać z implementacjami będącymi próbą utworzenia podobnego rozwiązania.

Niektóre z przytoczonych w niniejszym rozdziale sposobów przedstawienia struktur chemicznych leżą u podstawy wyznaczania wartości deskryptorów molekularnych, inne są bardzo pomocne w poszukiwaniach podobnych struktur chemicznych lub też na potrzeby przechowywania danych w bazach danych.

Przedstawiony w rozdziale przegląd można poszerzyć np. o artykuł Wigh i wsp. [1].

1.1 Notacje liniowe

Notacje liniowe charakteryzują się tym, że opis struktury składa się ze znaków alfanumerycznych zapisanych w jednej linii. Została ona zaproponowana w latach 1950 - 1970. Wówczas powstały notacje liniowe Wiswesser (WLN) [2] oraz ROSDAL [3]. W latach osiemdziesiątych powstały notacje SMILES [4–6] i SLN. Rysunek 1.1 przedstawia cząsteczkę acetonu i jej strukturę zakodowaną w czterech notacjach liniowych.



SMILES	CC(C)=O
InChI	InChI=1S/C3H6O/c1-3(2)4/h1-2H3
Klucz InChI	CSCPPACGZOOCGX-UHFFFAOYSA-N
SELFIES	[C][C][Branch1][C][C][=O]

RYSUNEK 1.1: Aceton i jego reprezentacje w różnych notacjach liniowych.

1.1.1 SMILES i SMARTS

Obecnie najbardziej popularna notacja liniowa. Charakteryzuje się dość zwartą konstrukcją i jest stosunkowo czytelna dla użytkownika. Ciąg SMILES oparty jest o następujące zasady:

1. atomy reprezentowane są przez symbole ich związków chemicznych, przy czym należy pamiętać, że jeżeli dany pierwiastek przedstawiany jest symbolem dwuliterowym, wówczas należy umieścić go w nawiasie kwadratowym,

2. atomy wodoru są pomijane w notacji SMILES - zakłada się, że są one dopełnieniem wartościowości atomów,

3. atomy sąsiadujące ze sobą umieszczone są obok siebie, dwie niepołączone ze sobą części cząsteczki oddzielone są kropką,

4. zaznacza się tylko wiązania podwójne symbolem „=” oraz potrójne symbolem „#”; pierścienie aromatyczne określone są poprzez symbole atomów występujących w tym pierścieniu napisane małymi literami,

5. rozgałęzienia zaznacza się nawiasami okrągłymi,

6. pierścienie zaznacza się liczbami symbolizującymi, które atomy danego pierścienia są ze sobą połączone.

Ponadto notacja SMILES pozwala na zaznaczenie stereochemii za pomocą symboli „@” lub „@@” oraz stereoizomerów względem wiązań podwójnych symbolami „\” oraz „/”. Tak zdefiniowana notacja liniowa nie zapewnia jednak całkowitej jednoznaczności, ponieważ ten sam związek może być zapisany na kilka różnych sposobów (np. aceton może być również zapisany jako CC(=O)C). W celu rozwiązania tego problemu stworzony został algorytm generujący tzw. kanoniczny SMILES.

Pierwotnie celem notacji SMILES był opis struktury cząsteczek chemicznych. Do bardziej specjalistycznych zastosowań powstały jej odmiany i rozszerzenia, np. SMIRKS do kodowania reakcji chemicznych, CHUCKLES [7] do kodowania struktury peptydów, SMARTS do precyzyjnego definiowania wzorów do poszukiwania fragmentów cząsteczek chemicznych i inne.

SMARTS jest rozszerzeniem notacji SMILES pozwalająca na dokładną deklarację poszukiwanych fragmentów. Wprowadzone zmiany obejmują znaki dopasowujące dowolny atom, operacje logiczne, deklaracje pierścieni o konkretnej liczbie atomów, a także znaki dotyczące dopasowania rodzaju wiązania. Dokładny opis wraz z przykładami można znaleźć na stronie internetowej Daylight Chemical Information Systems, Inc. [8]

1.1.2 SLN

Notacja liniowa SYBYL przeznaczona jest do określania struktury związków chemicznych, wzorów dopasowań fragmentów i reakcji przy użyciu jednej notacji [9, 10]. Jest w pewnym stopniu podobna do SMILES, występują jednak różnice w traktowaniu atomów wodoru (wszystkie muszą być zaznaczone w notacji), zapisu aromatyczności (kodowana przez znak „:”), początku i końca pierścieni oraz stereochemii. Ciągi SLN są dłuższe niż ich odpowiedniki zapisane w notacji SMILES.

1.1.3 InChI

Od 2000 roku IUPAC podjęło się opracowania wzajemnie jednoznacznego identyfikatora związków chemicznych, który z powodzeniem mógłby być stosowany do zastosowania np. w bazach danych. Stale rozwijany, otwartoźródłowy algorytm InChI [11–13] obejmuje następujące kroki:

1. normalizacja - związek chemiczny zamieniany jest w jednoznaczną strukturę,
2. kanonikalizacja - numeracja atomów w związku przebiega zgodnie z algorytmem Morgana, niezależnie od tego jak struktura została narysowana,
3. serializacja - zamiana na alfanumeryczny ciąg znaków - InChI.

Dzięki otwartości algorytmu może on być swobodnie implementowany w edytorach molekuł w ten sam sposób, co gwarantuje takie same identyfikatory niezależnie od zastosowanego narzędzia. Wyszukiwarka Google pozwala na podanie InChI jako zapytanie do wyszukania. Identyfikator InChI składa się z następujących po sobie, tzw. warstw oddzielonych od siebie znakami „/" i oznaczonymi literami. Są to kolejno:

1. warstwa główna - zawiera trzy podwarstwy: wzoru empirycznego, połączeń między atomami (/c), atomów wodoru (/h),
2. warstwa ładunków - składa się z dwóch podwarstw /q ładunków i /p protonów,
3. warstwa stereochemii,
4. warstwa izotopów,
5. warstwa niemobilnych atomów wodoru - w celu opisania jednej konkretnej formy tautomeru,
6. warstwa do oznaczania wiązań z metalami (np. w związkach organometalicznych).

Ze względu na obecność warstw w InChI dokonany został podział na standardowy InChI zawierający tylko warstwy 1 - 4 i niestandardowy, mogący zawierać warstwy 5 i 6. Identyfikator InChI nie jest tak łatwy do odczytania przez człowieka jak SMILES i zawiera dużo informacji, co powoduje, że wraz ze wzrostem cząsteczki

związku chemicznego jego długość rośnie, co może utrudniać ich zastosowanie. W celu skrócenia ciągu alfanumerycznego stworzono klucz InChI poprzez zastosowanie funkcji haszującącej. Klucz ten składa się z pięciu elementów:

1. 14 wielkich liter będących wynikiem haszowania podwarstwy połączeń pomiędzy atomami,
2. 8 wielkich liter otrzymanych z haszowania pozostałych podwarstw,
3. litera oznaczająca, czy klucz powstał ze standardowego identyfikatora InChI,
4. litera oznaczająca wersję algorytmu tworzącego InChI,
5. litera oznaczająca, czy cząsteczka posiada neutralną liczbę atomów wodoru, ich nadmiar lub niedomiar.

Klucz podzielony jest myślnikami na 3 części - pomiędzy elementami 1, a 2 oraz 4 i 5.

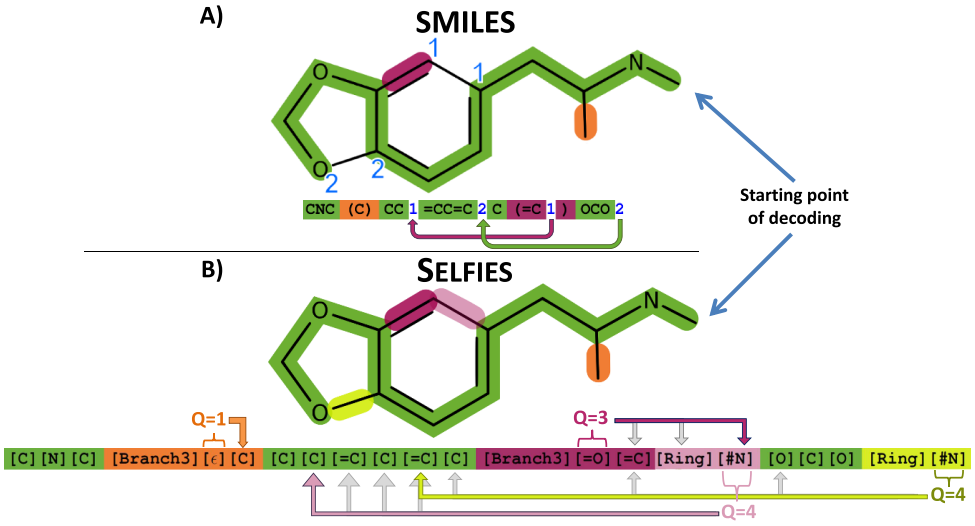
Opracowywane są również odmiany identyfikatora InChI - RInChI [14] do opisu reakcji chemicznych oraz MInChI do opisu mieszanin.

1.1.4 SELFIES

Krenn M i wsp. chcąc stworzyć reprezentację pozwalającą na aplikację wprost do modeli opartych o sieci neuronowe (więcej na ten temat znajduje się w rozdziale 2.3) zaproponowali nową notację liniową, którą nazwali SELFIES [15]. Podstawowym problemem, który badacze chcieli rozwiązać był fakt, że budując takie modele w oparciu o reprezentację SMILES, w dużej części otrzymywano błędne wyniki - SMILES związków łamiących podstawowe prawa chemiczne lub nie odpowiadające żadnemu grafowi molekularnemu.

Samoodnoszenie się SELFIES objawia się tym, że przestawienie lub usunięcie części notacji nie powoduje błędów (jak to się często zdarza w przypadku SMILES), lecz tworzy nową cząsteczkę. Można by powiedzieć, że notacja ta jest „świadoma" wartościowości kodowanych atomów - nawet jeśli w notacji zakodowane jest wiązanie potrójne i złamałoby to zasady walencyjności, to przełożenie SELFIES na graf molekularny poskutkuje taką zmianą struktury, aby cząsteczka spełniała te zasady. Ponadto atomy zamykające pierścienie zakodowane są względem innych atomów, a nie konkretnie wpisane w notację jak w przypadku SMILES.

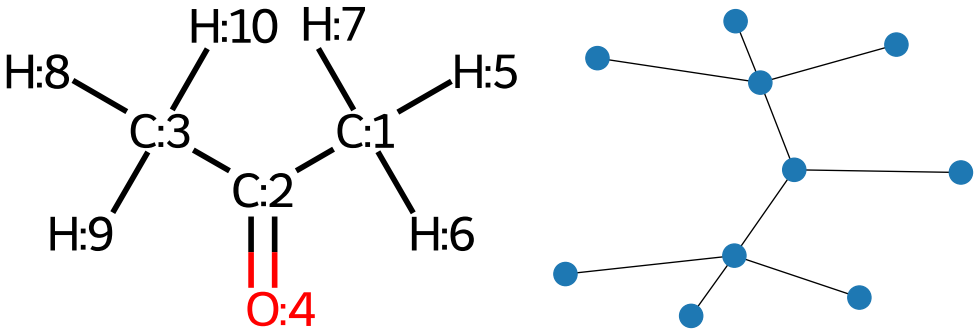
Porównanie SMILES i SELFIES przedstawia rysunek 1.2.



RYSUNEK 1.2: Porównanie SMILES i SELFIES. Ten sam związek przedstawiony w postaci A) SMILES i B) SELFIES. Rysunek pochodzi z [15].

1.2 Grafy molekularne

Widząc podobieństwo rysowanych przez chemików struktur związków chemicznych do znanych w matematyce grafów, zaadaptowano ich teorię do opisu struktur cząsteczek [16–20]. W wierzchołku grafu molekularnego znajdują się atomy tworzące związek chemiczny, a jego krawędzie to wiązania chemiczne. W grafach najczęściej pomija się atomy wodoru połączone z atomami węgla. Przyrównanie cząsteczek do grafów pozwala na matematyczne przetwarzanie związków chemicznych na liczby. Przykład cząsteczki chemicznej i jej graf molekularny pokazane są na rysunku 1.3. Zastosowanie teorii grafów w chemii pozwala np. na wyznaczenie szeregu deskryptorów molekularnych, które opisane są w rozdziale 2.



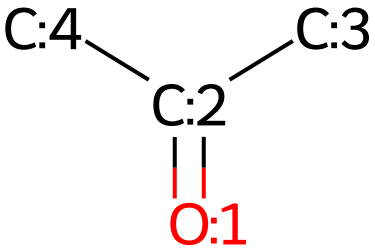
RYSUNEK 1.3: Cząsteczka acetonu i jej graf.

1.3 Postać macierzowa

Grafy można przedstawić w postaci macierzy, co pozwala na zastosowanie operacji macierzowych i algebry liniowej. W odniesieniu do grafów chemicznych w zależności od informacji zawartych w poszczególnych elementach macierzy można wymienić następujące ich rodzaje: macierz sąsiedztwa, incydencji, odległości, wiązań oraz wiązań i elektronów. Z wyjątkiem macierzy incydencji, macierze te są redundantne - informacje są podwojone, co pozwala na zachowanie tylko połowy macierzy, po jednej ze stron przekątnej. Możliwe jest również pominięcie atomów wodoru w reprezentacji macierzowej, ponieważ zakłada się, że w razie potrzeby zostaną one uzupełnione dzięki regułom walencyjności pozostałych atomów.

Macierz sąsiedztwa i macierz wiązań

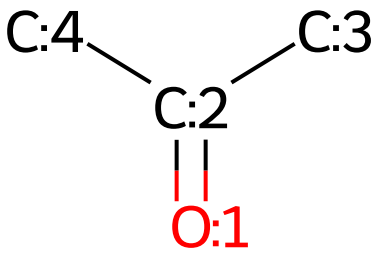
W macierzy sąsiedztwa kolumny i wiersze odpowiadają atomom, a ponadto dla związków n-atomowych jest macierzą kwadratową nxn. Elementy macierzy przyjmują wartości 1, kiedy atomy odpowiadające danej kolumnie i wierszowi są ze sobą połączone wiązaniem chemicznym lub 0, gdy takiego połączenia nie ma. Przekątna takiej macierzy zawsze zawiera wartości 0. Ze względu na brak informacji o krotności wiązań chemicznych, stereochemii oraz wolnych elektronach, ten rodzaj reprezentacji macierzowej jest mało przydatny. Macierz sąsiedztwa przedstawiona została na rysunku 1.4



	1	2	3	4
1	0	1	0	0
2	1	0	1	1
3	0	1	0	0
4	0	1	0	0

RYSUNEK 1.4: Macierz sąsiedztwa.

Odmianą macierzy sąsiedztwa jest przedstawiona na rysunku 1.5 macierz wiązań, w której niezerowe elementy niosą informację o krotności wiązania chemicznego.

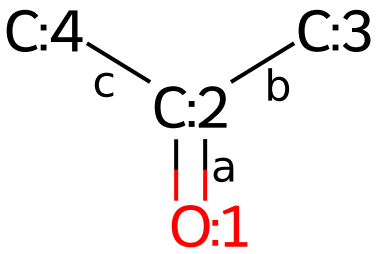


	1	2	3	4
1	0	2	0	0
2	2	0	1	1
3	0	1	0	0
4	0	1	0	0

RYSUNEK 1.5: Macierz wiązań.

Macierz incydencji

Przedstawiona na rysunku 1.6 macierz incydencji jest jedyną niekwadratową reprezentacją postaci macierzowych o wymiarach nxm, przedstawiająca n-atomowy związek, w którym występuje m wiązań chemicznych (krawędzi). Elementy macierzy przyjmują wartości 1, gdy na którymś z końców danego wiązania chemicznego znajduje się dany atom, lub 0, gdy takiego połączenia nie ma. W teorii grafów istnieje jeszcze wartość -1, w przypadku grafów skierowanych, co nie występuje w grafach molekularnych.

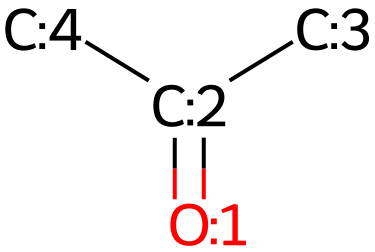


	1	2	3	4
a	1	1	0	0
b	0	1	1	0
c	0	1	0	1

RYSUNEK 1.6: Macierz incydencji.

Macierz odległości

Ten rodzaj macierzy kwadratowej w swoich elementach zawiera informacje o odległościach między poszczególnymi atomami. Istnieją dwie odmiany tej reprezentacji - podana jest odległość geometryczna w Å lub odległość topologiczna w liczbie krawędzi (wiązań) oddzielających dwa atomy. W odmianie topologicznej macierz odległości przedstawiona została na rysunku 1.7.

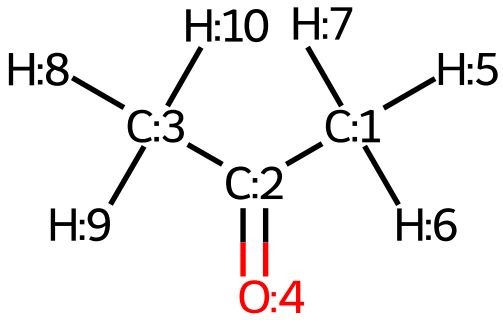


	1	2	3	4
1	0	1	2	2
2	1	0	1	1
3	2	1	0	2
4	2	1	2	0

RYSUNEK 1.7: Macierz odległości topologicznej.

Macierz wiązań i elektronów

Pochodząca z modelu Dugundji-Ugi [21] macierz kwadratowa *nxn* opisująca n atomową cząsteczkę. Jest to jedyna macierz, której elementy na przekątnej nie muszą być zerami, tylko przedstawiają liczbę wolnych elektronów walencyjnych danego atomu. Przykład takiej macierzy przedstawia rysunek 1.8.



	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	1	1	1	0	0	0
2	1	0	1	2	0	0	0	0	0	0
3	0	1	0	0	0	0	0	1	1	1
4	0	2	0	4	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0	0	0

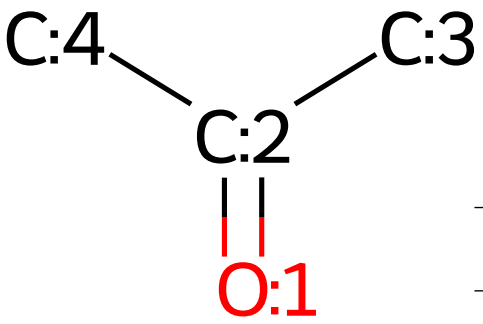
RYSUNEK 1.8: Macierz wiązań i elektronów.

Macierz ta posiada ciekawe właściwości matematyczne, które mają swoje odzwierciedlene w chemii:

- suma wszystkich elementów w kolumnie lub wierszu jest liczbą elektronów walencyjnych danego atomu,
- suma wszystkich elementów macierzy jest liczbą elektronów walencyjnych całej cząsteczki,
- jeśli suma elementów w kolumnie lub wierszu nie odpowiada właściwej wartości, wówczas na danym atomie w cząsteczce znajduje się ładunek,
- suma elementów w i-tej kolumnie i i-tym wierszu pomniejszona o element znajdujący się na ich przecięciu powinna dawać całkowitą liczbę elektronów walencyjnych danego atomu w cząsteczce (można sprawdzić, czy spełniona jest reguła oktetu).

1.4 Tablica połączeń

Reprezentacje macierzowe gwałtownie zwiększają swoje wymiary wraz ze wrostem liczby atomów w związku chemicznym. By zaradzić temu problemowi opracowano bardziej zwięzły sposób przechowywania informacji o strukturze chemicznej - tablicę połączeń. Taka tablica powstaje z grafu molekularnego i może być skonstruowana na 2 sposoby - jako dwie osobne tabele, z których jedna zawiera listę atomów z przydzielonymi im indeksami, a druga tabela z listą wiązań (wraz z krotnością) oraz indeksami atomów połączonych danym wiązaniem. Drugim sposobem konstrukcji tablicy połączeń jest jedna, kombinowana tabela zawierająca informacje o atomach w związku chemicznym wraz z indeksami atomów z którymi są połączone oraz krotnością tych wiązań. Tak skonstruowana reprezentacja cząsteczki rośnie w sposób liniowy z liczbą atomów w związku chemicznym, a nie z kwadratem liczby atomów, tak jak to ma miejsce w przypadku reprezentacji macierzowych. Rysunek 1.9 przedstawia przykładową tablicę połączeń acetonu.



Tablica Atomów

indeks	pierwiastek
1	O
2	C
3	C
4	C

Tablica Wiązań

indeks		krotność
pierw. 1	pierw. 2	wiązania
1	2	2
2	3	1
2	4	1

Atomy		Wiązania	
indeks	pierwiastek	indeks	krotność
1	O	2	2
2	C	4	1
3	C	2	1
4	C		

RYSUNEK 1.9: Tablica połączeń. U dołu znajduje się wersja kombinowana.

1.5 Problem braku unikatowości - kanonikalizacja

Większość spośród wymienionych już sposobów reprezentacji związków chemicznych jest jednoznaczna - z danej reprezentacji można wygenerować tylko jeden graf molekularny. Jednak istnieje problem unikatowości, czyli z jednego grafu molekularnego można wygenerować więcej niż jeden ciąg (w przypadku notacji liniowych) lub więcej niż jedną macierz lub tablicę połączeń. Jako przykład można przytoczyć tablicę połączeń, dla związku o n liczbie atomów istnieje $n!$ sposobów indeksowania atomów, co w przypadku związków o budowie niesymetrycznej pozwala na skonstruowanie $n!$ różnych tablic połączeń. Sprawia to, że bezpośrednie porównywanie dwóch tablic połączeń może być obarczone błędem, np. w poszukiwaniu duplikatów związków w bazie danych. W celu rozwiązania tego problemu stworzony został algorytm Morgana [22], który w szeregu kroków nadaje indeksy atomom w związku chemicznym. Proces taki nazywany jest kanonikalizacją.

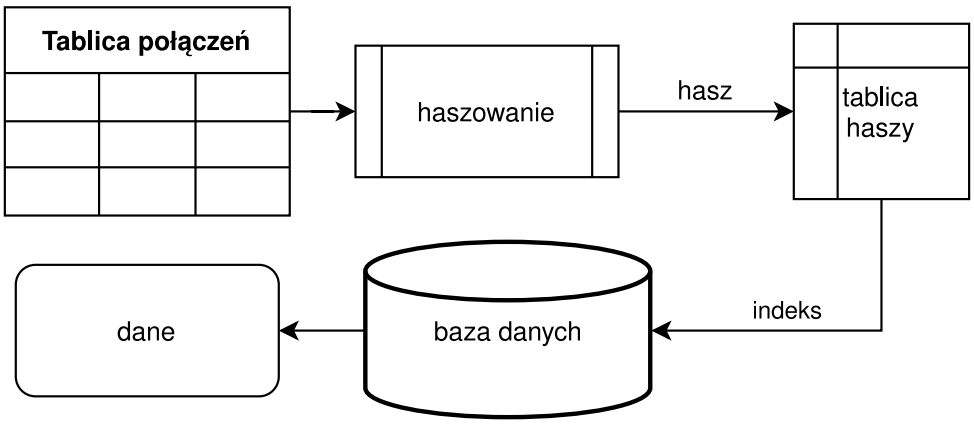
1.6 Kodowanie struktury w innych formach

Oprócz przechowywania konkretnej struktury związku chemicznego w bazie danych, użyteczne mogą okazać się również inne metody opisu cząsteczek związków chemicznych. Wymienione poniżej metody reprezentacji są metodami stratnymi - na ich podstawie nie można wywnioskować, z jakich cząsteczek zostały wygenerowane, choć prowadzone są badania, aby to umożliwić [23].

Hasze

Haszowanie jest procesem kodowania informacji o różnej wymiarowości do alfanumerycznych ciągów o ściśle określonej długości przy użyciu funkcji haszującej. Dobre funkcje haszujące powinny zapewnić jednokierunkowość procesu - z wygenerowanego ciągu nie można uzyskać zahaszowanych danych, co jest szczególnie przydatne w bazach haseł. Wśród zastosowań haszy można wymienić zastosowanie w bazach danych, do przechowywania indeksów, celem szybszego dostępu do danych [24], w kryptografii [25] i w ograniczaniu redystrybucji multimediów [26]. Dzięki funkcjom haszującym można w szybki sposób przeszukać dużą bazę danych chemicznych w poszukiwaniu związku chemicznego, bez porównywania wielu tablic połączeń. Schematycznie uzyskiwanie danych w ten sposób przedstawiono na

rysunku 1.10. Zapytanie w formie tablicy połączeń w pierwszej kolejności jest kodowane, otrzymany hasz porównuje się z tablicą haszy i z niej uzyskuje się indeks bazy danych odpowiadający elementowi tej bazy zawierającemu dane związku przedstawionego przez tablicę połączeń będącą zapytaniem. Jedną z wad haszy jest fakt, że zdarzają się zjawiska kolizji - różne dane wejściowe funkcji haszującej mogą dać ten sam hasz.



RYSUNEK 1.10: Zastosowanie haszowania do odzyskiwania danych z bazy danych.

Kody fragmentów

Jednym ze sposobów innego podejścia do opisu cząsteczki jest wygenerowanie wektora zer i jedynek oznaczającego występowanie jakiś konkretnych, zadeklarowanych wcześniej fragmentów. Taki sposób można porównać do słów - kluczy dotyczących np. artykułu naukowego. Ten sposób reprezentacji nie jest jednoznaczny, jednak pełni ważną rolę w wyszukiwaniu związków podobnych strukturalnie lub zawierających tą samą grupę funkcyjną (jeśli jest obecna w zbiorze fragmentów). Przykładami mogą być klucze MACCS [27] (spośród których 166 jest dostępnych publicznie) lub odciski palców PubChem [28, 29].

Odciski palców

Istnieje wiele rodzajów chemicznych odcisków palców, np. cieszące się dużą popularnością odciski palców rozszerzonych połączeń (ECFP) i odciski palców par atomów [30]. Charakterystyczną cechą odcisków palców jest to, że są one generowane ze struktury związków, a nie spośród z góry ustalonych fragmentów. Podobnie jak w przypadku kodowania fragmentów otrzymywana jest reprezentacja wektorowa.

Najczęściej spotykane odciski palców ECFP, zwane również odciskami palców Morgana lub też okrężnymi, powstają poprzez iteracyjne generowanie podstruktur o coraz większych zasięgach względem poszczególnych atomów. Ten sposób podobny jest do algorytmu Morgana, skąd wzięła się alternatywna nazwa. W czasie wyznaczania, kilkukrotnie stosowane są funkcje haszujące. Wyznaczone, niepowtarzające się podstruktury są ostatecznie haszowane do wektorów binarnych, by na koniec otrzymać jeden wektor poprzez zastosowanie logicznej operacji „lub" (choć implementacje mogą się różnić).

1.7 Formaty plików zawierających struktury chemiczne

Dynamicznie rosnąca liczba znanych związków chemicznych sprawia, że gromadzenie informacji o ich właściwościach i strukturach staje się wyzwaniem, a tak duża liczba postaci przedstawienia struktur cząsteczek przekłada się na mnogość sposobów reprezentacji cząsteczki w systemach informatycznych. Poniżej przedstawiono najczęściej występujące formaty plików, wraz z rozszerzeniami oraz krótkim opisem. W zasadzie wszystkie te pliki są zwykłymi plikami tekstowymi, jednak rozszerzenie przydaje się w celu identyfikacji, co znajduje się wewnątrz takiego pliku. Zazwyczaj pozwalają na zapis struktury wielu cząsteczek w jednym pliku. Większość z nich może przechowywać dane topografii struktury (struktura trójwymiarowa).

Pliki z notacjami liniowymi

Wymienione wcześniej notacje liniowe mają swoje dedykowane rozszerzenia plików:

- SMILES: .smi lub .smiles,
- SMARTS: .sma lub .smarts.

W związku z tym, że notacje liniowe kodują tylko topologię cząsteczki, w powyższych plikach nie ma informacji o topografii.

Format XYZ

Pliki z przyrostkiem .xyz przechowują współrzędne kartezjańskie trójwymiarowej struktury związku chemicznego. Schemat zawartości jest następujący: w pierwszej linijce znajduje się sumaryczna liczba atomów w związku, w drugiej linijce znajduje się komentarz, np. nazwa związku, dane jednostki naukowej, lub można tą linijkę wykorzystać do przekazania właściwości związku. Od trzeciej linijki pliku .xyz, począwszy od symbolu pierwiastka znajdują się współrzędne kartezjańskie poszczególnych atomów wyrażone w Å - po jednym atomie na linijkę.

Molfile

Sam format został sformułowany przez MDL Information Systems (obecnie BIOVIA), dlatego też często występuje pod nazwą MDL molfile. Plik ten występuje z rozszerzeniem .mol i przechowuje zmodyfikowaną tablicę połączeń pojedynczego związku chemicznego. Całość ma ściśle określony schemat dotyczący informacji, które znajdują się w poszczególnych liniach, występujący w dwóch różnych formatach V2000 lub V3000, opisane w specyfikacji [31]. Modyfikacja tablicy połączeń polega na tym, że oprócz symboli pierwiastków zawiera ona również ich współrzędne kartezjańskie. Format ten umożliwia przechowywanie topologii (wówczas jedna ze współrzędnych jest zerami) lub topografii związku chemicznego.

SDfile

Rozszerzenie .sdf wskazuje na plik zawierający strukturę i dodatkowe, opcjonalne dane związków chemicznych. Ten format plików jest swoistym opakowaniem co najmniej jednego pliku .mol. Po każdym wystąpieniu tablicy połączeń może być obecny blok danych niestrukturalnych. Wpisy dotyczące poszczególnych związków chemicznych oddzielone są ciągiem „$$$$”.

Format PDB

Format stworzony specjalnie na potrzeby kodowania struktury białek i kwasów nukeinowych w banku danych protein (PDB). Początkowo ta baza danych znajdowała się w Brookhaven National Laboratory [32], a następnie przeniesiona została do Research Collaboratory for Structural Bioinformatics [33, 34]. W związku z tym, że format ten wywodzi się z czasów kart dziurkowanych, szablon zapisu danych

ograniczony jest do 80 znaków w linii. Plik z rozszerzeniem .pdb może przechowywać informacje o topografii wszystkich atomów struktury zapisanej w pliku, pierwszo lub drugorzędową strukturę białek, dane bibliograficzne, krystalograficzne i widma NMR. Nie ma również ograniczenia co do liczby cząsteczek chemicznych, które są zapisane w pliku. Format ten nie jest już rozwijany, ze względu na zastąpienie go formatem mmCIF [35], jednak jest on szeroko dostępny w bazie PDB.

Bogaty świat deskryptorów molekularnych

Liczba deskryptorów molekularnych jest tak duża, że ich opis w książce Molecular Descriptors for Chemoinformatics napisanej przez Todeschini i Consonni [36] zajmuje ponad 1300 stron, opisuje 3000 deskryptorów molekularnych i zawiera 6400 cytowań. W niniejszym rozdziale przedstawiono różne grupy deskryptorów molekularnych oraz cechy, którymi powinny się charakteryzować, aby uznać je za użyteczne.

Deskryptor molekularny można zdefiniować jako ostateczny wynik operacji logicznej lub matematycznej, która zamienia informację chemiczna zawartą w symbolicznej reprezentacji cząsteczki w użyteczną liczbę lub wynik jakiegoś wystandaryzowanego eksperymentu [37].

Pomimo dużej ich liczby, zbiór deskryptorów molekularnych nie jest skończony i możliwe jest pojawianie się nowych wielkości, które będą kodowały jakieś informacje dotyczące struktury związków chemicznych.

Wymagania stawiane deskryptorom molekularnym [38] są następujące:

1. wartość nie może zależeć od numeracji atomów w cząsteczce,
2. niezmienność pod wpływem rotacji cząsteczki,
3. zdefiniowany jest przez jednoznaczny algorytm,
4. posiada dobrze zdefiniowaną aplikację.

Ponadto pożądane jest, aby deskryptor molekularny:

1. miał interpretację strukturalną,
2. posiadał dobrą korelację z przynajmniej jedną właściwością eksperymentalną,
3. nie był zależny od innego deskryptoru molekularnego,

4. nie opierał się na właściwościach eksperymentalnych,
5. był zmienną ciągłą,
6. wykazywał jak najmniejszą degenerację,
7. charakteryzował się prostotą,
8. można było go zastosować do szerokiej grupy związków chemicznych,
9. rozróżniał izomery,
10. posiadał wartości liczbowie w odpowiednim zakresie dla związków dla których został obliczony.

Niniejszy rozdział skupia się na ogólnym przedstawieniu szerokiej domeny deskryptorów molekularnych wraz z podziałami ze względu na typ zwracanej wartości oraz ze względu na wymiarowość. Pokazane również zostaną trendy w których świat deskryptorów molekularnych jest rozwijany.

2.1 Podział deskryptorów molekularnych

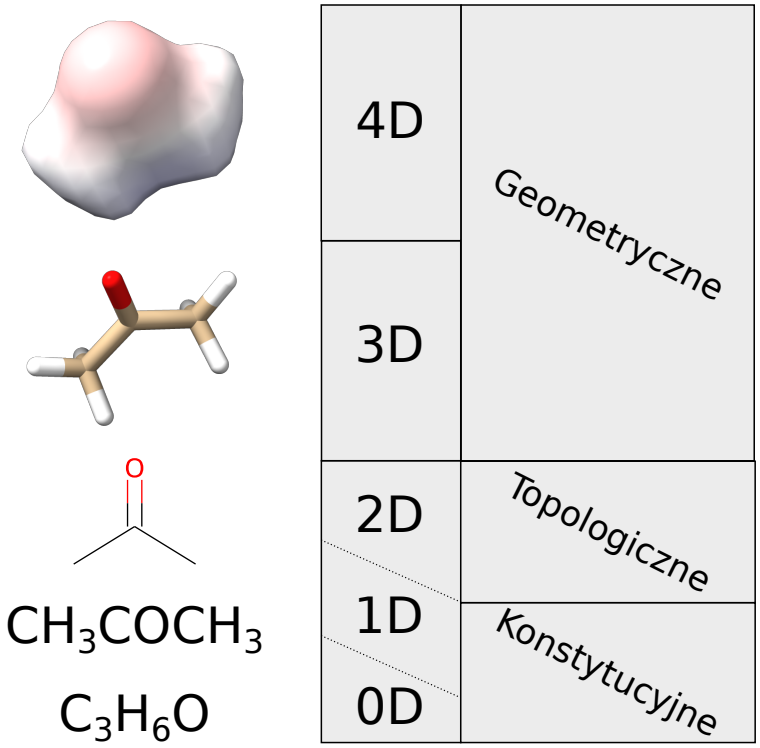
Pierwszym i najbardziej naturalnym podziałem deskryptorów jest ich podział ze względu na źródło pochodzenia, tj. eksperymentalne oraz teoretyczne. Do eksperymentalnych deskryptorów molekularnych możemy zaliczyć takie, które powstały w wyniku przeprowadzenia jakiegoś eksperymentu, z którego otrzymano wartość jakiejś właściwości fizykochemicznej, np. współczynnik podziału w układzie oktanol-woda. Natomiast deskryptory teoretyczne uzyskuje się w wyniku zastosowania jakiegoś chemoinformatycznego algorytmu, który po uzyskaniu struktury chemicznej na wejściu, zwraca wartość będącą deskryptorem molekularnym [39].

Innym możliwym podziałem deskryptorów może być podział ze względu na typ zwracanych danych. Tabela 2.1 wymienia kilka przykładowych deskryptorów molekularnych i jaki reprezentują rodzaj danych. Zauważyć można, że niekoniecznie są to proste typy danych, co może wymagać dość zaawansowanych metod ich późniejszego przetwarzania.

TABELA 2.1: Podział deskryptorów molekularnych ze względu na typ zwracanych danych.

Typ danych	Przykład
Boolean	Obecność przynajmniej jednego wiązania potrójnego
Liczba całkowita	Liczba pierścieni aromatycznych
Liczba rzeczywista	Indeksy κ [40]
Wektor	wektor autokorelacji 2D
Tensor	Tensor inercji
Pole skalarne	Potencjał elektrostatyczny
Pole wektorowe	Gradient potencjału elektrostatycznego

Kolejnym podziałem może być przedstawiony na rysunku 2.1 podział ze względu na to jakie wymiary cząsteczki są kodowane przez deskryptory molekularne, czyli tzw. wymiarowość. Podział ten bierze się ze sposobu przedstawienia, a co za tym idzie, ilości informacji o strukturze związku.



RYSUNEK 2.1: Podział deskryptorów molekularnych ze względu na wymiarowość.

Deskryptory 0D obejmują głównie te dotyczące bardzo podstawowych informacji strukturalnych - np. masę cząsteczkową, którą można uzyskać znając tylko wzór sumaryczny związku, liczby poszczególnych atomów, liczby wiązań o poszczególnych krotnościach. Wymagające więcej informacji, deskryptory 1D obejmują np. liczby atomów o danej hybrydyzacji, liczby występujących fragmentów lub grup

funkcyjnych. Jak zaznaczono na rysunku 2.1 granice między wymaganą reprezentacją do wygenerowania danego deskryptora zaliczanego do deskryptorów 0D, 1D i 2D nie są ostre. Można powiedzieć, że deskryptory 0D i 1D tworzą wspólnie grupę deskryptorów konstytucyjnych, czyli takich, które biorą pod uwagę tylko skład atomowy i rodzaje wiązań. Następną grupą są deskryptory 2D, czyli deskryptory topologiczne, które wymagają grafu molekularnego i uwzględniają wzajemne połączenia i ustawienia atomów. Wraz ze zwiększeniem wymiarowości reprezentacji, cząsteczkę można zacząć traktować jako obiekt w przestrzeni, który ma swoją powierzchnię i objętość. Z takiej reprezentacji powstają deskryptory geometryczne i steryczne określane jako deskryptory 3D. Możliwe jest dodanie jeszcze jednego wymiaru do reprezentacji geometrycznej, celem otrzymania deskryptorów 4D. W zależności od zastosowania może to być np. wymiar dotyczący dystrybucji elektronów walencyjnych cząsteczki, powstaje wówczas opis będący czymś w rodzaju pola skalarnego powiązanego z geometrią cząsteczki lub czas, w celu opisania dynamiki cząsteczki, jej elastyczności, zmian konformacyjnych i własciwości transportowych.

W tabeli 2.2 wymieniono przykładowe deskryptory molekularne z podziałem według ich wymiarowości. Jest to tylko mały wycinek możliwości. Warto zauważyć, że deskryptory molekularne kodują tylko część cech struktury danego związku chemicznego. Sprawia to, że nie ma żadnej złotej reguły dotyczącej tego, które deskryptory będą najlepsze w danym zastosowaniu.

TABELA 2.2: Przykładowe deskryptory molekularne.

Wymiarowość	Przykłady
0D	Masa molowa, liczba heteroatomów, liczba atomów węgla, liczba wiązań pojedynczych, podwójnych i potrójnych, suma atomowych objętości van der Waalsa.
1D	Liczba pierwszo-, drugo-, trzecio- i czwartorzędowych atomów węgla, liczba atomów węgla w pierścieniach, liczba donorów i akceptorów wiązania wodorowego, liczba pierścieni, liczby grup funkcyjnych.
2D	Deskryptory BCUT, całkowita powierzchnia polarna, wektor autokorelacji 2D, indeksy κ i chi, indeksy Balabana, Wiener i Zagreb.
3D	Promień bezwładności, ekscentryczność molekularna, deskryptory WHIM, molekularny potencjał elektrostatyczny, wektor autokorelacji 3D.
4D	Analiza podobieństwa molekularnego 4D

Dlatego stosuje się 2 możliwe podejścia: dobór podgrupy deskryptorów na podstawie np. odchylenia standardowego w danej bazie danych lub kondensację dużej

liczby deskryptorów w zdecydowanie mniejszą liczbę wielkości opisujących, np. analizę głównych składowych lub inne metody chemometryczne.

2.2 Inne typy deskryptorów molekularnych

Kwantowe

Metody chemii kwantowej cieszą się dużą popularnością wśród chemików. U ich podstaw leży rozwiązanie równania Schrődingera, choć w zasadzie jest to przybliżone rozwiązanie tegoż równania. Energie cząsteczek, orbitali molekularnych, polaryzowalność i moment dipolowy są tylko niektórymi właściwościami, których wartości można otrzymać jako wynik obliczeń kwantowych. Wyczerpujący przegląd zastosowania tych właściwości jako deskryptorów molekularnych w badaniach QSAR/QSPR przedstawili Karelson i wsp. [41]. Jednak problemem z zastosowaniem tego typu deskryptorów molekularnych jest długi czas obliczeń kwantowych, duże wymagania sprzętowe do ich przeprowadzenia oraz wymaga to doświadczonego badacza. Prawdopodobnie z tych względów wiele artykułów naukowych, w których stosowane są takie deskryptory molekularne, buduje modele QSAR lub QSPR w oparciu o stosunkowo mało liczne zbiory danych, zawężając w ten sposób domenę aplikowalności, np. do jednego typu związków organicznych. Jednak brak jest metody, która z powodzeniem by łączyła możliwości i dokładność obliczeniowej chemii kwantowej z szybkością i uniwersalnością chemoinformatyki. Prowadzone są natomiast badania w celu utworzenia modelu uczenia maszynowego, który by przewidywał właściwości kwantowe związków chemicznych, zastępując tym samym żmudne obliczenia. Jednakże opracowywane dotychczas narzędzia ograniczają się tylko do niektórych właściwości kwantowych i stosują zazwyczaj metody oparte o sieci neuronowe, którym zarzuca się brak interpretowalności. Potrzeba indywidualnego podejścia do każdych obliczeń, poziom skomplikowania zarówno przygotowania zadania dla pakietów obliczeniowych jak i interpretacji otrzymanych wyników sprawia, że nie ma możliwości utworzenia narzędzi typu czarna skrzynka pozwalających na obliczenia deskryptorów kwantowych. Istnieje zatem pewna nisza wśród deskryptorów molekularnych na takie, które by chociaż częściowo korzystały z osiągnięć chemii kwantowej, jednocześnie będąc stosunkowo uniwersalnymi.

Pary atomów

Zaproponowane w 1985 roku przez Carhart i wsp. [42] deskryptory molekularne powstające poprzez wygenerowanie listy par atomów, które można wyróżnić w danej cząsteczce. Każdy atom w parze opisany jest poprzez jego symbol, liczbę atomów o liczbie atomowej większej niż 1 z którymi jest związany oraz liczbę tworzonych wiązań typu π. Ponadto każda para zawiera informację, iloma wiązaniami rozdzielone są atomy w danej parze. Przykładem takiej pary atomów może być ((C, 1, 0), 3, (C, 2, 0)), co oznacza, że dana para atomów składa się z dwóch oddalonych od siebie o trzy wiązania atomów węgla, z których jeden atom związany jest z jednym atomem innym niż wodór i nie tworzy żadnego wiązania typu π. Drugi atom węgla połączony jest z dwoma atomami innymi niż wodór i również nie tworzy wiązań typu π. Ten typ deskryptorów molekularnych doczekał się szeregu pochodnych, np. zaproponowane przez Nilakantan i. in. [43] deskryptory torsji topologicznej będące listą szeregów czterech kolejno związanych ze sobą atomów, które opisane są w ten sam sposób co w parach atomów, czyli symbolem, liczbą wiązań z atomami innymi niż atomy wodoru oraz liczbą tworzonych wiązań typu π. Kolejna modyfikacja zarówno par atomów jak i deskryptorów torsji topologicznej [44] polegała na zastąpieniu symbolu pierwiastka w opisie poszczególnych atomów przypisaniem tego pierwiastka do jednej z siedmiu klas. Klasy te mają odwzorowywać właściwości pierwiastków, np. donor pary elektronowej.

Nieco innym podejściem jest metoda CATS [45], w której każdy atom w cząsteczce przyporządkowany jest do jednej z pięciu klas. Cała cząsteczka opisywana jest jako histogram częstości występowania par danych klas atomów (15 możliwych kombinacji), przy czym występuje podział ze względu na odległość od siebie atomów wyrażoną w liczbie wiązań oddzielających atomy. Zatem cząsteczka opisana jest jako wektor o długości $15n$, gdzie n jest maksymalną rozważaną odległością pomiędzy atomami w parze.

Jako odmianę par atomów i deskryptorów torsji topologicznej można wyróżnić

również metodę kluczy Simlog [46], w której każdy atom kodowany jest, w myśl zasady DABE, ze względu na cztery właściwości: donor wiązania wodorowego (Donor), akceptor wiązania wodorowego (Acceptor), promień van der Waalsa (Bulkiness) i elekrododatniość (Electropositivity). Cząsteczka związku chemicznego przedstawiana jest jako lista możliwych do wyróżnienia trójek atomów, gdzie każdy z atomów przedstawiony jest kodem DABE. Kody te oddzielone są od siebie liczbą wiązań pomiędzy atomami w danej trójce.

Odciski palców

Przedstawione w rozdziale 1 odciski palców mogą być z powodzeniem stosowane jako deskryptory molekularne [47]. Ponadto odciski palców są podstawą do przeszukiwania baz danych w poszukiwaniu podobnych cząsteczek. Poprzez określenie częstości występowania fragmentów cząsteczek generowanych dla cząsteczek znajdujących się w bazie danych PubChem powstała metryka przystępności syntetycznej [48].

Deskryptory oparte na tekście

Rozwój metod przetwarzania tekstu otwiera nowe drogi wykorzystania informacji znajdujących się w chemicznych bazach danych. Metody tokenizacji fragmentów tekstu mogą posłużyć do budowy modeli QSAR/QSPR [49]. Poprzez analizę notacji liniowych związków w bazach danych reakcji chemicznych modele przetwarzania języka naturalnego badane są pod kątem przewidywania produktów reakcji chemicznych [50], w celach retrosyntezy [51] oraz przewidywania mechanizmów reakcji i produktów pośrednich [52].

2.3 Poznawanie reprezentacji

W czasach ogromnej popularności uczenia maszynowego wyłania się alternatywa dla deskryptorów molekularnych. Zamiast podejmowania się zakodowania pewnych zjawisk i zależności strukturalnych w cząsteczkach, czasem poprzez dość zaawansowane operacje matematyczne i w ich oparciu budowanie modeli uczenia maszynowego, istnieje podejście polegające na pozwoleniu komputerowi samodzielnie skonstruować reprezentację cząsteczki na potrzeby danego zadania, co

nazywane jest poznawaniem reprezentacji (ang. representation learning). Takie podejście można określić jako metody sterowane danymi (ang. data driven). Jest to sposób, który zdecydowanie upraszcza wstępne przetwarzanie danych, ponieważ do algorytmu wprowadza się możliwie najbardziej surowe dane [53]. Na wejściu do takich metod można wprowadzić cząsteczkę w formie grafu, wówczas stosuje się metody grafowych sieci neuronowych (GNN) lub sieci neuronowych przekazywanych wiadomości (MPNN). Siatkowe sieci neuronowe (ang. grid neural network) pozwalają na stworzenie reprezentacji cząsteczki z trójwymiarowych współrzędnych kartezjańskich. Część wspomnianych metod opartych na tekście również jest poznawaniem reprezentacji z surowych danych (RNN). Ciekawą implementacją są generatywne sieci przeciwstawne (GAN), czyli zestawienie dwóch sieci neuronowych, z których jedna generuje syntetyczne, nowe instancje danych, a druga ocenia jak dobrze to zadanie zostało wykonane. Przykładowe badania zastosowań sieci neuronowych można znaleźć w bibliografii [54–56].

Metody te jednak wymagają dużych zbiorów danych, żeby schematy w nich występujące mogły zostać dobrze poznane. W przypadku mniej licznych baz danych, standardowe deskryptory molekularne mogą wykazać się większą dokładnością. Jiang D. i in. dokonali porównania grafowych sieci neuronowych z deskryptorami molekularnymi [57].

Aplikacje wyznaczające deskryptory molekularne

W rozdziale przedstawiono wybrane oprogramowanie pozwalające na obliczanie wartości deskryptorów molekularnych. Podstawową różnicą pomiędzy komercyjnymi i niekomercyjnymi aplikacjami jest ich złożoność. Komercyjne rozwiązania są zazwyczaj wyposażone w interfejs graficzny, ponadto często ich funkcjonalności nie ograniczają się tylko do zwracania wartości deskryptorów, tylko dodatkowo do wstępnej, a w niektórych przypadkach całościowej analizy i wizualizacji otrzymanych danych. Natomiast niekomercyjne rozwiązania są bezpłatne, co dla naukowców jest ogromną zaletą i najczęściej nie ma żadnych obostrzeń co do ich używania. Jeśli dodatkowo rozwiązanie takie jest otwartoźródłowe, to możliwe jest sprawdzenie poprawności kodu, a nawet jego poprawa lub modyfikacja.

3.1 Aplikacje niekomercyjne

Mordred

Najnowszy spośród niekomercyjnych aplikacji liczących deskryptory molekularne (ostatnia wersja z 2019 roku). Mordred [58] może być używany jako aplikacja CLI oraz moduł Pythona. Ponadto autorzy opracowali również aplikację z interfejsem w przeglądarce internetowej. Mordred pozwala na wyznaczenie wartości 1826 deskryptorów molekularnych, z czego 213 jest deskryptorami 3D. Sama instalacja jest łatwiejsza niż w przypadku innych kalkulatorów. Rozwiązania zastosowane w Mordred sprawiają również, że generowanie deskryptorów molekularnych jest zdecydowanie szybsze. Jednak aby można było skorzystać z deskryptorów 3D, należy posiadać wyznaczoną geometrię cząsteczki.

PaDEL-descriptor

Napisany w języku Java program do obliczeń deskryptorów molekularnych. PaDEL [59] pozwala na wyznaczenie 1875 deskryptorów molekularnych oraz odcisków palców. Ponadto gdy użytkownik nie posiada geometrii cząsteczki, program pozwala na jej wyznaczenie metodą MM2 lub MMFF94 i obliczenie deskryptorów 3D. Choć w czasie pisania ninejszej pracy stona internetowa, na której znajduje się kod tego programu była nieaktywna, to nadal możliwe jest jego pobranie z repozytorium będącym jego opakowaniem w formie modułu Pythona [60].

ChemoPy

ChemoPy [61] pozwala na wyznaczenie wartości 1135 deskryptorów molekularnych oraz 7 różnych grup odcisków palców. W połączeniu z oprogramowaniem MOPAC, jest w stanie wyznaczyć deskryptory 3D. Nie jest to samodzielny program, tylko moduł Pythona 2 (już nierozwijany), co wymaga znajomości tego języka. Mimo, że dostępny do pobrania w wersji 1.0, to niestety od 2013 roku nie został zaktualizowany.

PyDPI

Podobnie jak ChemoPy, jest to moduł Pythona, który umożliwia wyznaczenie 615 deskryptorów molekularnych i odcisków palców dla małych cząsteczek. Ponadto PyDPI [62] zawiera funkcjonalność obliczania deskryptorów molekularnych białek. Od wydania nieaktualizowany, choć możliwa jest instalacja za pomocą Python Package Index (PyPI, komenda pip).

3.2 Aplikacje komercyjne

alvaDesc 2.0

Produkt Kode Cheminformatics, oprogramowanie to pozwala na wyznaczenie 5666 deskryptorów molekularnych, które są zebrane w 33 grupach logicznych. Deskryptory obejmują również te, które wymagają na wejściu geometrii cząsteczki i odciski palców. Pod warunkiem przekazania wyników badań do domeny publicznej, można uzyskać darmową licencję akademicką, która obejmuje pomniejsze aktualizacje przez rok. alvaDesc posiada również wbudowane funkcje do wstępnej

analizy i wizualizacji danych takie jak analiza głównych składowych, t-SNE oraz analiza korelacji.

Poprzednikiem alvaDesc był program Dragon, który posiadał podobne możliwości, jednak liczba dostępnych deskryptorów molekularnych wynosiła 5270 pogrupowanych w 30 bloków. Tak jak następca, Dragon pozwalał na wstępną analizę i wizualizację danych. Progam ten nie jest już dystrybuowany przez producenta, jednak niektóre laboratoria na świecie wciąż mogą z niego korzystać.

ADMEWORKS ModelBuilder

Zgodnie ze specyfikacją producenta (Fujitsu) oprogramowanie to pozwala na wyznaczenie 400 deskryptorów molekularnych oraz „nielimitowaną" liczbę deskryptorów związanych z fragmentami cząsteczek. Wbudowane funkcje pozwalają nie tylko na obliczenie wartości deskryptorów, lecz również na analizę, wizualizację danych i zbudowanie własnego modelu QSAR/QSPR opartą o szereg metod statystycznych (uczenia maszynowego).

MOE - Molecular Operating Environment

Oprogramowanie zawarte w MOE pozwala na wyznaczenie wartości ponad 400 deskryptorów molekularnych i pozwala na deklarację własnych. Ponadto wbudowane są funkcje tworzenia modeli QSAR/QSPR. Jednak MOE to nie tylko kalkulator deskryptorów, tylko środowisko dla badaczy z różnych dziedzin, nie tylko chemoinformatyki. Środowisko to obejmuje również takie funkconalności jak np. modelowanie białek i dokowanie do nich ligandów oraz modelowanie z zakresu dynamiki molekularnej.

Część II

Część eksperymentalna

Metodyka oraz obiekt badań

4.1 Analiza danych

Analizy wykonywane były przy użyciu trzech różnych komputerów z systemem operacyjnym Linux (dystrybucja Xubuntu). Pod względem sprzętowym pierwsze prace wykonywano na komputerze z procesorem Intel i5 drugiej generacji oraz pamięcią 8 GB RAM, następnie czasowo na komputerze z procesorem Intel i7 ósmej generacji wyposażonym w 32 GB pamięci operacyjnej, ostatecznie na komputerze z procesorem Intel i5 czwartej generacji oraz 16 GB pamięci operacyjnej. Podstawowym narzędziem pracy był język Python [63]. Poniżej przedstawiono używane biblioteki/moduły Pythona spoza biblioteki standardowej:

- NumPy [64],
- Pandas [65, 66],
- Scikit-learn [67],
- Matplotlib [68],
- Imblearn [69],
- Seaborn [70],
- RDKit [71] - moduł w początkowym etapie realizacji pracy doktorskiej instalowany był z repozytorium dystrybucji Linuxa, a następnie za pomocą polecenia pip.

Stosowanymi środowiskami programistycznymi były Jupyter Lab [72], Visual Studio Code oraz PyCharm (wersja Community).

Spośród dostępnych algorytmów uczenia maszynowego w badaniach stosowano przede wszystkim regresję liniową lub logistyczną, które są jednymi z najprostszych alogrytmów. Ponadto stosowano metody złożone, oparte o drzewa decyzyjne - las

losowy oraz wzmocnienie gradientowe. Algorytmy te charakteryzują się, np. małą podatnością na brak standaryzacji zmiennych objaśniających oraz wieloma parametrami, które można zmieniać celem uzyskania jak najlepszego dopasowania do danych. Ponadto algorytmy mające u podstaw drzewa decyzyjne cenione są za interpretowalność.

Celem jak najlepszej oceny i porównania zdolności predykcyjnych budowanych modeli uczenia maszynowego stosowano metryki oceniające. W przypadku budowania modeli regresji najważniejsze były: kwadrat współczynnika Pearsona R^2 oraz średni błąd kwadratowy. Parametr R^2 mówi o tym jaka jest różnica pomiędzy predykcjami z modelu, a predykcjami z modelu o stałej wartości predykcji (zawsze zwraca tą samą wartość, zazwyczaj średnią arytmetyczną). Średni błąd kwadratowy (MSE) oraz jego pierwiastek (RMSE) pozwala na oszacowanie jaka jest średnia różnica pomiędzy wartością przewidzianą przez model, a wartością rzeczywistą, przy czym podniesienie tej wartości do kwadratu penalizuje wartości bardziej odległe od rzeczywistej. Można powiedzieć, że w stosunku do średniego błędu absolutnego (MAE), średni błąd kwadratowy uwzględnia rozrzut wartości przewidzianej względem wartości rzeczywistej - mniejszą wartość MSE będzie miał model zwracający średnio dokładne predykcje niż model zwracający dokładne i bardzo niedokładne predykcje.

Do oceny zdolności predykcyjnych modeli klasyfikujących wybrano zwłaszcza dokładność zbalansowaną, która jest średnią arytmetyczną czułości i specyficzności, czyli parametrem szacującym jak dobrze model klasyfikuje zarówno jedną jak i drugą klasę. W jednym z prowadzonych badań zastosowano F miarę, której wartość jest zależna od precyzji i czułości, czyli miar uwzględniających tylko poprawność klasyfikacji klasy określonej jako dodatnia. Innym, często stosowanym sposobem oceny poprawności modeli klasyfikujących jest pole pod krzywą ROC, która jest wykresem zależności czułości od specyficzności dla różnych progów dyskryminacji pomiędzy klasami. Miara ta została zastosowana w jednym z prowadzonych badań. Dobór odpowiednich metryk w przypadku klasyfikacji jest często zależny od kontekstu samych badań, zatem nie ma jednej uniwersalnej metody.

4.2 Obliczenia kwantowe

Wysokie zapotrzebowanie na moc obliczeniową w obliczeniach kwantowych spowodowało, że w celu wykonania niezbędnych obliczeń zastosowano maszynę wirtualną pozyskaną z Laboratorium Technologii Chmurowych Wydziału Cybernetyki Wojskowej Akademii Technicznej. Maszyna posiadała 32 procesory, 128 GB pamięci RAM oraz dysk o pojemności 1 TB. Systemem operacyjnym zainstalowanym na maszynie wirtualnej był Linux (dystrybucja Xubuntu, wersja 20.04). Obliczenia wykonywane były w programie Psi4 [73] (wersje od 1.4 do 1.7, w miarę dostępnych aktualizacji), wsparte skryptami automatyzującymi pracę, napisanymi w językach Python oraz Bash.

Wyniki obliczeń połączone z inną bazą danych udostępnione są na platformie Zenodo [74]

4.3 Prowadzone badania

W toku realizacji pracy doktorskiej dokonano trzech różnych definicji proponowanych deskryptorów kwantowych. Opierając się na tych definicjach budowano modele predykcyjne lub klasyfikacyjne w badaniach typu QSAR/QSPR. Otrzymywane wielkości oceniające porównywano do wielkości oceniających otrzymanych przy pomocy modeli predykcyjnych skonstruowanych w oparciu o powszechnie stosowane deskryptory molekularne lub inne ugrunowane w chemoinformatyce komputerowe reprezentacje związków chemicznych oraz tzw. linii bazowych, czyli losowego przypisania klasy w klasyfikacji lub predykcji stałej wartości w przypadku regresji. Porównanie do linii bazowej jest realizowane celem sprawdzenia, czy inne metody posiadają jakąkolwiek zdolność do predykcji.

Zaproponowane w niniejszej monografii kwantowe reprezentacje związków chemicznych można podzielić na 2 podgrupy:

1. reprezentacje oparte o właściwości kwantowe wykrytych fragmentów cząsteczek związków chemicznych, opisane w rozdziałach 5, 6 oraz 8,

2. kwantowo zmodyfikowane pary atomów opisane w rozdziale 9.

Pierwszą grupę nowych deskryptorów zastosowano do modelowania właściwości optycznych związków typu OLED oraz do klasyfikacji związków ze względu

na biokumulację w organizmach żywych. Kwantowo zmodyfikowane pary atomów sprawdzono na 10 różnych bazach danych eksperymentalnych, 6 zawierających zmienne ilościowe (regresja) i 4 zawierające zmienne jakościowe (klasyfikacja).

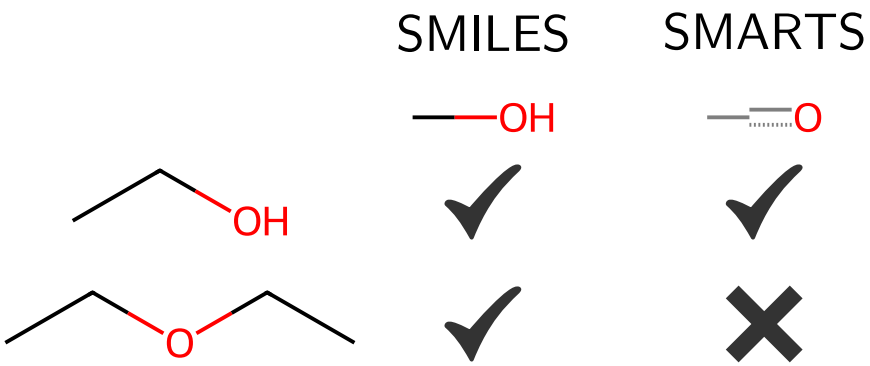
4.4 Powstałe oprogramowanie

Jednym z rezultatów przeprowadzonych badań było powstanie wielu skryptów i Jupyter Notebooków. Jupyter Notebooki związane z opublikowanymi artykułami dostępne są na platformie GitHub oraz dołączone jako materiały uzupełniające do artykułów. Ponadto, w celu umożliwienia innym badaczom zastosowania opracowanych deskryptorów kwantowych bez potrzeby własnej implementacji, na platformie GitHub opublikowano program z interfejsem graficznym pozwalający na wygenerowanie fragmentarycznych deskryptorów kwantowych i kwantowych par atomów.

Fragmentaryczne deskryptory kwantowe

Zastosowanie eksploracyjnej analizy danych w modelowaniu aktywności biologicznych lub właściwości związków chemicznych oparta jest na zastosowaniu jakiegoś opisu struktury tych związków i próbie znalezienia relacji pomiędzy tym opisem, a modelowaną właściwością. Opis struktury realizowany jest np. za pomocą deskryptorów molekularnych i odcisków palców. W niniejszej pracy doktorskiej podjęto wysiłek ku wzbogaceniu dostępnych metod opisu struktur związków chemicznych na podstawie podstruktur, które można w nich wyróżnić. Proponowane w pracy niekonwencjonalne deskryptory molekularne wyznaczane były na podstawie właściwości kwantowych tych podstruktur. W badaniach przyjęto założenie, że jeśli możliwe jest wyróżnienie danej struktury w związku chemicznym, to właściwości związku są wypadkową właściwości kwantowych wszystkich znalezionych podstruktur. Same podstruktury i ich właściwości kwantowe pochodziły ze z góry zdefiniowanej grupy.

W badaniach stosowano bazy danych cząsteczek, w których struktury zakodowane są w postaci notacji liniowej SMILES. Należy w tym miejscu wskazać, że do wyszukiwania fragmentów cząsteczek lepsze jest wyszukiwanie z zastosowaniem notacji SMARTS, ponieważ można w ten sposób zdefiniować konkretną strukturę, lub jej część, która będzie wyszukana bez żadnych odstępstw (np. grupa hydroksylowa). Jako przykład może posłużyć dopasowanie fragmentu grupy hydroksymetylowej do etanolu i eteru przedstawiony na rysunku 5.1. Zatem w badaniach brano pod uwagę podstruktury, które nie były zdefiniowane w sposób bardzo ścisły - np. podstruktura będąca alkoholem mogła być wykryta w cząsteczce eteru. Samo wyszukiwanie podstruktur wykonywane było za pomocą metody HasSubstructMatch obiektu klasy Mol biblioteki RDKit.

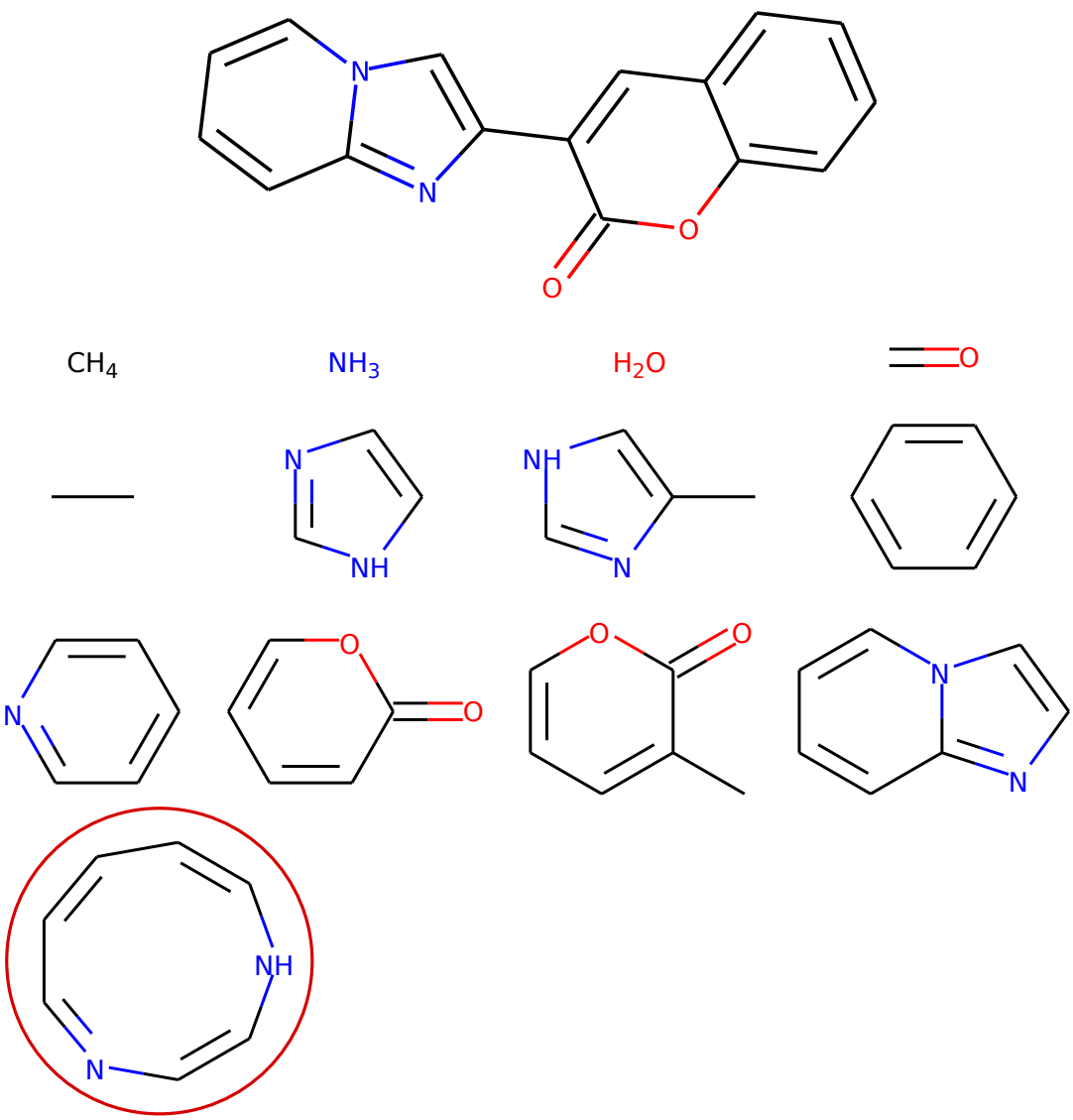


RYSUNEK 5.1: Dopasowanie struktury hydroksymetylowej za pomocą SMILES i SMARTS. Przy zdefiniowaniu fragmentu poprzez SMILES metanol jako podstruktura wyszukiwany jest zarówno w etanolu i eterze dietylowym. Natomiast po zastosowaniu SMARTS możliwe jest osiągnięcie tego dopasowania tylko dla etanolu.

Od początku podjętej tematyki zauważono, że jednym z potencjalnych ograniczeń opisu struktury cząsteczki związku przez jej podstruktury może być sam zbiór podstruktur. Stwierdzono, że związki chemiczne w których poszukuje się podstruktur powinny składać się z tych samych pierwiastków, co same podstruktury, ponieważ w przeciwnym razie deskryptory mogłyby być obarczone dużym błędem. Pochodzenie tego błędu jest takie, że np. heteroatomy mogą mieć wpływ na właściwości kwantowe, a brak podstruktur zawierających dane heteroatomy skutkuje nie uwzględnieniem tego wpływu. Z tego też powodu w toku realizacji pracy doktorskiej podjęto się rozszerzenia jednej z baz danych właściwości kwantowych.

Ponadto zdefiniowany zbiór podstruktur nie zawiera wszystkich możliwych kombinacji atomów danych pierwiastków, tylko wybrane związki chemiczne, co sprawia, że istnieją związki chemiczne, w których nie można wyróżnić żadnej z poszukiwanych podstruktur.

Analizując niektóre wyszukane podstruktury zauważono, że istnieją takie, których nie można było w prowadzonych badaniach i analizach wyróżnić w związku eksperymentalnym. Przykładem może być związek, którego strukturę i wykryte podstruktury przedstawiono na rysunku 5.2. Błędnym w przedstawionym przypadku jest zaznaczony czerwonym kołem związek, którego nie można wydzielić jako podstruktura w związku wyjściowym. Jednakże ze względu na brak skutecznego rozwiązania tego problemu, a sprawdzenie i ręczne usunięcie błędnych podstruktur byłoby wyjątkowo żmudne, dopuszczono ten błąd godząc się jednocześnie z potencjalnymi błędami wynikającymi z tego zjawiska.



RYSUNEK 5.2: Przykład wykrytych podstruktur w związku z eksperymentalnej bazy danych. Czerwonym okręgiem zaznaczono strukturę nie występującą w związku.

Predykcja długości fali dla maksimum emisji związków optycznie czynnych

Projektowanie nowych związków chemicznych o pożądanych właściwościach fizycznych jest procesem pracochłonnym, zwłaszcza gdy poszukiwanie takie prowadzone jest w sposób ściśle eksperymentalny. Z tego powodu do badań angażuje się komputer. Jedną z metod jest wykorzystanie oprogramowania do obliczeń kwantowych, które potrafią dać dokładny wynik. Jednak jest to narzędzie wymagające zarówno pod względem sprzętowym, czasowym oraz stopnia skomplikowania. Aby uzyskiwać precyzyjne wyniki z zastosowaniem metod chemii kwantowej wymgana jest wiedza oraz doświadczenie w tej dziedzinie. W opracowywaniu związków o pożądanych właściwościach pomocne mogą się okazać metody QSPR. Stworzenie modelu, który znane wartości parametrów fizycznych powiąże ze strukturą związków chemicznych pozwala na opracowanie narzędzi informatycznych, które każdy chemik mógłby wykorzystywać do wstępnej oceny projektowanych związków lub do badania przesiewowego wielu struktur i wybrania potencjalnych kandydatów do dalszego eksperymentu. W związku z powyższym podjęto próbę zastosowania fragmentarycznych deskryptorów kwantowych w celu zbudowania modelu regresji wybranej właściwości optycznej (długości fali przy której następuje maksimum emisji) związków organicznych typu OLED.

6.1 Baza danych właściwości kwantowych i nowe deskryptory kwantowe.

Do policzenia nowego rodzaju deskryptorów zastosowano bazę danych QM9 [75], która składa się z prawie 134 tysięcy teoretycznych związków chemicznych zawierających w swojej strukturze do 9 atomów spośród węgla, azotu, tlenu i fluoru. Związki te zostały wyselekcjonowane z bazy danych GDB17 [76], która zawiera 166 miliardów związków chemicznych, należących do dwóch grup związków, w tym rzeczywistych i teoretycznych. Należy nadmienić, że baza QM9 nie zawiera wszystkich możliwych struktur o danej liczbie atomów, np. wśród związków o dwóch atomach cięższych od wodoru możliwe jest 10 różnych kombinacji, natomiast w bazie znajduje się ich tylko 2. Samą bazę danych pobrano ze strony projektu MoleculeNet [77], ponieważ udostępniona tam jest w postaci pliku .csv, zwierającym struktury związków w postaci notacji SMILES i ich właściwości kwantowe, podczas gdy w oryginale każdy związek występuje w postaci pliku .xyz, po jednym na związek chemiczny, co wiązałoby się z dodatkowym wczytywaniem poszczególnych plików, konwersją struktury ze współrzędnych kartezjańskich na notację SMILES i wczytywaniem właściwości. Lista właściwości kwantowych zawartych w bazie danych QM9 znajduje się w tabeli 7.1.

Celem sprawdzenia tezy postawionej w pracy doktorskiej, zastosowano fragmentaryczne deskryptory kwantowe do modelowania QSPR. Deskryptory te w opisywanym badaniu powstały metodą częściowo kombinatoryczną - utworzone zostało 14 różnych zestawów deskryptorów, które były wynikiem operacji matematycznych na właściwościach kwantowych. Poniżej przedstawiono opis części zestawów deskryptorów molekularnych i ewentualnie wzory do obliczenia deskryptorów kwantowych. Z wyjątkiem zestawów nr 1 i 3, wszystkie zestawy były kombinacją deskryptorów molekularnych kwantowych i tradycyjnych.

Zestaw 1

Zawierał wyłącznie deskryptory zawarte w bibliotece RDKit. Pełnił rolę porównawczą.

Zestaw 2

Suma właściwości kwantowych wykrytych fragmentów pomnożonych przez liczbę wystąpień danego fragmentu w cząsteczce.

$$\sum_{i}^{N} n_i \epsilon_{HOMO_i}, \sum n_i \epsilon_{LUMO_i}, ... \tag{6.1}$$

gdzie i - indeks wykrytej podstruktury, n - ile razy dana podstruktura wysępuje w cząsteczce, ϵ_{HOMO}, ϵ_{LUMO} - energia HOMO i LUMO.

Zestaw 3

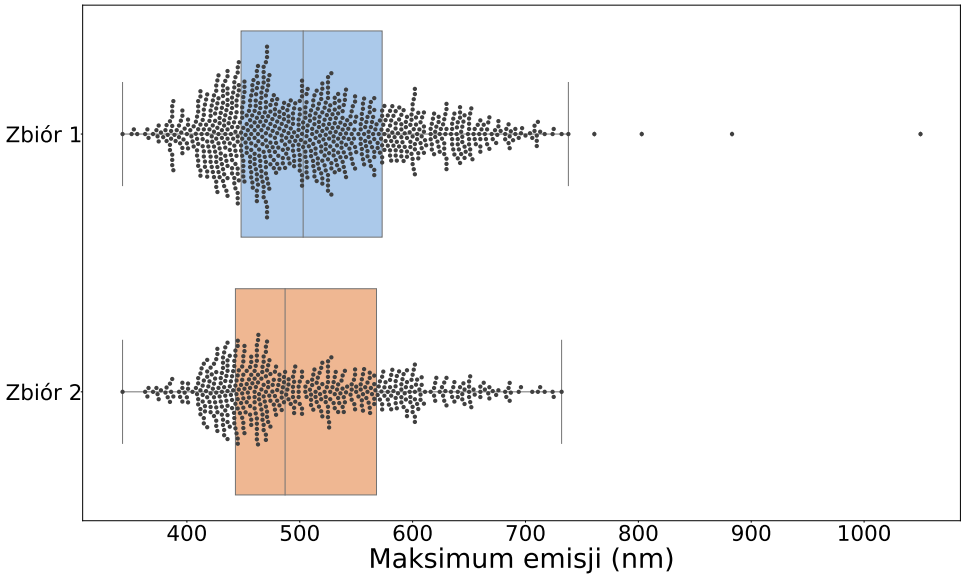
Zawierał tylko fragmentaryczne deskryptory kwantowe z zestawu 2 (nieobecne były deskryptory z biblioteki RDKit).

Wzory do obliczenia pozostałych zestawów (4 - 14) fragmentarycznych deskryptorów kwantowych zawarte są w dodatku A.

6.2 Metodyka

Modelowane właściwości eksperymentalne wzięto z bazy danych chromofo–rów [78] zawierającej ponad 20000 wierszy, które tworzone są przez kombinację 7016 związków organicznych występujących w 365 różnych rozpuszczalnikach lub w stałym stanie skupienia. Zbiór ten powstał w sposób manualny, poprzez wybranie związków chemicznych i ich właściwości z opublikowanych artykułów naukowych. Autorzy bazy danych narzucili ograniczenia w kwestii liczby atomów związków - maksymalnie 150 oraz składu pierwiaskowego - z wyjątkiem wodoru, związki w bazie danych składają się tylko z C, N, O, S, F, Cl, Br, I, Se, Te, Si, P, B, Sn, Ge. W celu uniknięcia efektów rozpuszczalnikowych na długość fali emitowanej przez związek optycznie czynny, zdecydowano się do ograniczenia badania do związków w fazie stałej. Zabieg ten ograniczył bazę danych do 956 rekordów, jednak 897 z nich posiada zaraportowaną wartość długości fali przy której następuje maksimum emisji. W ten sposób powstał jeden zbiór danych do badania (zbiór 1).

Ponadto powstał drugi zbiór danych poprzez ograniczenie pierwszego do związków składających się tylko z C, O, N i F (523 związki), oznaczony jako zbiór 2, w celu sprawdzenia, czy wystąpi różnica w wynikach regresji. Zakres wartości maksimum emisji w poszczególnych zbiorach widoczny jest na rysunku 6.1.



RYSUNEK 6.1: Rozstęp wartości modelowanej właściwości.

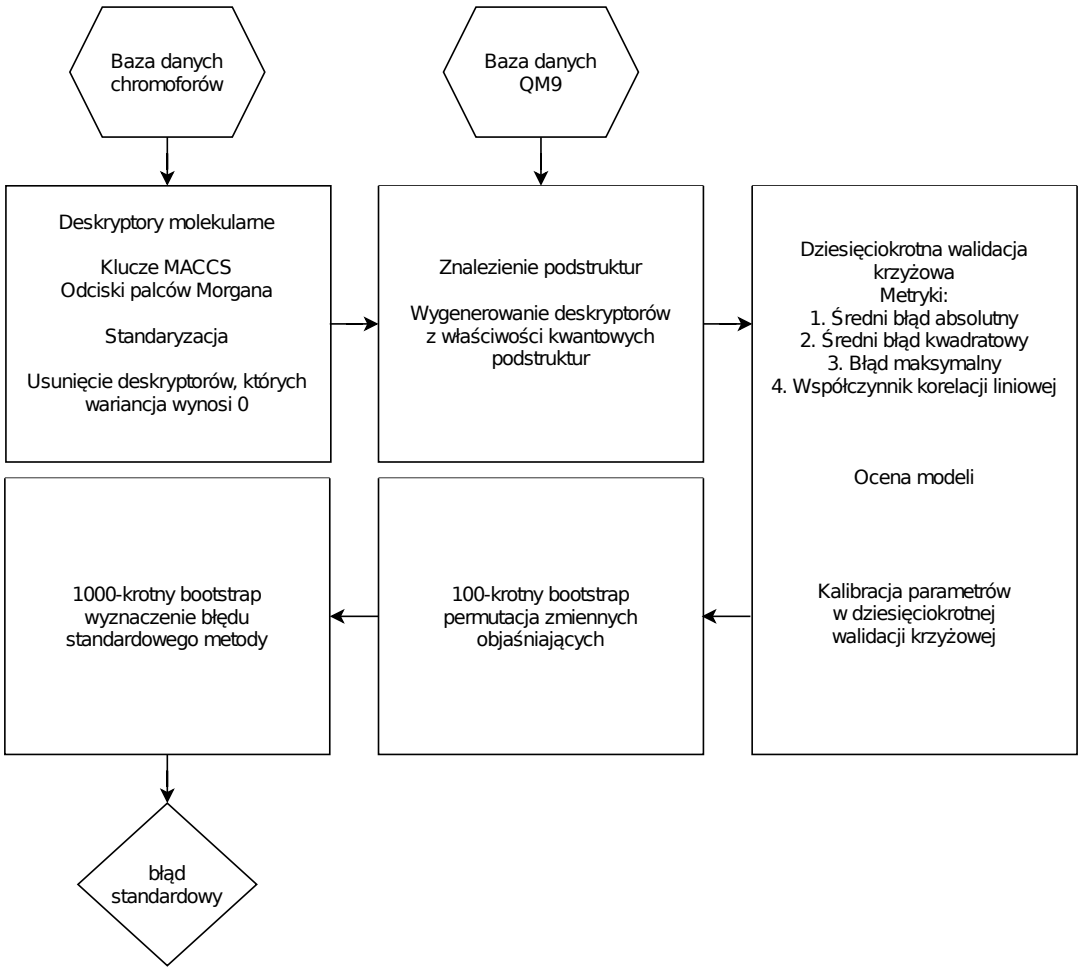
Można zauważyć, że zbiór 1 charakteryzuje się podobnym rozkładem długości fali, co zbiór 2, jednak w zbiorze 1 występują wartości odstające, co może mieć wpływ na zdolności predykcyjne modeli.

Spośród dostępnych w bibliotece RDKit deskryptorów molekularnych wygenerowano wszystkie dostępne deskryptory nie wymagające podania geometrii cząsteczki, klucze MACCS oraz odciski palca Morgana. Po wygenerowaniu dokonano ich standaryzacji oraz usunięto deskryptory, których wariancja wynosiła 0. Następnie wygenerowano deskryptory pochodzące z właściwości kwantowych (wszystkich dostępnych w bazie danych QM9) wykrytych podstruktur. W dziesięciokrotnej walidacji krzyżowej testowano algorytmy regresji wieloliniowej, lasu losowego oraz wzmocnienia gradientowego według implementacji modułu Scikit-learn Pythona. Jako wielkości testujące wybrane zostały średni błąd absolutny, średni błąd kwadratowy, błąd maksymalny oraz współczynnik determinacji R^2. Po dokonaniu oceny wybrano algorytm i zestaw deskryptorów, które dawały najlepsze rezultaty. Wybrany model podlegał kalibracji trzech hiperparametrów (liczba estymatorów, głębokość drzewa decyzyjnego oraz maksymalna liczba zmiennych) w dziesięciokrotnej walidacji krzyżowej. Dostosowawszy parametry, oceniono, które zmienne

są najważniejsze z punktu widzenia regresji. W tym celu przeprowadzono bootstrap danych i przy użyciu tak powstałych zbiorów, metodą permutacji oceniono, które zmienne mają największy wpływ na wyniki. Czynność tą powtórzono stukrotnie. Na koniec tysiąckrotnie powtórzono następujące czynności:

1. bootstrap z powtórzeniem,
2. dopasowanie danych uczących do algorytmu,
3. wykonanie predykcji,
4. wyznaczenie wartości błędu.

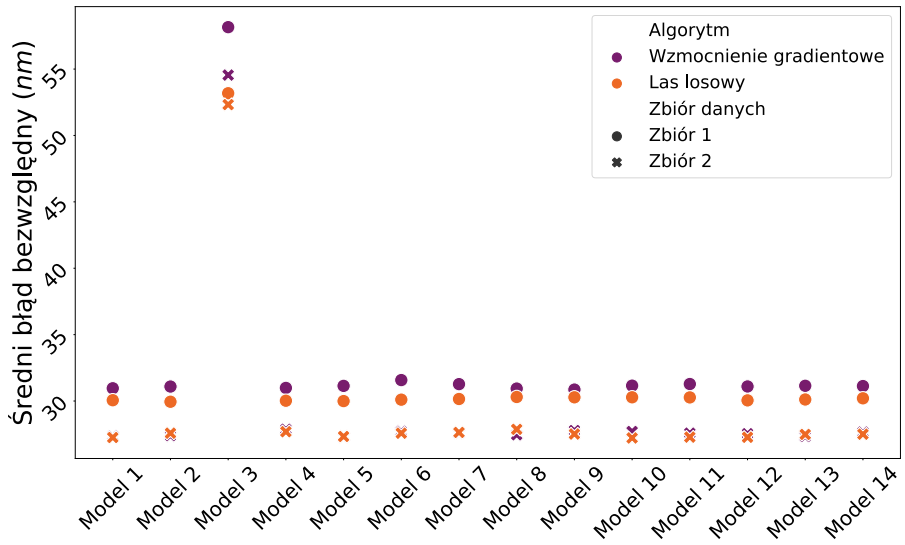
Otrzymawszy rozkład popełnianych błędów dokonano oceny odchylenia standardowego tego rozkładu. Schemat procesu analizy zobrazowany jest na rysunku 6.2.



RYSUNEK 6.2: Schemat analizy.

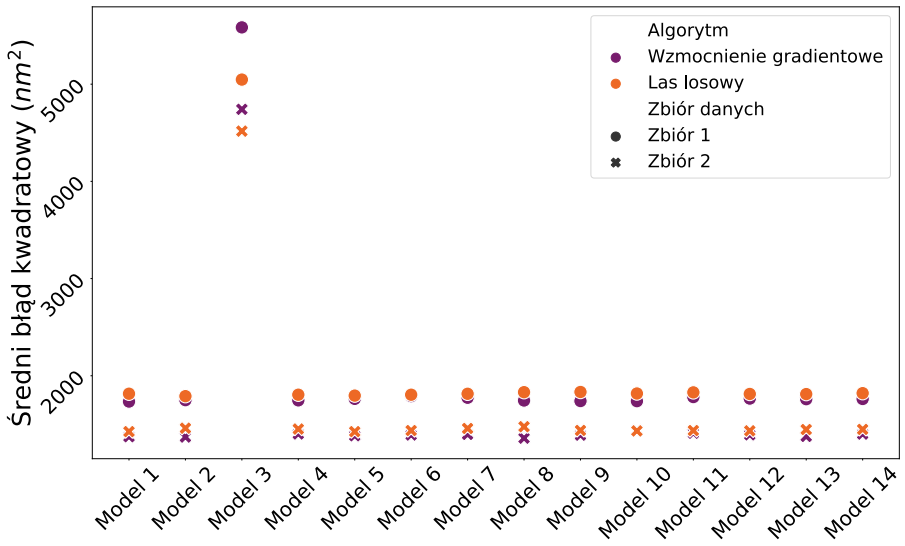
6.3 Wyniki i dyskusja

Pierwsza ocena zaprojektowanych modeli regresji przedstawiona jest na rysunkach 6.3-6.6. Modele 1-14 odpowiadają zestawom deskryptorów, które zostały w nich zastosowane. Osiągany średni błąd bezwzględny wyniósł ok. 31 nm, dla zbioru 1 oraz 27 nm dla zbioru 2. Od tych wartości znacząco, prawie dwukrotnie, odbiegały wyniki modelu, w którym deskryptory pochodzące od fragmentów cząsteczek zastosowanych samodzielnie w stosunku do wyników, gdzie zastosowano deskryptory z biblioteki RDKit. Jednoznacznie wynika, że dodanie właściwości kwantowych fragmentów cząsteczek do zmiennych modelu predykcyjnego nie wpływa na otrzymany wynik, niezależnie od tego, jak takie deskryptory zostały zdefiniowane. Zauważyć można dodatkowo, że algorytm wzmocnienia gradientowego charakteryzował się większym średnim błędem absolutnym w regresji zbioru 1 niż algorytm lasu losowego.



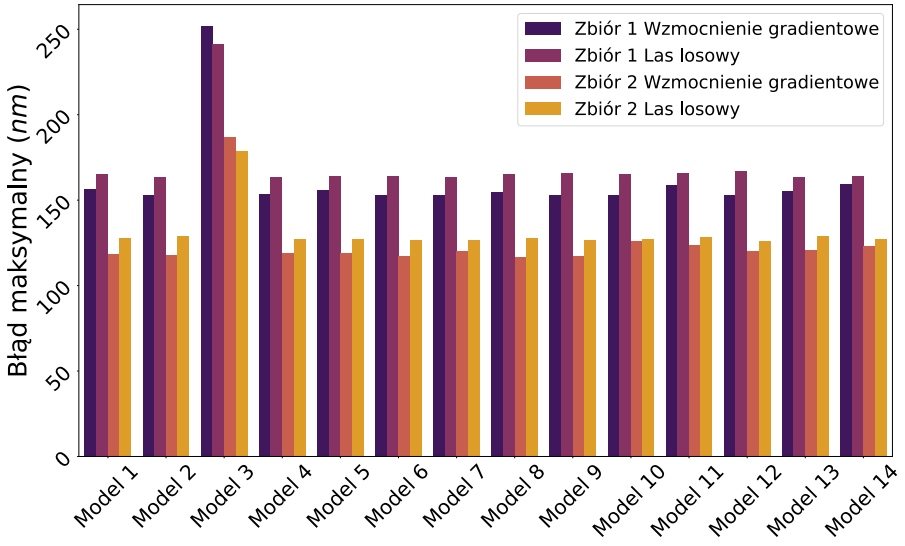
RYSUNEK 6.3: Średni błąd bezwzględny dla różnych modeli.

Wyniki średniego błędu kwadratowego miały podobny charakter do wyników średniego błędu bezwzględnego, przy czym model, w którym zastosowano tylko nowe deskryptory kwantowe osiągnął wartości tej miary ponad dwukrotnie większe od wyników, modeli, gdzie zastosowano tradycyjne deskryptory molekularne. Nie ma natomiast tak wyraźnej różnicy pomiędzy zastosowanymi algorytmami. Ponownie, błąd jest niższy, gdy jako dane w modelowaniu zastosowano zbiór 2.



RYSUNEK 6.4: Średni błąd kwadratowy dla różnych modeli.

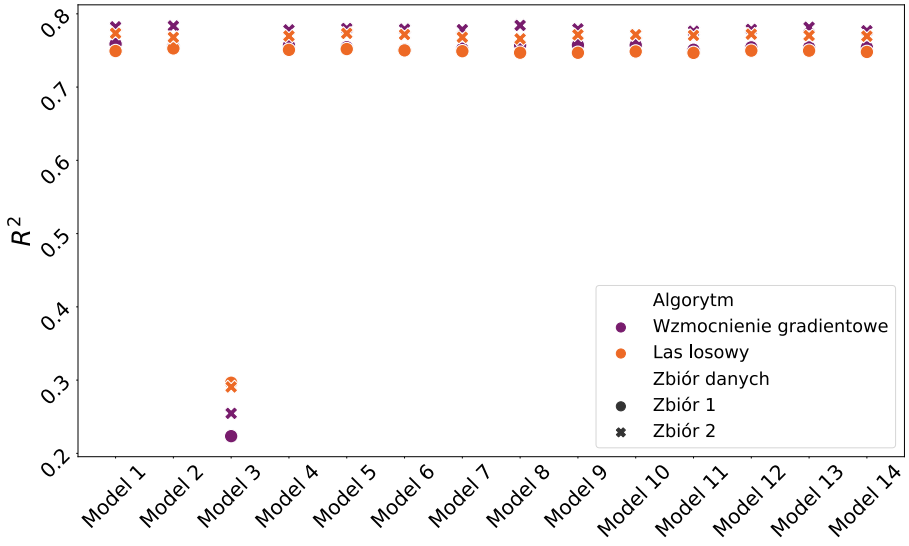
Maksymalny błąd osiągany przez modele wyniósł minimum 115 nm w czasie walidacji krzyżowej przeprowadzonej na zbiorze 2. W większości przypadków największy błąd algorytmu wzmocnienia gradientowego był mniejszy niż największy błąd algorytmu lasu losowego. Dla zbioru 1 błędy maksymalne były wyższe niż dla zbioru 2. Tak samo jak w przypadku poprzednich metryk oceniających, model 3 odstawał pod względem błędu maksymalnego.



RYSUNEK 6.5: Osiągnięty błąd maksymalny przez poszczególne modele.

Wartości współczynnika determinacji w walidacji krzyżowej wyniosły pomiędzy 0,75, a 0,79. Wyjątkiem jest model, w którym zastosowano tylko nowe deskryptory kwantowe, ponieważ wartości tego współczynnika wynoszą od 0,2 do

0,3, w zależności od algorytmu uczenia maszynowego i zbioru danych na których przeprowadzono ocenę. Natomiast, gdy zastosowanym algorytmem był las losowy, nie ma znaczącej różnicy w wartości współczynnika determinacji dla modelu 3 pomiędzy zbiorem 1 i zbiorem 2.



RYSUNEK 6.6: Współczynnik determinacji R^2 wyliczony dla predykcji wykonanej przez różne modele.

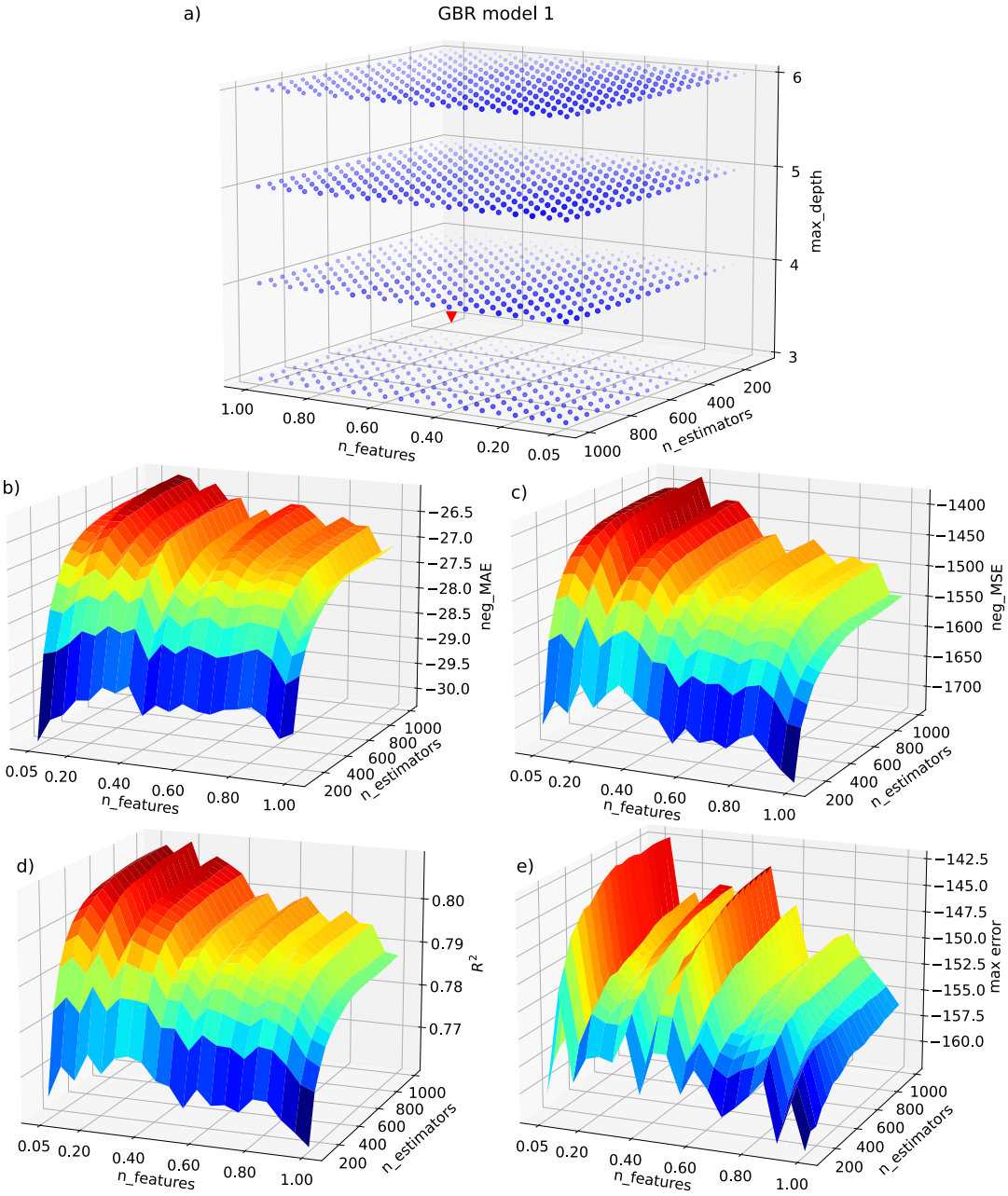
Wyniki regresji wielorakiej znacząco odbiegały od wyników osiąganych z użyciem innych algorytmów uczenia maszynowego. W związku z tym, zdecydowano się na porównanie wyników zestawów deskryptorów, które stosowane były dotychczas oraz tylko deskryptorów kwantowych. Otrzymane wyniki miar oceniających przedstawiono w tabeli 6.1. Oznaczenie LM1 w tabeli odnosi się do dotychczas zdefiniowanych zestawów deskryptorów, natomiast LM2 oznacza zestawy deskryptorów bez deskryptorów z biblioteki RDKit. Gdy zastosuje się tradycyjne deskryptory molekularne, wiele spośród współczynników determinacji jest ujemna lub bliska 0, co oznacza, że wyniki predykcji nie są znacząco lepsze od przypisania każdej predykcji wartości średniej. Na podstawie tabeli można wywnioskować, że w przypadku regresji wielorakiej lepsze rezultaty otrzymuje się stosując tylko deskryptory kwantowe, a spośród nich najlepszą relację deskryptory-właściwość można zbudować na podstawie zwykłej sumy właściwości kwantowych fragmentów. Należy jednak wyraźnie zaznaczyć, że regresja wieloraka daje gorsze rezultaty niż algorytmy lasu losowego oraz wzmocnienia gradientowego.

TABELA 6.1: Wyniki regresji wielorakiej. BM - błąd maksymalny, ŚBB - średni błąd bezwzględny, ŚBK - średni błąd kwadratowy. LM1 - modele trenowane i testowane na dotychczasowych deskryptorach, LM2 - modele trenowane i testowane tylko na nowych deskryptorach.

Zbiór 1								
	LM1				LM2			
Metryka Model	BM	ŚBB	ŚBK	R^2	BM	ŚBB	ŚBK	R^2
1	>500	>100	>10000	<-1	–	–	–	–
2	>500	49,8	>10000	<-1	242,3	58,1	5543	0,220
3	242,3	58,1	5543	0,220	242,3	58,1	5543	0,220
4	>500	46,9	>10000	<-1	249,9	68,0	7025	0,032
5	>500	47,1	>10000	<-1	366,1	69,5	9253	-0,339
6	>500	46,8	>10000	<-1	254,9	68,7	7165	0,011
7	>500	47,5	>10000	<-1	252,4	66,2	6789	0,064
8	>500	46,6	>10000	<-1	282,7	67,1	7059	0,020
9	>500	47,9	>10000	<-1	255,9	62,0	6211	0,133
10	>500	48,9	>10000	<-1	254,2	66,7	6835	0,052
11	>500	48,9	>10000	<-1	255,3	67,2	6910	0,042
12	>500	48,7	>10000	<-1	259,0	69,0	7196	0,007
13	423,9	47,7	7625	-0,157	471,8	71,0	>10000	<-1
14	>500	48,6	>10000	<-1	238,4	62,3	6061	0,155
Zbiór 2								
1	>500	>100	>10000	<-1	–	–	–	–
2	>500	53,3	>10000	<-1	222,2	57,3	5315	0,188
3	222,2	57,3	5315	0,188	222,2	57,3	5315	0,188
4	427,5	46,0	8824	-0,430	220,0	65,2	6593	-0,012
5	416,7	46,1	8986	-0,481	310,7	68,9	>10000	-0,477
6	464,6	47,5	9932	-0,610	197,5	64,3	6074	0,060
7	372,1	45,4	6622	-0,047	199,5	62,0	5793	0,102
8	405,0	46,4	7754	-0,257	194,9	62,4	5725	0,112
9	316,4	45,1	5595	0,125	206,0	57,0	5138	0,204
10	366,0	45,7	6553	-0,045	201,3	63,2	5920	0,085
11	370,8	45,7	6747	-0,079	201,1	63,8	6003	0,071
12	476,4	47,7	>10000	-0,739	198,2	65,8	6232	0,036
13	>500	65,9	>10000	<-1	389,2	70,3	>10000	<-1
14	454,0	47,6	>10000	-0,745	209,2	59,7	5426	0,159

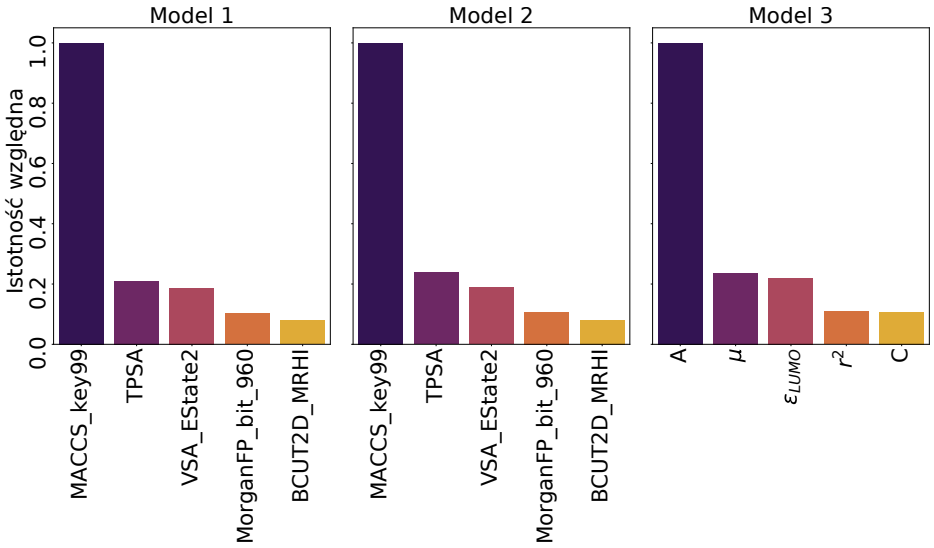
Powyższe wyniki wskazują, że najlepsze modele regresji oparte są o algorytmy lasu losowego i wzmocnienia gradientowego. Biorąc ten fakt pod uwagę, kalibracja parametrów algorytmów uczenia maszynowego została przeprowadzona dla obu algorytmów. W świetle otrzymanych wyników walidacji krzyżowej, postanowiono przeprowadzić ten proces dla zestawów deskryptorów nr 1 i 2 dla obu zbiorów danych. Na podstawie sporządzonych symulacji można zauważyć, że wydajność predykcji do pewnego stopnia rosła ze wzrostem liczby estymatorów (drzew decyzyjnych) w modelu. W przypadku liczby zmiennych losowanych przy podziale węzłów w drzewach decyzyjnych widać różnicę pomiędzy algorytmami - w przypadku lasu losowego wartości większości z zastosowanych metryk rosną, gdy zwiększamy maksymalną liczbę zmiennych. Odwrotnie sytuacja przedstawia się dla algorytmu wzmocnienia gradientowego. Najbardziej korzystna głębokość drzew decyzyjnych w zastosowanych algorytmach przyjmuje wartości z przedziału 4 do 9. Rysunek 6.7 przedstawia wykres wartości oceniających wydajność predykcji w zależności od zastosowanych różnych wartości parametrów algorytmu. Wartości niektórych metryk na wykresie przedstawione są jako wartości ujemne, aby ułatwić czytelność poprzez ujednolicenie w myśl zasady „im więcej, tym lepiej". Wykres a) przedstawia wyniki średniego błędu bezwzględnego w zależności od maksymalnej głębokości drzewa decyzyjnego, liczby drzew w algorytmie i jaka część zmiennych objaśniających była używana przez każde z drzew. Im ciemniejszy punkt, tym wynik lepszy. Wykresy b) - d) prezentują wartości oceniające przy stałej wartości maksymalnej głębokości równej 6, zmiennej liczbie drzew decyzyjnych oraz zmiennej części zmiennych objaśniających wprowadzanych do każdego drzewa. Zauważyć można, że najlepsze wyniki osiągano przy większej liczbie drzew decyzyjnych oraz mniejszej liczbie zmiennych objaśniających, które uwzględniały poszczególne drzewa decyzyjne.

Przedostatnim krokiem w badaniach było sprawdzenie metodą permutacji, które zmienne objaśniające mają największą istotność w modelowaniu długości fali emisji. Metoda ta polega na tasowaniu wartości zmiennych i sprawdzaniu jaki ma to wpływ na błędy predykcji - jeśli błąd się zwiększa, wówczas dana zmienna jest istotna dla modelu. Wyniki tej operacji przedstawiono na rysunku 6.8 jako wyniki względem wyniku osiągniętego przez najbardziej istotną zmienną. Na szczególne zaakcentowanie zasługuje fakt, że względem najbardziej istotnej zmiennej, istotność pozostałych jest niższa niż 20%.



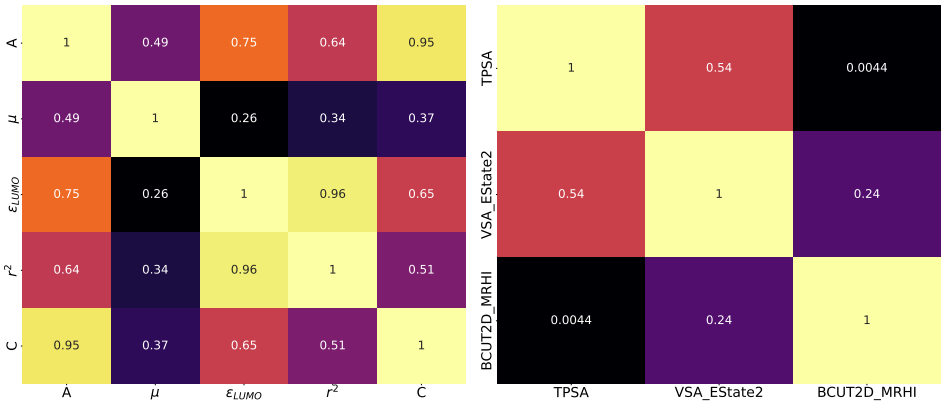
RYSUNEK 6.7: Wykres kalibracji parametrów algorytmu wzmocnienia gradientowego. Czerwonym trójkątem zaznaczono przypadek domyślnych wartości parametrów.

Wśród tradycyjnych deskryptorów molekularnych największy wkład w modelowanie emisji ma obecność wiązań podwójnych w cząsteczce (klucz MACCS nr 99), drugim istotnym parametrem jest całkowita powierzchnia polarna (TPSA), o wartości równej 0,21, powierzchnia Van der Waalsa, 0,18 (VSA_EState2) oraz jeden z deskryptorów BCUT.



RYSUNEK 6.8: Względne istotności zmiennych objaśniających w zależności od modelu.

Chcąc rzucić więcej światła na przyczyny słabych wyników osiąganych przez deskryptory kwantowe sprawdzono, jaka jest korelacja pomiędzy pięcioma najbardziej istotnymi zmiennymi (rysunek 6.9).

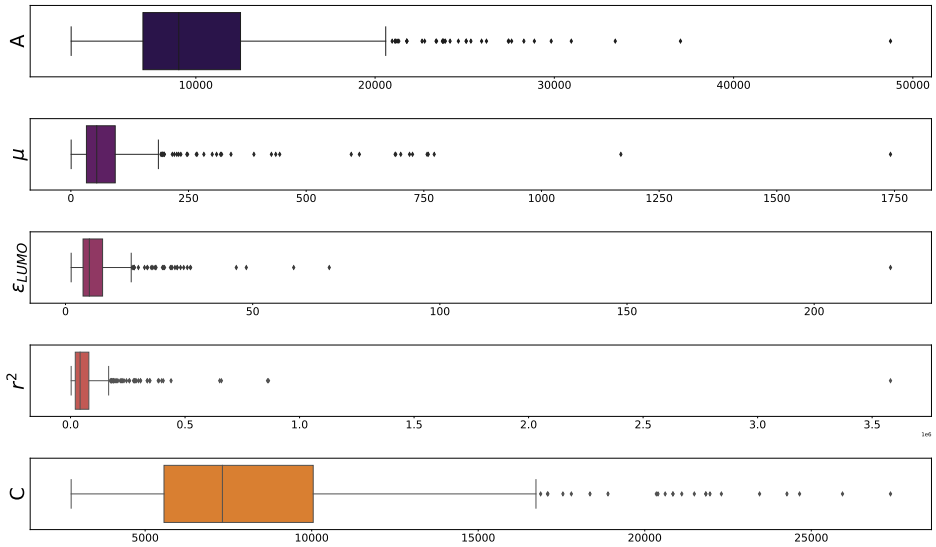


RYSUNEK 6.9: Współczynniki korelacji najbardziej istotnych zmiennych objaśniających. W tradycyjnych deskryptorach pominięto deskryptory przyjmujące wartości binarne (0 lub 1).

Bardzo silną korelacją charakteryzują się stałe rotacyjne A i C oraz energia LUMO i r^2. Stała rotacyjna A jest silnie skorelowana z energią LUMO. Obie stałe rotacyjne

A i C były skorelowane z r^2 w stopniu umiarkowanym. Pozostałe korelacje były słabe. W przypadku tradycyjnych deskryptorów molekularnych w stopniu umiarkowanym skorelowane ze sobą były VSA_EState2 i TPSA.

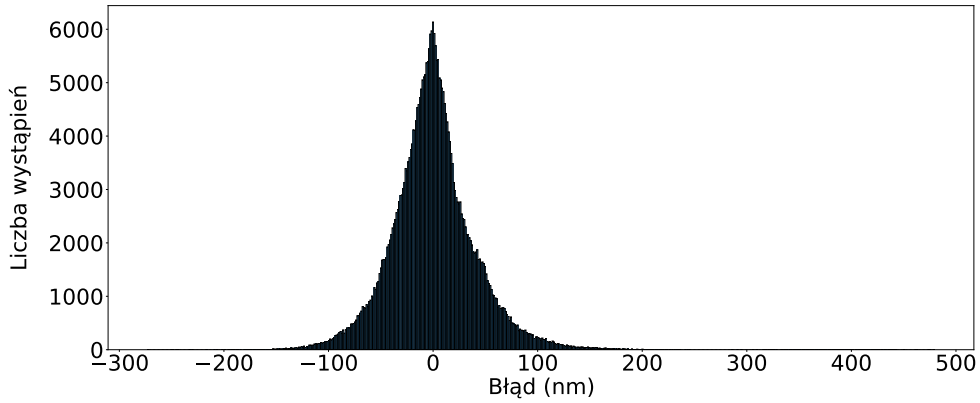
Wykreślono również zakresy wartości zastosowanych deskryptorów kwantowych (rysunek 6.10).



RYSUNEK 6.10: Zakres wartości najbardziej istotnych deskryptorów kwantowych.

Jak można zauważyć wartości stałej rotacyjnej A dla badanych związków wynoszą od 0 do 20500. Podobnie szeroki zakres wartości obserwowany jest dla drugiej stałej rotacyjnej, C przyjmującej wartości od 0 do 17000. Większość wartości momentu dipolowego mieści się w zakresie od 0 do 200. Zdecydowanie mniejszą wariancją cechuje się energia LUMO - większość wartości pochodzi z przedziału od 2 do 20. Najmniej różnią się od siebie wartości przyjmowane przez właściwość r^2, której wartości wynoszą głównie od 0 do 0,2. Warto zauważyć, że występują wartości odstające. W dodatku A na rys. A.1 zamieszczone zostały struktury związków chemicznych, które charakteryzują się odstającymi wartościami powyższych deskryptorów kwantowych.

Na koniec przeprowadzono tysiąckrotny bootstrap, w celu przeprowadzenia dużej liczby predykcji i wyznaczenia rozkładu osiąganych błędów. Przykładowy rozkład ukazany został na rysunku 6.11. Wartości osiągniętych błędów standardowych przedstawiono w tabeli 6.2.



RYSUNEK 6.11: Rozkład błędu.

TABELA 6.2: Błąd standardowy modeli 1 do 3.

Model	Zbiór 1	Zbiór 2
1	41,6	39,2
2	41,8	39,3
3	77,4	73,5

6.4 Rozwinięcie badania dzięki bazie danych QM-symex

Baza danych QM9 zawiera właściwości kwantowe związków w stanie podstawowym, natomiast właściwości optyczne wiążą się z przejściami cząsteczek do stanu wzbudzonego. Dlatego podjęto analizy mające na celu wygenerowanie deskryptorów z właściwości kwantowych fragmentów pochodzących z bazy danych QM-symex [79] zawierającej 173 tysiące związków chemicznych. Jest ona rozwinięciem bazy danych QM-sym [80]. Obie bazy danych charakteryzują się tym, że zawierają struktury symetryczne. Natomiast ważną cechą QM-symex, szczególnie interesującą z punktu widzenia badań jest to, że zawiera obliczone właściwości kwantowe związków chemicznych również w stanie wzbudzonym.

Bazę danych QM-symex tworzą pliki .xyz, po jednym na związek chemiczny, zawierające właściwości w stanie podstawowym, geometrię cząsteczki oraz 10 singletowych i tripletowych przejść elektronowych. Niestety napotkano szereg problemów ze skorzystaniem z bazy danych.

Pierwszym problemem, który wystąpił w trakcie wczytywania bazy danych QM-symex była zamiana geometrii cząsteczek na ich reprezentację SMILES za pomocą biblioteki RDKit. W celu przezwyciężenia tego problemu zastosowano bibliotekę Open Babel [81], a w szczególności jej API Pythona o nazwie Pybel [82]. Napotkano również problemy wynikające z tego, że nie wszystkie pliki reprezentujące

cząsteczki miały homogeniczną strukturę - np. brak lub dodatkowe znaki nowej linii sprawiały problem w automatyzacji wczytania całości bazy danych.

Ostatecznie udało się wczytać 105 tysięcy struktur chemicznych i ich właściwości kwantowe.

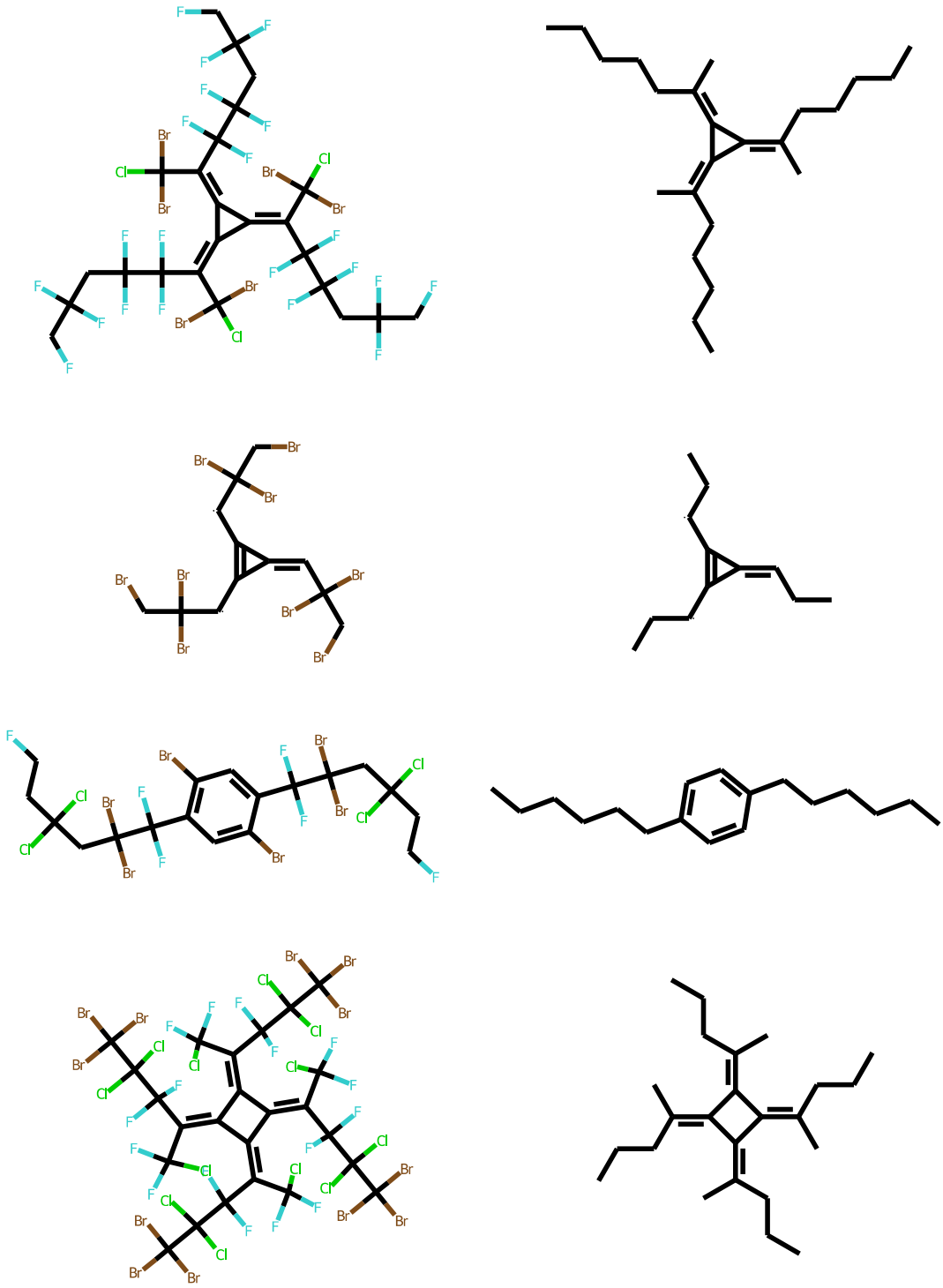
Największym i równocześnie dyskwalifikującym z dalszego badania problemem ze związkami chemicznymi z bazy danych QM-symex był wysoki poziom skomplikowania struktur samych molekuł. Znalezienie tak rozbudowanych podstrutkur w związkach eksperymentalnych było mało prawdopodobne. Chcąc zaradzić temu problemowi podjęto próbę zmniejszenia skomplikowania struktur poprzez usunięcie atomów halogenów i przyjęciem właściwości podstruktur nieuproszczonych jako właściwości podstruktur uproszczonych, jednak nie rozwiązało to problemu ze znalezieniem takich podstruktur w związkach eksperymentalnych. Przykładowe struktury wraz z uproszczeniem zobrazowano na rys 6.12.

6.5 Podsumowanie

Zaprezentowany sposób modelowania wybranej właściwości optycznej daje przybliżone wyniki, które mogłyby w pewnym stopniu wspomóc chemika projektującego związki typu OLED. Można znaleźć publikacje, gdzie raportowane są lepsze osiągnięcia w predykcji właściwości optycznych związków organicznych [83, 84].

Nietradycyjne deskryptory molekularne pochodzące od właściwości kwantowych podstruktur związków nie sprawdziły się w zastosowaniu do modelowania długości fali, przy której następuje maksimum emisji światła. Zastosowanie ich wspólnie z tradycyjnymi deskryptorami molekularnymi powoduje niepotrzebną komplikację i złożoność obliczeniową, opóźniając otrzymanie wyniku.

Wpływ na osiągnięcia predykcyjne mogła mieć również przyjęta metodyka badań - nastąpiła duża redukcja zbioru danych, co spowodowało utratę informacji. Również ograniczenie deskryptorów molekularnych do deskryptorów pomijających geometrię cząsteczki mogło mieć wpływ na otrzymywane wyniki. Istnieją zastosowania, w których zastosowanie takich deskryptorów molekularnych poprawia otrzymywane wyniki [85]. Ponadto, w przyjętej metodyce wartości deskryptorów kwantowych nie były standaryzowane, co mogło mieć wpływ na wyniki modelowania.



RYSUNEK 6.12: Przykładowe związki z bazy danych QM-symex oraz wypróbowane uproszczenie struktury. Po lewej znajdują się oryginalne związki, a po prawej związki po usunięciu wszystkich atomów chlorowców.

Ponadto zastosowana baza danych QM9 zawiera właściwości kwantowe związków w stanie podstawowym, nie bierze więc pod uwagę sytuacji wzbudzenia cząsteczki. Może mieć to fundamentalne znaczenie w prognozowaniu takich właściwości jak długość fali emisji. Podjęto próbę zastosowania bazy danych QM-symex zawierającej właściwości cząsteczek w stanie wzbudzonym. Niestety jednak struktury cząsteczek, które się w niej znajdują są zbyt rozbudowane i skomplikowane aby mogły posłużyć do celów poszukiwania fragmentów.

Można sobie wyobrazić również takie skonstruowanie szeregu predykcyjnego, w którym pierwszym etapem byłaby redukcja treningowej bazy danych na podstawie miar podobieństwa (np. rozszerzone indeksy podobieństwa [86, 87]) lub metod klasteryzacji i budowanie modelu predykcyjnego „w locie".

Rozszerzenie bazy QM9

7.1 Dobór związków do obliczeń

Wspomniane ograniczenie fragmentarycznych deskryptorów kwantowych, sprawiło, że podjęto pracę nad rozszerzeniem bazy danych właściwości kwantowych QM9 o związki zawierające w swej strukturze inne halogeny, niż tylko fluor. W celu zachowania możliwie podobnych struktur zdecydowano się na znalezienie związków zawierających atom fluoru i jego zamianę na chlor, brom i jod. Takie rozszerzenie bazy danych kwantowych pozwalało na mniejsze ograniczanie bazy danych eksperymentalnych. Korzystając z biblioteki RDKit wygenerowano wstępne geometrie związków za pomocą metody EmbedMolecule. Metoda ta jest implementacją metody ETKDG [88]. Za pomocą skryptu Pythona generowano pliki wsadowe do obliczeń, w których zawierano wygenerowaną początkową geometrię atomów w cząsteczce. Osobny skrypt w języku Bash służył do automatyzacji uruchamiania kolejnych obliczeń bez ingerencji człowieka.

7.2 Obliczenia

Celem zachowania spójności między bazą rozszerzoną, a oryginałem, obliczenia wykonywano metodą B3LYP, która jest jedną z metod teorii funkcjonału gęstości elektronowej (DFT), w której zastosowany jest korelacyjno-wymienny funkcjonał hybrydowy Becke'go, Lee-Yanga-Parra. Jednak ze względu na to, że baza funkcyjna 6-31G (2p,2d) zastosowana w bazie danych QM9, nie występuje w programie Psi4 oraz jednym z atomów związków wziętych do obliczeń był jod, zdecydowano się na bazę funkcyjną def2-svp, będącą bazą funkcyjną powstała poprzez ulepszenie bazy rodziny Dunning'a opisującej korelację elektronową, funkcje te są spolaryzowane, typu split-valence. Polaryzowalność i moment dipolowy obliczano metodą CCSD, czyli sprzężonych klasterów ze wzbudzeniami pojedynczymi i podwójnymi. Same obliczenia wykonywano początkowo w trzech, a później w pięciu iteracjach,

w przypadku związków zawierających chlor oraz czterech iteracjach dla związków zawierających pozostałe chlorowce, gdzie do kolejnej iteracji brano związki, których obliczenia nie zakończyły się sukcesem. W oryginalnej bazie danych związków zawierających fluor było 2163, więc planowano więc wykonanie obliczeń dla sumarycznie 6489 związków o długościach od 2 do 9 atomów cięższych od wodoru. W pierwszej kolejności wykonywano obliczenia dla związków zawierających w swojej strukturze atomy chloru. Wyniki wczytywano przy użyciu skryptu Pythona. Skrypt ten dodatkowo korzystał z biblioteki AaronTools [89], która pozwala na automatyczne zastosowanie poprawki RRHO (sztywny rotator-oscylator harmoniczny). Jest to przybliżenie ruchu atomów w cząsteczce stosowane do wprowadzenia poprawek w obliczonych właściwościach termodynamicznych - w przypadku omawianych obliczeń, do obliczonych wartości energii wewnętrznej, entalpii i energii swobodnej.

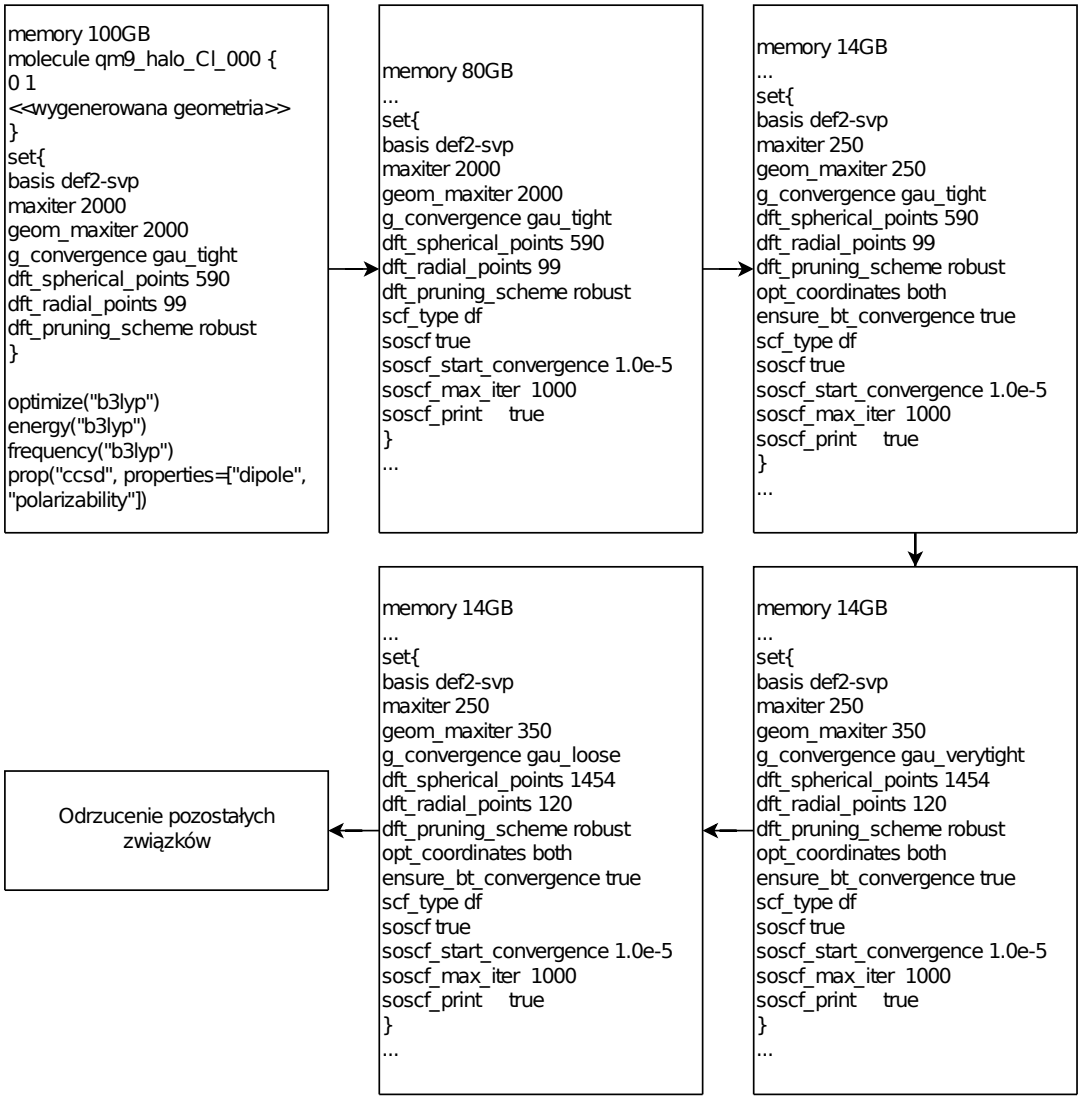
7.3 Optymalizacja procesu obliczeniowego

Obliczenia z dziedziny chemii kwantowej są procesem bardzo złożonym i nie ma jednej uniwersalnej metody ani przepisu na wykonanie obliczeń które zakończyłyby się sukcesem. Dlatego też zazwyczaj w razie pierwszego niepowodzenia obliczenia powtarza się ze zmienionymi parametrami i ustawieniami. W szczególnie trudnych przypadkach obliczenia powtarza się wielokrotnie. Dlatego też w toku realizacji pracy doktorskiej, początkowe podejście do wykonania niezbędnych obliczeń zakładało 3 iteracje, przy czym pierwsza okazała się być procesem długotrwałym - czas trwania ok. pół roku. Obliczenia uruchamiane były po kolei, każdemu wywołaniu przydzielano po 32 procesory do obliczeń równoległych oraz 100GB pamięci operacyjnej. Liczba iteracji optymalizacji geometrycznej cząsteczki ustawiona została na 2000. W drugim kroku obliczeniowym zmieniono rozdzielczość siatki sferycznej i próg zbieżności geometrycznej na bardziej wymagający, by w trzecim kroku pozostawić rozdzielczość siatki, ale zmniejszyć wymagania dotyczące zbieżności geometrycznej. W przedstawiony sposób, z różnych powodów 275 obliczeń zakończyło się porażką.

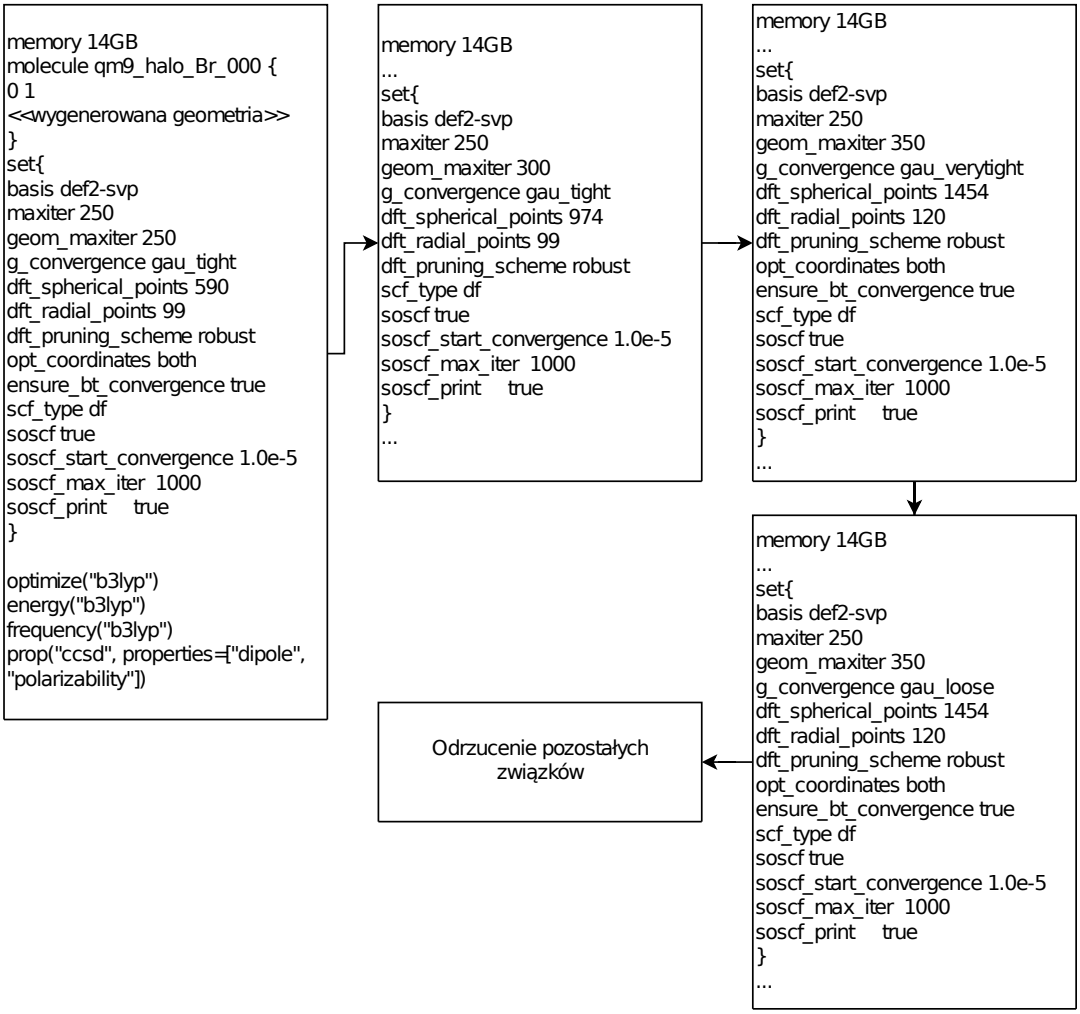
Celem zwiększenia liczby poprawnie zakończonych obliczeń zdecydowano się na powtórne wykonanie zadania w przypadku związków, które nie dały rezultatu w pierwszej iteracji, odrzucając wyniki dwóch pozostałych iteracji. W związku

z tym, że szybkość obliczeń nie rośnie liniowo z liczbą procesorów w obliczeniach równoległych, badziej adekwatnym rozwiązaniem w celu przyspieszenia wykonania obliczeń jest uruchomienie kilku obliczeń na mniejszej liczbie procesorów. Mając to na uwadze, aby przyspieszyć pracę, w drugiej iteracji podzielono obliczenia na 3 serie, gdzie każdej z serii przydzielono po 10 procesorów i 80 GB pamięci operacyjnej, a w każdej następnej na 8 serii, po 4 procesory i 14 GB pamięci na serię. Zmniejszono również liczbę iteracji optymalizacji geometrycznej do 250. Osiągnięto w ten sposób znaczne przyspieszenie procesu obliczeń. Schemat plików wsadowych w poszczególnych krokach obliczeniowych przy obliczeniach dla związków chlorowanych przedstawiono na rysunku 7.1. Zmieniono pierwotny plan z wykonania trzech iteracji, na pięć iteracji. Istotnym okazał się wprowadzony w trzeciej iteracji parametr *opt_coordinates both*, który poleca programowi Psi4 traktowanie podanej geometrii jako współrzędne kartezjańskie lub współrzędne wewnętrzne. Czwarta iteracja zwiększała liczbę punktów w siatce sferycznej i radialnej z jednoczesnym zwiększeniem wymagań dotyczących zbieżności geometrycznej. Piąta iteracja obniżała wymagania dotyczące zbieżności geometrycznej. Ostatecznie sukcesem zakończyły się obliczenia właściwości 2062 związków chlorowanych, czyli porażką zakończyło się 101.

Na podstawie doświadczenia w obliczeniach związków chlorowanych, zdecydowano się wykonać obliczenia dla związków bromowanych w czterech krokach iteracyjnych (schemat na rysunku 7.2). W tym przypadku już od pierwszej iteracji zastosowano opcję *opt_coordinates both*. W drugiej zwiększano rozdzielczość siatki sferycznej i maksymalną liczbę iteracji optymalizacji geometrycznej. Trzecia iteracja obejmowała dalsze zwiększenie rozdzielczości siatki sferycznej, zwiększenie rozdzielczości siatki radialnej, zwiększenie wymagań dotyczących zbieżności geometrycznej oraz maksymalnej liczby iteracji optymalizacji geometrycznej. W ostatniej iteracji zmniejszono wymagania dotyczące zbieżności geometrycznej. Taki sam sposób planowano zastosować w obliczeniach dla związków jodowanych.



RYSUNEK 7.1: Poszczególne kroki obliczeniowe dla związków chlorowanych.



RYSUNEK 7.2: Poszczególne kroki obliczeniowe dla związków bromowanych.

7.4 Baza danych QM9-extended

Niezależnie od powyższego, Lim i in. opublikowali bazę danych QM9-extended [90], będącą rozszerzeniem bazy danych QM9. Badacze, którzy podjęli się tego zadania wybrali z GDB17 związki zawierające Cl i S, a następnie wykonali dla nich obliczenia kwantowe. Dostępność związków zawierających chlor oraz siarkę w tej bazie danych sprawiła, że zdecydowano się na kombinację nowo opublikowanej bazy danych z własnym rozszerzeniem bazy danych QM9. W związku z tym, że metoda wyboru związków rozszerzających bazę danych QM9 była inna w przypadku pracy Lim i in. niż w przypadku własnego rozszerzenia, istniała szansa, że w obu rozszerzeniach znajdują się te same związki. Poszukiwania pokrywających się związków chemicznych wykonano metodą współczynnika Tanimoto. W 290 przypadkach współczynnik ten wynosił 1, co oznacza, że porównywane związki były identyczne. W tych przypadkach pierwszeństwo postanowiono dać właściwościom kwantowym pochodzącym z bazy danych QM9-extended.

7.5 Baza danych QM9-extended-plus

W toku realizacji niniejszej pracy doktorskiej baza QM9-extended powiększona została o wyniki obliczeń 1781 związków zawierających przynajmniej jeden atom chloru oraz 2020 związków zawierających przynajmniej jeden atom bromu. Połączenie bazy danych QM9-extended i wyników własnych obliczeń nazwano bazą danych QM9-extended-plus. Zawiera ona 157488 związków chemicznych i 11 właściwości kwantowych wymienionych w ostatniej kolumnie tabeli 7.1. W stosunku do pierwowzoru jest to o 4 właściwości kwantowe mniej. Stałe rotacyjne są nieobecne w bazie danych QM9-extended, a program Psi4 nie oblicza rozmiaru chmury elektronowej.

Nadmienić należy, że praca nad obliczeniami rozpoczęła się w grudniu 2021 roku, a koniec obliczeń przypadł na 26 czerwca 2023 roku. Wygenerowane dane (pliki wsadowe, logów i pliki wyjściowe) zajmują 16,3 GB pamięci komputera.

TABELA 7.1: Właściwości kwanowe w omawianych bazach danych. Bazy danych: A - QM9, B - QM9-extended, C - QM9-extended-plus. + - właściwość obecna w bazie danych, x - brak w bazie danych.

Właściwość	Jednostka	Opis	A	B	C
A	GHz	Stała rotacyjna A	+	x	x
B	GHz	Stała rotacyjna B	+	x	x
C	GHz	Stała rotacyjna C	+	x	x
mu	Debye	Moment dipolowy	+	+	+
alpha	$Bohr^3$	Polaryzowalność	+	+	+
homo	Hartree	Energia HOMO	+	+	+
lumo	Hartree	Energia LUMO	+	+	+
gap	Hartree	Różnica pomiędzy LUMO i HOMO	+	+	+
r2	$Bohr^2$	Rozmiar chmury elektronowej	+	+	x
zpve	Hartree	Energia wibracyjna punktu zerowego	+	+	+
U0	Hartree	Energia wewnętrzna w temperaturze 0 K	+	+	+
U	Hartree	Energia wewnętrzna w temperaturze 298.15 K	+	+	+
H	Hartree	Entalpia w temperaturze 298.15 K	+	+	+
G	Hartree	Energia swobodna w temperaturze 298.15 K	+	+	+
Cv	cal/(mol K)	Pojemność cieplna w temperaturze 298.15 K	+	+	+

Pomimo początkowych planów, zdecydowano, że w związku z bardzo dużym zapotrzebowaniem na zasoby obliczeniowe, długim czasem trwania samych obliczeń i polskim miksem energetycznym, ślad węglowy procesu obliczeniowego jest zbyt duży. W razie szczególnego wzrostu zainteresowania zbiorem danych, możliwe jest powtórne uruchomienie procesu obliczeniowego, jednak wskazana jest dalsza optymalizacja tego procesu, aby skrócić czas wykonywania obliczeń. Ponadto, celem racjonalnego korzystania z zasobów Ziemi, dobrą decyzją byłoby korzystanie z zasobów obliczeniowych zlokalizowanych w krajach o zdecydowanie mniejszej emisyjności dwutlenku węgla z sektora energetycznego.

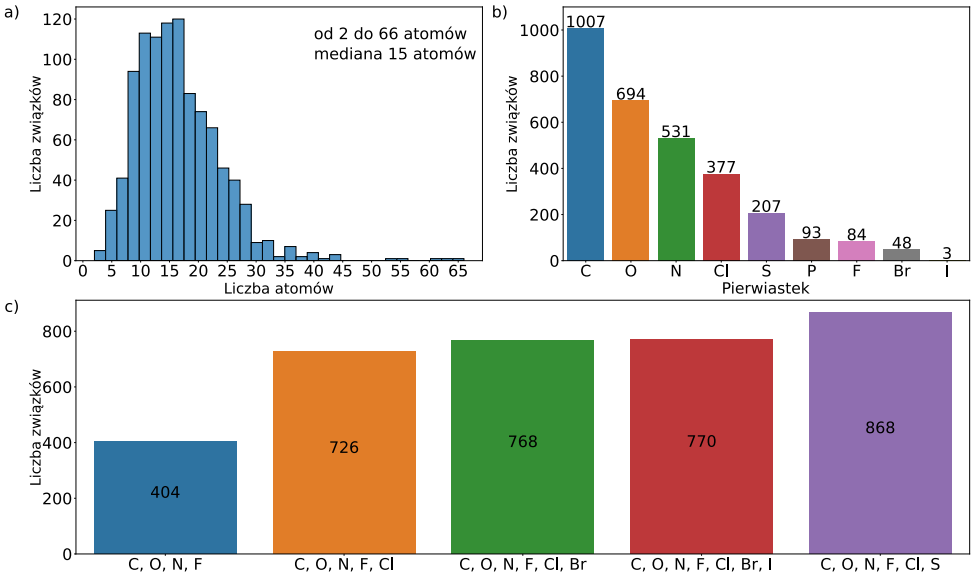
Rozbudowana baza danych właściwości kwantowych umożliwiła poszerzenie domeny aplikowalności fragmentarycznych deskryptorów kwantowych w pierwszej kolejności o związki chemiczne zawierające atomy chloru, a następnie bromu. Jednak czas wykonania obliczeń sprawił, że w badaniu opisanym w następnym rozdziale zastosowano tylko bazę danych poszerzoną o związki chlorowane do budowy klasyfikatora związków chemicznych ze względu na potencjał do biokumulacji w organizmach żywych. Z pełnej wersji bazy danych QM9-extended-plus wygenerowane zostały kwantowe pary atomów opisane w rozdziale 9.

Klasyfikacja związków chemicznych ze względu na biokumulację w ogranizmach żywych

8.1 Metodyka

Baza danych biokumulacji

Dane eksperymentalne do modelowania pobrano z repozytorium qsardb.org [91]. Pozyskana baza danych biokumulacji [92, 93] zawiera 1007 związków chemicznych z wyznaczoną eskperymentalnie wartością współczynnika biokumulacji. Na podstawie tych wartości autorzy dokonali klasyfikacji na związki ulegające lub nie ulegające kumulacji w organizmach żywych (logBCF < 3). Celem oceny możliwości zastosowania fragmentarycznych deskryptorów kwantowych, wiedząc, że ich domena aplikowalności ograniczona jest składem pierwiastkowym dokonano analizy składu pierwiastkowego w bazie danych biokumulacji. Jej wyniki przedstawiono na rysunku 8.1. Można zauważyć, że większość zwiazków w bazie danych składa się z nie więcej niż 30 atomów, cięższych od wodoru, a najmniejsze zbudowane są z dwóch atomów. Ze względu na małą liczbę cząsteczek zawierających atomy bromu i jodu w bazie danych, pomimo początkowych planów włączenia ich do symulacji, nie zostały one wzięte pod uwagę. Dzięki takiemu podejściu nie było potrzeby ukończenia wszystkich czasochłonnych obliczeń kwantowych.



RYSUNEK 8.1: Analiza składu atomowego związków w bazie danych eksperymentalnych. a) rozkład liczby związków w zależności od liczby atomów, z których składa się dany związek; b) liczba związków zawierających przynajmniej jeden atom danego pierwiastka; c) liczba związków składających się z atomów tylko wybranej grupy pierwiastków.

Ponadto ograniczając się tylko do związków złożonych z atomów C, O, N, F, baza danych musiałaby być zredukowana do tylko 404 związków. Dodanie chloru do listy pierwiastków, pozwala objąć w symulacjach dodatkowe 322 związki chemiczne, a następne chlorowce dałyby kolejno 42 związki dla bromu i 2 związki dla jodu. Uwzględnienie w badaniu związków zawierających atomy siarki pozwalaja uwzględnić kolejne 242 związki chemiczne. Oryginalną bazę danych zredukowano zatem do 868 związków zbudowanych tylko z atomów C, O, N, S, F i Cl. Wśród nich 681 było zaklasyfikowanych jako związki nie kumulujące się, a 187 jako biokumulujące się, więc dane musiały być zaklasyfikowanego jako dane niezbalansowane, co jest częstym problemem w chemicznych bazach danych.

Jednak ze względu na brak wykrytych podstruktur z bazy danych QM9-extended-plus, baza danych eksperymentalnych musiała zostać zredukowana o kolejne 12 związków. Zatem ostatecznie do badania wzięto 856 związków chemicznych.

Fragmentaryczne deskryptory kwantowe

Do obliczania deskryptorów kwantowych zastosowano bazę danych QM9-extended-plus, o której wspomniano w rozdziale 7. Deskryptory te w opisywanym badaniu

zdefiniowano w odmienny sposób w stosunku do opisanych w rozdziale 6. Dokonane zostało grupowanie ze względu na liczbę atomów fragmentów, spodziewając się, że fragmenty o jednych długościach mogą nieść więcej informacji niż inne. Ogólnie mówiąc, deskryptory te są sumą właściwości kwantowych fragmentów o danej długości, podzielone przez liczbę wykrytych fragmentów. Wygenerowano 2 serie deskryptorów pochodzących z właściwości kwantowych - jakościową i ilościową. Poniższy wzór przedstawia sposób wyznaczania wartości deskryptorów. W deskryptorach jakościowych parametr N_{occ} przyjmował maksymalną wartość równą 1. W ten sposób przechowują one tylko informację o tym, że dany fragment występuje w cząsteczce. Pełną informację dotyczącą liczby wystąpień danego fragmentu zawierają deskryptory ilościowe.

$$
\begin{aligned}
&FQD^{i}_{prop} = \frac{\sum_{j(i)} prop_j * N_{occ}}{n}, \\
&prop - \text{właściwość kwantowa fragmentu } j \\
&i - \text{liczba atomów w j-tym fragmencie} \\
&j(i) - \text{fragment zawierający } i \text{ atomów} \\
&n - \text{liczba wykrytych fragmentów o liczbie atomów } i \\
&N_{occ} - \text{liczba ile razy występuje dany fragment w cząsteczce}
\end{aligned}
\tag{8.1}
$$

Ponadto w badaniu wyznaczono również wartości deskryptorów przypadające na jeden atom związku poprzez podzielenie wartości deskryptoru przez liczbę atomów w związku chemicznym.

Przygotowanie danych

Pierwszym krokiem w tworzeniu modelu klasyfikującego było wygenerowanie deskryptorów. Zdecydowano się na sprawdzenie trzech grup zmiennych objaśniających - fragmentaryczne deskryptory kwantowe, deskryptory z biblioteki RDKit wraz z kluczami MACCS oraz odciskami palców Morgana, trzecią grupą była kombinacja poprzednich. W związku z występującymi brakami w danych spowodowanymi brakiem dopasowania podstruktur o danej liczbie atomów dokonano ich uzupełnienia. Następnie klasy biokumulacji zostały zakodowane jako 0 i 1, gdzie 1 oznaczało, że związek kumuluje się w organizmach żywych. Zgodnie z dobrymi praktykami w badaniach obejmujących uczenie maszynowe [94, 95] wydzielono

również z bazy danych eksperymentalnych 330 związków, które utworzyły zbiór danych testowych, zachowując stosunek reprezentacji klas taki sam jak w całej bazie danych. Reszta danych przeznaczona została do opracowania modelu klasyfikującego. W celu zmniejszenia wpływu nadreprezentacji jednej z klas na projektowane modele, dane treningowe poddawano losowemu nadpróbkowaniu klasy mniejszościowej. W walidacjach krzyżowych dane walidacyjne nie były poddawane temu procesowi. Następnie dokonywano standardyzacji zmiennych opisujących. Tak przygotowane dane były wprowadzane na wejściu do algorytmów uczenia maszynowego.

Ze względu na dużą wymiarowość danych sprawdzono również wariant przygotowania danych, w którym dodano również analizę najważniejszych składowych (PCA) z uwzględnieniem tylko tych składowych, które zachowują 95% wariancji danych. Ten sposób redukcji wymiarowości macierzy zmiennych objaśniających był stosowany tylko dla zmiennych ciągłych. Zmienne dyskretne (klucze MACCS i odciski palców Morgana) pozostawiono niezmienione.

Brakujące dane

W celu wybrania najlepszego rozwiązania problemu brakujących danych, sprawdzono 3 warianty działania:

- zostawienie brakujących danych,
- podstawienie wartościami za pomocą algorytmu k najbliższych sąsiadów,
- wstawienie zer, z dodaniem dodatkowych zmiennych objaśniających, w których zakodowano, czy dane zostały podstawione.

Pierwszy wariant ograniczał możliwości zastosowania algorytmów uczenia maszynowego do tych, które pozwalają na istnienie brakujących danych w macierzy zmiennych opisujących. Drugi sposób wydaje się być nieodpowiedni do tego rodzaju brakujących danych, ponieważ nie były to dane brakujące w sposób przypadkowy. Sam mechanizm generowania deskryptorów dopuszcza otrzymywanie brakujących danych. Sprawdzenia dokonano na etapie prototypowania algorytmów uczenia maszynowego w pięciokrotnej walidacji krzyżowej.

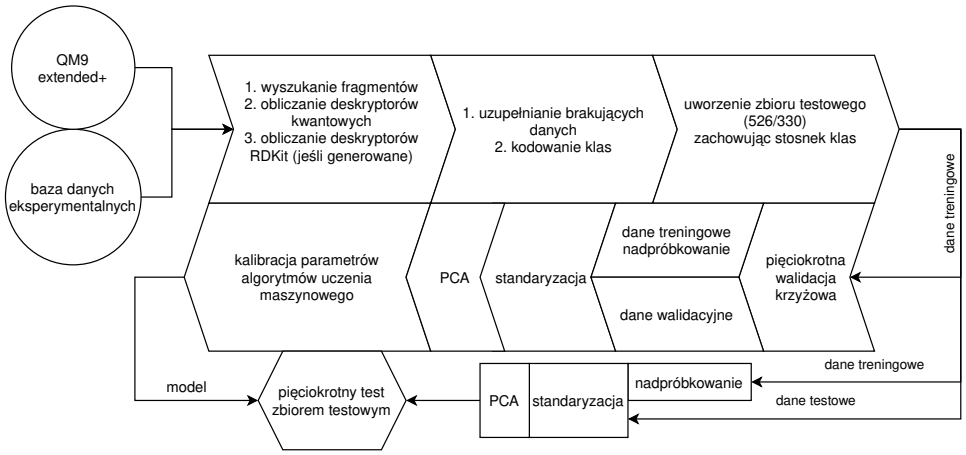
Modelowanie

Po przygotowaniu danych podjęto się wyboru odpowiedniego algorytmu uczenia maszynowego do zbudowania modelu klasyfikującego. Wybierano spośród następujących algorytmów: regresji logistycznej, lasu losowego, wzmocnienia gradientowego, k najbliższych sąsiadów (KNN), wspartych wektorów, XGBoost [96] oraz LightGBM [97]. Wstępne modelowanie, którego celem było wybranie algorytmu uczenia maszynowego cechującego się najlepszym dopasowaniem do danych, wykonywano w pięciokrotnej walidacji krzyżowej. Wyniki otrzymane za pomocą deskryptorów kwantowych porównano również do przypadku losowego przypisywania klas. Ocena modeli polegała na porównaniu otrzymywanych metryk - F miary oraz dokładności zbalansowanej, jednak za istotniejszy parametr przyjęto F miarę, ponieważ na tą metrykę składają się precyzja i czułość, które nie biorą pod uwagę wyników prawdziwie ujemnych, których było dużo ze względu na brak zbalansowania danych. Po wyborze algorytmów (po jednym na zestaw deskryptorów), przeprowadzono kalibrację ich parametrów, aby zmaksymalizować wynik F miary.

Kalibracja ta wykonywana była poprzez ocenę F miary otrzymywanej w pięciokrotnej walidacji krzyżowej wybranego algorytmu uczenia maszynowego z parametrami będącymi kombinacją wartości z zadeklarowanych zakresów. Na tej podstawie wybrano zestaw parametrów, który najlepiej klasyfikował zbiór treningowy.

Ostatnim elementem procesu budowy modeli było sprawdzenie ich przy zastosowaniu wspominanego wcześniej, składającego się z 330 emelentów zbioru testowego. Proces ten wykonano pięciokrotnie, zmieniając za każdym razem wartość ziarna (seed) generatora pseudolosowości danego algorytmu.

Rysunek 8.2 stanowi przedstawienie metodyki badań w formie graficznej.



RYSUNEK 8.2: Graficzne przedstawienie metodyki badań obejmujących fragmentaryczne deskryptory kwantowe.

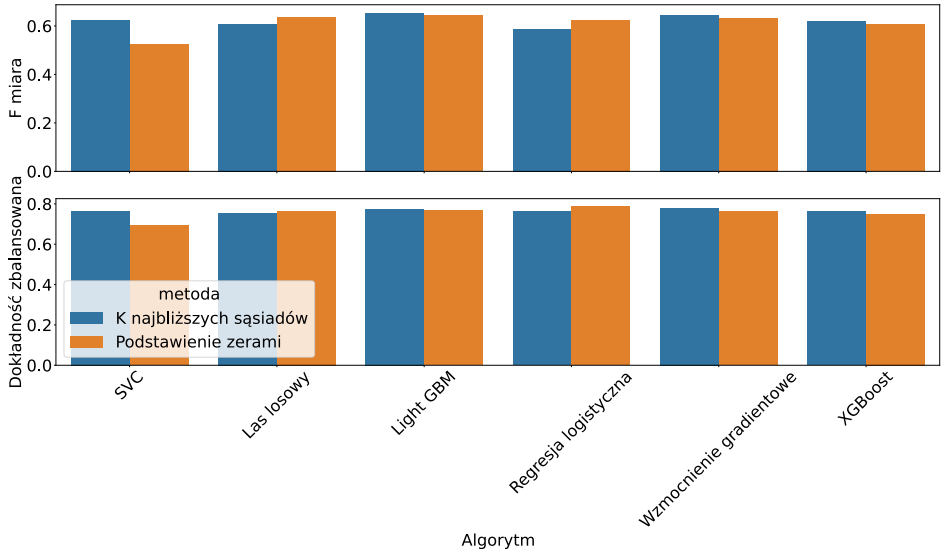
8.2 Wyniki i dyskusja

Fragmentaryczne deskryptory kwantowe i dobór metody uzupełniania brakujących danych

W związku z dużym podobieństwem w sposobie zdefiniowania jakościowych i ilościowych fragmentarycznych deskryptorów kwantowych dokonano sprawdzenia, czy w przypadku analizowanego zbioru danych jest różnica w rozkładzie ich wartości. Do porównania rozkładów zastosowano współczynnik zmienności zdefiniowany jako stosunek odchylenia standardowego do wartości średniej arytmetycznej. Analiza taka wykazała, że wartości tych współczynników dla odpowiadających sobie deskryptorów jakościowych i ilościowych są sobie równe. Na tej podstawie zdecydowano się na ograniczenie zastosowania nowych deskryptorów molekularnych tylko do deskryptorów jakościowych.

Rysunek 8.3 przedstawia wyniki walidacji krzyżowej przy zastosowaniu dwóch różnych metod uzupełniania brakujących danych. Trzecia metoda ograniczała wachlarz dostępnych algorytmów uczenia maszynowego do XGBoost i LightGBM, które same uzupełniają brakujące dane poprzez podstawienie zerami.

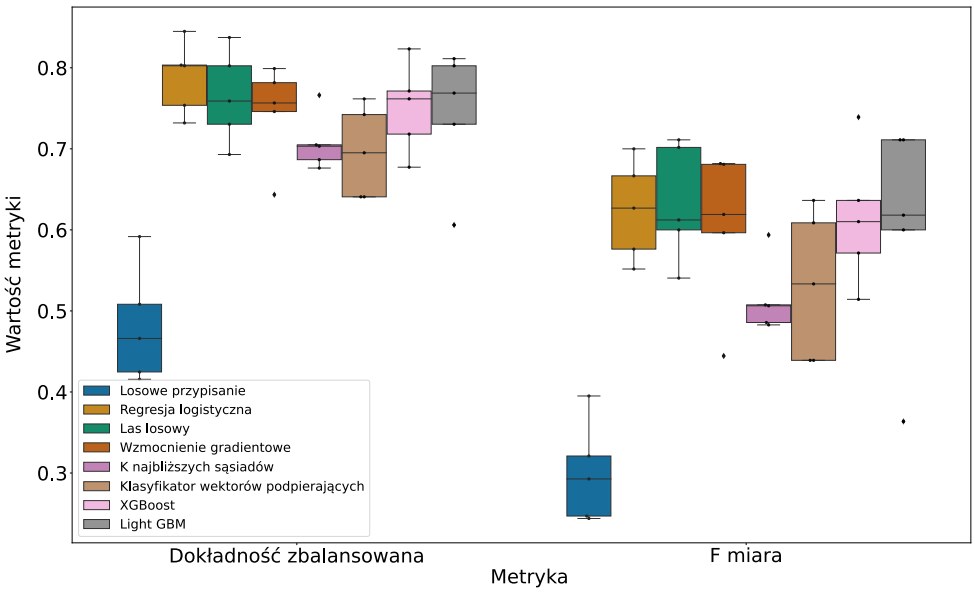
Zauważyć można, że wyniki kształtują się podobnie. Sposób podstawienia brakujących wartości najbardziej wpływa na wyniki otrzymywane w przypadku modelowania algorytmami wspartych wektorów oraz k najbliższych sąsiadów. Mając na uwadze niewielkie różnice w otrzymanych wynikach oraz fakt, że brakujące wartości powstają w sposób nielosowy, zdecydowano się na podstawienie zerami z zaznaczeniem, czy wartość była brakująca.



RYSUNEK 8.3: Walidacja krzyżowa dwóch metod uzupełnienia brakujących danych. SVC - klasyfikator wektorów podpartych.

Ocena algorytmów

Wstępny wybór algorytmu uczenia maszynowego, który pozwalał na osiągnięcie najlepszych wyników klasyfikacji dokonany został na podstawie przeprowadzonej walidacji krzyżowej. Rysunek 8.4 przedstawia wyniki walidacji krzyżowej dla różnych algorytmów, z zastosowaniem fragmentarycznych deskryptorów kwantowych.



RYSUNEK 8.4: Wyniki pięciokrotnej walidacji krzyżowej klasyfikacji związków ze względu na potencjał do biokumulacji w organizmach żywych.

Wszystkie algorytmy osiągnęły lepsze rezultaty niż klasyfikacja losowa. Przez wzgląd na występowanie odstających wartości, porównania dokonano w oparciu o średnią z rozstępu międzykwartylowego. Na tej podstawie określono, że najlepsze w klasyfikacji okazały się być algorytmy oparte na drzewach decyzyjnych, spośród których najwyższe wyniki osiągnął algorytm LightGBM. Osiągnął on wyniki lepsze od innych algorytmów o od 0,8% do 28% pod względem F miary. Podobnie było, gdy zastosowano kombinację deskryptorów tradycyjnych i fragmentarycznych deskryptorów kwantowych - najlepszym algorytmem okazał się algorytm LightGBM, którego F miara była wyższa o ponad 2% względem innych algorytmów. W przypadku zasosowania samych deskryptorów molekularnych wygenerowanych z biblioteki RDKit, najlepszymi wynikami charakteryzował się model oparty o algorytm XGBoost. W tym wypadku F miara była wyższa o minimum 3% względem innych algorytmów.

Zastosowanie analizy głównych składowych w procesie przygotowania danych wywołało spadek osiąganych wartości oceniających dopasowanie klasyfikatorów. W takim przypadku podczas wstępnej oceny algorytmów najlepsze wyniki osiągał las losowy, gdy cząsteczki reprezentowano FQD oraz tradycyjnymi deskryptorami molekularnymi. Dla kombinowanego zestawu wielkości objaśniających najlepszym dopasowaniem do danych charakteryzował się algorytm wzmocnienia gradientowego. Jednak dokładność zbalansowana i F miara osiągane przez las losowy po zastosowaniu PCA w stosunku do tych samych parametrów osiąganych przez algorytm LightGBM bez stosowania PCA były niższe o odpowiednio 4 i 9%.

Analiza głównych składowych

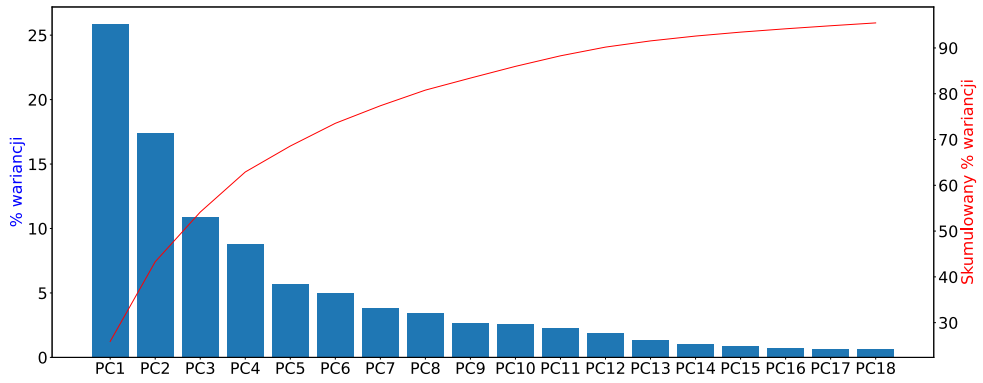
Zastosowanie analizy głównych składowych zredukowało liczbę zmiennych objaśniających w przypadku FQD ze 176 do 18 zmiennych. Daje to zmniejszenie wymiarowości o ok. 90%. Wykres 8.5 przedstawia liczbę głównych składowych, jaki procent wariancji opisują poszczególne składowe oraz kumulatywną sumę.

Analiza trzech pierwszych głównych składowych uzyskanych z FQD wykazała, że składają się one przede wszystkim z właściwości fragmentów o następujących długościach

1. pierwsza składowa - fragmenty czteroatomowe,
2. druga składowa - fragmenty składające się z siedmiu i ośmiu atomów,

3. trzecia składowa - fragmenty dwuatomowe.

Natomiast wśród pierwszych siedmiu najważniejszych składowych najwięcej udziałów przypadło energii wewnętrznej, entalpii i energii swobodnej w temperaturze 298,15 K, polaryzowalności, energii wibracyjnej punktu zerowego, pojemności cieplnej, energii HOMO, różnicy w energiach HOMO i LUMO oraz momentowi dipolowemu.



RYSUNEK 8.5: Redukcja wymiarowości deskryptorów kwantowych.

Jednak spadek zdolności predykcyjnych projektowanych modeli po zastosowaniu PCA wskazuje na fakt, że ewentualne grupowanie pochodzące z głównych składowych nie jest tożsame z klasami biokumulacji. Aby uzyskać większy wgląd w rzeczywiste współzależności pomiędzy wielkością objaśnianą, a zmiennymi objaśniającymi, dokonano oceny informacji wzajemnej. Uzyskane wyniki wskazują, że najwięcej informacji o klasie biokumulacji pochodzi ze zmiennych odpowiadających fragmentom siedmio, dziewięcio i ośmioatomowych. Natomiast w kwestii właściwości kwantowej są to: moment dipolowy, energia wewnętrzna w temperaturze 0 K, entalpia i energia wewnętrzna w temperaturze 298,15 K, różnica pomiędzy energiami LUMO i HOMO, energia wibronowa punktu zerowego, pojemność cieplna i polaryzowalność.

Wyznaczono również informację wzajemną pomiędzy klasą biokumulacji, a otrzymanymi głównymi składowymi. Zaskakująco największą współzależnością ze zmienną objaśnianą cechowała się piąta główna składowa (PC5), która odpowiada za 5% wariancji w zbiorze danych. Wartość informacji wzajemniej dla PC1 była dwukrotnie mniejsza.

Następnym krokiem było wykonanie kalibracji parametrów wybranych algorytmów uczenia maszynowego. Zakres testowanych parametrów w zależności od zastosowanych deskryptorów przedstawiono w tabeli B.1 w dodatku B. Ponadto znajdują się tam również wybrane parametry zastosowane w klasyfikacji zbioru testowego.

Klasyfikacja zbioru testowego

Utworzony na początku badania zbiór testowy składający się z 330 związków posłużył w końcowym etapie badań do ostatecznego sprawdzenia zbudowanego modelu. Wyniki przedstawiono w tabelach 8.1 i 8.2 odpowiednio dla przypadku, gdy wstępne przetwarzanie danych obejmowało PCA oraz bez PCA. Wszystkie deskryptory pozwoliły na uzyskanie wyników lepszych niż klasyfikator losowy. Jednak deskryptory kwantowe osiągnęły wyniki gorsze niż tradycyjne deskryptory z biblioteki RDKit. Przy zastosowaniu kombinacji obu rodzajów deskryptorów również osiągnęto niższe wyniki wartości oceniających. Podobnie jak w czasie oceny wstępnych modeli, dodanie PCA jako jednego z kroków przetwarzania danych spowodowało spadek skuteczności klasyfikacji.

TABELA 8.1: Wyniki klasyfikacji zbioru testowego po zastosowaniu PCA.

Metryka	Numer testu	Klasyfikator losowy	Deskryptory		
			FQD	RDKit	Kombinacja
F miara	I	0.444	0.594	0.735	0.682
	II	0.444	0.586	0.711	0.672
	III	0.444	0.594	0.725	0.652
	IV	0.444	0.588	0.720	0.657
	V	0.444	0.603	0.715	0.652
Dokładność zbalansowana	I	0.458	0.751	0.834	0.789
	II	0.458	0.747	0.829	0.788
	III	0.458	0.751	0.838	0.777
	IV	0.458	0.746	0.836	0.779
	V	0.458	0.759	0.842	0.777

Pod względem F miary oraz dokładności zbalansowanej najlepszym modelem okazał się ten zbudowany na podstawie tradycyjnych deskryptorów molekularnych. Natomiast pod względem precyzji i specyficzności wyznaczonych z wyników klasyfikacji testowego zbioru danych lepszym modelem okazał się być model stosujący kombinację FQD oraz tradycyjnych deskryptorów molekularnych.

Wyniki precyzji, specyficzności, czułości oraz dokładności klasyfikacji zamieszczono w dodatku B.

TABELA 8.2: Wyniki klasyfikacji zbioru testowego bez zastosowania PCA.

Metryka	Numer testu	Klasyfikator losowy	Deskryptory		
			FQD	RDKit	Kombinacja
F miara	I	0.444	0.620	0.759	0.729
	II	0.444	0.620	0.759	0.729
	III	0.444	0.620	0.759	0.729
	IV	0.444	0.620	0.759	0.729
	V	0.444	0.620	0.759	0.729
Dokładność zbalansowana	I	0.458	0.759	0.875	0.829
	II	0.458	0.759	0.875	0.829
	III	0.458	0.759	0.875	0.829
	IV	0.458	0.759	0.875	0.829
	V	0.458	0.759	0.875	0.829

8.3 Podsumowanie

Pomimo uzyskania wyników lepszych niż wyniki losowe, fragmentaryczne deskryptory kwantowe nie wywołały polepszenia skuteczności klasyfikacji ze względu na zdolność do biokumulacji, w stosunku do znanych już deskryptorów molekularnych. Poszukiwanie fragmentów związków chemicznych wśród ponad 150 tysięcy możliwych krótkich związków jest procesem czasochłonnym. Obliczanie fragmentarycznych deskryptorów kwantowych dla wielu związków eksperymentalnych znacząco wydłuża cały proces.

Metodyka badań nie obejmowała ograniczenia bazy danych eksperymentalnych ze względu na liczbę atomów w związku. Fakt, że w bazie danych eksperymentalnych znajdowały się związki krótsze niż najdłuższe poszukiwane fragmenty mógł mieć wpływ na otrzymywane wyniki.

Nie do końca rozwiązanym problemem jest również zjawisko otrzymywania brakujących wartości deskryptorów. Należy tak zdefiniować same deskryptory, aby nie otrzymywać brakujących wartości lub uzupełnić bazę danych fragmentów.

Kwantowo informowane pary atomów

Zidentyfikowawszy problemy występowania brakujących danych i stosunkowo długiego czasu poszukiwania podstruktur oraz mając na uwadze, że FQD nie pozwalają na osiągnięcie leszych zdolności predykcyjnych niż tradycyjne deskryptory molekularne, zdecydowano o daleko idącej zmianie definicji proponowanych deskryptorów kwantowych. Opisane w niniejszym rozdziale kwantowo informowane pary atomów są podejściem, którego celem nie jest zdefiniowanie nowych deskryptorów kwantowych od podstaw, lecz „ukwantowienie“ metody już istniejącej.

9.1 Metodyka

Kwantowo informowane pary atomów

Celem wzbogacenia tradycyjnego opisu cząsteczki w postaci par atomów, wygenerowano zbiór par atomów występujących w cząsteczkach z bazy danych QM9-extended-plus, którą opisano w rozdziale 7. Wybrano pary atomów odległych od siebie o nie więcej niż 4 wiązania. Proces ten wykazał, że wśród 157488 związków chemicznych zbudowanych z maksymalnie 9 atomów 7 różnych pierwiastków można wyróżnić 1089 różnych par atomów. Każdej z tak wyznaczonych par atomów przyporządkowano wartości będące pochodną właściwości kwantowych cząsteczek, w których występują dane pary atomów. Wartości te zdefiniowano na 2 sposoby.

Pierwsza definicja polegała na przypisaniu każdej parze atomów średniej arytmetycznej właściwości kwantowych cząsteczek, w której ta para atomów występuje. W przypadku gdy w cząsteczce dana para atomów występowała więcej niż raz,

przeliczano wartość właściwości kwantowej na jedno wystąpienie. Takie zdefiniowanie skutkuje przyporządkowaniem każdej parze atomów 11 właściwości kwantowych.

Drugą definicją było wyznaczenie dla każdej pary atomów szeregów rozdzielczych (histogramów) o dziesięciu przedziałach dla każdej właściwości kwantowej. Otrzymano w ten sposób zbiór, w którym każdej parze atomów przyporządkowano 110 zmiennych opisujących, po 10 na każdą z 11 właściwości kwantowych.

Powyższe definicje posłużyły do zaprojektowania trzech rodzajów deskryptorów opartych o wartości przyporządkowane parom atomów wykrytych w cząsteczkach eksperymentalnych obecnych w analizowanych bazach danych. Strukturę cząsteczek kodowano jako:

- 11 zmiennych będących sumą wartości z pierwszej definicji (sumQAP),
- 110 zmiennych będących sumą wartości z drugiej definicji (hQAP - histogram QAP),
- 11979 zmiennych będących listą wszystkich 11 wartości pochodzących z pierwszej definicji wszystkich par atomów wyróżnionych z bazy właściwości kwantowych (1089). W przypadku, gdy dana para atomów była nieobecna w cząsteczce, jej 11 wartości wynosiło 0. (spQAP - sparse QAP).

Należy mieć na uwadze, że wygenerowanie reprezentacji spQAP powoduje powstanie macierzy rzadkiej, która nastręcza problemów z przechowywaniem w pamięci komputera podczas wykonywania operacji. Problem ten wystąpił również w omiawianym badaniu - w przypadku najliczniejszej bazy danych eksperymentalnych, gdzie 16GB pamięci operacyjnej komputera okazało się niewystarczające, co powodowało przerwanie modelowania. Problem ten rozwiązano poprzez zwiększenie partycji wymiany (swap) systemu operacyjnego.

Analogicznie do deskryptorów kwantowych przedstawionych w rozdziałach 6 i 8, kwantowo informowane pary atomów charakteryzują się ograniczoną domeną aplikowalności. W badaniu wszystkie bazy danych eksperymentalnych ograniczano do związków, których zbiór par atomów o maksymalnej odległości 4 wiązań był podzbiorem 1089 par atomów wyznaczonych z bazy właściwości kwantowych.

Oceny zdolności do opisywania struktury związków chemicznych przez QAP dokonywano metodą porównania z wybranymi, ugruntowanymi w chemoinformatyce metodami kodowania struktury chemicznej. Poniżej przedstawiono listę wybranych metod:

- odciski palców Morgana,
- pary atomów,
- odciski palców RDKit,
- odciski palców torsji topologicznej,
- deskryptory molekularne z biblioteki RDKit.

Ponadto oceniano wyniki w stosunku do linii bazowej, którą w przypadku regresji przyjęto za wartość średnią z wartości eksperymentalnych, a w przypadku klasyfikacji, losowe przypisanie do klasy z uwzględnieniem liczebności każdej z klas.

Celem redukcji wymiarowości oraz odrzucenia szumów mogących negatywnie wpływać na zdolności predykcyjne projektowanego modelu, zdecydowano się na wybranie tylko tych zmiennych objaśniających, których odchylenie standardowe było wyższe niż 0,05.

Modelowanie wykonywano metodą lasu losowego, o liczbie estymatorów równej 500 oraz uwzględniającego po 30% zmiennych objaśniających na każde drzewo decyzyjne. Celem możliwie największej generalizacji otrzymanych wyników stosowano losową, dziesięciokrotnie powtórzoną dziesięciokrotną walidację krzyżową korzystając z klasy RepeatedKFold modułu Scikit-learn Pythona. Jakość otrzymywanych predykcji oceniano, w przypadku regresji na podstawie kwadratu współczynnika Pearsona, R^2 oraz pierwiastka kwadratowego z sumy błędu kwadratowego (RMSE), a w przypadku klasyfikacji na podstawie dokładności zbalansowanej oraz pola pod krzywą ROC.

Bazy danych eksperymentalnych

Oceny zdolności QAP do kodowania struktury związków chemicznych dokonano na 11 dostępnych bazach danych - 7 baz pozwalających za zbudowanie modelu regresji oraz 4 odpowiednich do klasyfikacji. Tabela 9.1 stanowi opis charakterystyki baz danych zastosowanych w badaniach. Najmniej liczna baza danych

zawierała 285 związków chemicznych będących w domenie aplikowalności, a najbardziej liczna zawierała ich 7316. Bazy danych do modeli klasyfikacyjnych zawierały po 2 klasy. Dwie z tych baz danych charakteryzowały się silnym brakiem zbalansowania klas. Większość zbiorów danych służących do budowy modeli regresji charakteryzowały się odchyleniami standardowymi zmiennej objaśnianej od 1,04 do 2,18. Zmienna objaśniana w bazie danych temperatury topnienia posiadała największe odchylenie standardowe, które wynosiło 94,09°C.

TABELA 9.1: Bazy danych zastosowane w badaniu. Górna tabela przedstawia bazy danych do budowy modeli regresji, a dolna do budowy modeli klasyfikacji. Liczba związków w bazach danych (n) przedstawia liczbę związków będących w zastosowanej domienie aplikowalności.

Baza danych	n	Min.	Maks.	Źródło
logP	5574	-4.64	8.27	[98]
BACE-1	285	2.699	10.523	[77, 99]
Rozpuszczalność	803	-11.6	1.58	[77]
Lipofilowość	2237	-1.50	4.48	[77]
Energia jonizacji	1575	1.04	13.94	[100]
Temperatura topnienia[1]	7316	-196.0	492.5	[101, 102]

[1]Baza danych temperatury topnienia jest kombinacją dwóch baz danych.

Baza danych	n	Stosunek klas	Źródło
BBBP	1089	161/928	[103]
Ames mutagenność	3801	1779/2022	[104]
hERG	2452	1068/1385	[105]
ClinTox	707	56/651	[77]

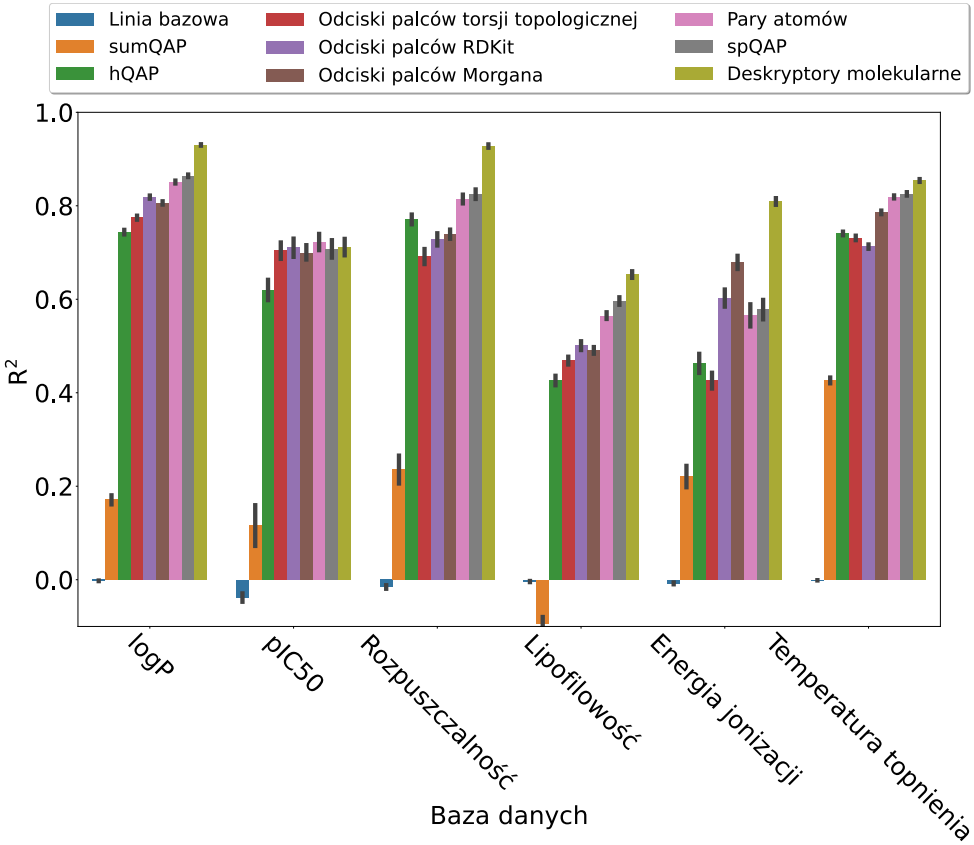
9.2 Wyniki i dyskusja

Regresja

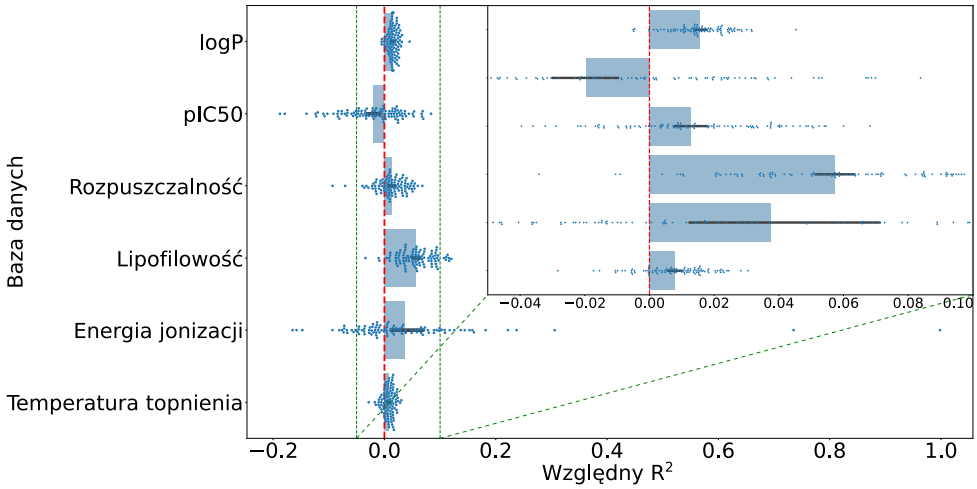
W większości przypadków, z wyjątkiem regresji lipofilowości, wszystkie zastosowane deskryptory i odciski palców osiągnęły wyniki lepsze niż linia bazowa. Gorsze wyniki od linii bazowej w predykcji lipofilowości osiągnęły deskryptory sumQAP. W szerszej perspektywie deskryptory te wykazały najniższą skuteczność w budowaniu modeli regresji - współczynnik R^2 w każdej testowanej bazie danych był niższy niż 0,5 (rys 9.1). Można również powiedzieć, że w przypadku większości baz danych, najlepszym dopasowaniem do danych cechują się predykcje, gdzie zmiennymi objaśniającymi były deskryptory molekularne, następnie spQAP oraz

pary atomów. Wskazuje to na fakt, że wzbogacenie par atomów o informacje pochodzące z chemii kwantowej spowodowały polepszenie zdolności predykcyjnych.

Rysunek 9.2 przedstawia pomniejszony o 1 stosunek wartości metryk otrzymywanych dzięki zastosowaniu spQAP do metryk otrzymywanych na podstawie par atomów w poszczególnych podziałach walidacji krzyżowej. Daje się zauważyć wzrost R^2 średnio o 0,5 do 6%. Tylko w przypadku bazy danych zawierającej pIC50 daje się zauważyć spadek R^2 o 2%.

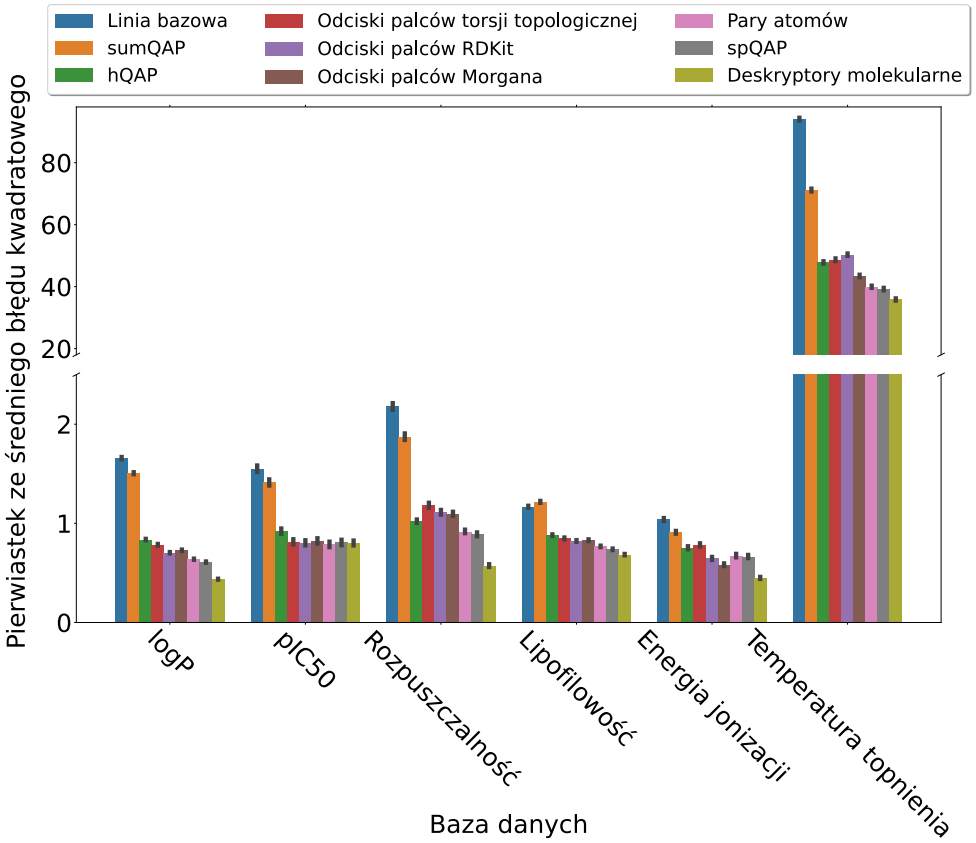


RYSUNEK 9.1: Wyniki kwadratu współczynnika Pearsona R^2 w walidacji krzyżowej. Słupki przedstawiają wartość średnią, a czarne linie 95% przedział ufności.

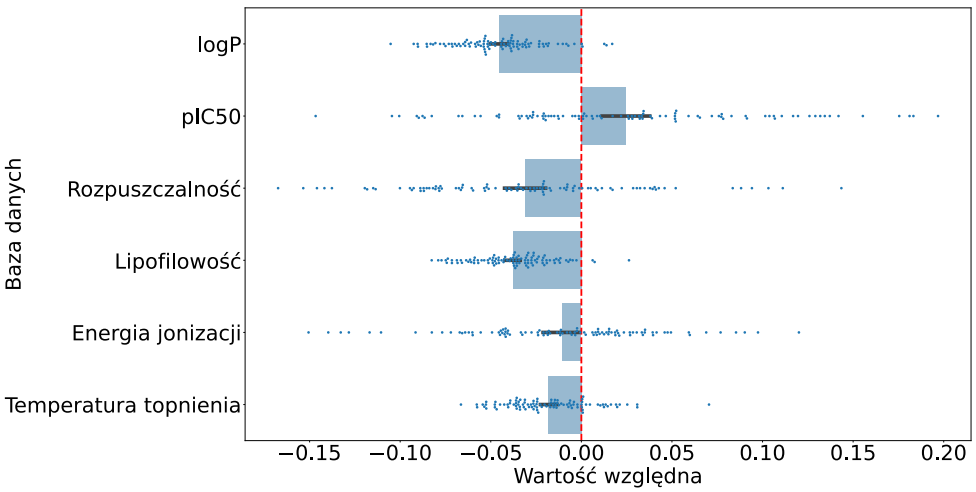


RYSUNEK 9.2: Względny współczynnik R^2 otrzymany w poszczególnych podziałach walidacji krzyżowej. Punkty przedstawiają wartości otrzymane na podstawie spQAP względem par atomów. Wartość względna pomniejszona została o 1. Słupki pokazują wartość średnią, a czarne linie 95% przedział ufności.

Wartości pierwiastka ze średniego błędu kwadratowego przedstawiono na rys. 9.3. Ogólny obraz jest podobny jak w przypadku współczynnika R^2 - średnio najmniejszym błędem charakteryzują się predykcje dokonane na podstawie deskryptorów molekularnych. Jedynym wyjątkiem są modele uczone i walidowane na bazie danych pIC50 - w tym wypadku najlepiej dopasowane okazały się być predykcje dokonane na podstawie par atomów. Zasadniczo, po deskryptorach molekularnych najlepszymi predykcjami cechują się modele wytrenowane na podstawie spQAP, z wyjątkiem opisanej wyżej wartości pIC50 oraz energii jonizacji, w przypadku której lepsze dopasowanie osiąga się za pomocą odcisków palców Morgana. Porównując ze sobą wartości metryk w poszczególnych podziałach walidacji krzyżowej, można zauważyć, że zastosowanie spQAP pozwala na osiągnięcie predykcji średnio o od ok. 0,5% do 5% lepszych niż w przypadku zwykłych par atomów. W przypadku regresji logP, lipofilowości i temperatury topnienia zdecydowana większość podziałów w walidacji krzyżowej, dawała rezultat w którym średni błąd kwadratowy dla spQAP był niższy niż przy użyciu standardowych par atomów (rysunek 9.4.



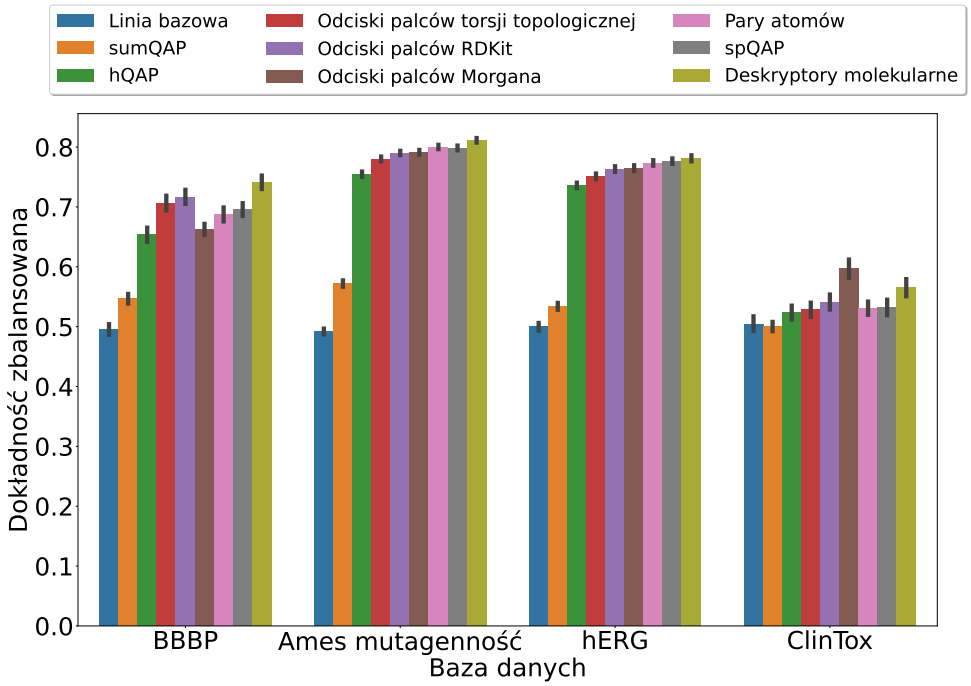
RYSUNEK 9.3: Wykres pierwiastka ze średniego błędu kwadratowego. Słupki przedstawiają wartość średnią z wszystkich podziałów walidacji krzyżowej, a czarne linie 95% przedział ufności.



RYSUNEK 9.4: Porównanie wartości pierwiastka ze średniego błędu kwadratowego pomiędzy deskryptorami spQAP i tradycyjnymi parami atomów. Poszczególne punkty ukazują pomniejszony o 1 stosunek metryka dla spQAP/metryka dla par atomów. Słupki pokazują wartość średnią, a czarne linie 95% przedział ufności.

Klasyfikacja

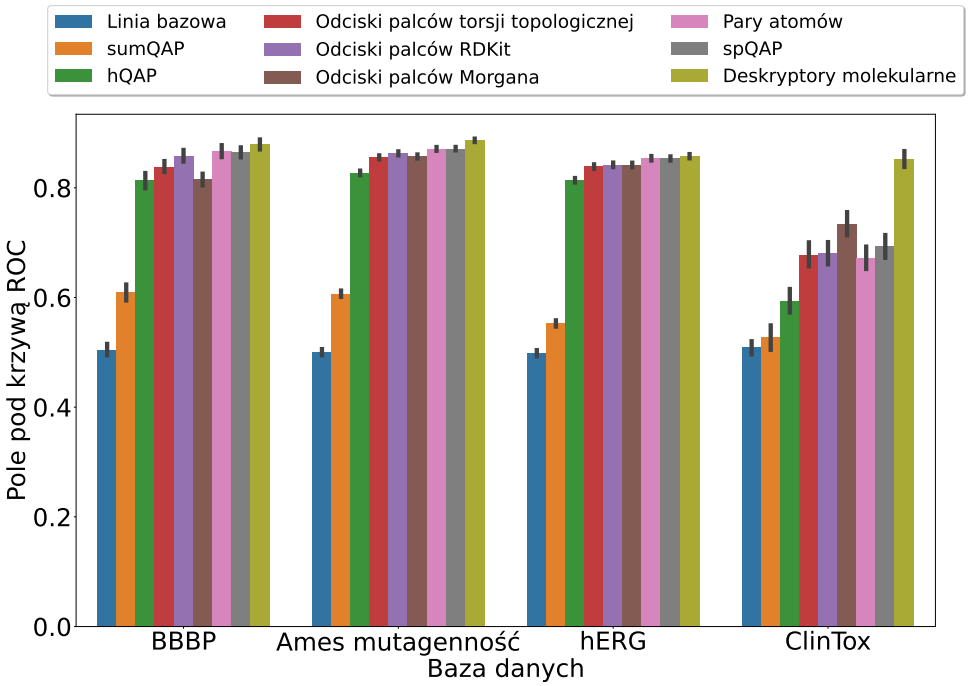
Zadanie klasyfikacji realizowane było na podstawie baz danych, w których występowały dwie klasy. Dwie spośród czterech wybranych baz danych charakteryzowała nadreprezentacja jednej z klas. Zauważalnie, w większości baz danych wyniki przedstawiają się w sposób podobny do modelowania regresji - najlepszymi zmiennymi objaśniającymi są deskryptory molekularne, następnie spQAP i pary atomów oraz inne reprezentacje cząsteczek. Należy nadmienić, że testowane modele posiadają słabe zdolności do rozdziału klas w przypadku bazy danych ClinTox, ponieważ osiągnięta dokładność zbalansowana (rysunek 9.5) nieznacznie przekracza wartość dla linii bazowej - 0,6 wobec 0,5. Takie zjawisko spowodowane jest silną nadreprezentacją jednej z klas w bazie danych, ponieważ stosunek ich liczebności wynosi prawie 1:12. Drugą bazą danych o wysokiej nadreprezentacji jednej z klas jest baza danych BBBP (stosunek niecałe 1:6). Otrzymane wyniki dokładności zbalansowanej w przypadku tej bazy danych wskazują, że do opisu struktury związków lepsze są odciski palców torsji topologicznej, odciski palców RDKit i deskryptory molekularne niż spQAP oraz pary atomów.



RYSUNEK 9.5: Dokładności zbalansowane klasyfikacji na podstawie różnych baz danych.

Jednak wyniki dokładności zbalansowanej nie w pełni pokrywają się z wynikami pola pod krzywą ROC przedstawionymi na rysunku 9.6. Wartości tego parametru

wskazują, że deskryptory molekularne pozwalają zbudować najbardziej dokładne modele klasyfikujące, jednak kwantowe pary atomów i ich pierwowzór nie odbiegają od nich w znaczący sposób. Wyjątkiem jest baza danych ClinTox, gdzie deskryptory molekularne osiągnęły wyraźnie większą wartość parametru ROC AUC. Ponadto pod względem tej metryki większość pozostałych reprezentacji pozwoliła osiągnąć zdecydowanie lepsze wyniki niż linia bazowa. Różnice te wynikają prawdopodobnie ze względu na niezbalansowanie danych. Istnieją doniesienia, że ROC AUC może być mało dokładne przy klasyfikacji takich danych [106].



RYSUNEK 9.6: Pole pod krzywą ROC modeli klasyfikujących.

Ze względu na podobne wyniki metryk pomiędzy spQAP oraz standardowymi parami atomów, przeprowadzono testy statystyczne celem sprawdzenia, czy różnice wyników powtarzanej walidacji krzyżowej są statystycznie istotne. Początkowo sprawdzono warunek normalności rozkładów wyników przy pomocy testu Shapiro-Wilka oraz równości wariancji przy pomocy testu Levene'a. Testy te pozwoliły na podjęcie decyzji który test statystyczny, t-Studenta, czy U Manna-Whitneya należy zastosować do oceny wyników powtarzanych walidacji krzyżowych.

Wyniki pokazały, że dla danych obecnych w rozpatrywanych bazach danych wariancje metryk otrzymanych w walidacjach krzyżowych modelując właściwości oraz aktywność są podobne. Jednak w przypadku połowy baz danych eksperymentalnych wyniki te nie spełniają warunku normalności. Zdecydowano zatem na

przeprowadzenie testu U Manna-Whitneya.

Analiza wyników testu wykazała, że przy założeniu poziomu istotności statystycznej $\alpha = 0,05$ istnieją statystycznie istotne różnice w wynikach walidacji krzyżowych pomiędzy wynikami otrzymanymi modelując właściwości na podstawie spQAP, a wynikami otrzymanymi przy zastosowaniu standardowych par atomów dla trzech baz danych eksperymentalnych. Wartość $p < \alpha$ otrzymano dla bazy danych logP, lipofilowości oraz temperatury topnienia. Są to bazy danych, które posłużyły do sformułowania zadania regresji właściwości. Jednocześnie zauważyć można, że są to również trzy najbardziej liczne bazy danych wśród sześciu zastosowanych do modelowania regresji.

9.3 Podsumowanie

Na podstawie powyższych wyników, przedstawiona w rozdziale metoda opisu struktury związków chemicznych poprzez właściwości kwantowe par atomów pozwala na zwiększenie skuteczności modelowania w stosunku do opisu przy pomocy standardowych par atomów. Obserwacja ta dotyczy jednak przede wszystkim modeli regresji. Największy średni spadek pierwiastka ze średniego kwadratu błędu wyniósł ok. 6%. Zastosowane testy statystyczne dodatkowo potwierdzają, że w połowie przeanalizowanych przypadków różnice w predykcjach pomiędzy predykcjami otrzymanymi na podstawie zmodyfikowanych par atomów, a zwykłymi parami atomów są statystycznie istotne.

W przypadku modeli klasyfikujących, wyniki prezentują się podobnie lub nieznacznie gorzej pomiędzy klasyfikacjami dokonanymi na podstawie różnych sposobów reprezentacji cząsteczek chemicznych. Można zatem wyciągnąć wniosek, że nadanie parom atomów wartości liczbowej innej niż zwykła krotność wystąpienia pozwala na polepszenie zdolności predykcyjnych, gdy modelowana zmienna ma charakter ilościowy.

Jednak w przedstawionych zastosowaniach to deskryptory molekularne pozwalają zbudować najlepsze modele uczenia maszynowego. Przyczyną takiego stanu rzeczy jest najprawdopodobniej fakt, że deskryptory molekularne w pewnym stopniu opisują właściwości związków chemicznych.

Podsumowanie i wnioski z pracy doktorskiej

Praca doktorska miała na celu zbadanie możliwości budowania prostych i dokładnych modeli predykcyjnych lub klasyfikacyjnych w oparciu o właściwości kwantowe fragmentów cząsteczek chemicznych. Wychodząc od bazy danych właściwości kwantowych opracowano deskryptory pochodzące od fragmentów cząsteczek, które posłużyły próbom modelowania właściwości optycznych. Na podstawie wniosków z tego modelowania oraz zidentyfikowanych ograniczeń dokonano redefinicji proponowanych deskryptorów kwantowych, a także takiego rozszerzenia bazy danych właściwości kwantowych, aby powiększyć zbiór fragmentów o związki zawierające inne heteroatomy.

W większości testów, niezależnie od sposobu definicji fragmentarycznych deskryptorów kwantowych otrzymywano lepsze wyniki niż ustalenie predykcji jako wartości średniej w przypadku zadania regresji lub przypisanie losowe w przypadku klasyfikacji. Nawet jeśli zaproponowane deskryptory niosą informację, to wyzwaniem jest zbudowanie lepszej relacji struktura - właściwość w stosunku do tradycyjnie stosowanych w chemoinformatyce deskryptorów molekularnych. Jednak podjęto się zbudowania takich relacji na podstawie ograniczonej liczby baz danych i jest szansa, że istnieje taka właściwość związków chemicznych, którą można by w przedstawiony sposób modelować z wysoką dokładnością.

Wyznaczenie deskryptorów kwantowych przedstawionych w rozdziałach 6 i 8 jest procesem długotrwałym, co może wydłużać czas prototypowania i budowy modeli predykcyjnych. Ponadto, jak wspominano w rozdziale 5, baza danych stosowana jako baza fragmentów nie zawiera wszystkich możliwych kombinacji atomów, co może skutkować brakiem dopasowania jakiegokolwiek fragmentu dla danej cząsteczki eksperymentalnej.

Wobec zidentyfikowanych problemów, a w szczególności definicyjnej możliwości występowania brakujących wartości fragmentarycznych deskryptorów kwantowych podjęto się zdecydowanej zmiany poszukiwanych fragmentów cząsteczek. Jako punkt wyjścia zastosowano znane w chemoinformatyce pary atomów. Na trzy różne sposoby parom tym nadano wielkości kwantowe na podstawie występowania w bazie danych właściwości kwantowych.

Na podstawie dużej liczby prób w walidacjach krzyżowych udowodniono, że utworzone w ten sposób kwantowe pary atomów w części modelowanych właściwości poprawiły zdolności predykcyjne w stosunku do swoich pierwowzorów.

Domena aplikowalności fragmentarycznych deskryptorów kwantowych jest również ograniczona przez wzgląd na bazę danych fragmentów. Początkowo obejmowała ona tylko związki zbudowane z atomów czterech pierwiastków, w wyniku pracy własnej oraz innych zespołów badawczych poszerzyła się do siedmiu pierwiastków - C, O, N, F, S, Cl, Br. W badaniach kwantowych par atomów zdecydowano się na dodatkowe ograniczenie związków do tych, w których można wyróżnić tylko pary atomów wygenerowanie z bazy danych QM9-extended-plus. Ponadto kwantowe pary atomów, w swej niosącej najwięcej informacji definicji, tworzą macierz rzadką, która wymaga dużej ilości pamięci komputera celem jej przetworzenia.

Praca doktorska zaowocowała również napisaniem i udostępnieniem programu umożliwiającego wyznaczenie wartości deskryptorów kwantowych przedstawionych w rozdziałach 5 oraz 9. Dzięki temu inni zainteresowani mogą z łatwością podjąć próbę budowania własnych modeli predykcyjnych opartych o przedstawione w pracy nowe deskryptory kwantowe.

Możliwe kierunki rozwoju badań

1. Stworzenie lub uzupełnienie bazy danych właściwości kwantowych aby zawierała fragmenty o jak największej liczbie kombinacji atomów.

2. Zastosowanie metody z fragmentami wygenerowanymi z cząsteczek eksperymentalnych np. metodą BRICS, RECAP lub metodami stosowanymi w metodach chemii kwantowej dużych układów, np. molecules-in-molecules [107], a nie ze z góry określonymi fragmentami.

3. Poprawa metodyki wykrywania fagmentów pod kątem uwzględniania tylko fragmentów, które się na siebie nie nakładają.

4. Uwzględnienie nie tylko istnienia danych fragmentów lecz również ich ewentualnego oddziaływania ze sobą.

Udostępnianie danych badawczych

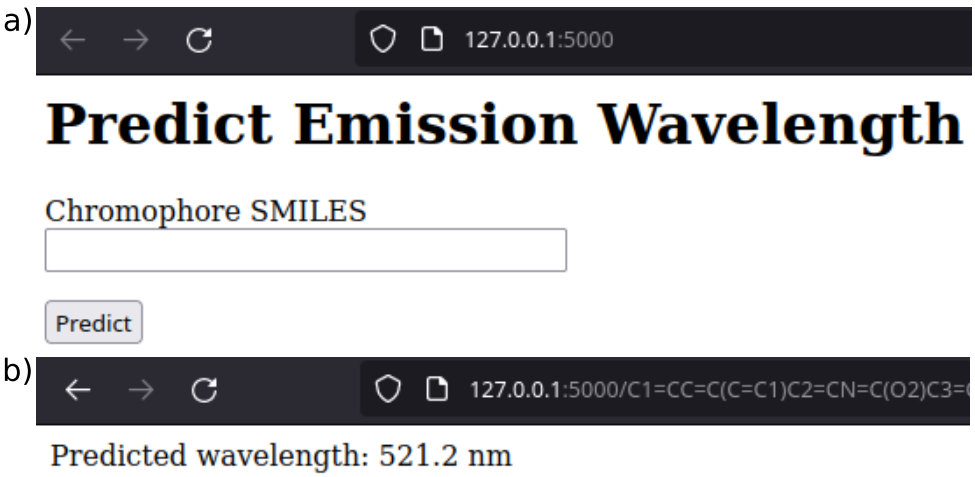
Niniejsza praca doktorska została zrealizowana dzięki udostępnionym danym badawczym. Pokazuje to jak ważne jest udostępnianie danych wygenerowanych w czasie badań. Takie postępowanie powinno być, jeśli nie wymagane, to promowane przez jednostki prowadzące działalność badawczą. Utworzone w związku z niniejszą pracą dane w postaci kodu oraz wyników obliczeń również zostały udostępnione. Ponadto publikacje naukowe powstałe w toku realizacji pracy były publikowane w formie preprintów, a następnie w formie otwartego dostępu.

Część aplikacyjna

Stworzone aplikacje udostępione są na platformie GitHub.

11.1 Aplikacja webowa do przewidywania długości fali maksimum emisji

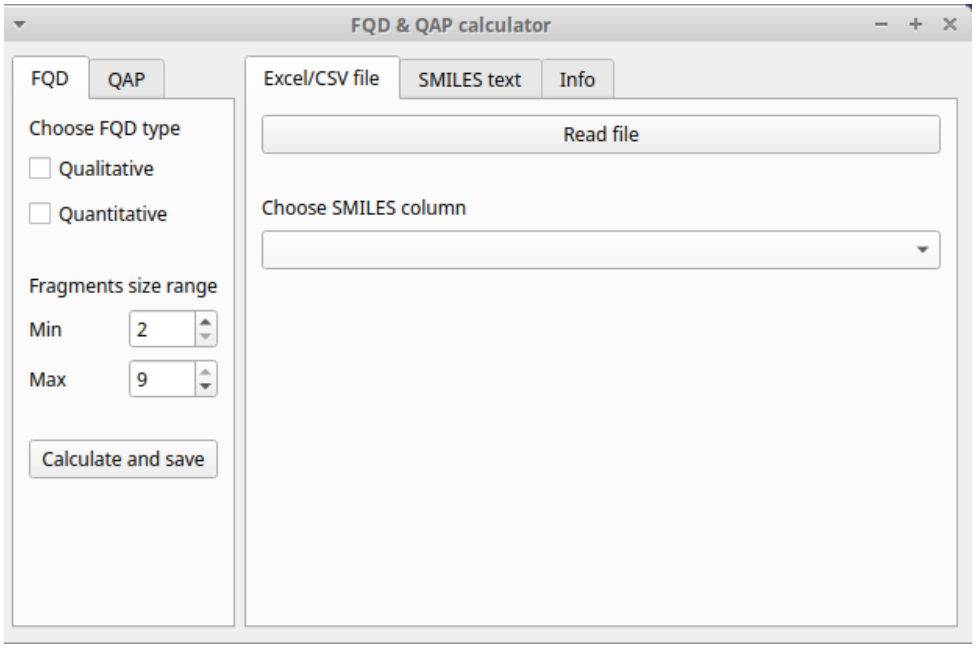
Z myślą o ewentualnym zastosowaniu metody opisanej w rozdziale 6 do udostępnienia dla każdego pracownika laboratorium, stworzona została aplikacja webowa. Dostęp do predykcji można osiągnąć na dwa sposoby - poprzez formularz w przeglądarce, gdzie wpisuje się strukturę związku chemicznego w notacji SMILES i zatwierdza przyciskiem lub bezpośrednio poprzez umieszczenie SMILES w pasku adresowym przeglądarki, z innej aplikacji poprzez ten sam adres lub z wiersza poleceń (np. komendą curl). Rysunek 11.1 przedstawia zrzuty ekranów aplikacji. Sama aplikacja jest na tyle mało skomplikowana, że można ją uruchomić np. na mikrokomputerze Raspberry Pi.



RYSUNEK 11.1: Interfejs webowy aplikacji do przewidywania długości fali maksimum emisji. a) formularz do wprowadzenia struktury; b) bezpośrednie przejście do predykcji po wpisaniu SMILES w pasku adresowym przeglądarki.

11.2 Aplikacja z interfejsem graficznym do generowania fragmentarycznych deskryptorów kwantowych i kwantowych par atomów

Na podstawie badań przedstawionych w ninejszej pracy doktorskiej utworzone zostały moduły Pythona służące do obliczania fragmentarycznych deskryptorów kwantowych oraz kwantowych par atomów. W oparciu o te moduły powstała aplikacja z interfejsem graficznym do obliczania FQD zdefiniowanych w rozdziale 8 oraz QAP zdefiniowanych w rozdziale 9. Zrzut okna programu ukazany jest na rysunku 11.2.



RYSUNEK 11.2: Okno aplikacji.

W przypadku FQD aplikacja pozwala na wybranie jaki typ deskryptorów ma zostać wygenerowany oraz jaki ma być zakres długości poszukiwanych fragmentów. Po wybraniu zakładki QAP możliwe jest wybranie rodzaju QAP. Deklarację związków dla których obliczone mają być deskryptory kwantowe można zrealizować na dwa sposoby: poprzez wczytanie pliku Excel lub CSV i deklarację w której kolumnie znajduje się zakodowana struktura związków w postaci notacji SMILES. Drugim sposobem jest wpisanie w polu tekstowym ciągów SMILES oddzielonych znakami nowej linii. Po wciśnięciu przycisku „Calculate and save" liczone są deskryptory, co jest dość czasochłonnym procesem, gdy program generuje FQD. Wyznaczanie QAP jest zdecydowanie szybsze. Następnie wygenerowane tablice danych zapisywane są

w postaci pliku CSV, w którym pierwszą kolumną jest SMILES związku, z którego struktury deskryptory zostały policzone.

Dodatek do rozdziału 6

Wzory zastosowanych deskryptorów

Zestaw 4

Od tego zestawu podjęto próbę uwzględnienia faktu, że fale elektromagnetyczne oddziałują z chmurą elektronową, zatem mnożono niektóre właściwości kwantowe przez moment dipolowy, polaryzowalność, ich sumę lub podniesione do kwadratu.

$$\begin{aligned}
&\sum_i^N n_i \epsilon_{HOMO} \mu && \sum_i^N n_i \epsilon_{HOMO} \alpha \\
&\sum_i^N n_i \epsilon_{LUMO} \mu && \sum_i^N n_i \epsilon_{LUMO} \alpha \\
&\sum_i^N n_i \epsilon_{gap} \mu && \sum_i^N n_i \epsilon_{gap} \alpha \\
&\sum_i^N n_i \langle r^2 \rangle \mu && \sum_i^N n_i \langle r^2 \rangle \alpha \\
&\sum_i^N n_i zpve \mu && \sum_i^N n_i zpve \alpha
\end{aligned} \tag{A.1}$$

gdzie μ - moment dipolowy, α - polaryzowalność, ϵ_{gap} - różnica energii między LUMO a HOMO, $\langle r^2 \rangle$ - rozmiar chmury elektronowej.

Zestaw 5

$$\begin{aligned}
&\sum_i^N n_i \epsilon_{HOMO}(\mu+\alpha)\\
&\sum_i^N n_i \epsilon_{LUMO}(\mu+\alpha)\\
&\sum_i^N n_i \epsilon_{gap}(\mu+\alpha)\\
&\sum_i^N n_i zpve(\mu+\alpha)\\
&\sum_i^N n_i(\mu+\alpha)\\
&\sum_i^N n_i\langle r^2\rangle(\mu+\alpha)
\end{aligned} \tag{A.2}$$

gdzie $zpve$ - energia wibracyjna punktu zerowego.

Zestaw 6

$$\begin{aligned}
&\sum_i^N n_i \epsilon_{HOMO}\mu^2 && \sum_i^N n_i \epsilon_{HOMO}\alpha^2\\
&\sum_i^N n_i \epsilon_{LUMO}\mu^2 && \sum_i^N n_i \epsilon_{LUMO}\alpha^2\\
&\sum_i^N n_i \epsilon_{gap}\mu^2 && \sum_i^N n_i \epsilon_{gap}\alpha^2\\
&\sum_i^N n_i\langle r^2\rangle\mu^2 && \sum_i^N n_i\langle r^2\rangle\alpha^2\\
&\sum_i^N n_i zpve\mu^2 && \sum_i^N n_i zpve\alpha^2
\end{aligned} \tag{A.3}$$

Zestaw 7

Podzielono właściwości kwantowe przez sumaryczną liczbę wystąpień wykrytych podstruktur, n_i.

$$\begin{array}{ll} \frac{\sum_i^N n_i \epsilon_{HOMO} \mu^2}{\sum_i^N n_i} & \frac{\sum_i^N n_i \epsilon_{HOMO} \alpha^2}{\sum_i^N n_i} \\ \frac{\sum_i^N n_i \epsilon_{LUMO} \mu^2}{\sum_i^N n_i} & \frac{\sum_i^N n_i \epsilon_{LUMO} \alpha^2}{\sum_i^N n_i} \\ \frac{\sum_i^N n_i \epsilon_{gap} \mu^2}{\sum_i^N n_i} & \frac{\sum_i^N n_i \epsilon_{gap} \alpha^2}{\sum_i^N n_i} \\ \frac{\sum_i^N n_i zpve \mu^2}{\sum_i^N n_i} & \frac{\sum_i^N n_i zpve \alpha^2}{\sum_i^N n_i} \\ \frac{\sum_i^N n_i \langle r^2 \rangle \mu^2}{\sum_i^N n_i} & \frac{\sum_i^N n_i \langle r^2 \rangle \alpha^2}{\sum_i^N n_i} \end{array} \tag{A.4}$$

Zestaw 8

Podzielono właściwości kwantowe przez sumaryczną liczbę wykrytych podstruktur, N.

$$\begin{array}{ll} \frac{\sum_i^N n_i \epsilon_{HOMO} \mu^2}{N} & \frac{\sum_i^N n_i \epsilon_{HOMO} \alpha^2}{N} \\ \frac{\sum_i^N n_i \epsilon_{LUMO} \mu^2}{N} & \frac{\sum_i^N n_i \epsilon_{LUMO} \alpha^2}{N} \\ \frac{\sum_i^N n_i \epsilon_{gap} \mu^2}{N} & \frac{\sum_i^N n_i \epsilon_{gap} \alpha^2}{N} \\ \frac{\sum_i^N n_i zpve \mu^2}{N} & \frac{\sum_i^N n_i zpve \alpha^2}{N} \\ \frac{\sum_i^N n_i \langle r^2 \rangle \mu^2}{N} & \frac{\sum_i^N n_i \langle r^2 \rangle \alpha^2}{N} \end{array} \tag{A.5}$$

Zestaw 9

Właściwości kwantowe pomnożono przez liczbę atomów w podstrukturze i podzielono przez sumaryczną liczbę wykrytych podstruktur.

$$\begin{array}{ll} \frac{\sum_i^N \epsilon_{HOMO}\mu^2 c_{at}}{N} & \frac{\sum_i^N \epsilon_{HOMO}\alpha^2 c_{at}}{N} \\ \frac{\sum_i^N \epsilon_{LUMO}\mu^2 c_{at}}{N} & \frac{\sum_i^N \epsilon_{LUMO}\alpha^2 c_{at}}{N} \\ \frac{\sum_i^N \epsilon_{gap}\mu^2 c_{at}}{N} & \frac{\sum_i^N \epsilon_{gap}\alpha^2 c_{at}}{N} \\ \frac{\sum_i^N zpve\mu^2 c_{at}}{N} & \frac{\sum_i^N zpve\alpha^2 c_{at}}{N} \\ \frac{\sum_i^N \langle r^2\rangle\mu^2 c_{at}}{N} & \frac{\sum_i^N \langle r^2\rangle\alpha^2 c_{at}}{N} \end{array} \tag{A.6}$$

gdzie c_{at} - liczba atomów w podstrukturze, N - liczba wykrytych podstruktur.

Zestaw 10

$$\begin{array}{ll} \sum_i^N \epsilon_{HOMO}\mu^2 c_{at} & \sum_i^N \epsilon_{HOMO} c_{at}\alpha^2 c_{at} \\ \sum_i^N \epsilon_{LUMO}\mu^2 c_{at} & \sum_i^N \epsilon_{LUMO}\alpha^2 c_{at} \\ \sum_i^N \epsilon_{gap}\mu^2 c_{at} & \sum_i^N \epsilon_{gap}\alpha^2 c_{at} \\ \sum_i^N zpve\mu^2 c_{at} & \sum_i^N zpve\alpha^2 c_{at} \\ \sum_i^N \langle r^2\rangle\mu^2 c_{at} & \sum_i^N \langle r^2\rangle\alpha^2 c_{at} \end{array} \tag{A.7}$$

Zestaw 11

$$\begin{aligned}
&\sum_i^N \epsilon_{HOMO}\mu^2 c_{e_{val}} && \sum_i^N \epsilon_{HOMO} c_{e_{val}}\alpha^2 c_{e_{val}} \\
&\sum_i^N \epsilon_{LUMO}\mu^2 c_{e_{val}} && \sum_i^N \epsilon_{LUMO}\alpha^2 c_{e_{val}} \\
&\sum_i^N \epsilon_{gap}\mu^2 c_{e_{val}} && \sum_i^N \epsilon_{gap}\alpha^2 c_{e_{val}} \\
&\sum_i^N zpve\mu^2 c_{e_{val}} && \sum_i^N zpve\alpha^2 c_{e_{val}} \\
&\sum_i^N \langle r^2\rangle \mu^2 c_{e_{val}} && \sum_i^N \langle r^2\rangle \alpha^2 c_{e_{val}}
\end{aligned} \tag{A.8}$$

gdzie $c_{e_{val}}$ - liczba elektronów walencyjnych w podstrukturze

Zestaw 12

$$\begin{aligned}
&\sum_i^N \epsilon_{HOMO}\mu^2\langle r^2\rangle && \sum_i^N \epsilon_{HOMO}\alpha^2\langle r^2\rangle \\
&\sum_i^N \epsilon_{LUMO}\mu^2\langle r^2\rangle && \sum_i^N \epsilon_{LUMO}\alpha^2\langle r^2\rangle \\
&\sum_i^N \epsilon_{gap}\mu^2\langle r^2\rangle && \sum_i^N \epsilon_{gap}\alpha^2\langle r^2\rangle \\
&\sum_i^N zpve\mu^2\langle r^2\rangle && \sum_i^N zpve\alpha^2\langle r^2\rangle
\end{aligned} \tag{A.9}$$

Zestaw 13

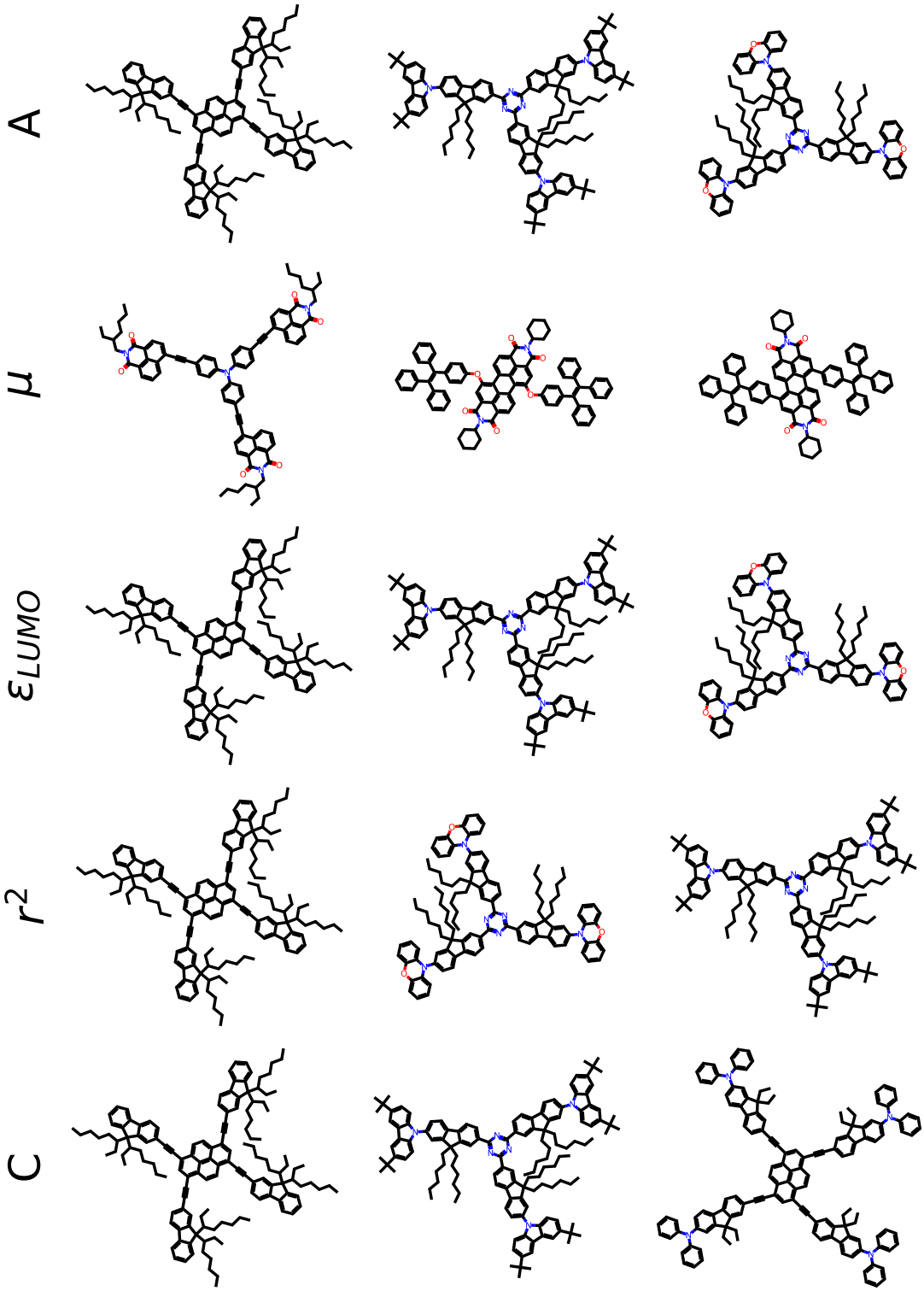
W zestawie 13 i 14 mnożono właściwości kwantowe przez liczbę elektronów walencyjnych w związku optycznie czynnym.

$$\begin{aligned}
&\sum_i^N n_i \epsilon_{HOMO} \mu^2 C_{e_{val}} && \sum_i^N n_i \epsilon_{HOMO} \alpha^2 C_{e_{val}} \\
&\sum_i^N n_i \epsilon_{LUMO} \mu^2 C_{e_{val}} && \sum_i^N n_i \epsilon_{LUMO} \alpha^2 C_{e_{val}} \\
&\sum_i^N n_i \mu^2 \langle r^2 \rangle C_{e_{val}} && \sum_i^N n_i \alpha^2 \langle r^2 \rangle C_{e_{val}} \\
&(\sum_i^N n_i \langle r^2 \rangle) C_{e_{val}}
\end{aligned} \tag{A.10}$$

gdzie $C_{e_{val}}$ - liczba elektronów walencyjnych w chromoforze.

Zestaw 14

$$\begin{aligned}
&\frac{\sum_i^N \epsilon_{HOMO} \mu^2 C_{e_{val}}}{N} && \frac{\sum_i^N \epsilon_{HOMO} \alpha^2 C_{e_{val}}}{N} \\
&\frac{\sum_i^N \epsilon_{LUMO} \mu^2 C_{e_{val}}}{N} && \frac{\sum_i^N \epsilon_{LUMO} \alpha^2 C_{e_{val}}}{N} \\
&\frac{\sum_i^N \mu^2 \langle r^2 \rangle C_{e_{val}}}{N} && \frac{\sum_i^N \alpha^2 \langle r^2 \rangle C_{e_{val}}}{N} \\
&\frac{\sum_i^N \langle r^2 \rangle) C_{e_{val}}}{N}
\end{aligned} \tag{A.11}$$



RYSUNEK A.1: Cząsteczki o najbardziej odstających wartościach deskryptorów kwantowych. Cząsteczki podzielone są według właściwości kwantowych - po 3 na daną właściwość. Warto zauważyć, że dwie struktury powtarzają się w czterech różnych właściwościach kwantowych.

Dodatek do rozdziału 8

TABELA B.1: Zakres testowanych parametrów.

	Zakres			
Parametr	Las losowy	Wzmocnienie gradientowe	XGBoost	LightGBM
learning_rate	-	0.01, 0.05, 0.1, 0.15		
max_depth		3, 4, 5, ..., 12		
n_estimators		200, 350, 500, 1000		
max_features		0.25, 0.3, 0.35, 0.4	-	-
reg_alpha	-	-	0, 1, 2, ..., 15	
reg_lambda	-	-	0, 1, 2, ..., 15	

TABELA B.2: Wybrane parametry modeli do klasyfikacji zbioru testowego. Wyniki obejmują modele, w których zastosowano PCA. RF - algorytm lasu losowego, GB - algorytm wzmocnienia gradientowego.

	Deskryptory		
	FQD	RDKit	FQD + RDKit
Algorytm	RF	RF	GB
learning_rate	-	-	0.1
max_depth	8	6	4
n_estimators	500	200	200
max_features	0.4	0.3	0.4

TABELA B.3: Wybrane parametry modeli do klasyfikacji zbioru testowego. Wyniki obejmują modele, w których nie zastosowano PCA.

	Deskryptory		
	FQD	RDKit	FQD + RDKit
Algorytm	LGBM	XGB	LGBM
learning_rate	0.05	0.01	0.05
max_depth	4	4	7
reg_alpha	3	0	0
reg_lambda	14	4	2
n_estimators	350	350	200

TABELA B.4: Specyficzność, czułość, dokładność i precyzja w klasyfikacji zbioru testowego.

	Z PCA			Bez PCA		
	Reprezentacja					
Metryka	FQD	RDKit	Kombinowane	FQD	RDKit	Kombinowane
Specyficzność	0.836	0.889	0.926	0.881	0.881	0.918
	0.828	0.893	0.910	0.881	0.881	0.918
	0.836	0.889	0.902	0.881	0.881	0.918
	0.840	0.902	0.906	0.881	0.881	0.918
	0.836	0.885	0.902	0.881	0.881	0.918
Czułość	0.667	0.763	0.652	0.638	0.870	0.739
	0.667	0.783	0.667	0.638	0.870	0.739
	0.667	0.783	0.652	0.638	0.870	0.739
	0.652	0.783	0.652	0.638	0.870	0.739
	0.681	0.783	0.652	0.638	0.870	0.739
Dokładność	0.799	0.863	0.866	0.827	0.879	0.879
	0.792	0.869	0.856	0.827	0.879	0.879
	0.799	0.866	0.847	0.827	0.879	0.879
	0.799	0.875	0.850	0.827	0.879	0.879
	0.802	0.863	0.847	0.827	0.879	0.879
Precyzja	0.535	0.662	0.714	0.603	0.674	0.718
	0.523	0.675	0.676	0.603	0.674	0.718
	0.535	0.667	0.652	0.603	0.674	0.718
	0.536	0.692	0.662	0.603	0.674	0.718
	0.540	0.659	0.652	0.603	0.674	0.718

Lista publikacji i wystąpień konferencyjnych

W toku realizacji pracy doktorskiej opublikowano następujące publikacje i wystąpienia konferencyjne:

1. B. Fliszkiewicz, M. Sajdak, Towards quantum informed atom pairs. Manuskrypt wysłany do czasopisma oraz opublikowany jako preprint na platformie ChemRxiv, DOI: 10.26434/chemrxiv-2023-pcbrw,

2. B. Fliszkiewicz, M. Sajdak, Fragments quantum descriptors in classification of bio-accumulative compounds, Journal of Molecular Graphics and Modelling, 2023,

3. B. Fliszkiewicz, M. Sajdak, QM9-extended-plus database, Zenodo DOI: 10.5281/zenodo.7021447 (2022 - pierwsza wersja, 2023 - druga i trzecia wersja),

4. B. Fliszkiewicz, A study of boosting molecular descriptors with quantum-de–rived features in prediction of maximum emission wavelengths of chromophores, Chemical Data Collections 37, 2022,

5. B. Fliszkiewicz, Badania nad zastosowaniem uczenia maszynowego w przewidywaniu maksimów emisji optycznie aktywnych związków organicznych. - wystąpienie na 63 Zjeździe Polskiego Towarzystwa Chemicznego, 2021.

Publikacje i wystąpienia konferencyjne nie związane bezpośrednio z pracą doktorską:

1. S. Neffe, J. Lemańska, B. Fliszkiewicz, S. Jednoróg, Radiation impact of ashes from the combustion of bottom sediments in a municipal sewage treatment plant, plakat na konferencji NUTECH, Kraków, 2023,

2. B. Fliszkiewicz, O. Anttalainen, Z. Safaei, M. Wiśnik-Sawka, E. Budzyńska, J. Puton, Determination of the mobility coefficient based on data obtained with differential mobility spectrometer, ISIMS 2019 poster, Hannover 2019,

3. B. Fliszkiewicz, O. Anttalainen, Z. Safaei, M. Wiśnik-Sawka, E. Budzyńska, J. Puton, Identyfikacja analitów na podstawie danych otrzymywanych za pomocą różnicowych spektrometrów ruchliwości jonów, plakat w czasie 62 Zjazdu PTChem, Warszawa, 2019,

4. E. Budzyńska, M. Grabka, J. Kopyra, M. Maziejuk, Z. Safaei, B. Fliszkiewicz, M. Wiśnik, J. Puton, Ion mobility spectrometers and electron capture detector – A comparison of detection capabilities, Talanta 2019.

Bibliografia

[1] Daniel S. Wigh, Jonathan M. Goodman i Alexei A. Lapkin. „A review of molecular representation in the age of machine learning". W: *WIREs Computational Molecular Science* 12.5 (2022), e1603. DOI: 10.1002/wcms.1603.

[2] William J. Wiswesser. *A line-notation chemical formula*. Nowy Jork: Thomas Y. Crowell Company, 1954.

[3] John M. Barnard, Clemens J. Jochum i Stephen M. Welford. „A Universal Structure/Substructure Representation for PC-Host Communication". W: t. 400. ACS Symposium Series. 0. American Chemical Society, lip. 1989, s. 76–81. ISBN: 9780841216648. DOI: 10.1021/bk-1989-0400.ch008.

[4] David Weininger. „SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". W: *Journal of Chemical Information and Computer Sciences* 28.1 (lut. 1988), s. 31–36. ISSN: 0095-2338. DOI: 10.1021/ci00057a005.

[5] David Weininger, Arthur Weininger i Joseph L. Weininger. „SMILES. 2. Algorithm for generation of unique SMILES notation". W: *Journal of Chemical Information and Computer Sciences* 29.2 (maj 1989), s. 97–101. ISSN: 0095-2338. DOI: 10.1021/ci00062a008.

[6] David Weininger. „SMILES. 3. DEPICT. Graphical depiction of chemical structures". W: *Journal of Chemical Information and Computer Sciences* 30.3 (sierp. 1990), s. 237–243. ISSN: 0095-2338. DOI: 10.1021/ci00067a005.

[7] Michael A. Siani, David Weininger i Jeffrey M. Blaney. „CHUCKLES: A method for representing and searching peptide and peptoid sequences on both monomer and atomic levels". W: *Journal of Chemical Information and Computer Sciences* 34.3 (maj 1994), s. 588–593. ISSN: 0095-2338. DOI: 10.1021/ci00019a017.

[8] Inc. Daylight Chemical Information Systems. *SMARTS - A Language for Describing Molecular Patterns*. URL: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (term. wiz. 27. 02. 2023).

[9] Sheila Ash i in. „SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation". W: *Journal of Chemical Information and Computer Sciences* 37.1 (sty. 1997), s. 71–79. ISSN: 0095-2338. DOI: 10.1021/ci960109j.

[10] R. Webster Homer i in. „SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries". W: *Journal of Chemical Information and Modeling* 48.12 (grud. 2008), s. 2294–2307. ISSN: 1549-9596. DOI: 10.1021/ci7004687.

[11] Stephen Heller i in. „InChI - the worldwide chemical structure identifier standard". W: *Journal of Cheminformatics* 5.1 (sty. 2013), s. 7. ISSN: 1758-2946. DOI: 10.1186/1758-2946-5-7.

[12] Stephen R. Heller i in. „InChI, the IUPAC International Chemical Identifier". W: *Journal of Cheminformatics* 7.1 (maj 2015), s. 23. ISSN: 1758-2946. DOI: 10.1186/s13321-015-0068-4.

[13] Jeffery Leigh. „InChIs and Registry Numbers". W: *Chemistry International - Newsmagazine for IUPAC* 34.6 (2012), s. 23–23. DOI: 10.1515/ci.2012.34.6.23.

[14] Guenter Grethe i in. „International chemical identifier for reactions (RInChI)". W: *Journal of Cheminformatics* 10.1 (maj 2018), s. 22. ISSN: 1758-2946. DOI: 10.1186/s13321-018-0277-8.

[15] Mario Krenn i in. „Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". W: *Machine Learning: Science and Technology* 1.4 (paź. 2020), s. 045024. DOI: 10.1088/2632-2153/aba947.

[16] Alexandru T Balaban. „Chemical graphs: looking back and glimpsing ahead". W: *Journal of Chemical Information and Computer Sciences* 35.3 (1995), s. 339–350. DOI: 10.1021/ci00025a001.

[17] Harry P Schultz. „Topological organic chemistry. 1. Graph theory and topological indices of alkanes". W: *Journal of Chemical Information and Computer Sciences* 29.3 (1989), s. 227–228. DOI: 10.1021/ci00063a012.

[18] Harry P Schultz, Emily B Schultz i Tor P Schultz. „Topological organic chemistry. 2. Graph theory, matrix determinants and eigenvalues, and topological indexes of alkanes". W: *Journal of Chemical Information and Computer Sciences* 30.1 (1990), s. 27–29. DOI: 10.1021/ci00065a007.

[19] Stephan Wagner i Hua Wang. *Introduction to Chemical graph theory*. CRC press, 2018.

[20] Nenad Trinajstic. *Chemical graph theory*. CRC press, 2018.

[21] James Dugundji i Ivar Ugi. „An algebraic model of constitutional chemistry as a basis for chemical computer programs". W: *Computers in Chemistry*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1973, s. 19–64. ISBN: 978-3-540-38510-3. DOI: 10.1007/BFb0051317.

[22] H. L. Morgan. „The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." W: *Journal of Chemical Documentation* 5.2 (maj 1965), s. 107–113. ISSN: 0021-9576. DOI: 10.1021/c160017a018.

[23] Tuan Le i in. „Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures". W: *Chemical Science* 11 (38 2020), s. 10378–10389. DOI: 10.1039/D0SC03115A.

[24] Alan G. Konheim. *Hashing in Computer Science: Fifty Years of Slicing and Dicing*. John Wiley & Sons, Inc., 2010.

[25] RK Gupta. „A Review paper on concepts of cryptography and cryptographic hash function". W: *European Journal of Molecular & Clinical Medicine* 7.7 (2020), s. 3397–408. URL: https://ejmcm.com/issue-content/a-review-paper-on-concepts-of-cryptography-and-cryptographic-hash-function-11128.

[26] Hany Farid. „An Overview of Perceptual Hashing". W: *Journal of Online Trust and Safety* 1.1 (paź. 2021). DOI: 10.54501/jots.v1i1.24.

[27] Joseph L. Durant i in. „Reoptimization of MDL Keys for Use in Drug Discovery". W: *Journal of Chemical Information and Computer Sciences* 42.6 (list. 2002), s. 1273–1280. ISSN: 0095-2338. DOI: 10.1021/ci010132r.

[28] Sunghwan Kim. „Exploring Chemical Information in PubChem". W: *Current Protocols* 1.8 (2021), e217. DOI: 10.1002/cpz1.217.

[29] U.S. National Library of Medicine PubChem National Center for Biotechnology Information. *PubChem Substructure Fingerprint*. URL: https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf (term. wiz. 01. 03. 2023).

[30] Raymond E Carhart, Dennis H Smith i R Venkataraghavan. „Atom pairs as molecular features in structure-activity studies: definition and applications". W: *Journal of Chemical Information and Computer Sciences* 25.2 (1985), s. 64–73. DOI: 10.1021/ci00046a002.

[31] Dassault Systemes. *CTFILE FORMATS BIOVIA DATABASES 2020*. URL: https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf (term. wiz. 02. 03. 2023).

[32] Frances C. Bernstein i in. „The protein data bank: A computer-based archival file for macromolecular structures". W: *Journal of Molecular Biology* 112.3 (1977), s. 535–542. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(77)80200-3.

[33] Helen M. Berman i in. „The Protein Data Bank". W: *Nucleic Acids Research* 28.1 (sty. 2000), s. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[34] Dassault Systemes. *CTFILE FORMATS BIOVIA DATABASES 2020*. URL: https://discover.3ds.com/sites/default/files/2020-08/biovia_ctfileformats_2020.pdf (term. wiz. 02. 03. 2023).

[35] Sydney Hall i Brian McMahon. „International tables for crystallography, Vol. G: Definition and exchange of crystallographic data." W: *International Union Of Crystallography* (2005).

[36] Viviana Consonni i Roberto Todeschini. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References*. John Wiley & Sons, 2009.

[37] Roberto Todeschini i Viviana Consonni. *Handbook of molecular descriptors*. John Wiley & Sons, 2008.

[38] Andrea Mauri, Viviana Consonni i Roberto Todeschini. „Molecular descriptors". W: *Handbook of computational chemistry*. Springer, 2017, s. 2065–2093.

[39] T. Puzyn, J. Leszczynski i M.T. Cronin. *Recent Advances in QSAR Studies: Methods and Applications*. Challenges and Advances in Computational Chemistry and Physics. Springer Netherlands, 2010. ISBN: 9789048132416.

[40] Lemont B. Kier. „Indexes of molecular shape from chemical graphs". W: *Medicinal Research Reviews* 7.4 (1987), s. 417–440. DOI: 10.1002/med.2610070404.

[41] Mati Karelson, Victor S. Lobanov i Alan R. Katritzky. „Quantum-Chemical Descriptors in QSAR/QSPR Studies". W: *Chemical Reviews* 96.3 (sty. 1996), s. 1027–1044. ISSN: 0009-2665. DOI: 10.1021/cr950202r.

[42] Raymond E. Carhart, Dennis H. Smith i R. Venkataraghavan. „Atom pairs as molecular features in structure-activity studies: definition and applications". W: *Journal of Chemical Information and Computer Sciences* 25.2 (1985), s. 64–73. DOI: 10.1021/ci00046a002.

[43] Ramaswamy Nilakantan i in. „Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors". W: *Journal of Chemical Information and Computer Sciences* 27.2 (1987), s. 82–85. DOI: 10.1021/ci00054a008.

[44] Simon K. Kearsley i in. „Chemical Similarity Using Physiochemical Property Descriptors". W: *Journal of Chemical Information and Computer Sciences* 36.1 (1996), s. 118–127. DOI: 10.1021/ci950274j.

[45] Ansgar Schuffenhauer i in. „Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins". W: *Journal of Chemical Information and Computer Sciences* 43.2 (2003). PMID: 12653501, s. 391–405. DOI: 10.1021/ci025569t.

[46] Ansgar Schuffenhauer i in. „Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins". W: *Journal of Chemical Information and Computer Sciences* 43.2 (2003). PMID: 12653501, s. 391–405. DOI: 10.1021/ci025569t.

[47] Jingbo Yang i in. „Concepts and applications of chemical fingerprint for hit and lead screening". W: *Drug Discovery Today* 27.11 (2022), s. 103356. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2022.103356.

[48] Peter Ertl i Ansgar Schuffenhauer. „Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". W: *Journal of Cheminformatics* 1.1 (czer. 2009), s. 8. ISSN: 1758-2946. DOI: 10.1186/1758-2946-1-8.

[49] Shachar Fite, Omri Nitecki i Zeev Gross. „Custom Tokenization Dictionary, CUSTODI: A General, Fast, and Reversible Data-Driven Representation and

Regressor". W: *Journal of Chemical Information and Modeling* 61.7 (lip. 2021), s. 3285–3291. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00563.

[50] Philippe Schwaller i in. „Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction". W: *ACS Central Science* 5.9 (wrz. 2019), s. 1572–1583. ISSN: 2374-7943. DOI: 10.1021/acscentsci.9b00576.

[51] Igor V. Tetko i in. „State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis". W: *Nature Communications* 11.1 (list. 2020), s. 5575. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19266-y.

[52] Xu J i in. *New Application of Natural Language Processing (NLP) for Chemist Predicting Intermediate and Providing an Effective Direction for Mechanism Inference*. Preprint. ChemRxiv. Wrz. 2021. DOI: 10.26434/chemrxiv-2021-1bhnc.

[53] Hugh M Cartwright. *Machine Learning in Chemistry: The Impact of Artificial Intelligence*. The Royal Society of Chemistry, lip. 2020. ISBN: 978-1-78801-789-3. DOI: 10.1039/9781839160233.

[54] Kohulan Rajan, Achim Zielesny i Christoph Steinbeck. „DECIMER 1.0: deep learning for chemical image recognition using transformers". W: *Journal of Cheminformatics* 13.1 (sierp. 2021), s. 61. DOI: 10.1186/s13321-021-00538-8.

[55] Hayley Weir i in. „ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning". W: *Chemical Science* 12 (31 2021), s. 10622–10633. DOI: 10.1039/D1SC02957F.

[56] Yuemin Bian i in. „Deep Convolutional Generative Adversarial Network (dcGAN) Models for Screening and Design of Small Molecules Targeting Cannabinoid Receptors". W: *Molecular Pharmaceutics* 16.11 (list. 2019), s. 4451–4460. ISSN: 1543-8384. DOI: 10.1021/acs.molpharmaceut.9b00500.

[57] Dejun Jiang i in. „Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models". W: *Journal of Cheminformatics* 13.1 (lut. 2021), s. 12. ISSN: 1758-2946. DOI: 10.1186/s13321-020-00479-8.

[58] Hirotomo Moriwaki i in. „Mordred: a molecular descriptor calculator". W: *Journal of Cheminformatics* 10.1 (lut. 2018), s. 4. ISSN: 1758-2946. DOI: 10.1186/s13321-018-0258-y.

[59] Chun Wei Yap. „PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints". W: *Journal of Computational Chemistry* 32.7 (2011), s. 1466–1474. DOI: 10.1002/jcc.21707.

[60] Ecrl. *ECRL/padelpy: A python wrapper for Padel-Descriptor Software*. URL: https://github.com/ecrl/padelpy.

[61] Dong-Sheng Cao i in. „ChemoPy: freely available python package for computational biology and chemoinformatics". W: *Bioinformatics* 29.8 (mar. 2013), s. 1092–1094. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt105.

[62] Dong-Sheng Cao i in. „PyDPI: Freely Available Python Package for Chemoinformatics, Bioinformatics, and Chemogenomics Studies". W: *Journal of Chemical Information and Modeling* 53.11 (list. 2013), s. 3086–3096. ISSN: 1549-9596. DOI: 10.1021/ci400127q.

[63] Guido Van Rossum i Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[64] Charles R. Harris i in. „Array programming with NumPy". W: *Nature* 585.7825 (wrz. 2020), s. 357–362. DOI: 10.1038/s41586-020-2649-2.

[65] The pandas development team. *pandas-dev/pandas: Pandas*. Wer. latest. Lut. 2020. DOI: 10.5281/zenodo.3509134.

[66] Wes McKinney. „Data Structures for Statistical Computing in Python". W: *Proceedings of the 9th Python in Science Conference*. Red. Stéfan van der Walt i Jarrod Millman. 2010, s. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[67] F. Pedregosa i in. „Scikit-learn: Machine Learning in Python". W: *Journal of Machine Learning Research* 12 (2011), s. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[68] J. D. Hunter. „Matplotlib: A 2D graphics environment". W: *Computing in Science & Engineering* 9.3 (2007), s. 90–95. DOI: 10.1109/MCSE.2007.55.

[69] Guillaume Lemaître, Fernando Nogueira i Christos K. Aridas. „Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". W: *Journal of Machine Learning Research* 18.17 (2017), s. 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[70] Michael L. Waskom. „seaborn: statistical data visualization". W: *Journal of Open Source Software* 6.60 (2021), s. 3021. DOI: 10.21105/joss.03021.

[71] *RDKit: Open-source cheminformatics*. http://www.rdkit.org. [Online; accessed 11-February-2021].

[72] Thomas Kluyver i in. „Jupyter Notebooks - a publishing format for reproducible computational workflows". W: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Red. Fernando Loizides i Birgit Scmidt. Netherlands: IOS Press, 2016, s. 87–90. DOI: 10.3233/978-1-61499-649-1-87.

[73] Daniel G. A. Smith i in. „PSI4 1.4: Open-source software for high-throughput quantum chemistry". W: *The Journal of Chemical Physics* 152.18 (2020), s. 184108. DOI: 10.1063/5.0006002.

[74] Bartłomiej Fliszkiewicz i Marcin Sajdak. *QM9-extended-plus database*. Zenodo, list. 2023. DOI: 10.5281/zenodo.10184793.

[75] Raghunathan Ramakrishnan i in. „Quantum chemistry structures and properties of 134 kilo molecules". W: *Scientific Data* 1.1 (sierp. 2014), s. 140022. ISSN: 2052-4463. DOI: 10.1038/sdata.2014.22.

[76] Lars Ruddigkeit i in. „Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17". W: *Journal of Chemical Information and Modeling* 52.11 (2012). PMID: 23088335, s. 2864–2875. DOI: 10.1021/ci300415d.

[77] Zhenqin Wu i in. „MoleculeNet: a benchmark for molecular machine learning". W: *Chemical Science* 9 (2 2018), s. 513–530. DOI: 10.1039/C7SC02664A.

[78] J F Joung i in. „Experimental database of optical properties of organic compounds". W: *Scientific Data* 7 (wrz. 2020), s. 295. DOI: 10.1038/s41597-020-00634-8.

[79] Jiechun Liang i in. „QM-symex, update of the QM-sym database with excited state information for 173 kilo molecules". W: *Scientific Data* 7.1 (list. 2020), s. 400. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00746-1.

[80] Jiechun Liang i in. „QM-sym, a symmetrized quantum chemistry database of 135 kilo molecules". W: *Scientific Data* 6.1 (paź. 2019), s. 213. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0237-9.

[81] Noel M. O'Boyle i in. „Open Babel: An open chemical toolbox". W: *Journal of Cheminformatics* 3.1 (paź. 2011), s. 33. ISSN: 1758-2946. DOI: 10.1186/1758-2946-3-33.

[82] Noel M. O'Boyle, Chris Morley i Geoffrey R. Hutchison. „Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit". W: *Chemistry Central Journal* 2.1 (mar. 2008), s. 5. ISSN: 1752-153X. DOI: 10.1186/1752-153X-2-5.

[83] Zong-Rong Ye i in. „Predicting the emission wavelength of organic molecules using a combinatorial QSAR and machine learning approach". W: *RSC Advances* 10 (40 2020), s. 23834–23841. DOI: 10.1039/D0RA05014H.

[84] Cheng-Wei Ju i in. „Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields". W: *Journal of Chemical Information and Modeling* 61.3 (mar. 2021), s. 1053–1065. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c01203.

[85] Kaifu Gao i in. „Are 2D fingerprints still valuable for drug discovery?" W: *Physical Chemistry Chemical Physics* 22 (16 2020), s. 8373–8390. DOI: 10.1039/D0CP00305K.

[86] Ramón Alain Miranda-Quintana i in. „Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics†". W: *Journal of Cheminformatics* 13.1 (kw. 2021), s. 32. ISSN: 1758-2946. DOI: 10.1186/s13321-021-00505-3.

[87] Ramón Alain Miranda-Quintana i in. „Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection". W: *Journal of Cheminformatics* 13.1 (kw. 2021), s. 33. ISSN: 1758-2946. DOI: 10.1186/s13321-021-00504-4.

[88] Sereina Riniker i Gregory A. Landrum. „Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation". W: *Journal of Chemical Information and Modeling* 55.12 (grud. 2015), s. 2562–2574. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00654.

[89] Victoria M. Ingman i in. „QChASM: Quantum chemistry automation and structure manipulation". W: *WIREs Computational Molecular Science* 11.4 (2021), e1510. DOI: 10.1002/wcms.1510.

[90] Megan A. Lim i in. „Exploring Deep Learning of Quantum Chemical Properties for Absorption, Distribution, Metabolism, and Excretion Predictions". W: *Journal of Chemical Information and Modeling* (czer. 2022). ISSN: 1549-9596. DOI: 10.1021/acs.jcim.2c00245.

[91] V Ruusmann, S Sild i U Maran. „QSAR DataBank repository: open and linked qualitative and quantitative structure-activity relationship models". W: *Journal of Cheminformatics* 7 (2015), s. 32. DOI: 10.1186/1758-2946-6-25.

[92] G. Piir, S. Sild i U. Maran. „Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model". W: *SAR and QSAR in Environmental Research* 25.12 (2014). PMID: 25482723, s. 967–981. DOI: 10.1080/1062936X.2014.969310.

[93] G Piir, S Sild i U Maran. „Data for: Classifying bio-concentration factor with random forest algorithm, influence of the bio - accumulative vs. non - bio - accumulative compound ratio to modelling result, and applicability domain for random forest model." W: (2014). QsarDB repository, QDB.116. DOI: 10.15152/QDB.116.

[94] Nongnuch Artrith i in. „Best practices in machine learning for chemistry". W: *Nature Chemistry* 13.6 (czer. 2021), s. 505–508. ISSN: 1755-4349. DOI: 10.1038/s41557-021-00716-z.

[95] Anthony Yu-Tung Wang i in. „Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices". W: *Chemistry of Materials* 32.12 (2020), s. 4954–4965. DOI: 10.1021/acs.chemmater.0c01907.

[96] Tianqi Chen i Carlos Guestrin. „XGBoost: A Scalable Tree Boosting System". W: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, s. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.

[97] Guolin Ke i in. „LightGBM: A Highly Efficient Gradient Boosting Decision Tree". W: *Advances in Neural Information Processing Systems*. Red. I. Guyon i in. T. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[98] Tiejun Cheng i in. „Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model with Knowledge". W: *Journal of Chemical Information and Modeling* 47.6 (2007). PMID: 17985865, s. 2140–2148. DOI: 10.1021/ci700257y.

[99] Govindan Subramanian i in. „Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches". W: *Journal of Chemical Information and Modeling* 56.10 (2016). PMID: 27689393, s. 1936–1949. DOI: 10.1021/acs.jcim.6b00290.

[100] Yufeng Liu i Zhenyu Li. „Predict Ionization Energy of Molecules Using Conventional and Graph-Based Machine Learning Models". W: *Journal of Chemical Information and Modeling* 63.3 (2023). PMID: 36683339, s. 806–814. DOI: 10.1021/acs.jcim.2c01321.

[101] Kamel Mansouri i in. „OPERA models for predicting physicochemical properties and environmental fate endpoints". W: *Journal of Cheminformatics* 10.1 (mar. 2018), s. 10. ISSN: 1758-2946. DOI: 10.1186/s13321-018-0263-1.

[102] Alla P. Toropova, Andrey A. Toropov i Emilio Benfenati. „The self-organizing vector of atom-pairs proportions: use to develop models for melting points". W: *Structural Chemistry* 32.3 (czer. 2021), s. 967–971. ISSN: 1572-9001. DOI: 10.1007/s11224-021-01778-y.

[103] Ines Filipa Martins i in. „A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling". W: *Journal of Chemical Information and Modeling* 52.6 (2012). PMID: 22612593, s. 1686–1697. DOI: 10.1021/ci300124c.

[104] Katja Hansen i in. „Benchmark data set for in silico prediction of Ames mutagenicity". W: *Journal of Chemical Information and Modeling* 49.9 (2009), s. 2077–2081. DOI: 10.1021/ci900161g.

[105] Rani Nilima i in. „Machine-learning technique, QSAR and molecular dynamics for hERG–drug interactions". W: *Journal of Biomolecular Structure and Dynamics* 0.0 (2023), s. 1–26. DOI: 10.1080/07391102.2023.2193641.

[106] Takaya Saito i Marc Rehmsmeier. „The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". W: *PLOS ONE* 10.3 (mar. 2015), s. 1–21. DOI: 10.1371/journal.pone.0118432.

[107] Nicholas J. Mayhall i Krishnan Raghavachari. „Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials". W: *Journal of Chemical Theory and Computation* 7.5 (2011). PMID: 26610128, s. 1336–1343. DOI: 10.1021/ct200033b.