Uniwersytet Śląski w Katowicach
Wydział Matematyki, Fizyki i Chemii
Instytut Chemii

Rozprawa doktorska

ANALIZY QSPR WIELKICH BIBLIOTEK ZWIĄZKÓW CHEMICZNYCH NA PRZYKŁADZIE BAZY DANYCH ABAMACHEM

mgr Urszula Kucia

Rozprawa doktorska napisana
w Zakładzie Chemii Organicznej
Instytutu Chemii pod kierunkiem
Prof. dr. hab. inż. Jarosława Polańskiego

Katowice, 2019

Składam serdeczne podziękowania

Panu Prof. dr. hab. inż. Jarosławowi Polańskiemu

za wszelką pomoc, zrozumienie i opiekę naukową

Spis treści

Wykaz skrótów:

P- właściwość,

DE- deskryptor molekularny,

S- deskryptor,

CS- przestrzeń chemiczna,

VCS- wirtualna przestrzeń chemiczna,

FCS- zbiór cząsteczek,

MW- masa cząsteczkowa,

MW bin- binowana masa cząsteczkowa,

WBM- miara wagowa mas,

MBM- miara molowa mas,

SAS1- syntetyczna dostępność,

SAS1 bin- binowana dostępność syntetyczna,

QSPR- metoda badania ilościowej zależności między właściwością a strukturą,

QSAR- metoda badania ilościowej zależności między aktywnością biologiczną a strukturą,

QSER- metoda badania ilościowej zależności między ekonomią a strukturą,

SAR- metoda zależności struktura-aktywność,

BD- baza danych Beilstein,

ND- liczba Avogadro,

AC- całkowita liczba atomów,

MS- spektrometria mass,

Metody jonizacji cząsteczek w spektrometrach mas:

EI - Jonizacja elektronami,

ESI - Elektrorozpylanie,

MALDI - Desorpcja laserowa z udziałem matrycy,

APCI- Atmospheric Pressure Chemical Ionization.

Wstęp

Rozwój nauki uzależniony jest zarówno od wyników badań, które uzyskujemy przeprowadzając naukowe eksperymenty, jak również od teorii i hipotez, które próbują wyniki te tłumaczyć i modelować. Dlatego w dzisiejszych czasach komputer stał się powszechnym narzędziem badań chemicznych. Miało to wpływ na powstanie dyscypliny naukowej jaką jest chemoinformatyka. Chemoinformatyka (ang. chemoinformatics) to interdyscyplinarna dziedzina łącząca ze sobą dwa kierunki - chemię oraz informatykę[1]. Do zadań chemoinformatyki należy: tworzenie baz danych i ich eksploracja, jak również wyszukiwanie, analiza, rozpowszechnianie, wizualizacja oraz wykorzystanie informacji chemicznej. Przedmiotem badań chemoinformatyki są m.in. reprezentacje związków chemicznych in silico[2], operacje na molekułach in silico, obliczenia deskryptorów molekularnych dla wirtualnych molekuł, prognozowanie właściwości substancji chemicznych oraz projektowanie ścieżek syntez chemicznych. Geneza chemoinformatyki związana jest z projektowaniem leków, przy czym termin „projektowanie leków" (ang. drug design and discovery) definiowany jest zwykle nieco szerzej i oznacza poszukiwanie możliwości wytworzenia nowych farmaceutyków (inaczej leków) czyli substancji, które będą wykazywać pożądany profil aktywności biologicznej[3].

Integralną częścią chemoinformatyki jest projektowanie molekularne, definiowane w literaturze jako konstruowanie nowych molekuł o określonym profilu aktywności chemicznej bądź biologicznej. Jej zadaniem jest konstruowanie i poszukiwanie takich molekuł, które będą wykazywać pożądaną właściwość oraz odpowiednio skuteczne działanie. Do konstruowania projektowanych molekuł wykorzystuje się chemię organiczną, ponieważ większość nowych leków stanowią związki syntetyczne otrzymane drogą syntezy organicznej. W metodach projektowania molekularnego wykorzystuje się matematykę i techniki obliczeniowe. Przy ich pomocy tworzy się odpowiednie modele, przy czym model matematyczny reprezentuje zbiór danych i faktów, tłumacząc modelowane efekty molekularne. Parametry modelu mogą wyjaśniać lub prognozować pojedyncze fakty, korzystając z bardzo złożonych danych. W wielu przypadkach trudność stworzenia modelu wynika ze złożoności modelowanych obiektów, jak również z braku odpowiednich danych, przez co konieczne staje się uproszczenie modelu za

cenę jego mniejszej wiarygodności. Dlatego zastosowanie metod komputerowych w chemii zwiększa efektywność przetwarzania danych w porównaniu z obliczeniami dokonywanymi bez ich udziału. Metody komputerowe wykorzystują matematykę in silico polegającą na przetwarzaniu przez komputer dużej ilości prostych operacji[3].

Chemia in silico obejmuje trzy dyscypliny naukowe:

- ✓ Chemię kwantową, której przedmiotem badań są m.in. atomy lub małe cząsteczki;
- ✓ Chemometrię, której przedmiotem badań są m.in. statystyczne i numeryczne metody analizy danych;
- ✓ Chemoinformatykę, której przedmiotem badań są m.in. duże (bio) systemy chemiczne[3].

Badana zależność między strukturą a właściwością związku chemicznego stanowi istotny problem w chemii, jest także ważnym elementem w procesie projektowania leków. Jednak aby w pełni zrozumieć mechanizm wprowadzania leków na rynek, niezbędne jest także zrozumienie ekonomicznych uwarunkowań projektowania molekularnego. Ekonomia jest decydującym czynnikiem określającym komercyjny aspekt obecności leku na rynku farmaceutycznym. W tym miejscu warto postawić pytanie - czy możliwe jest modelowanie efektów ekonomicznych na rynku związków chemicznych? Niniejsza rozprawa doktorska dotyczy eksploracji relacji między strukturą a właściwościami QSRP (ang. Quantiative Structure - Property Relationship), w tym także ceną dla dużej komercyjnej biblioteki bloków budulcowych (ang. building blocs) zawierającej związki chemiczne[4].

CEL PRACY

Analiza wielkich danych stanowi istotne wyzwanie badawcze ostatnich lat. Metoda ta ma wiele zalet. Pomimo tego, zarówno w naukach chemicznych, jak i projektowaniu leków spotyka się ją rzadko, ponieważ istotnym problemem jest ograniczony zakres dostępności takich danych. Celem pracy była analiza dostępności, akwizycji i przetwarzania zbiorów danych tego typu. Jako przykład takiej analizy wykorzystano katalog związków chemicznych Abamachem. W badaniach po raz pierwszy przeprowadzono analizę zależności zachodzącej pomiędzy właściwościami ekonomicznymi a deskryptorami molekularnymi dla komercyjnego katalogu związków chemicznych Abamachem zawierającego ceny[4].

CZEŚĆ LITERATUROWA

1. Wybrane problemy projektowania leków in silico

Do głównych zastosowań dyscypliny naukowej jaką jest chemoinformatyka w procesie projektowaniu leków zaliczamy m.in.:

- poszukiwanie korelacji między strukturą a właściwościami cząsteczki QSPR (ang. quantitative structure-property relationship),
- analiza ilościowej zależności między strukturą a aktywnością QSAR (ang. quantitative structure-activity relationship),
- analiza zależności między budową leku a jego działaniem SAR (ang. structure - activity relationships),
- tworzenie i obliczanie deskryptorów molekularnych,
- klasyfikacja związków ze względu na podobieństwo,
- tworzenie baz danych chemicznych,
- wirtualny skrining związków chemicznych[5].

1.1. Koncepcje obliczeniowe stosowane w metodach in silico

Ze względu na rozmiar danych, jakie przetwarza chemoinformatyka niezbędne jest zastosowanie metod in silico. Metoda ta od wielu lat wykorzystywana jest do projektowania nowych leków. Jednym z zastosowań metod in silico jest obliczanie deskryptorów molekularnych. W ostatnim czasie stosowane one były głównie do identyfikacji i przewidywania różnic między lekami a biologicznie aktywnymi cząsteczkami branymi pod uwagę jako potencjalne wzorce w procesie projektowania leków[6]. Ponadto od wielu lat w przemyśle farmaceutycznym i chemicznym wykorzystuje się tzw. racjonalne projektowanie leków (ang. rational drug desing) w celu odkrywania nowych substancji leczniczych. Jedną z głównych koncepcji w projektowaniu leków jest tzw. lekotypia (ang. drug-likeness) lub lekopodobieństwo. Koncepcja ta w najprostszym ujęciu opiera się na założeniu, że leki stanowią

pewną klasę podobnych do siebie związków chemicznych o zbliżonych właściwościach. Umożliwia to projektowanie ich właściwości farmaceutycznych i farmakokinetycznych, takich jak na przykład: rozpuszczalność cząsteczki, trwałość, biodostępność, czy profil dystrybucji[6]. Najbardziej znanym kryterium w koncepcji lekotypii jest reguła Lipińskiego pozwalająca na dobór cząsteczek spełniających największe prawdopodobieństwo przedostania się leku do komórki, a następnie jego aktywacji poprzez odziaływanie z receptorem. Inną koncepcją jest koncepcja struktur uprzywilejowanych (ang. privileged structures), której ideą jest wybór tylko pewnych powtarzających się elementów, motywów strukturalnych wpływających na tworzenie kompleksu ligand-receptor[7].

2. Deskryptory molekularne

Deskryptory molekularne (ang. molecular descriptor) relatywnie definiuje się jako dowolną numeryczną reprezentację cząsteczki chemicznej. Oblicza się je przez transformację informacji chemicznej, która koduje cząsteczkę czasami także w postaci symbolicznego obrazu. Wykorzystuje się w tym celu procedury matematyczne, często opracowane w postaci algorytmów numerycznych realizowanych przez programy komputerowe. Deskryptorem molekularnym może być na przykład liczba atomów węgla czy wodoru w molekule. Do wyznaczenia takiego deskryptora potrzebne jest zliczenie konkretnych elementów strukturalnych[3].

2.1. Podział deskryptorów molekularnych

Dokonanie podziału deskryptorów jest trudne ze względu na różnorodność reprezentacji molekularnych. Deskryptory można podzielić ze względu na typ danych, jaki reprezentują oraz ze względu na tzw. wymiarowość danych opisujących cząsteczki czy wymiaru obliczanego deskryptora. Wyróżniany jest również inny rodzaj podziału ze względu na zakres zastosowania, do którego zaliczamy deskryptory kodujące pełniące funkcję, którą w najprostszy sposób określić można jako definicja konstytucji i stereochemii molekuły oraz deskryptory niekodujące niosące różnorodne informacje o budowie cząsteczki. Do deskryptorów kodujących można zaliczyć deskryptory typu notacji liniowej. Ich podstawową cechą jest jednoznaczność odwzorowania budowy cząsteczki. Natomiast deskryptory niekodujące, opisując wybrane

cechy molekularne, pełnią funkcję informacyjną. Deskryptory te zwykle uniemożliwiają odtworzenia cząsteczki, ponieważ w czasie obliczania deskryptora traci się część informacji konstytucyjnej i/lub stereochemicznej, w wyniku czego te same wartości często odwzorowuje wiele cząsteczek [3,8,9].

2.2.1. Kodujące deskryptory konstytucyjne

Do kodujących deskryptorów konstytucyjnych zaliczamy systemy notacji liniowej cząsteczek oraz macierzowe systemy kodowania konstrukcyjnego.

Przykłady systemów notacji liniowej to między innymi:

- SMILES (ang. Simplified Molecular Input Line Entery System), pełniące funkcję kodowania cząsteczki. Opracowany przez Davida Weiningera kod SMILES jest najbardziej rozpowszechnionym systemem kodowania opartym na teorii wiązań walencyjnych, w którym znaki alfanumeryczne kodu ASCII są używane do kodowania cząsteczki oraz reakcji chemicznych. Główną cechą tych deskryptorów jest możliwość odtworzenia cząsteczki w jednoznaczny sposób[2].
- WNL (ang. Wiswessr), pełni funkcję kodowania cząsteczki, gdzie konkretnemu symbolowi przypisany jest atom lub zbiór atomów. Notacja ta stosuje symbole pierwiastków chemicznych, a elementy strukturalne, takie jak na przykład grupy funkcyjne czy pozycje podstawników, są reprezentowane przez litery. Choć notacja ta ma wiele zalet, jest skomplikowana[3].
- RPSDAL pełni funkcję kodowania cząsteczki. Została ona opracowana w Instytucie Beilstein dla systemu DIALOG. Struktury chemiczne kodowane są tu przez znaki alfanumeryczne, co polega na przetworzeniu informacji strukturalnej w system bazodanowy. Notacja ta stosuje unikatowe numery i symbole, a elementy strukturalne, takie jak podstawniki czy łańcuchy boczne są kodowane przy pomocy przecinków[3].

Macierzowe systemy kodowania molekuł

Innym typem deskryptorów kodujących należących do macierzowych systemów kodowania konstrukcyjnego są desktyptory toplogiczne (2D) numerycznej reprezentacji grafu, w której cząsteczka jest złożona z atomów, czyli wierzchołków, a wiązania są definiowane jako krawędzie grafu. Macierzowe zapisy grafów molekularnych przedstawiane są na kilka sposobów w zależności od sposobów kodowania atomów połączonych wiązaniami chemicznymi. Wyróżniamy zapisy w postaci macierzy: sąsiedztwa, odległości, częstości, wiązań i elektronów wiążących[2,10].

2.2.2. Deskryptory obliczane na podstawie atomowej reprezentacji cząsteczki

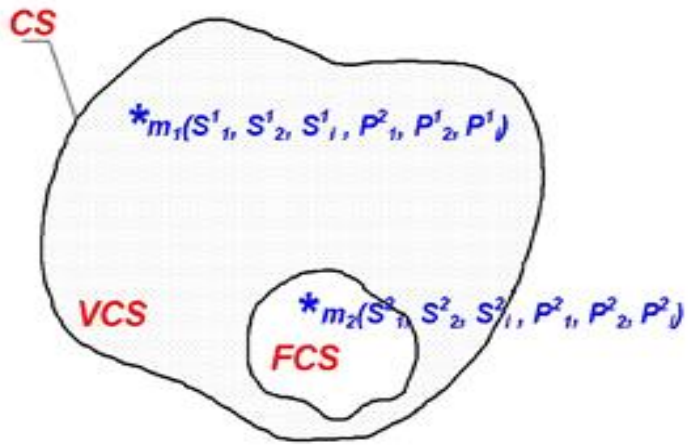
Przykładem tego typu deskryptorów jest m.in. masa cząsteczkowa, całkowita liczba atomów, liczby poszczególnych atomów: węgla, wodoru, heteroatomów, liczby wiązań, liczby wiązań donorowych, liczby wiązań akceptorowych, asymetrycznych atomów, logP, TPSA, promienia atomowego itp. Na podstawie wzoru strukturalnego lub cząsteczkowego można obliczyć bardzo dużo deskryptorów tego typu[3].

2.2.3. Deskryptory daktyloskopowe cząsteczki

Daktylogramy molekularne (ang. molecular fingerprinters) należą do metod zapisu (sub)strukturalnych elementów cząsteczki (ang. structural analysis). Metoda ta polega na obliczeniu elementów strukturalnych cząsteczki przedstawionej w reprezentacji dwu- lub trójwymiarowej. Taki rodzaj deskryptora przedstawiany jest przy użyciu tablic binarnych, których zadaniem jest obliczenie wybranych cech strukturalnych i wyświetlenie odpowiednich wartości. Jest to wektor, którego poszczególne elementy określają występowanie w cząsteczce określonej substruktury. Wartości kodują: 1-obecność lub 0-brak obliczonego elementu strukturalnego. Deskryptory daktyloskopowe to matematyczna forma reprezentacji danych, a jej postać graficzną stanowi histogram[3,11].

2.2.4. Deskryptory skorelowane z właściwościami

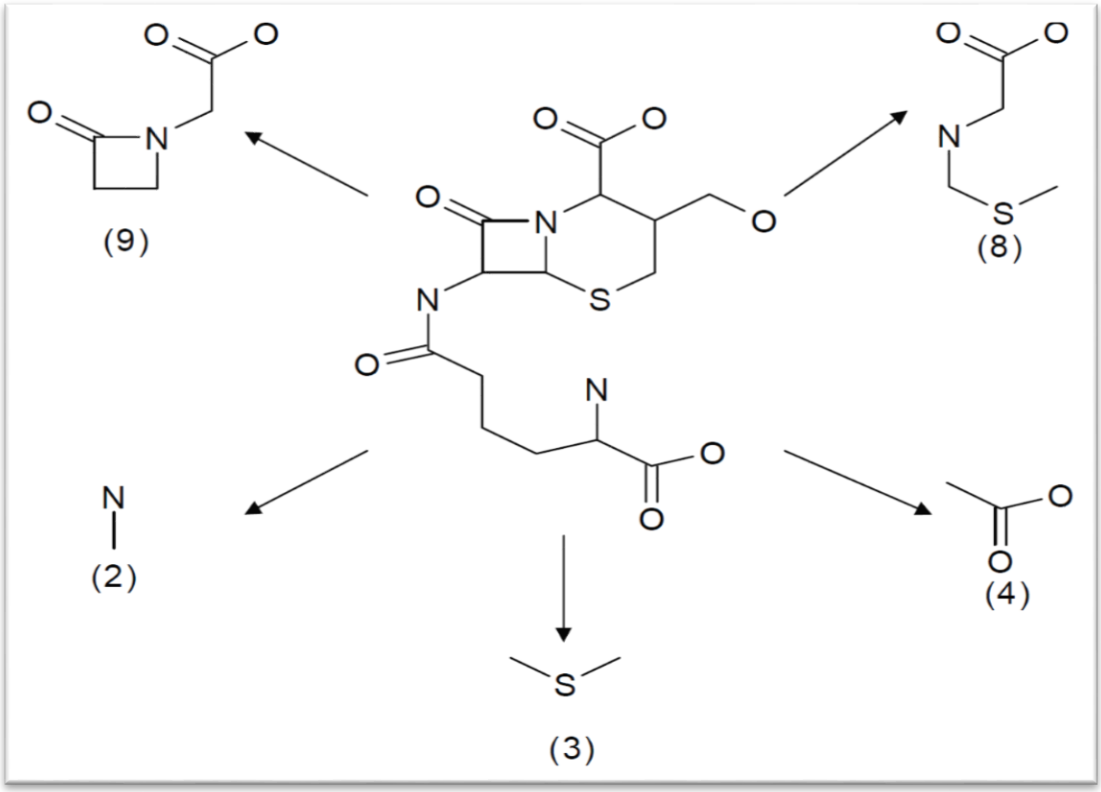
Prognozowanie właściwości chemicznych, fizycznych i biologicznych cząsteczek oraz reaktywności cząsteczki jest ważnym elementem projektowania molekularnego. Taki rodzaj modelowania najczęściej wykorzystuje substancję (molekułę) w kontekście przetwarzania molekularnego korelacji pomiędzy wybranymi deskryptorami molekularnymi a wybranymi właściwościami. Ważnymi parametrami opisującymi związki chemiczne są następujące właściwości: lipofilowość, aktywność IC_{50} oraz stałe dysocjacji kwasów pka. Dla wielu związków zostały one zmierzone w eksperymencie. Właściwości te odgrywają znaczącą rolę w metodach projektowania molekularnego, ponieważ są stosowane do obliczania deskryptorów strukturalnych, które określić można jako korelaty właściwości[3]. W metodologii tej wyznacza się model regresyjny mający na jej podstawie służyć prognozowaniu właściwości (P) dla nowych nieopisanych jeszcze cząsteczek - cząsteczki zbioru wirtualnego VCS[12], co zilustrowano na rysunku 1.



Rysunek 1. Odwzorowanie cząsteczki w przestrzeni chemicznej (CS) i wirtualnej przestrzeni chemicznej (VCS) dla zbioru cząsteczek (FCS) stosowane w metodach in silico[2]*.* m_1, m_2 *– cząsteczki reprezentowane przez deskryptory S lub właściwości P.*

2.2.5. Deskrypory obliczane na podstawie fragmentów molekularnych

Deskryptory obliczane na podstawie fragmentów molekularnych są reprezentowane przez zbiór wartości definiujących określone fragmenty (podstruktury) cząsteczki, jak również sposoby ich łączenia. Jako przykład może służyć wektorowa reprezentacja wszystkich możliwych fragmentów (podstruktury) o zdefiniowanej liczbie atomów. Deskryptory tego typu umożliwiają definiowanie wybranych fragmentów struktury badanego związku, takich jak na przykład: liczbę -pierwszo -drugo -trzecio -czwarto rzędowych węgli C (sp^3), liczbę wiązań wodorowych, liczbę grup hydroksylowych, amidowych, indeksy nienasycenia, liczbę donorowych i akceptorowych atomów wodoru oraz wiele innych ugrupowań. Technika kodowania fragmentacyjnego stosowana jest na przykład w analizie HQSAR[13].

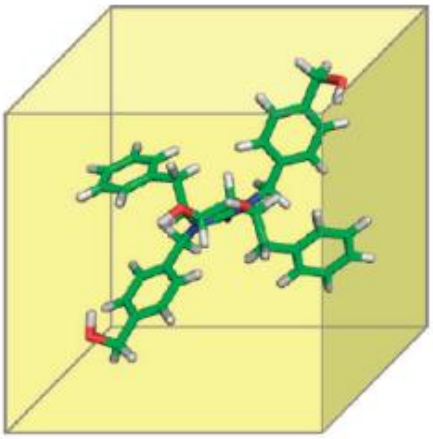


Rysunek 2. Deskryptory obliczane na podstawie fragmentów molekularnych (w nawiasach podano liczbę atomów występujących w poszczególnych podstrukturach)[13].

2.2.6. Deskryptory geometryczne i pola oddziaływań cząsteczkowych

Kształt cząsteczki jest pojęciem trudnym i jego opis jest skomplikowany, jednakże ma on duże znaczenie, gdyż zawiera ważne informacje wpływające na korelację z różnymi deskryptorami molekularnymi. Ma także wpływ na potencjalne połączenia z innymi strukturami, w wyniku czego dochodzi do aktywowania lub blokowania, na przykład leku czy ligandu łączącego się w miejscu aktywnym[14]. Aby precyzyjnie określić kształt i strukturę molekuły, wykorzystuje się deskryptory geometryczne (ang. geometrical descriptors, molecular shape descriptors), które dzieli się na proste i złożone. Proste deskryptory geometryczne używane są do opisu powierzchni, objętości i sił Van der Vaalsa, a także potencjału elektrostatycznego, hydrofobowego oraz wiązania wodorowego. Natomiast złożone deskryptory geometryczne dostarczają bardziej szczegółowych informacji o modelowaniu efektów molekularnych. Wykorzystuje się je w momencie, gdy proste deskryptory nie wystarczają do wytłumaczenia efektów molekularnych. Definiowane są w celu modelowania efektów fizycznych, chemicznych czy biologicznych, które są wytwarzane przez cząsteczki oraz otoczenie, na które one oddziałują[15].

Deskryptory pola odziaływań cząsteczkowych MIF (ang. Molecular Interaction Fields) to deskryptory wykorzystywane do obliczenia wartości energii pól o charakterze elektrostatycznym i sterycznym oraz wzajemnych oddziaływań cząsteczki z sondami atomowymi (ang. probe atom) umieszczanymi w węzłach sieci (ang. grid). Do obliczenia pól sterycznych i elektrostatycznych wokół cząsteczek wykorzystuje się potencjał Lenarda-Jonesa i Kulomba. W rzeczywistości regularna sieć otaczająca cząsteczkę to złożona tablica punktów zawierająca współrzędne sieci (x,y,z). MIF to deskryptory pozwalające na zidentyfikowanie wybranych, interesujących badacza fragmentów cząsteczki; zazwyczaj jednak analizowane są wszystkie atomy cząsteczki oraz otaczające ją węzły sieci[16].



Rysunek 3. Regularna sieć otaczająca cząsteczkę stosowna do projekcji deskryptora MIF[16].

2.2.7. Deskryptory profilu konformacyjnego i wirtualnego miejsca receptorowego

Deskryptory profilu konformacyjnego to zbiór deskryptorów obsadzenia komórek jednostkowych GCODs (ang. Grid Cell Ocupancy Descriptors), służący do wygenerowania przestrzeni konformacyjnej cząsteczki. Proces tworzenia deskryptora konformacyjnego polega na określeniu zbioru konformerów wraz z ich przestrzennym umiejscowieniem w sieci komórek. Na kolejnym etapie oblicza się częstotliwości występowania poszczególnych atomów. Jednostkowy deskryptor opisujący cząsteczki w sieci jest definiowany przy użyciu schematu obsadzania konkretnych jednostkowych komórek przez atomy badanych cząsteczek w trakcie trwania analizy konformacyjnej. Atomy określają przede wszystkim rodzaje wybranych przestrzennych ugrupowań, które powinna mieć cząsteczka, aby została rozpoznana na przykład przez receptor. Graficzna wizualizacja obsadzenia określonych komórek jednostkowych tworzy widmo kształtu molekularnego MSS (ang. Molecular Shape Spectrum)[17,18].

Deskryptory wirtualnego miejsca receptorowego QUASAR (ang. Qusasi-Atomistic Receptor Surrogate) są stosowane w modelach 5D-QSAR i 6D-QSAR. W modelach 6D-QSAR należy uwzględnić efekty solwatacyjne, a otrzymany deskryptor nazywany jest wówczas deskryptorem efektów rozpuszczalnikowych[19,20,21]. Deskryptory QUASAR mogą być analizowane jako reprezentacja atomowa (ang. atomistic receptor site model), jak również jako reprezentacja powierzchniowa (ang. receptor surface model). Wirtualny receptor to receptor wiążący na swojej powierzchni ligandy. Powierzchnię receptora można wyznaczyć przy użyciu

zbioru konformerów, czyli aktywnych ligandów. Symulacja geometryczna polega na dopasowaniu powierzchni liganda do powierzchni wirtualnego receptora. Graficzna wizualizacja deskryptora QUASAR to powierzchnia reprezentująca każdy atom konformeru zakodowany przy użyciu kolorów[22,23,24] .

W literaturze przedmiotu spotyka się także deskryptory złożonych systemów cząsteczkowych ligand-receptor nazywanych zazwyczaj deskryptorami odziaływań międzycząsteczkowych COMBINE (ang. Comparative Molecular Binding Energy), które służą do obliczenia energii oddziaływań każdej pary atomów liganda oraz receptora[25].

3. Problemy pomiarów efektów fizykochemicznych, biologicznych i ekonomicznych

3.1. Efekty fizykochemiczne

3.1.1. Równanie Hammetta

Równie Hammeta opisuje w sposób ilościowy wpływ efektów elektronowych podstawników na reaktywność cząsteczki według poniższego wzoru[26].

$$log\frac{k}{k_0} = \sigma * \rho \tag{1.1}$$

gdzie:

logk- logarytm stałej szybkość reakcji hydrolizy,

$logk_0$- logarytm stałej szybkość reakcji dysocjacji (0 odnosi się do podstawnika referencyjnego),

σ- stała określająca podstawnik Hammeta,

p- stała określająca dany typ reakcji.

Stała Hammetta σ opisuje reaktywność związków chemicznych poprzez analizę efektów elektronowych, uwzględnia wpływ indukcyjny i rezonansowy podstawnika na równowagę reakcji chemicznej. Równanie Hammeta nie znajduje zastosowanie do modelowania wiązania się ligandu z receptorem. Hammett w swoich badaniach zajmował się w szczególności

reakcjami dysocjacji pochodnych kawasów benzoesowych oraz ich hydrolizą w pozycjach meta i para. Stałe Hammeta σ będące miarą efektu elektronowego podstawnika poprawnie opisują efekty elektronowe dla analogów podstawionych w pozycji meta oraz para, natomiast wartości stałych Hammeta σ w pozycji orto nie są wiarygodne, co jest spowodowane silnym wpływem sferycznym i rezonansowym.

Podstawniki w pierścieniu aromatycznym wpływają na szybkość reakcji, zwiększają bądź zmniejszą gęstość elektronową wywołując aktywację bądź dezaktywację pierścienia.

Podstawniki o ujemnych wartościach σ są donorami elektronów, a podstawniki o dodatnich wartościach σ są akceptorami elektronów[26]; znaczenie ma także rozmiar podstawnika. Do opisu efektów przestrzennych podstawnika wykorzystuje się Stałą Tafta E_s wyrażoną za pomocą poniższego wzoru[27]:

$$E_S = \log\left(\frac{k}{k_0}\right) \tag{1.2}$$

gdzie: k- stała szybkości badanej reakcji,

k_o– stała szybkości reakcji referencyjnej (dla podstawnika metylowego).

3.2. Efekty biologiczne

3.2.1. Metoda Hanscha

Metoda Hanscha opisuje wpływ liopofilowości na efekty biologiczne. Metoda ta wymaga ilościowej reprezentacji właściwości fizykochemicznych analizowanych związków[28]. Właściwością fizykochemiczną substancji decydującą o przenikaniu leku przez błony komórkowe jest lipofilowość (hydrofobowość); ze względu na łatwość pomiaru można ją zmierzyć doświadczalnie. Miarą lipofilowość (hydrofobowości) jest logP, będący logarytmem współczynnika podziału P w mieszaninie n–oktanol/woda.

$$logP = log\frac{stężenie\ substancji\ w\ n-oktanol}{stężenie\ substancji\ w\ wodzie} \tag{1.3}$$

Związki hydrofobowe o wysokiej wartość współczynnika P silnie wiążą się z fazami lipidowymi, przez co mogą zostać zatrzymane przez tkankę tłuszczową i częściowo przejść przez błonę komórkową, utrudniając tym samym dotarcie leku do receptora. Natomiast związki hydrofilowe o niskiej wartości współczynnika P mogą pozostać w fazie wodnej i w konsekwencji zostać usunięte z organizmu. Dlatego poszukiwane są związki, dla których wartość współczynnika podziału P byłaby optymalna i umożliwiłaby transport związku przez błonę komórkową[29].

Wartość współczynnika P w dużej mierze zależy od wielkości oraz budowy chemicznej rozpuszczanego związku. Hansch i jego współpracownicy zaproponowali model, który umożliwia obliczenie logP na podstawie wartości cząstkowych logP fragmentów molekularnych. Wyznacza się je eksperymentalnie, obliczając współczynnik podziału P dla związku z podstawnikiem i bez podstawnika według poniższego wzoru[28].

$$\pi = logP_X - logP_H \tag{1.4}$$

gdzie:

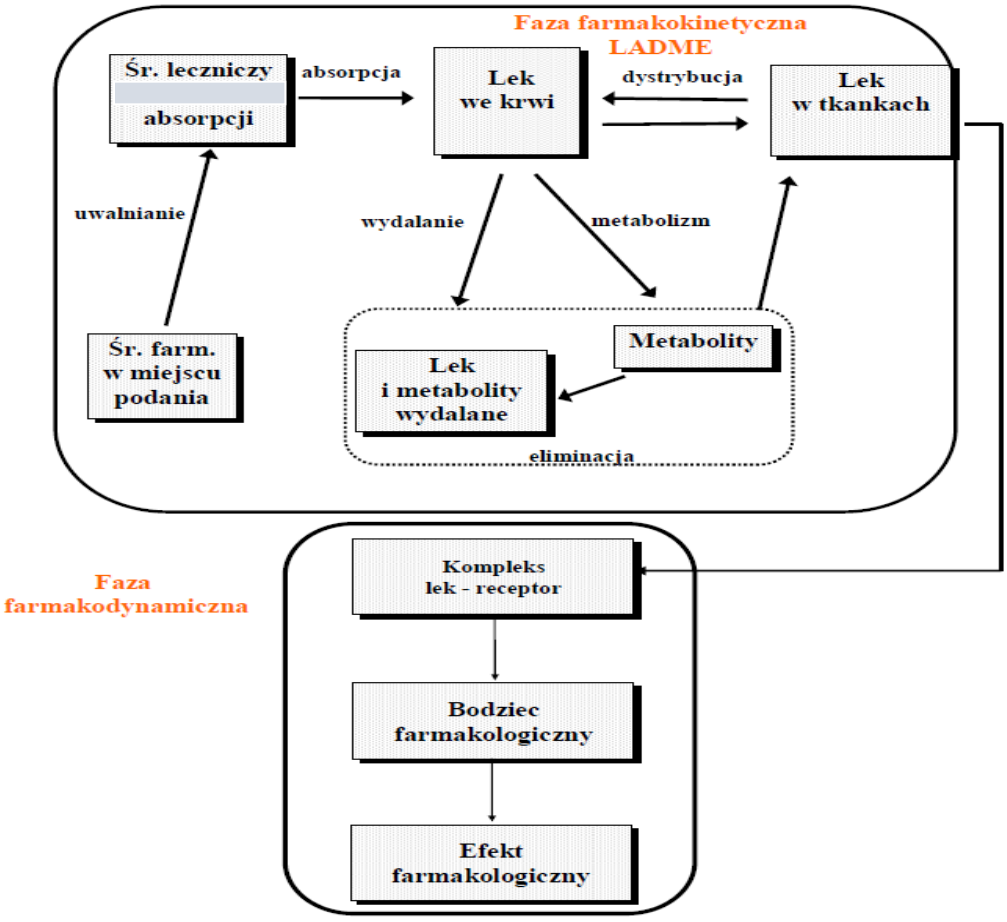
π – stała Hanscha

$logP_X$ - logarytm współczynnika podziału dla związku (X- z podstawnikiem),

$logP_H$- logarytm współczynnika podziału dla związku standardowego (H– podstawionego wodorem).

W 1964 roku Hansch i Fuijta niezależnie opracowali podstawy teoretyczne, wyróżniając dwa niezależne od siebie procesy transportu leku. Etapem wstępnym jest droga, którą przebywa lek od momentu podania do miejsca działania, czyli tzw. faza farmakokinetyczna. Kolejnym procesem jest reakcja chemiczna w miejscu działania leku, czyli najczęściej w komórce; jest to tzw. faza farmakodynamiczna (rysunek 4.)[29]. W tym samym roku badacze opublikowali równanie matematyczne opisujące ilościową zależność parametrów hydrofobowych, elektronowych i sterycznych od aktywności biologicznej[30]. Model ten zakłada, że zarówno transport cząsteczki leku do komórki, jak i utworzenie kompleksu aktywny lek-receptor zależy od wielu czynników, takich jak równowaga lipidowo-hydrofobową, rozkład gęstości

elektronowej w cząsteczce oraz kształt i wielkości cząsteczki. Matematyczny uogólniony model Hansach opisuje równanie 1.5.



Rysunek 4. Etapy wędrówki leku: faza farmakokinetyczna i farmakodynamiczna[31].

$$log\frac{1}{C} = -k_1\pi^2 + k_2\pi + k_3\sigma + k_4E_S + k_5 \quad (1.5)$$

gdzie:

1/C - aktywność związku chemicznego;

k - stałe otrzymane w wyniku analizy regresji;

π - stała Hanscha - parametr hydrofobowy podstawników;

σ - stała Hammetta - parametr elektronowy podstawników;

Es - stała Tafta - parametr sterczyny podstawników.

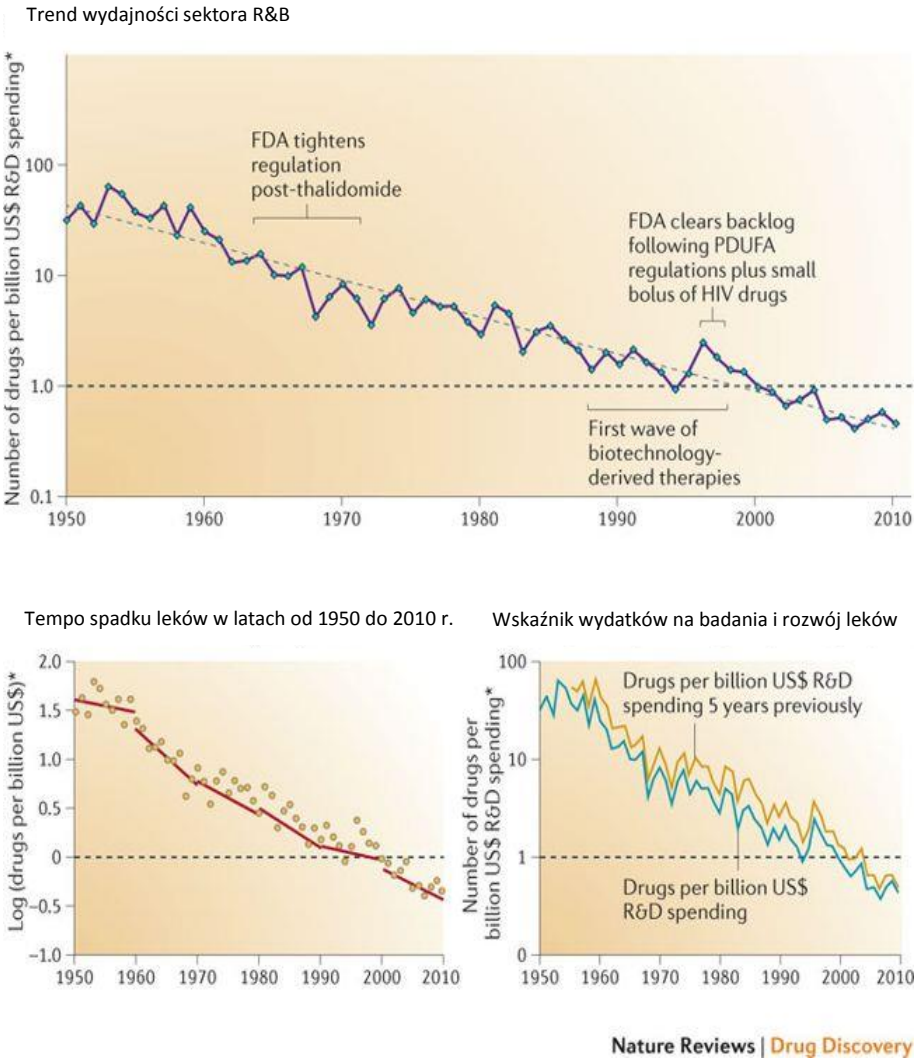
3.3. Efekty ekonomiczne

3.3.1. Prawo Erooma

Prawo Erooma opisuje efektywność ekonomiczną badań R&D w przemyśle farmaceutycznym. Poszukiwanie nowych leków staje się coraz mniej wydajne oraz coraz bardziej kosztowne.

Przyczyny spadku wydajności badań i rozwoju można wyjaśnić poniższymi modelami jakościowymi[11]:

- Lepszy niż Beatles - niezwykła popularność zespołu the Beatles spowodowała, że tworzenie nowych utworów muzycznych, których sukces komercyjny byłby większy, niż piosenek skomponowanych przez tę grupę stało się niemal niemożliwe. Podobne zjawisko zaobserwowano również na rynku farmaceutycznym w odniesieniu do leków odnoszących sukces w sprzedaży; prawdopodobieństwo „przebicia" ich popularności przez nowe specyfiki było również niewielkie.
- Zaostrzenie regulacji prawnych – spowodowało, że na rynku farmaceutycznym leki niespełniające norm prawnych zostały wycofane, a wprowadzenie na ich miejsce nowych leków spełniających wymogi prawne wiązało się z wyższymi nakładami finansowymi na niezbędne badania oraz dłuższym okresem ich prowadzenia.
- Tendencja wydawania pieniędzy – złe zarządzanie budżetem poprzez niewłaściwe oszacowanie kosztów często prowadziło do przekraczania zakładanego limitu środków.
- Przecenianie możliwości – ograniczenia wynikające ze zrozumienia mechanizmów biochemicznych wpływających na jakość oraz skuteczność nowoczesnych metod in silico i badań przesiewowych HTS–wykorzystywanych w projektowaniu leków nie okazały się tak skuteczne, jak oczekiwano[11].

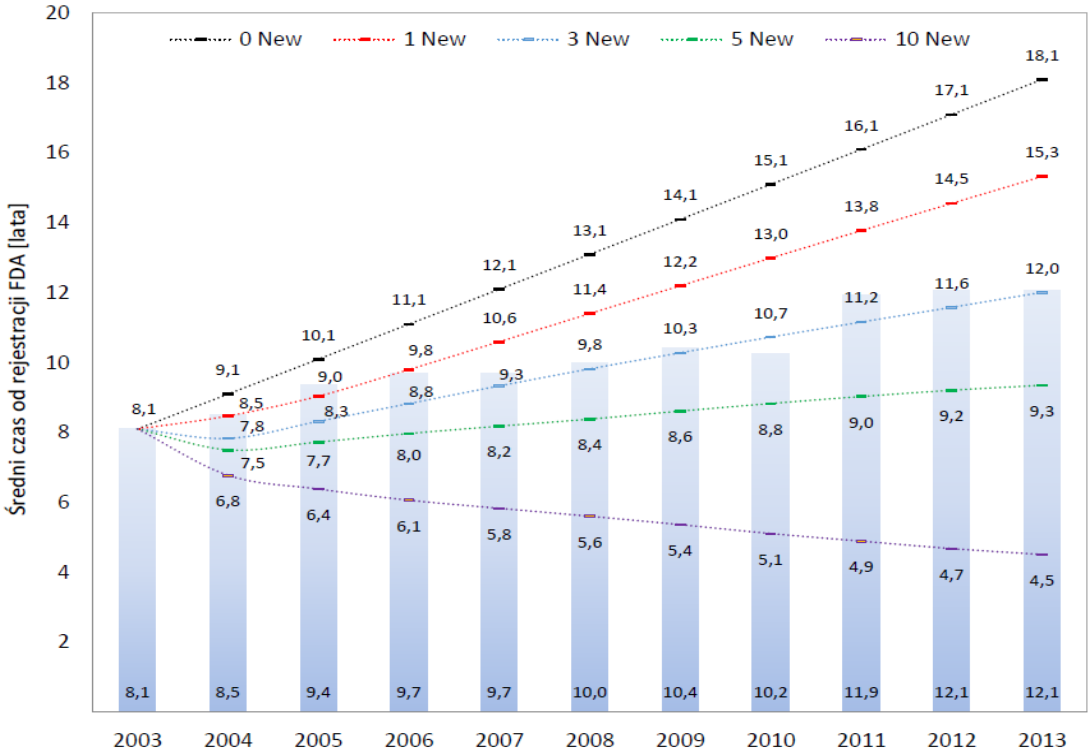


Rysunek 5. Prawo Erooma w przemyśle farmaceutycznym[11].

3.3.2. Wiek leku

Polanski i Bogacz zaproponowali, aby stagnację na rynku farmaceutycznym zilustrować wiekiem bestsellerów farmaceutycznych, a wiek ten określić za pomocą ilości lat, która upłynęła od jego rejestracji przez FDA[32].

Na rysunku 6. przedstawiono histogram ilustrujący średni wiek leku dla bestsellerów z listy top100. Analiza wykazuje, że leki się starzeją.



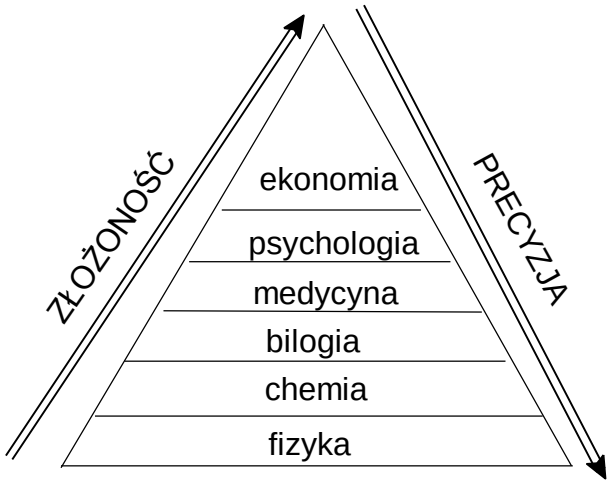
Rysunek 6. Średni czas od rejestracji FDA w latach - wiek dla leków z listy top 100 w latach 2003 - 2013. Linie ilustrują odpowiednio, jak zmieniałby się wiek leku, gdyby corocznie lista była uzupełniana odpowiednią liczbą nowych leków[32].

4. Analizy big data w chemii i projektowaniu leków

Analiza wielkich danych (big data) stanowi nowe podejście o potencjalnym zastosowaniu w projektowaniu leków. Na przykład Slezak i in. twierdzą, że wykorzystanie big data w amerykańskim systemie opieki zdrowotnej pozwoliłoby na zredukowanie rocznych kosztów związanych z jego funkcjonowaniem o 300 mld dolarów[33].

Co ciekawe, analiza danych typu big data nie jest szeroko rozpowszechniona w chemii i projektowaniu leków. Ten typ analiz spotyka się z szerszym zainteresowaniem w naukach społecznych, m.in. w psychologii oraz ekonomii Te dziedziny wiedzy przetwarzają znacznie większą ilość danych. Polański zaproponował, aby fakt ten wytłumaczyć korzystając z porównania liczby poznanych dotychczas związków chemicznych, czyli ok. 150 milionów

z liczbą ludzi zamieszkałych na naszej planecie, a więc z ok. 7 miliardami[1]. Gromadzenie danych oraz tempo wzrostu informacji zależy od stopnia złożoności tej ostatniej. Chemia potrzebuje fizyki, a biologia - chemii, aby zrozumieć i wytłumaczyć podstawy metod badawczych. W tym kierunku przybierają też redukcji w znanej metodzie badawczej tzn. redukcjonizmem[1]. Używając tego typu argumentacji, psychologia potrzebuje biologii i medycyny. Największą złożoność danych obserwuje się w ekonomii, która bada materialne, psychologiczne, społeczne, poznawcze i emocjonalne efekty ludzkich zachowań. Tak więc, chociaż dane z dziedziny nauk chemicznych wydają się nam złożone, ich porównanie z innymi naukami pozwala zorientować się, że tak naprawdę to interakcje zachodzące między ludźmi tworzą prawdziwie złożone dane. Na rysunku 7. zilustrowano wzrost złożoności informacji dyscyplin naukowych[34].



Rysunek 7. Wzrost złożoności informacji dyscyplin naukowych[1].

Aby w pełni zrozumieć mechanizm wprowadzania leków i związków chemicznych na rynek farmaceutyczny, niezbędna jest znajomość ekonomicznych uwarunkowań projektowania molekularnego[34]. Nowoczesne leki powstają w wyniku tego właśnie projektowania, a koszty wejścia leku na rynek szacowane są na miliardy dolarów. Ostatecznie to rynek określa, czy lek odniesie sukces czy poniesie porażkę[35].

Big data w chemii jest generowana poprzez:

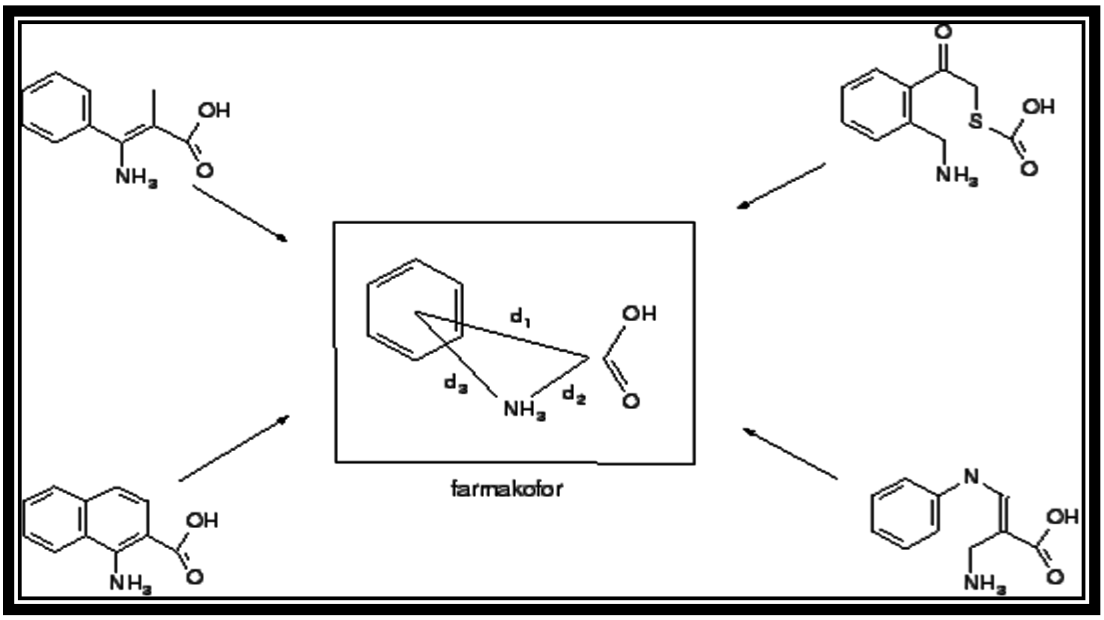
- Zwiększenie liczby deskryptorów molekularnych (DE) lub analizę wszystkich dostępnych deskryptorów.
- Zwiększenie liczby mierzonych właściwości (P) lub analizę wszystkich dostępnych właściwości[34].

4.1. Ekspansja danych w wielowymiarowych analizach QSAR

Poniżej opisałam typowe przykłady generowania dużych bibliotek danych stosowanych w analizach QSAR. Metoda sprowadza się do zwiększania (ekspansji) wymiarowości deskryptorów opisujących układy molekularne. Liczba mierzonych właściwości nie ulega przy tym zmianie.

4.4.1. Metoda COMFA

Metoda CoMFA (ang. Cofmparative Molecular Field Analysis), porównawcza analiza pola cząsteczkowego, należy do metod modelowania 3D QSAR[36]. Opiera się ona na technikach statystycznych, takich jak PCA czy PLS i opisuje ilościowe relacji pomiędzy aktywnością biologiczną, a strukturą wygenerowaną w 3D. Do modelowania metodą CoMFA niezbędne jest przygotowanie zestawu cząsteczek o jednakowym profilu aktywności biologicznej. Wybrany zestaw cząsteczek zostaje poddany optymalizacji geometrii (tzn. minimalizacji energii)[29], a następnie oblicza się cząstkowe ładunki na atomach związków, głównie za pomocą metod półempirycznych, takich jak AM1, AM3 lub metodą Gasteigera-Marsillego. Kolejnym etapem tego procesu jest odpowiednie nałożenie wszystkich analizowanych struktur 3D na jeden wybrany związek. Do nakładania cząsteczek wymagany jest wspólny motyw strukturalny. Zakłada się, że wybrany związek, na który nakładane są wszystkie cząsteczki, wykazuje najlepsze dopasowanie do miejsca receptorowego. Metoda CoMFA jest techniką AAA (ang. Active Analog Approach), która polega na badaniu serii ligandów[37,38]. Późniejszy element badań to tworzenie farmakofora, który ilustruje przestrzenne i elektronowe rozmieszczenie cech strukturalnych podstawników potrzebnych do aktywacji bądź dezaktywacji receptora.



Rysunek 8. Tworzenie farmakofora[39].

Analizowany zestaw cząsteczek umieszczany jest następnie w wirtualnej, trójwymiarowej przestrzennej siatce punków o ściśle skategoryzowanych wymiarach. Najczęściej odległości siatki wynoszą 2Å. Po superpozycji cząsteczki w węzłach siatki obliczane są wartości pól molekularnych będące numerycznym ujęciem cząsteczek w przestrzeni trójwymiarowej; z reguły obliczane są wartości pól elektrostatycznych lub sterycznych. W zależności od użytego pola molekularnego w węzłach sieci, umieszcza się konkretne sondy atomowe. Przykładowe sondy atomowe to m.in. H^+, CH_3^+, CH_3^0. Sondy te służą do obliczeń energii ich oddziaływania z danymi cząsteczkami[40].

W zależności od rozmiaru komórki, jak również od liczby obliczonych pól molekularnych, cząsteczka reprezentowana może być przez tysiące zmiennych wielowymiarowych, które są wzajemnie ze sobą skorelowane[41]. Zmienne te w metodzie CoMFA określane są jako deskryptor MIF (ang. molecular interaction field), które do obliczenia wartości potencjałów elektrostatycznego i sterycznego wykorzystują funkcje Kulomba (równanie1.6) lub Lennarda-Jonesa (równanie1.7)[3].

$$E = \sum_{i=1}^{n} \frac{q_i q_j}{D r_{ij}} \quad (1.6)$$

$$E_{vdW} = \sum_{i=1}^{n} (A_{ij} r_{ij}^{-12} - C_{ij}\, r_{ij}^{-6}) \quad (1.7)$$

gdzie:

q_i- ładunek atomu;

q_j- ładunek sondy atomowej;

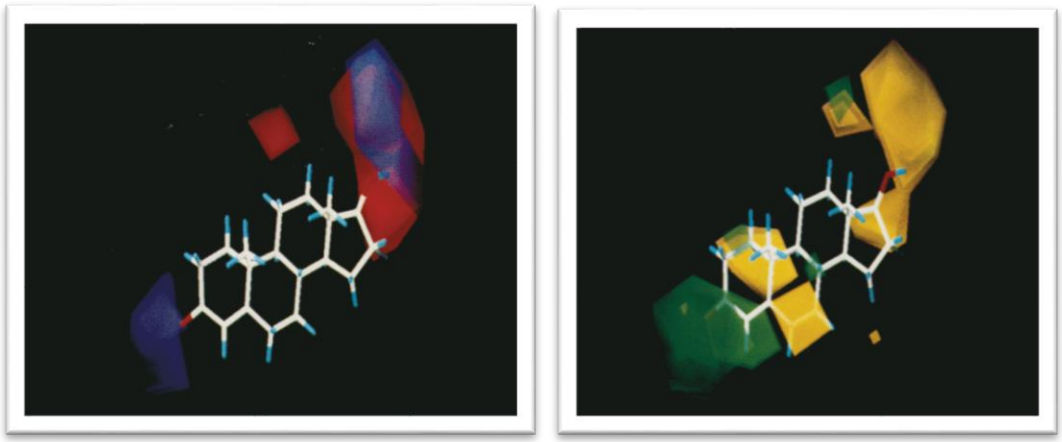
D- stała dielektryczna;

r_{ij}- odległość między atomem i a sondą j;

A_{ij},C_{ij}- współczynniki zależne od promieni van der Walsa.

Modelowanie zmiennej zależnej przy użyciu skorelowanych zmiennych niezależnych wymaga zastosowania metody PLS z walidacją krzyżową w modelowaniu CoMFA. Metoda ta jest wykorzystywana do wyznaczania równania wiążącego aktywność biologiczną z wartościami obliczonych deskryptorów (czyli wartościami potencjałów pól elektronowych lub sterycznych otaczających cząsteczkę). PLS (ang. Partial Least Squares) to metoda częściowych najmniejszych kwadratów służąca do modelowania danych wielowymiarowych. Analizę wzajemnych zależności pomiędzy badanymi obiektami (deskryptorami) ułatwia zastosowanie jednej z technik walidacyjnych, na przykład walidacji krzyżowej (ang. cross-validation), w wyniku czego dla zbioru analizowanych cząsteczek tworzona jest macierz składająca się z kilku tysięcy kolumn będących zmiennymi opisującymi deskryptory (czyli wartości niewiążących pól elektronowych lub/i sterycznych) oraz wierszy, które odpowiadają kolejnym cząsteczkom[41].

Wynikiem analizy CoMFA jest wizualizacja obliczonych modeli za pomocą przestrzennej mapy oddziaływań, która otacza cząsteczkę. Daje to możliwość identyfikacji odpowiednich obszarów wpływających korzystnie bądź niekorzystnie na wiązanie się liganda z receptorem. Interpretację wyników ułatwia kodowanie kolorów obszarów pól steryczych i elektrostatycznych, gdzie obszary zaznaczone są – odpowiednio: korzystne elektrostatycznie kolorem niebieskim, a niekorzystne kolorem czerwonym. Natomiast obszary korzystne sterycznie zaznaczone są kolorem zielonym, a kolorem żółtym obszary niekorzystne[40].



Rysunek 9. Przykładowe wyniki analizy CoMFA dla steroidów o powinowactwie TBG[40].

Wyniki analizy CoMFA w dużej mierze zależą od sposobu nałożenia cząsteczek, ponieważ nawet niewielka zmiana nałożenia jednej cząsteczki wpływa na zmianę wartości odziaływań sondy z atomami cząsteczki.

4.4.2. Metoda CoMSIA

Metoda CoMSIA (ang. Comparative Molecular Similarity Indices Analysis) porównawcza analiza cząsteczkowych indeksów podobieństwa jest rozwinięciem metody CoMFA. Metoda ta polega na porównaniu cząsteczek posiadających podobne właściwości, co umożliwia znalezienie ogólnych cech mających znaczenie w przypadku wiązania się receptorem. W metodzie CoMSIA definiowane są indeksy podobieństwa (ang. similarity indices), które są obliczane w węzłach siatki dla każdej cząsteczki. Wykorzystanie właściwych sond atomowych daje możliwość obliczenia odziaływań: elektrostatycznych, hydrofobowych, wodorowych czy sterycznych według równania (1.8)[42].

$$A_F = -\sum_{i=1}^{m}\sum_{i=1}^{n} w_{ij}\, e^{-\alpha r_{ij}} \qquad (1.8)$$

gdzie:

A_F- wartość indeksu w węźle;

m- całkowita liczba sondy atomowej;

n- całkowita liczba atomów dla konkretnej cząsteczki;

rij – odległość między atomami i a węzłem j;

w_{ik} – wartość współczynnika, który zależy od rodzaju oddziaływania;

α – współczynnik szerokości funkcji odległości.

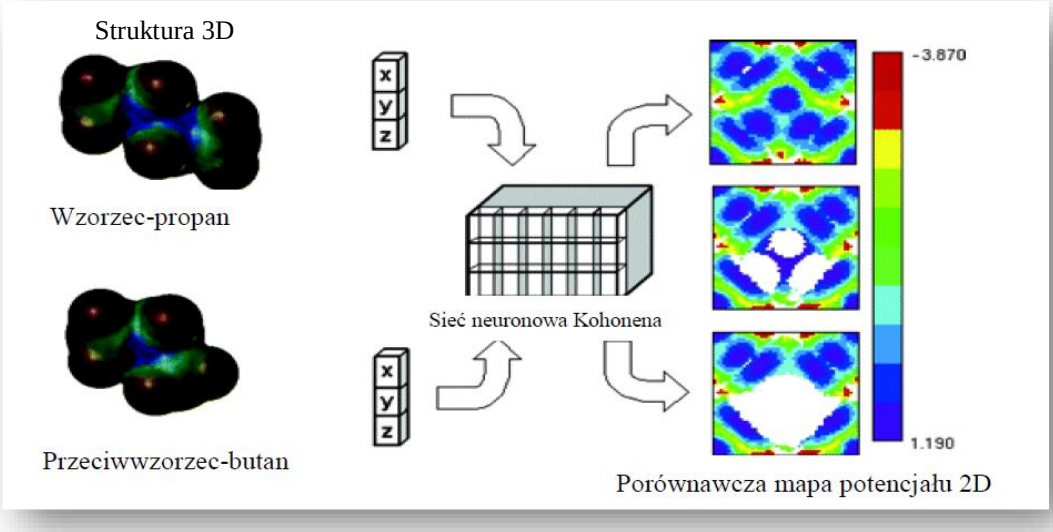
Potencjał Lenarda-Johnsa i/lub Kulomba w metodzie CoMFA zostaje zastąpiony w metodzie CoMSIA przez funkcję odległości typu Gaussa. Często poprawia to wyniki modelowania[43].

4.4.4. Metoda CoMSA

Metoda CoMSA (ang. Comparative Molecular Surface Analysis) - porównawcza analiza powierzchni cząsteczkowej[24] wykorzystująca technikę samoorganizującej się sieci neuronowej Kohonena SOM (ang. Self Organizing Maps) oraz analizę PLS opracowana została w zespole Polańskiego[44,45,46,47,48]. Wynikiem analizy jest określenie ilościowej zależności pomiędzy kształtem powierzchni cząsteczki a aktywnością biologiczną.

W metodzie CoMSA za istotną cechę badanych związków chemicznych uznano geometrię będącą sumą współrzędnych powierzchni van der Waalsa. W celu porównania powierzchni cząsteczkowej analizowanego zestawu związków, należy za pomocą sieci neuronowej przekształcić trójwymiarowe powierzchnie cząsteczkowe w dwuwymiarowe mapy porównawcze (ang. comparative maps). Mapy te reprezentują potencjał elektrostatyczny cząsteczki, dlatego ważne jest, aby topologia obiektu reprezentowana przez sygnały wejściowe, które są pobierane z powierzchni cząsteczki została zachowana. Następnie trójwymiarowe struktury zostają nałożone na wzorzec będący wspólnym motywem strukturalnym dla zbioru badanych cząsteczek. Istotnym etapem w analizie CoMSA jest tzw. uczenie sieci neuronowej z wykorzystaniem powierzchni cząsteczki wzorcowej w taki sposób, aby mogła on przechowywać informację o geometrii cząsteczki wzorcowej. Do sieci neuronowej wprowadza się współrzędne punktów d(x,y,z), które zostały wcześniej pobrane z powierzchni cząsteczki wzorcowej. Co ważne, każdej współrzędnej przypisane są określone wartości potencjału elektrostatycznego. Do tak wytrenowanej sieci

wprowadzane są kolejno kolejne współrzędne punktów z powierzchni innych cząsteczek analizowanego zestawu będące cząsteczkami przeciwwzorca. Wytrenowanie sieci umożliwia analizę podobieństwa powierzchni zbioru cząsteczek. Końcowym etapem jest przekształcenie sieci neuronowej w zestaw dwuwymiarowych map porównawczych powierzchni cząsteczkowych. Otrzymane mapy porównawcze są poddane modelowaniu PLS[24,45,46,47,48].



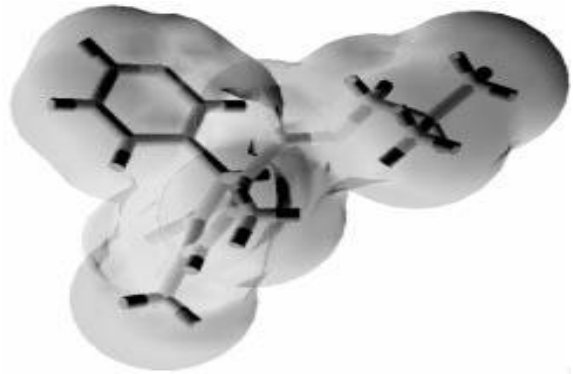
Rysunek 10. Porównawcza analiza powierzchni cząsteczkowej z zastosowaniem techniki samoorganizującej się sieci neuronowej Kohonena[24].

Przykład takiej analizy przedstawiono na rysunku 10. Zastosowane w metodzie CoMSA deskryptory porównawcze powierzchni cząsteczkowej dla cząsteczki propanu są wzorcem dla cząsteczki butanu będącej przeciwwzorcem. Wynikiem CoMSA są obrazy map potencjału elektrostatycznego cząsteczki otrzymane techniką sieci neuronowej Kohonena (Rysunek 10.). Kolory w mapach neuronowych kodują uśrednioną wartość potencjału i są przyporządkowane do konkretnego neuronu w mapie. Białe obszary opisują puste neurony, które w czasie analizy nie otrzymały sygnału z powierzchni cząsteczkowej przeciwwzorca. Określa to niezgodność topologiczna dla analizowanego zestawu cząsteczek[24].

4.4.5. Metoda CoRSA

Metoda CoRSA (ang. Comparative Receptor Surface Analysis) - porównawcza analiza powierzchni receptora, opierająca się na porównaniu obrazów powierzchni analizowanego zestawu cząsteczek. Analizę rozpoczyna się analogicznie jak w przypadku metod 3D QSAR od optymalizacji geometrii (minimalizacji energii) wybranego zestawu cząsteczek. W kolejnym etapie dokonuje się selekcji zbioru cząsteczek poprzez wybór od jednej do pięciu cząsteczek o najwyższej aktywności (hipoteza aktywnego analogu). Zakłada się tym samym, iż wybrane cząsteczki charakteryzują najlepiej geometrię miejsca receptorowego; z wybranego zestawu cząsteczek tworzony jest wirtualny receptor RGS (ang. Receptor Generation Set). W metodzie CoRSA istotne jest założenie, że tworzony obraz pseudoreceptora jest zbliżony do obrazu receptora rzeczywistego[23].

Do wygenerowania powierzchni wirtualnego receptora (pseudoreceptora) w metodzie CoRSA stosuje się na przykład algorytm Hahna-Rogersa[50]. Obraz pseudoreceptora opisany jest przez zestaw punktów próbkowych z jego powierzchni. Następnie w każdym z tych punktów zostaje obliczony szereg wartości takich jak: hydrofobowość, potencjał elektrostatyczny czy ładunek cząstkowy. W kolejnym etapie oblicza się we wszystkich punktach wygenerowanego pseudoreceptora energię oddziaływania dla każdej analizowanej cząsteczki. Wynikiem takiego modelowania jest wektor opisujący oddziaływanie cząsteczki z pseudoreceptorem. Otrzymane dane poddane są analizie PLS[49,50].



Rysunek 11. Model powierzchni wirtualnego receptora symulowany w metodzie CoRSA[23].

4.2. Analizy dużych populacji danych w chemii i projektowaniu leków

Jedną z takich analiz jest analiza wszystkich dostępnych właściwości. Klasycznym przykładem są analizy aktywności biologicznej i mas cząsteczkowych dla wszystkich leków i kandydatów leków (drug candidates). Wyniki takich analiz prowadzą do następujących wniosków:

- Leki dostępne na rynku farmaceutycznym mają zazwyczaj niższe MW niż potencjalne leki; obserwacja ta jest fundamentem koncepcji otyłości molekularnej oraz szczupłej farmacji (*molecular obesity* oraz *slim pharma*)[51,52].
- Nie istnieje zależność pomiędzy IC_{50} leków a ich dawką terapeutyczną[53].
- Większość leków komercyjnych nowych klas jest utworzona przez skrining fenotypowy (nowe klasy) bądź przez modele oparte na strukturze receptora[54].
- Aktywność biologiczna leków jest zbliżona do rozkładu normalnego[55].

Inną metodą zwiększania populacji wielowymiarowych danych wykorzystywaną w chemii jest zwiększenie liczby zmiennych reprezentujących właściwości, np. dzięki zastosowaniu koncepcji polifarmakologii. Jednym z ciekawszych rozwiązań tego typu jest metoda zespołu Gabriele Cruciani inspirowana polifarmakologią metoda tzw. lipidografii (*lipidomic scheme*)[53]. Metoda ta opiera się na niedrogiej i szybkiej ekstrakcji bibliotek lipidów oraz ich badaniu metodą spektrometrii mas MS (ang. Mass Spectrometry), w odniesieniu do średniej wielkości populacji lipidów opisujących reakcję organizmu na podanie leku (potencjalnego leku). W ciągu 20 minut otrzymujemy w ten sposób informację o stężeniu 1000 lipidów w surowicy krwi. Tego typu dane mogą zostać wykorzystane jako wielowymiarowy profil lipidu typu daktylogramu (*fingerprint-like)*, który badany jako zmienna w czasie, pozwala na opisanie aktywności i toksyczności użytych ksenobiotyków[56].

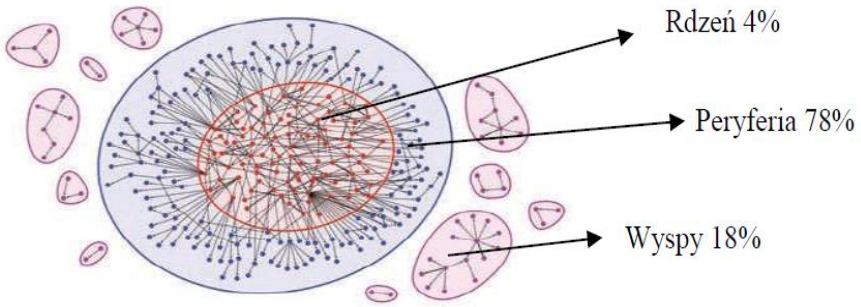
Podsumowując, duże zbiory danych w projektowaniu leków wymagają dużej liczby mierzonych właściwości. Niestety w praktyce dostępność danych jest wciąż ograniczona. Oznacza to, że populacja danych P zazwyczaj zwiększa się nie przez dodanie nowej właściwości P, lecz

przez zwiększenie liczby związków chemicznych opisywanych daną właściwością lub obliczalnych deskryptorów molekularnych. A zatem efektywność badań w zakresie projektowania leków warunkowana jest nie wynikami badań statystycznych, lecz opracowaniem i zastosowaniem nowych niedrogich metod pomiaru właściwości[2]. Tak więc wdrożenie nowych metod HTS (ang. high throughput screening*)* pomiaru właściwości może skutkować uzyskaniem nowej jakości w analizach big data w procesie projektowania leków.

5. Badania architektury chemii organicznej

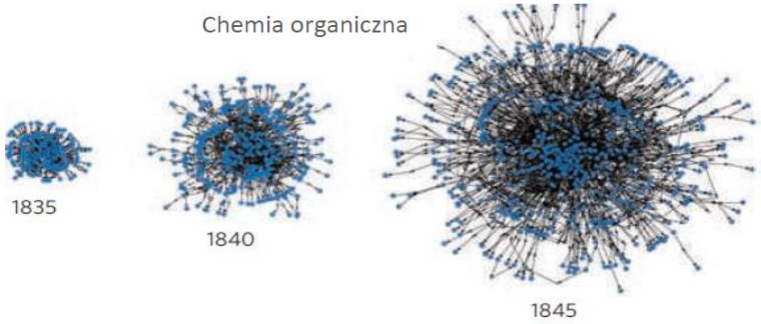
Istotnym typem analiz danych są metody sieciowe[57]. Metody tego typu np. do opisu związków organicznych i ich właściwości wykorzystują grafy. Kwestią istotną jest sposób przedstawienia złożonych zbiorów molekularnych oraz relacje, w jakie związki takie wchodzą. Ciekawą metodą ilustracji wszystkich związków organicznych oraz ich reakcji jest sieć nazwana *universe world of organic chemistry.* Sieć taką tworzą cząsteczki i reakcje, jakie zachodzą pomiędzy nimi. Badania struktury chemii organicznej, w szczególności prowadzone za pomocą tej sieci, zmierzają do identyfikacji cząsteczek organicznych, które cieszą się największym zainteresowaniem, będąc najbardziej przydatnymi surowcami lub produktami syntez. Wiedzę tę można wykorzystać do projektowania nowych syntez i otrzymywania nowych związków chemicznych[57].

Na rysunku 12. zilustrowano topologię chemii organicznej reprezentowanej przez taką sieć, reprezentowaną reaktywnością związków chemicznych. W jej centrum, rdzeniu występują najpopularniejsze cząsteczki, na peryferiach zaś znajdują się związki stanowiące wciąż wyzwanie dla syntezy. Są to cząsteczki, które mogą być syntetyzowane, wychodząc z powierzchni rdzenia. Natomiast na wyspach lokują się izolowane zbiory cząsteczek o wyspecjalizowanej (egzotycznej) budowie. Są one złożonymi produktami lub indywidualnymi klonami produktów naturalnych lub innymi substancjami, na przykład izotopami[57].



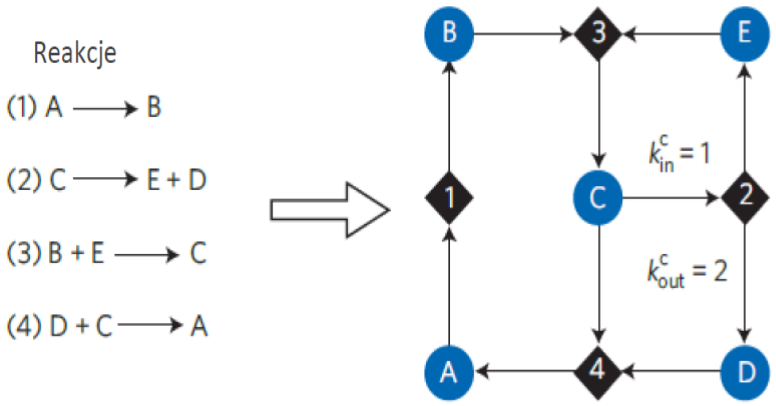
Rysunek 12. Główne elementy topologii sieci[57].

Na rysunku 13. zilustrowano cząsteczki organiczne połączone siecią reakcji oraz jej wzrost na przestrzeni lat. Analizując ten rysunek, można dojść do wniosku, że od 1835 do 1845 roku sieć zwiększyła swoje rozmiary 10000 razy. Wynika to z rosnącej liczby nowych syntez chemicznych.



Rysunek 13. Wzrost liczby związków organicznych połączonych siecią reakcji[57].

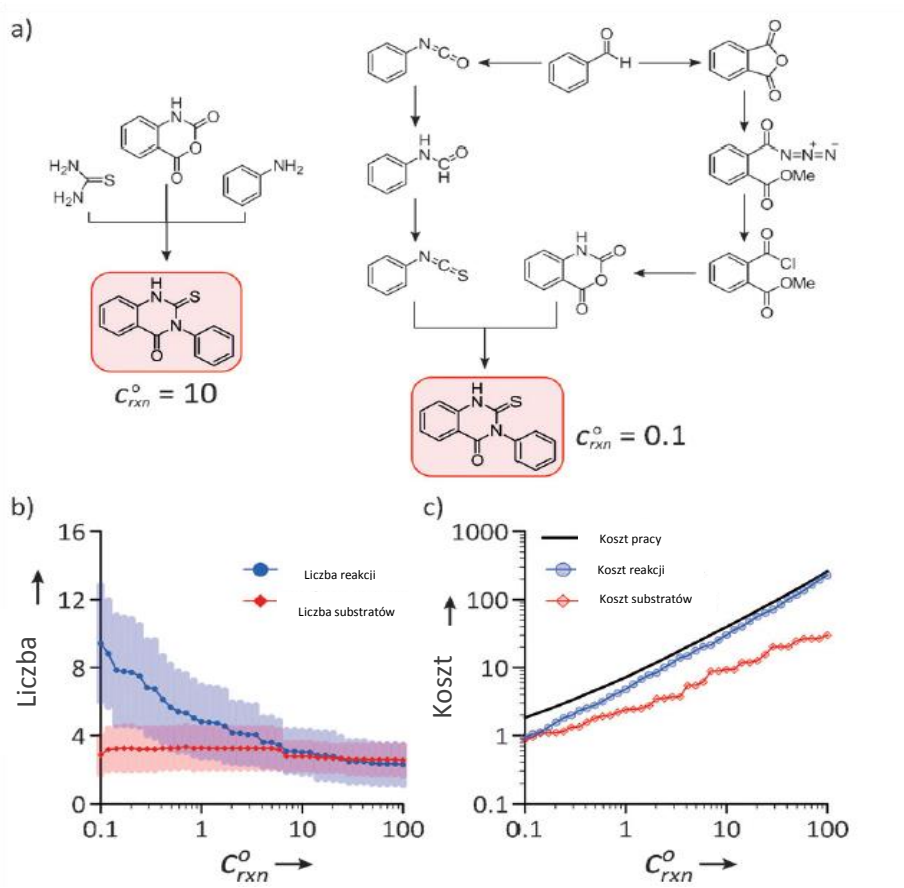
Na rysunku 14. zilustrowano inną przykładową sieć reakcji związków chemicznych, gdzie związkom chemicznym przypisano litery od A do D, a reakcjom chemicznym liczby od (1) do (4). Łączność każdej substancji chemicznej jest opisana przez liczbę wejściowych k_{in}^{c} i wyjściowych k_{out}^{c} strzałek, które są połączone reakcjami chemicznymi. Substancje chemiczne przedstawione są jako niebieskie okrągłe węzły, a reakcje chemiczne - jako czarne kwadraty. Plan ten przedstawia wszystkie niezbędne szlaki i związki syntetyczne, w których każda reakcja może mieć wiele reagentów lub produktów[57,58].



Rysunek 14. Sieci reakcji związków chemicznych[57].

5.1. Znaczenie ekonomii w syntezie organicznej

Analiza architektury sieci chemii organicznej prowadzi do poszukiwania aktywnych związków i optymalizacji struktury wiodącej w kontekście projektowania leków, wykorzystując jako przykład algorytm równoległej optymalizacji szlaków syntetycznych (ang. parallel optimization of synthetic pathways) stworzony przez grupę prof. Kyle J.M Bishopa z Columbia University. Algorytm ten szybko przetwarza syntetyczny plan syntezy z uwzględnieniem minimalnego kosztu wytwarzania. Jego działanie składa się z dwóch etapów. W pierwszym etapie algorytm ten analizuje wszystkie reakcje prowadzone do otrzymania produktu. Drugim etapem jest obliczenie minimalnego kosztu, przy czym jego obliczenie zależy od minimalnych kosztów wyszukanych substratów, które mogą być zakupione lub syntetyzowane[59]. Omawiany syntetyczny plan został wykorzystany przez firmę ProChima sprzedającą substancje, które zostały użyte do planowania syntezy, co przedstawia rysunek 15.



Rysunek 15. Rożne plany syntezy dihydrochinazoliny z uwzględnieniem kosztów[59].

gdzie:

a) dwa różne optymalne plany syntezy *dihydrochinazoliny*;

b) liczba reakcji i substratów użyta w planie optymalizacji syntezy z uwzględnieniem kosztu reakcji;

c) koszty: reakcji, substancji i pracy przy tworzeniu planu optymalizacji syntezy z uwzględnieniem kosztu reakcji[59].

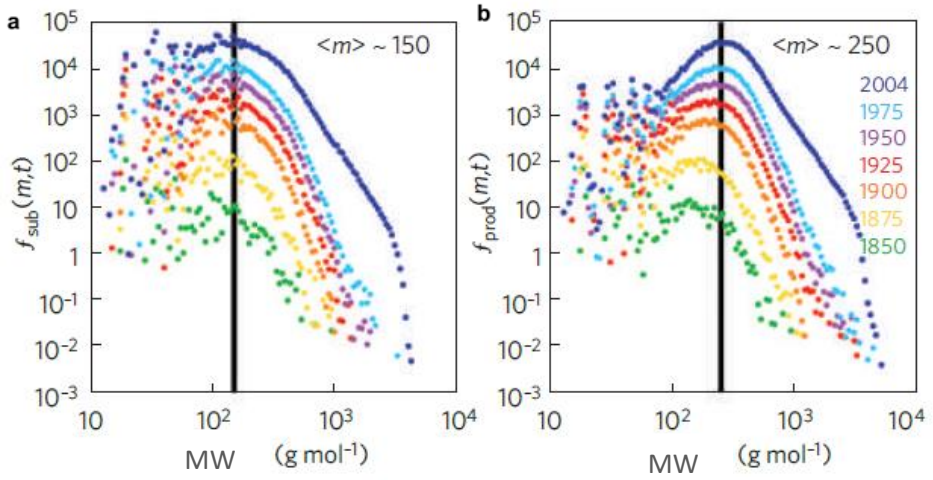
Na rysunku 15. przedstawiono plany syntezy dihydrochinazoliny z uwzględnieniem kosztów. Rysunek 15a ilustruje dwa optymalne sposoby syntezy dihydrochinazoliny (pochodnej naturalnego produktu), gdzie zmiana paramentów kosztów prowadzi do różnych optymalnych syntez otrzymania tego samego produktu. Koszt pierwszej proponowanej syntezy wynosi c^{0}_{rxn}=10, natomiast koszt drugiej proponowanej syntezy z różnych substratów wynosi c^{0}_{rxn}=0.1.

Rysunek 15b przedstawia liczbę syntez oraz liczbę substratów użytych w planie syntezy w stosunku do rosnącego kosztu reakcji c^0_{rxn}. Z interpretacji wykresu wynika, że dla 51 różnych substancji optymalne syntezy zmniejszają się wraz z rosnącym kosztem reakcji c^0_{rxn}. Rysunek 15c ilustruje koszty: reakcji, substancji i pracy w stosunku do kosztu reakcji c^0_{rxn}. Z opisu wykresu wynika, że wraz ze zmniejszającym się kosztem syntezy użyte substancje są tańsze, natomiast drogi syntetyczne stają się droższe lub dłuższe[59]. Innym opisanym w literaturze algorytmem z wykorzystaniem sieci chemicznej jest algorytm identyfikacji ewentualnych ścieżek reakcji jednogarnkowych (ang. one-pot reactions), który na podstawie znanej informacji nt. substancji syntetycznych szacuje, czy przewidywane ścieżki syntezy powinny rzeczywiście sprawdzić się w rzeczywistym eksperymencie, dając odpowiednio wysokie wydajności[60].

5.2. Masy cząsteczkowe MW a inne deskryptory molekularne

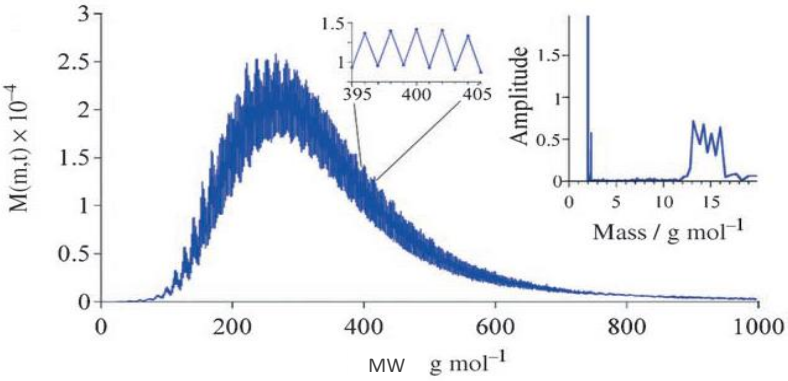
Warto podkreślić, że informacja o masach cząsteczek (MW) jest prostym parametrem opisującym jej ewolucję, ponieważ masa jest najprostszym z możliwych deskryptorów molekularnych, który można łatwo obliczyć. Informacja o niej jest także łatwo dostępna w bazach danych.

Na rysunku 16. przedstawiono rozkłady częstotliwości występowania mas cząsteczkowych w bazie Beilstein. Baza danych Beilstein (BD), która jest największym repozytorium reakcji organicznych zawierała (do kwietnia 2004 r.) listę 9550398 substancji chemicznych i 9293250 reakcji, w których te substancje biorą udział. Analiza rozkładu mas cząsteczkowych prowadzi do kilku interesujących wniosków, m. in. do wniosku, że pomimo znacznego postępu w metodologii syntez maksima MW substratu i produktu wynoszą odpowiednio MW_{sub}= 150 Da, MW_{prod}= 250 Da. Co ciekawe, kształty obu rozkładów krzywych i ich maksima nie zmieniają się w czasie, lecz tylko przesuwają się odpowiednio ku górze[57].



Rysunek 16. Rozkład częstotliwości mas dla bazy Beilstein, które były wykorzystywane jako (a) substraty (b) produkty w reakcjach zgłoszonych między 1850 a 2004 r.[57, 61].

Na rysunku 17. zilustrowano histogram rozkładu mas cząsteczkowych dla bazy Beilstein. Zilustrowany rozkład mas M (m, t=2004) o wysokiej częstotliwości mieści się w przedziale od ok. 250 do 300 Da. Przedstawione na wykresie powiększenie pokazuje widmo Fouriera obszaru dominującego. Ostre maksimum występuje w zakresie ok. 2, natomiast reszta lokalnych maksimów skupiona jest w przedziale 14-15 wskazując, że 48% cząsteczek odpowiada masom z powszechnie występujących bloków budulcowych (ang. building blocks), z których złożone są cząsteczki[61].



Rysunek 17. Częstotliwość rozkładu mas cząsteczkowych w bazie Beistein[61].

5.3. Ekonomia atomowa

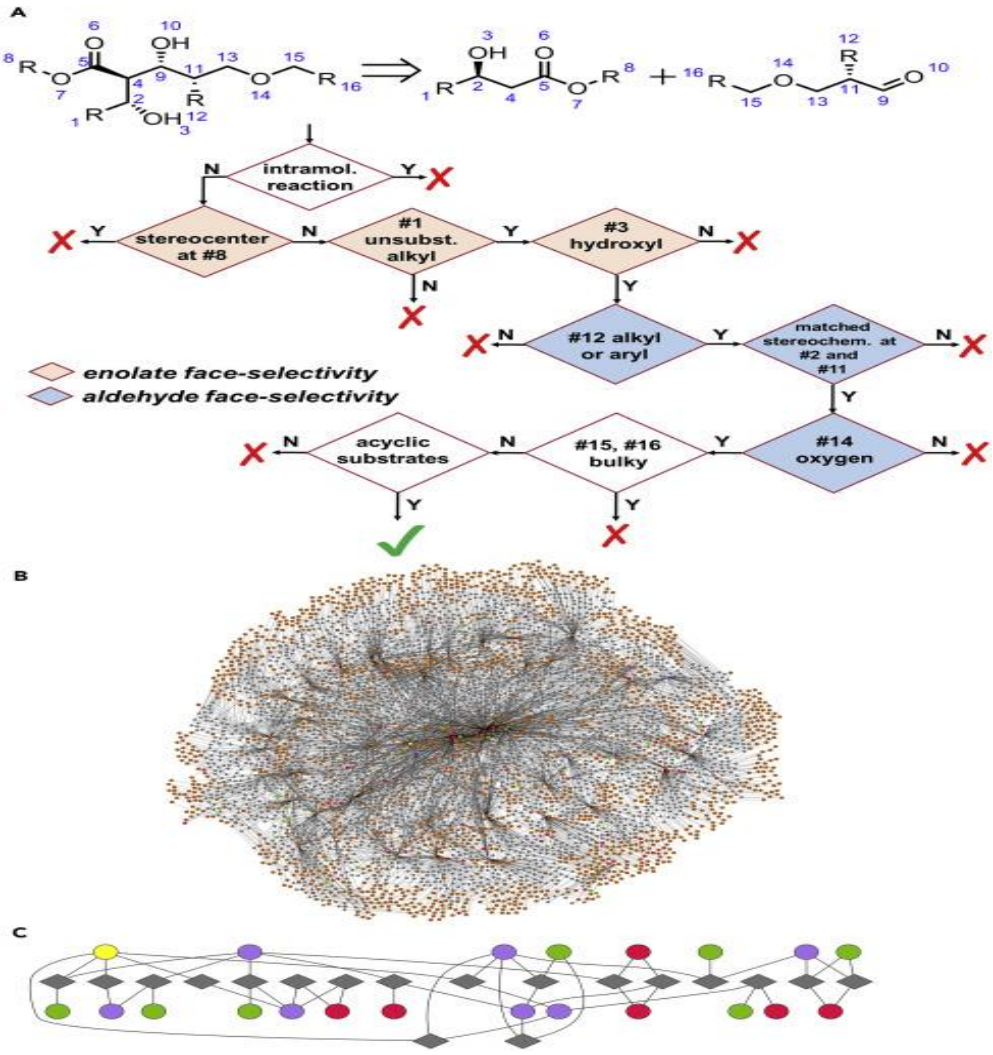
Ekonomia atomowa stanowi stosunkowo nowe pojęcie związane z planowaniem syntez chemicznych. Pozwala zrozumieć, że wydajność reakcji chemicznej nie jest jedynym parametrem decydującym o efektywności syntezy. Istotnymi parametrami jest stosunek atomów, które wchodzą w skład interesującego nas produktu, oraz produktów (odpadów), które są balastem[62]. W klasycznym znaczeniu ekonomia atomowa ma więc znaczenie czysto chemiczne. W rozszerzonym znaczeniu może jednak także uwzględniać czynnik ekonomiczny. Stworzenie programu komputerowego umożliwiającego planowanie ścieżek syntez chemicznych było do niedawna jednym z największych wyzwań współczesnej chemii organicznej/obliczeniowej. Pomimo przeprowadzenia wielu badań i wielu prób, nie odnotowano w literaturze przedmiotu informacji o zaprojektowaniu kompletnych ścieżek przez komputer, które następnie mogłoby zostać powtórzone w laboratoriach. Istniejące programy komputerowe posiadały ograniczoną wiedzę o przemianach chemicznych oraz odzwierciedlały brak strategii określającej, w jaki sposób należy dokonać połączenia poszczególnych kroków, aby stworzyć ekonomicznie optymalną ścieżkę syntezy[63]. Badania nad sieciami chemicznymi[57,59,60] doprowadziły do odkrycia nowych modułów retrosyntetycznych de novo w ramach oprogramowania Chematica, które wykorzystuje algorytmy i zbiorczą wewnętrzną bazę danych. Baza ta zawiera informacje z zakresu nauk chemicznych, zgromadzone w ciągu 250 lat ich istnienia. Oprogramowanie to łączy teorię sieci, nowoczesne obliczenia dużej mocy, sztuczną inteligencję oraz specjalistyczną wiedzę chemiczną. Pozwala projektować syntetyczne ścieżki prowadzące do wcześniej zsyntetyzowanych lub nowych celów syntezy oraz łączyć długie ścieżki syntezy w krótsze i bardziej ekonomiczne[63]. Rozwój oprogramowania prowadzony był przez Bartosza A. Grzybowskiego i został opublikowany w sierpniu 2012 roku. W 2017 r. oprogramowanie i baza danych zostały licencjonowane firmie Merck[64].

Pierwszym etapem w reprezentacji związków chemicznych in silico było zastosowanie odpowiedniego formatu danych, który byłby zrozumiały dla komputera i definiowałby cząsteczki oraz reakcje chemiczne. Drugim etapem była odpowiednia reprezentacja reakcji organicznej. W Chematica molekułę można określić na kilka sposobów, w tym przez

przeszukanie według numeru rejestru Beilsteina, numeru rejestru CAS, nazwy chemicznej, struktury SMILES/SMART lub poprzez narysowanie samego diagramu molekularnego. Chematica przeprowadza optymalizację reakcji według kosztów, wykorzystując funkcję RSF (ang. Reaction Scoring), która oblicza dla każdego węzła koszt reakcji oraz trudność wykonania (analizując koszt substratów, trudne etapy reakcji, dużą liczbę zabezpieczeń). Rozmiary węzłów można skalować masą cząsteczkową, występowanie produktu i występowanie reagentów. Program obsługuje również modelowanie 3D poszczególnych cząsteczek, a także etykietowanie grup funkcjonalnych[65].

Podstawą każdej z 50 000 reguł, którym posługuje się program Chemiatica jest drzewo decyzyjne. Reguły te, wykorzystujące retrosyntezę, są zasadami opisującymi różne typy reakcji. Reguły w obrębie drzewa określają zakres dopuszczalnych podstawników lub typów reakcji[63].

Na rysunku 18. przedstawiono przykładowe drzewo decyzyjne dla jednej z 50 tysięcy reakcji zaprojektowanych przez Chematika dla podwójnej stereoróżnicującej kondensacji estrów z aldehydami.



Rysunek 18. Przykładowe drzewo decyzyjne dla podwójnej stereoróżnicującej - kondensacji estrów z aldehydami[63].

gdzie:

A) Reguły w drzewie decyzyjnym dla każdej rozpatrywanej reakcji/substancji w każdym kroku uwzględniają:

- zakres dopuszczalnych oraz możliwych atomów lub podstawników,
- efekty elektronowe i sterczyne cząsteczki,
- informację na temat grup zabezpieczających,
- informację na temat warunków reakcji,

- informację na temat selektywności substancji, czyli zdolności reagowania w określonych warunkach z określoną grupą związków,
- przykłady podobnych reakcji opisanych w literaturze.

B) Reguły reakcji, z których budowane są ścieżki syntetyczne wykorzystują inteligentne algorytmy do wyszukiwania najbardziej wydajnych sekwencji reakcji. Zaimplementowane funkcje scoringowe oceniają zestawy substratów i sekwencje reakcji, które zostały użyte do osiągniecia określonego etapu.

C) Algorytm konstruuje i analizuje tysiące sieci, a następnie wyodrębnia z reprezentacji sieci tylko możliwe trasy syntezy[63].

Podpunkt A. Reguły, które są pokazane w podpunkcie A są wykorzystywane do badania liczby możliwych syntez. Przedstawiony obraz dotyczy wczesnego etapu planowania syntezy. Drzewo decyzyjne rozpoczyna się od stanu reakcji międzycząsteczkowej. Aby zapewnić selektywność enolanu, rozważne są odpowiednie warunki dla reakcji dla podstawników w pozycjach #8, #1, #3. Aby zapewnić selektywność aldehydu, rozważane są warunki dla podstawników #12, #2, #1, #14. Warunki dla podstawników w pozycjach #15, #16 są wspólne dla obu substratów, co zapewnia pożądaną diastereoselektywność.

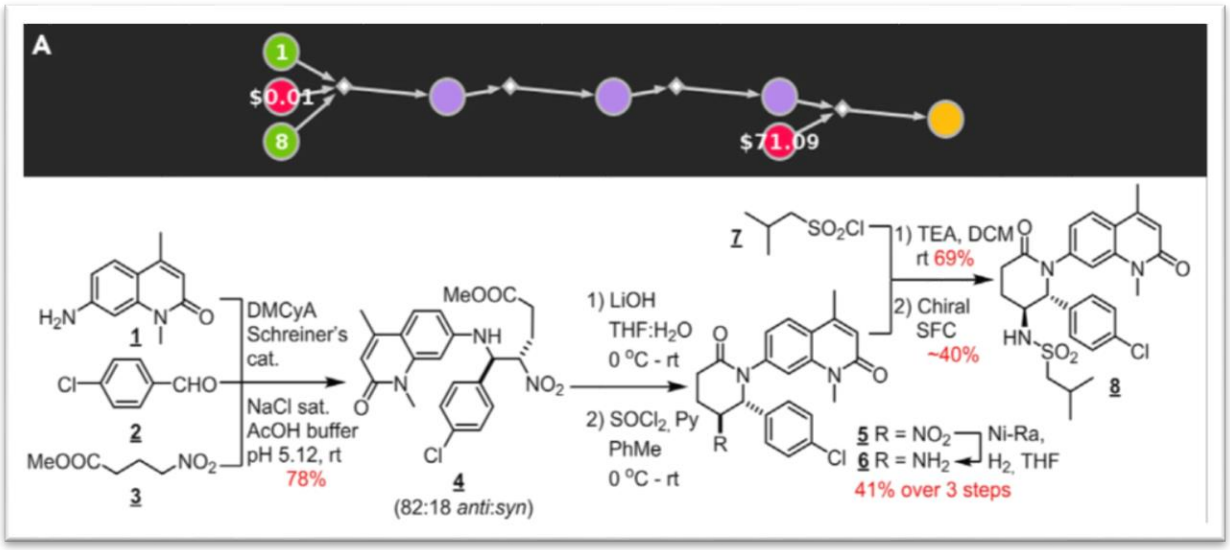
Podpunkt B. Reguły, które są opisane w podpunkcie B są wykorzystywane do konstrukcji i analizy sieci. Wybierane są tylko bardziej prawdopodobne ścieżki syntetyczne oraz dostępne handlowo (czerwone węzły) lub znane substancje (zielone węzły). Reguły reakcji są jedynie podstawowymi operacjami, z których mają zastać zbudowane kompletne ścieżki syntetyczne. Ponieważ liczba wyboru w każdym kroku retrosyntezy wynosi ~100, do przeszukania tak dużej przestrzeni potencjalnych syntez niezbędne są algorytmy inteligentne, które przeszukują ekonomicznie najwydajniejsze sekwencje kroków.

Podpunkt C. Reguły, które są pokazane w podpunkcie C są wykorzystywane do znalezienia najbardziej wydajnych ekonomicznie syntez. Następnie program wyodrębnia je z wewnętrznej syntetycznej reprezentacji sieci i wyświetla rzeczywiste ścieżki syntezy[63].

Algorytmy i metody opisane powyżej wykorzystano do zaprojektowania syntez ośmiu celi zróżnicowanych strukturalnie. Syntezy te zostały zaprojektowane po raz pierwszy przy użyciu oprogramowania Chematica, bez nadzoru chemika, gdzie głównym celem było zaprojektowanie ścieżek najbardziej opłacalnych pod względem kosztów. Następnie wszystkie syntezy zostały wykonane w laboratorium i zakończyły się sukcesem. Dla każdego celu wybrano ścieżkę syntezy ocenianą najwyżej przez system informatyczny.

- ✓ Pierwsze sześć celi zostało dostarczonych przez firmę MilliporeSigma-MS (dawniej SigmaAldrich). Wszystkie były biologicznie czynnymi związkami o wysokiej wartości handlowej (>100 $/mg), dla których poprzednie liczne próby planowania syntez przez MilliporeSigam -MS nie powiodły się.
- ✓ Siódmym był lek halucynogenny dronedaron firmy Sanofi-Aventis. Wybór ten został dokonany przez zespół Grzybowskiego, ponieważ synteza tego leku jest chroniona przez liczne patenty.
- ✓ Ósmym był naturalny produkt engelheptanoxid C, który został niedawno wyizolowany, ale nie został jeszcze zsyntetyzowany. Wybór ten został dokonany przez zespół Mrksicha w celu zaprojektowania syntezy dla produktu naturalnego.

Na rysunku 19. zilustrowano kompletną ścieżkę syntezy zaplanowaną przez oprogramowanie Chematica dla Inhibitora BRD 7/9. Synteza ta została następnie wykonana w laboratorium przez zespół chemików firmy MilliporeSigam-MS.



Rysunek 19. Ścieżka syntezy inhibitora BRD 7/9 zaprojektowana przez oprogramowanie Chematica[63].

gdzie w podpunkcie **A** kolory węzłów oznaczają kolejno:

- ✓ czerwone węzły - komercyjne związki chemiczne ceny podano w [$/g];
- ✓ zielone węzły – znane substancje (cyfra oznacza popularność substancji w syntezie);
- ✓ fioletowe węzły – nieznane substancje;
- ✓ żółte węzły - produkt syntezy.

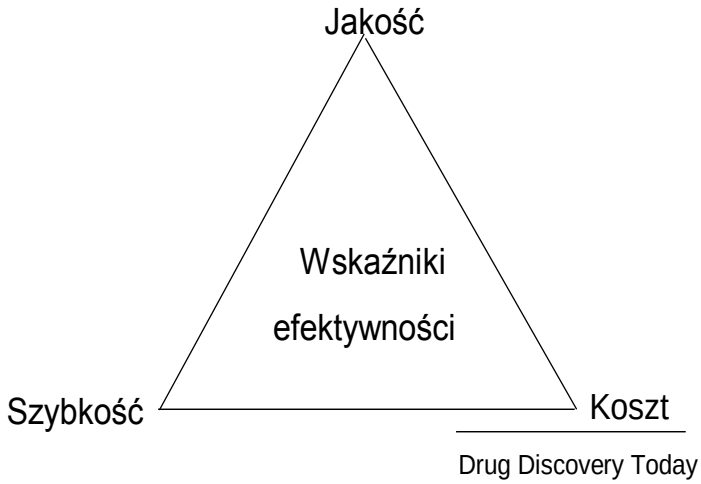
Inhibitor BRD 7/9 jest to niedawno odkryty silny, selektywny inhibitor białek zawierających bromodomenę BRD7 i BRD9 zaangażowanych w rozwój nowotworów[66]. W literaturze próba syntezy tego inhibitora opisana jest jako 8-etapowa ścieżka o niskiej wydajności. Ponadto synteza ta wymaga zastosowania chromatografii kolumnowej typu (FCC) aż w 7 etapach. Oprogramowanie Chematica pozwoliło na zaprojektowanie krótszej drogi rozpoczynającej się od trójskładnikowej reakcji aza-Henry'ego. Synteza taka umożliwiła poprawę wydajności oraz skrócenie drogi syntezy[63].

6. Strategie wpływające na decyzje w odkrywaniu nowych leków

Głównym celem działania firm farmaceutycznych jest odkrywanie i wprowadzanie na rynek nowych leków, które okażą się skuteczne i bezpieczne, a tym samym będą wpływać pozytywnie na jakość życia ludzi. Z danych zatwierdzonych przez Agencję Żywności i Leków FDA wynika, że liczba powstałych nowych leków jest stała od 10 lat[67], dlatego firmy farmaceutyczne inwestują w narzędzia poprawiające efektywność procesu projektowania i/lub poszukiwania nowych leków. Na rysunku 20. przestawiono przegląd procesu wykrywania i rozwoju leku. Proces ten składa się z kilku kroków, które są punktami decyzyjnymi lub działaniami i które można powtórzyć; ponadto każda faza procesu może składać się z kilku równoległych i/lub kolejnych podprocesów.

Rysunek 20. Proces poszukiwania nowych (research) leków (kolor fioletowy) oraz ich zaawansowanego testowania (development) (kolor różowy)[68].

Na rysunku 21. zilustrowano trzy główne wskaźniki efektywności prowadzące do sukcesu i rentowności, takie jak: szybkość, jakość i koszty czyniące badania i rozwój nowych leków bardziej wydajnymi. Z analizy rysunku 21. można wywnioskować, że niższe koszty związane z rozwojem nowego leku mogą doprowadzić do obniżenia jakości nowych leków i rentowności ekonomicznej.



Rysunek 21. Wskaźniki efektywności prowadzące do sukcesu i rentowności oraz wpływające na rozwój nowych leków to: szybkość, jakość i koszt[68].

Do dominujących strategii/optymalizacji wpływających na podjęcie właściwych decyzji w zakresie badań nad nowymi lekami należą m.in:

- Wybór celu chemicznego (białka docelowego) farmakologicznego – jest to kluczowy krok w procesie podejmowania decyzji służących odkrywaniu leków[69].
- Próba osiągnięcia akceptowalnego profilu procesów farmakokinetycznych (ADME) poprzez optymalizację właściwości fizycznych i chemicznych, które mogą obniżać lub zwiększać siłę działania dziennej dawki leku, ponieważ wykazano, że wysokie dawki leków są bardziej toksyczne od niskich dawek leku[70].
- Zwiększenie stężenia leku w (biodostępności) poprzez optymalizację profilu farmakokinetycznym związku[71].
- Optymalizacja kinetyki interakcji lek-cel chemicznego farmakologicznego (tzn. czasu przebywania leku w miejscu docelowym) poprzez odpowiednią farmakokinetyką oraz zależności farmakodynamiczne (PK/PD) opisujące zależność stężenia leku w płynach ustrojowych[72,73].
- Optymalizacja czasu przebywania leku w miejscu docelowym, a nie jego maksymalizacja, ponieważ w wielu przypadkach skuteczne leki posiadają kinetykę nierównowagową,

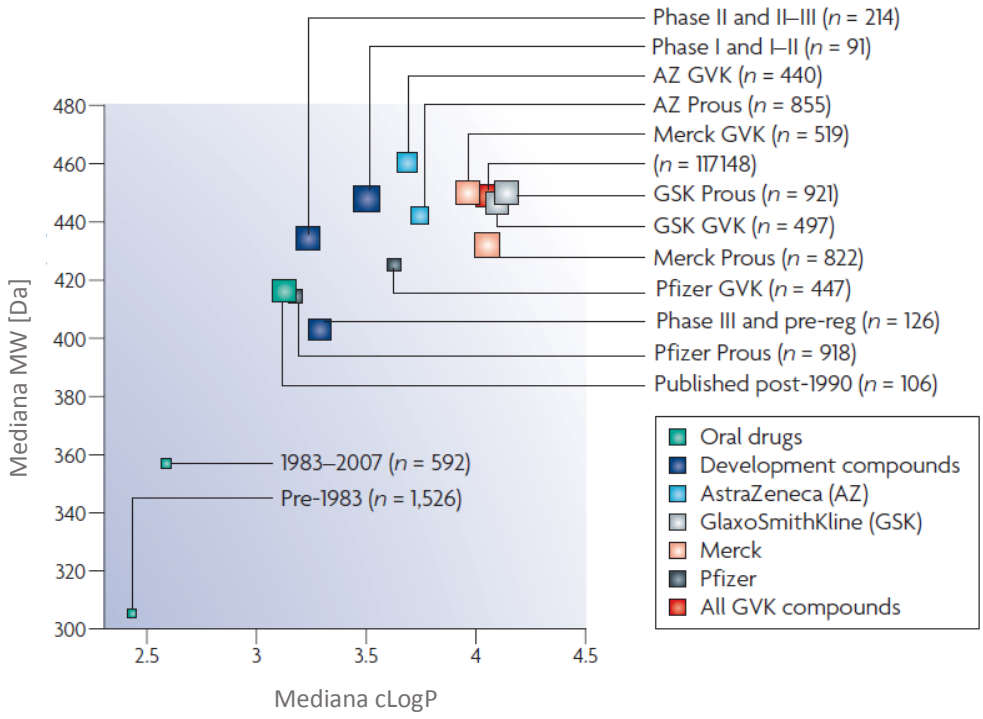
mającą profil termodynamiczny bardziej determinowany czynnikiem entalpi (wkładem entalpi i entropii do wiązanie)[74].

Historyczne strategie oparte na farmakologii in vivo skupiały się na konkretnych celach molekularnych wybranych w związku z leczeniem określonych chorób. Strategie te jednak na przestrzeni lat nie zwiększyły tempa badań nad lekami[75,76]. W związku z tym przemysł farmaceutyczny poszukiwał nowych technologii mających na celu zwiększenie efektywności procesu wykrywania leków. Jedna z takich technologii wykorzystuje prowadzenie wysoko wydajnych badań przesiewowych HTS (ang. high-throughput screening) polegających na testowaniu dużej liczby związków w zakresie biochemicznym w testach opartych na komórkach. Kluczowym wyznacznikiem wpływającym na podjęcie decyzji o wyborze serii wiodącej poddawanej testom w trakcie badań przesiewowych, była biodostępność i toksyczność syntetyzowanych związków chemicznych[75,77].

Ponadto, opublikowane w literaturze wyniki badań nad lekami prowadzonych przez firmy farmaceutyczne nie pozostawiają wątpliwości, że zwiększenie liczby mierzonych właściwości fizycznych badanych/syntetyzowanych związków jest kluczowym problemem w procesie poszukiwania nowych leków[78,79,80]. W literaturze przedmiotu dominuje pogląd, że właściwości fizykochemiczne, takie jak: lipofilowość - wyrażana ilościowo jako współczynnik podziału w skali logarytmicznej (logP) i masa cząsteczkowa (MW), mają szczególne znaczenie dla aktywności związków i parametrów ADMET (takich jak: wchłanianie, dystrybucja, metabolizm, wydalanie i toksyczności)[77,81]. Z przeprowadzonych badań nad lekami wynika, że związki, które przeszły do badań farmakokinetycznych firm farmaceutycznych, takich jak: GlaxoSmithKline[82] i Abbott[83] miały odpowiednio średnie wartości cLogP wynoszące 4.3 i 3.9 a wartości masy cząsteczkowej równe 480 i 434 Da.

Na rysunku 20. przedstawiono wyniki badań związków chemicznych będących w fazie rozwoju i patentami, a lekami odkrytymi w latach 90. W przeprowadzonych badaniach porównano wartości właściwości fizykochemicznych, takie jak: cLogP, MW dla leków z lat 90' z wartościami cLogP, MW związków będących w fazie rozwoju oraz aktualnymi patentami dla czterech międzynarodowych firm farmaceutycznych takich jak: AstraZeneca,

GlaxoSmithKline, Merck, Pfizer. Z analizy rysunku 22. wynika, że mediana patentowanego związku posiada wartość cLogP =4.1 a MW= 450 Da, podczas gdy wartość mediany leków doustnych odkrytych po 1990 roku, posiada odpowiednio wartość cLogP 3.1 oraz MW=432 Da[78].



Rysunek 22. Porównanie wartości właściwości fizykochemicznych takich jak clogP oraz MW dla leków z lat 90 z związkami będącymi w fazie rozwoju i aktualnymi patentami[78].

gdzie:

- cLogP oraz MW oznaczono kwadratem i oznakowano kolorem według źródła;
- n = liczba związków lub liczba patentów, otrzymanym odpowiednio z baz patentowych: Prous Science Integrity obejmującą lata 2001–2007, oraz GVK Bio obejmującą lata 2003–2007[78].

7. Wydajność ligandu LE jako miara stosowana w projektowaniu leków

W literaturze szeroko opisana jest miara wydajności ligandu LE (ang. Ligand Efficiency) wykorzystywana w procesie projektowania leków[84]. Miara LE została wprowadzona w celu normalizacji powinowactwa związków w odniesieniu do wielkości cząsteczek[85]. Wydajność ligandu (LE) jest definiowana jako stosunek energii swobodnej Gibbsa do liczby atomów ciężkich HAC (ang. Heavy Atom Counts)[84]. Zależność tę przedstawia poniższe równanie:

$$LE=([-2.303RT]/HAC)*logKd \quad (1.9)$$

gdzie:
HAC - liczba atomów ciężkich,
R - stała gazowa,
T - temperatura w stopniach Kelvina,
Kd - stała równowagi.

Ponadto w standardowych warunkach w roztworze wodnym w T= 300K, pH=7, stężeniu 1 M, wartość [-2,303RT] można przeliczyć do wartości -1,37 kcal/mol. W przybliżeniu równanie definicyjne spotyka się w literaturze niemal wyłącznie w postaci uproszczonej[84]:

$$LE = 1,4* pIC_{50} / HAC \quad (2.0)$$

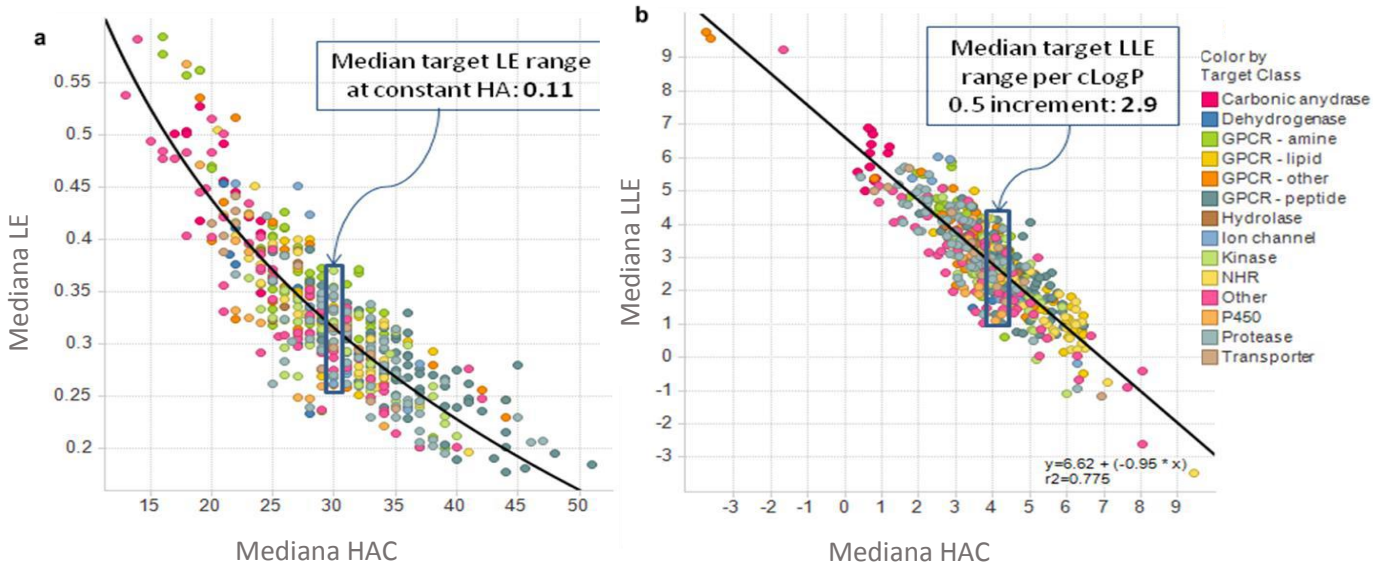
gdzie:
pIC_{50} - aktywność biologiczna,
HAC - liczba atomów ciężkich.

W ostatnim czasie zaproponowano wiele pokrewnych miar efektywności, które obejmują badanie wpływu takich właściwości fizykochemicznych jak lipofilowość czy masa cząsteczkowa[84,86,87].

Przegląd literatury dostarcza następujących informacji:

- Średnia lipofilność leków wprowadzanych na rynek mierzona współczynnikiem podziału (logP) lub współczynnikiem dystrybucji (logD) zmieniła się nieznacznie w ciągu ostatnich kilku lat. Oznacza to, że lipofilowość jest podstawową właściwością wpływającą na postęp programów odkrywania leków i możliwości rozwoju zidentyfikowanych kandydatów na leki.
- Mediana i/lub średnia masa cząsteczkowa zatwierdzonych leków wzrosła o około 50 Da (15 %) w ciągu ostatnich 3 dekad, podczas gdy mediana i/lub średnia masa cząsteczkowa zsyntetyzowanych związków eksperymentalnych wzrosła o ponad 100 Da (30%)[88,89,90,91,92].

Na rysunku 23. zilustrowano zależność wydajności ligandów od właściwości fizykochemicznych dla 480 par testów docelowych obejmujących 329 różnych celów reprezentujących 201 041 cząsteczek otrzymanych z bazy danych GVK BIO[93]. Klasy docelowe to m.in.: GPCR – receptor sprzężony z białkiem C, kinazy, proteazy, NHR - jądrowe receptory hormonalne.



Rysunek 23. Zależności wydajności ligandów od właściwości fizykochemicznych[52]*. LLE –lipofilowa wydajność liganda.*

gdzie:

a) LE vs. HAC [LE = (1.37 / HAC). pIC50 (or pKi)]
b) LLE vs. cLogP [LLE= pIC50 (or pKi) – cLogP (or LogD)]

Z analizy obu rysunków wynika, że:

- Mediana wydajności ligandu dla 480 par testów docelowych obejmuje szeroki zakres: LE~ 0,2–0,6, natomiast dla LLE ~ -3–9 , gdzie wartości mediany wynoszą - odpowiednio - dla: LE = 0,32, LLE = 2,83, HAC = 29,75, cLogP = 3,89[52].
- Mediana LE maleje wraz ze wzrostem liczby atomów ciężkich. Czynnik przyczyniający się do spadku LE może być uwarunkowany wpływem entropii konformacyjnej na wiązanie się większych ligandów, z czego można wnioskować, że duże cząsteczki są bardziej ograniczone konformacyjnie (rysunek 23a)[52,94].
- Mediana LLE maleje ze wzrostem liofilowości (cLogP). Czynnikiem przyczyniającym się do obserwowanej zależności może być wpływ entalpii i entropii na wiązania ligandu, które mają tendencję do zmniejszania się wraz ze wzrostem liofilowości (rysunek 23b)[52,95].

Zastosowanie miar wydajności ligandów w optymalizacji leków prowadzi do następujących wniosków:

- LE i LLE jest ważnym czynnikiem w analizie i ocenie właściwości związków będących kandydatami na leki.
- Kontrolowanie liofilowości jest podstawą udanej optymalizacji.
- Zwiększenie powinowactwa wiązania wymaga dodania polarnych grup funkcyjnych do cząsteczki, czego wynikiem będzie najprawdopodobniej powstanie nowych polarnych oddziaływań ligand–receptor (na przykład wiązań wodorowych), które zmienią powinowactwa do struktury i w konsekwencji zapewnią nowe możliwości optymalizacji[52,96].

BADANIA WŁASNE

W literaturze opisano niewiele modeli odwzorowujących strukturę chemiczną w zbiory danych ekonomicznych: prawo Erooma i wiek leku, które omówiono w rozdziale 3.3. części literaturowej. W przypadku analiz struktura – efekty ekonomiczne problemem jest zarówno dobór odpowiednich deskryptorów molekularnych, jak i konstruowanie modelu zależności. Przeprowadzona przeze mnie analiza ekonomicznych właściwości związków chemicznych, tzn. zależność ceny rynkowej od wybranych deskryptorów molekularnych, jest pierwszą tego typu analizą publikowaną w literaturze[97,98]. Katalog bloków budulcowych firmy Abamachem specjalizującej się w dziedzinie syntetycznej chemii organicznej zawiera wykaz ponad 2.2 miliona związków chemicznych dostępnych na rynku[4].

1. Dobór właściwości i deskryptorów dla zależności struktura-ekonomia

Efekty ekonomiczne

Dostępność "właściwości ekonomicznych" opisujących zbiory leków jest niewielka. Uniwersalnym miernikiem efektów ekonomicznych jest cena, ale nawet informacja o cenach leków jest trudno dostępna.

Deskryptory strukturalne

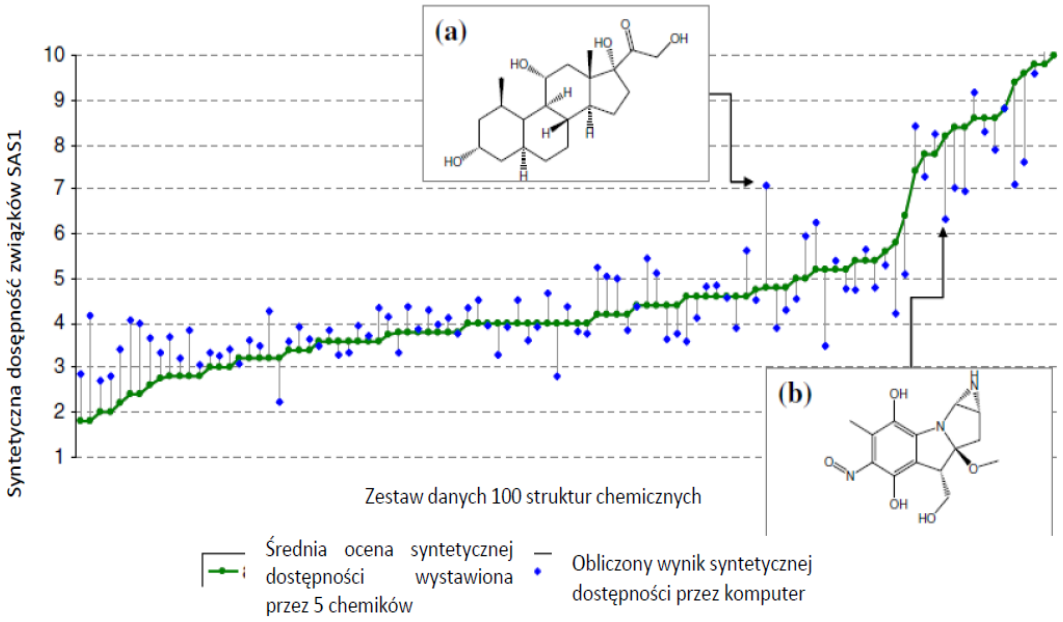
Prostym miernikiem zmienności strukturalnej stosowanym w analizach wielkich danych jest MW, HAC.

Wydaje się, że dostępność syntetyczna danego związku również powinna determinować jego cenę rynkową. Deskryptory syntetycznej dostępności zostały ostatnio zaproponowane przez Gasteigera.

1.1. Deskryptory dostępności syntetycznej

Syntetyczna dostępność SAS1 (ang. Synthetic Accessibility Score) jest deskryptorem, którego celem jest określenie trudności syntezy związków organicznych w skali od 1 do 10, gdzie 1 oznacza związki łatwe do zsyntetyzowania, natomiast 10 - związki trudne do zsyntetyzowania. Metoda ta szybko ocenia syntetyczną dostępność struktur w oparciu o kilka istotnych kryteriów, takich jak: złożoność struktury molekularnej i układów pierścieniowych oraz liczby stereo centrów. Badanie syntetycznej dostępności związku służy uzyskaniu informacji, na ile łatwo można zsyntetyzować cząsteczkę na podstawie analizy struktury molekularnej, a następnie otrzymany wynik porównuje z reakcjami organicznymi rejestrowanymi w bazach danych[99].

Na rysunku 24. według Gasteigera[99] przedstawiono porównanie wyniku oszacowania syntetycznej dostępności SAS1 przez komputer ze średnim wynikiem oszacowania syntetycznej dostępności przez pięciu chemików reprezentujących trzy różne firmy farmaceutyczne dla 100 struktur wybranych z Journal of Medicinal Chemistry. Ocena wystawiana przez chemików dotyczyła stopnia trudności zsyntetyzowania 100 struktur w skali od 1 do10 (gdzie:1- prosta synteza związku, 10 – trudna synteza związku). Z analizy rysunku 24. wynika, że na skutek oszacowania syntetycznej dostępności przez komputer otrzymuje się porównywalne wyniki, jak ma to miejsce w przypadku oszacowania trudności syntetyzowania 100 analizowanych struktur przez chemików. Potwierdza to wiarygodność metody oceny syntetycznej dostępności struktur opracowanej przez zespół prof. Gasteigera. Jej obliczenie in silico zajmuje 25 sekund. Wyjątek stanowią tylko dwie struktury (a) i (b), w których obliczone przez komputer SAS1 różnią się od oceny SAS1 wystawionej przez chemików. Różnice te według Gasteigera najprawdopodobniej wynikają z wysokiej złożoności struktury molekularnej oraz liczby i ustawienia stereo centrów. Potwierdzeniem powyższych konkluzji są wysokie i porównywalne wartości współczynników korelacji (tabela 1 i 2) mieszczące się w zakresie od 0.73 do 0.89[100].

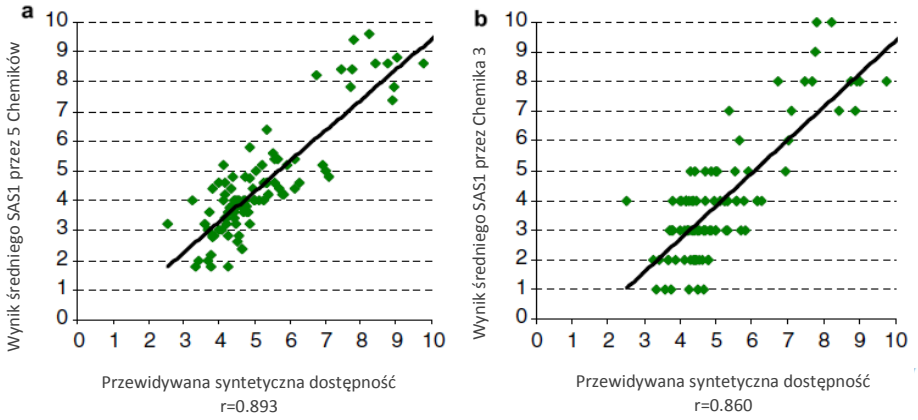


Rysunek 24. Porównanie wyniku oszacowania syntetycznej dostępności SAS1 przez komputer ze średnim wynikiem oszacowania syntetycznej dostępności przez pięciu chemików dla 100 struktur pobranych z Journal of Medicinal Chemistry, według[100].

W tabeli 1. przedstawiono współczynniki korelacji pomiędzy oceną SAS1 wystawioną przez pięciu chemików, a wynikiem SAS1 obliczonym przez komputer. Relacje obserwowane na rysunku 25. przedstawiają najwyższe wartości obliczonych współczynników korelacji według[100].

Tabela 1. Współczynniki korelacji pomiędzy oceną SAS1 wystawioną przez pięciu chemików, a wynikiem SAS1 obliczonym przez komputer według[100].

	Ocena SAS1 Chemik 1	Ocena SAS1 Chemik 2	Ocena SAS1 Chemik 3	Ocena SAS1 Chemik 4	Ocena SAS1 Chemik 5	Średnia
Obliczony wynik SAS 1	0.810	0.791	**0.860**	0.840	0.789	**0.893**



Rysunek 25. Prognozowanie SAS1 według[100].

gdzie:

(a) średniego oszacowania SAS1 przez chemików, a wynikiem SAS1 obliczonym przez komputer,

(b) średniego oszacowania SAS1 przez trzeciego chemika, a wynikiem SAS1 obliczonym przez komputer[100].

W tabeli 2. przedstawiono współczynniki korelacji pomiędzy oceną SAS1 szacowaną przez pięciu różnych chemików dla zestawu danych 100 związków według[100].

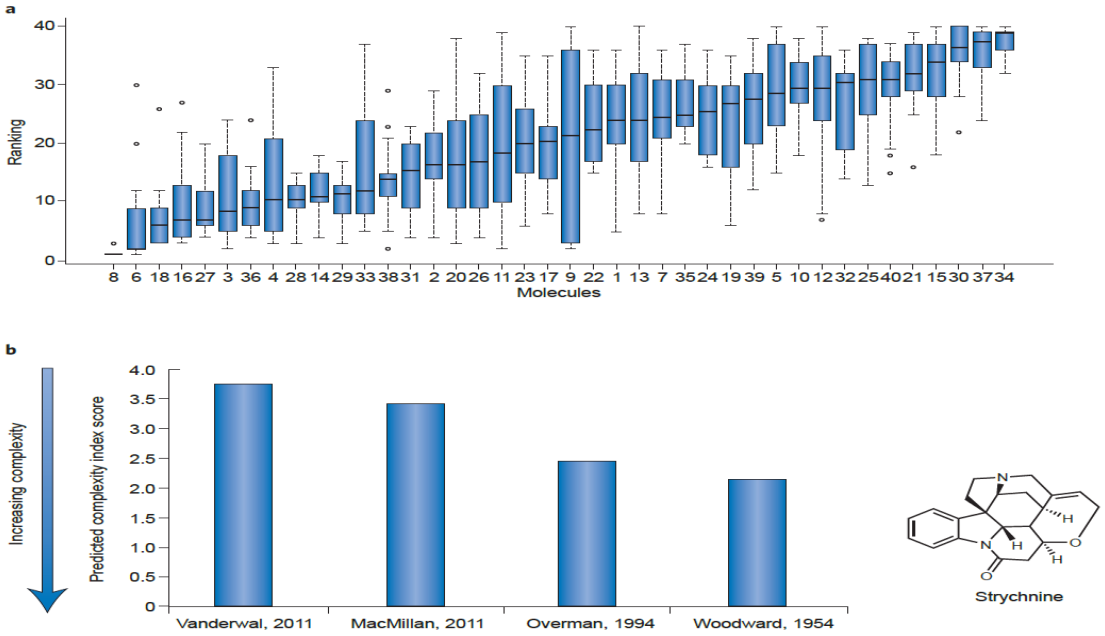
Tabela 2. Współczynniki korelacji pomiędzy deskryptorami SAS1 oraz wartościami szacowanymi przez pięciu różnych chemików według[100].

	Ocena SAS1 Chemik 1	Ocena SAS1 Chemik 2	Ocena SAS1 Chemik 3	Ocena SAS1 Chemik 4	Ocena SAS1 Chemik 5
Ocena SAS1 Chemik 1	-	0.75	0.77	0.84	0.74
Ocena SAS1 Chemik 2		-	0.78	0.73	0.74
Ocena SAS1 Chemik 3			-	0.82	0.75
Ocena SAS1 Chemik 4				-	0.81
Ocena SAS1 Chemik 5					-

1.2. Wskaźnik złożoności cząsteczek organicznych

Wskaźnik złożoności cząsteczek organicznych został opracowany przez Li i Eastgate. Wskaźnik ten łączy ze sobą wewnętrzną złożoność molekularną opartą na analizie struktury molekularnej oraz zewnętrzną złożoność syntetyczną opartą na technologii reakcji. Wskaźnik złożoności cząsteczek organicznych pozwala na porównanie cząsteczek, analizę trendów oraz umożliwia ocenę i porównanie nowych syntetycznych związków ze znanymi związkami, na podstawie złożoności struktury cząsteczki, która zmienia się w czasie[101].

Na rysunku 26a przedstawiono ranking oceny złożoność struktury molekularnej dla 40 cząsteczek. Ocenę złożoności struktury molekularnej w skali od 1 do 40 (gdzie: 1- złożona cząsteczka, 40– prosta cząsteczka) dla 40 cząsteczka wykonało osiemnastu chemików zajmujących się syntezą. Przy zastosowaniu takiego probabilistycznego podejścia ocena rankingu oparta jest na wiedzy i doświadczeniu chemików biorących udział w badaniu. Natomiast na rysunku 26b przedstawiono analizę poziomu złożoności syntezy i struktury molekularnej strychniny w wybranych latach. Rysunek 26. według[99,101].



Rysunek 26. Złożoność cząsteczek organicznych[99].

Z analizy rysunku 26b wynika, że poziom złożoności struktury molekularnej dla strychniny jest zmienna w czasie. Obserwujemy postęp w syntezie i technologii prowadzenia reakcji oraz metod syntezy organicznej. Strychnina jest klasycznym przykładem wpływu zmiany złożoności struktury molekularnej w czasie. Pierwotnie została wyizolowana w 1818 roku i dopiero po 130 latach dzięki ulepszeniu technologii reakcji syntez została zsyntetyzowana przez:

- Woodwardsa w 1954 roku - w 30 krokach
- Overmana w 1994 roku - w 25 krokach
- MacMillana w 2011 roku - w 13 krokach
- Vanderwala w 2011 roku - w 10 krokach[99,101].

Doskonalenie procesu syntez przeprowadzanych przez chemików stanowi ważny krok w ocenie złożoności cząsteczek organicznych, ponieważ umożliwia chemikom śledzenie złożoności struktury molekularnej w czasie.

Podsumowując, deskryptory SAS1 wydają się być łatwo dostępną metodą szacowania chemicznej kompleksowości struktur, dając wiarygodny obraz dostępności odpowiednich substancji. Ponieważ intuicyjnie wydaje się, że im mniej dostępna substancja, tym wyższa powinna być cena deskryptory SAS wykorzystywałam jako potencjalne korelaty cen biblioteki Abamachem.

2. Problemy modelowania – odwzorowanie efektów ekonomicznych w zbiór deskryptorów molekularnych - statystyka molekularna

Przeprowadzone przeze mnie badania stanowią przykład analizy danych typu big data. Całość opisanych dotąd związków chemicznych stanowi zbiór zawierający około 150 mln związków, tak więc analizowana grupa to około 2% ogólnej liczby związków. Analizy big data różnią się od typowych modeli QSAR, gdzie z reguły poddaje się analizie kilkadziesiąt związków lub analiz typy skriningu wirtualnego gdzie przegląda się (przesiewa) znacznie większe zbiory danych. W ostatnim przypadku z reguły nie buduje się modeli ilościowych, lecz obserwuje potencjalnie interesujące struktury danych[102]. Zaproponowaliśmy, aby dla

odróżnienia takie modele określać mianem statystyk molekularnych[97]. Statystyki molekularne polegają na masowej analizie przestrzeni chemicznej poprzez próbkowanie związków chemicznych. Stanowią one rodzaj analizy typu struktura - aktywność SAR (ang. Structure-Activity Relationship), z tą jednak różnicą, że zamiast pojedynczych związków badana jest szersza klasa połączeń. Klasy takie są mniej sprecyzowane, niż ma to miejsce w klasycznych analizach QSAR. Mogą być to na przykład typy związków zarejestrowanych przez FDA (ang. Food and Drug Administration)[97]. Z drugiej strony analiza dużej populacji molekularnej daje nadzieję na identyfikację domen przestrzeni chemicznej typowych dla określonych klas związków lub godnych uwagi ze względu na obiecujące zastosowania, np. w farmacji.

Ponadto w przypadku wielkich danych trudno jest oczekiwać precyzyjnych zależności ilościowych, typowych dla modeli QSAR. Analiza dużej populacji związków różni się od klasycznego podejścia QSAR, w której poddaje się badaniu kilkadziesiąt związków o znanej aktywności biologicznej. W analizach typu big data poddaje się analizie miliony związków, które są reprezentowane/zastępowane przez średnie maksymalne i minimalne wartości opisujące serie związków lub klas związków. W związku z tym potrzebne są dalsze badania, aby wykorzystać statystyki molekularne w modelowaniu QSAR[7]. Jedną z metod modelowania danych, którą zaproponowałam w mojej pracy, jest binowanie danych.

3. Binowania danych

Binowanie danych to operacja, która umożliwia zastąpienie zbioru wartości poprzez wyznaczenie jego mediany lub średniej oraz wizualizację obliczonych wartości, na przykład w postaci histogramu. Binowanie danych to także sposób, w którym pierwotne wartości danych zastępowane są przez ich średnie liczone w małych przedziałach. Metoda ta daje możliwość znacznego uproszczenia struktury danych oraz obserwacji ukrytych trendów dla wielkich danych[97].

W literaturze zajmującej się problematyką projektowania molekularnego zwracają uwagę prace Petera W. Kenny'ego oraz Carlosa A. Montanariego poświęcone metodzie binowania danych chemicznych. Obaj uczeni opisują ją jako metodę statystyczną, która przyczynia się do

wzrostu korelacji między danymi/zmiennymi[103]. W szczególności Kenny oraz Montanari twierdzą, że binowanie danych jest operacją ryzykowną (rysunek 27.).

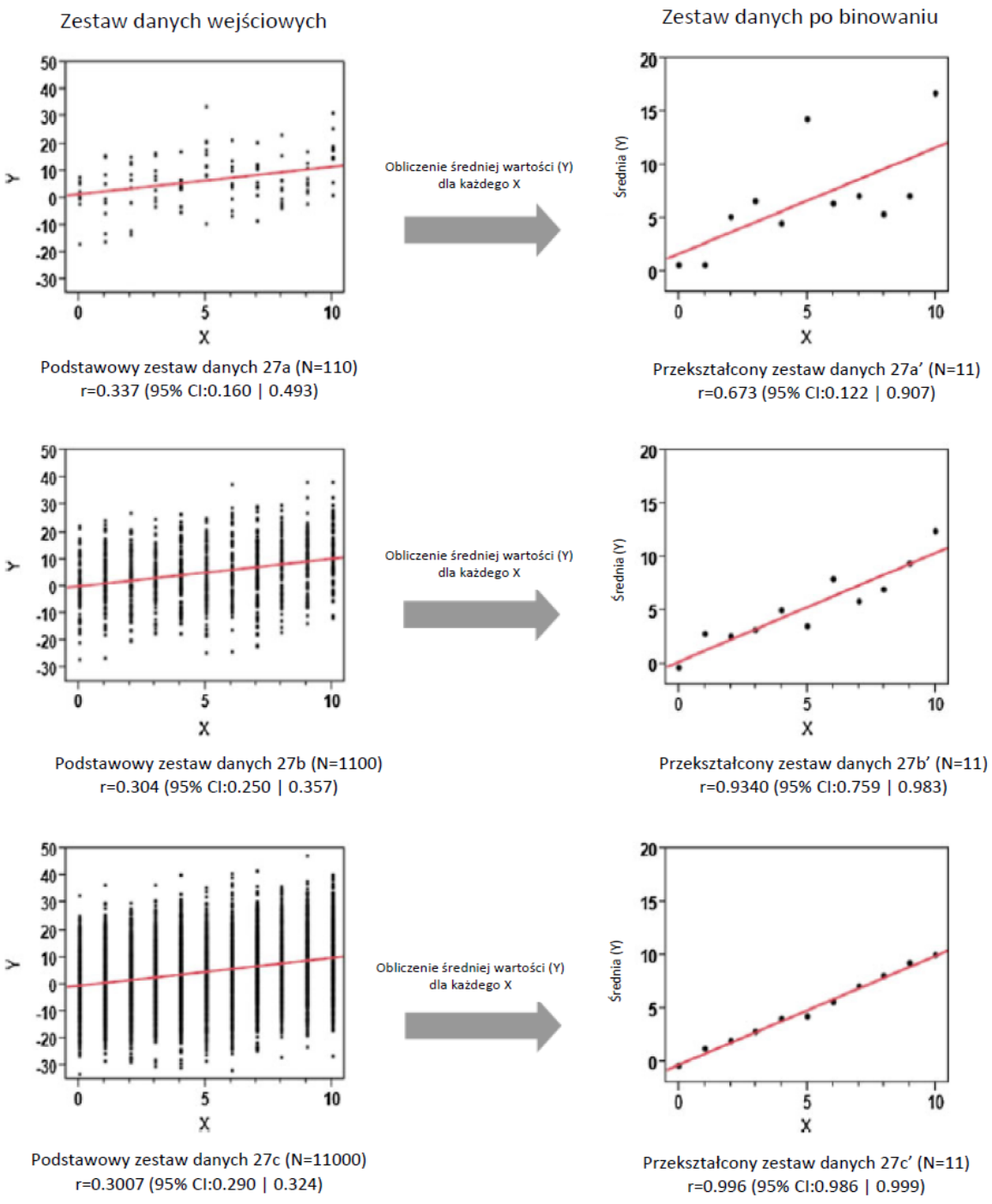
Na rysunku 27. zilustrowano schemat binowania danych dla trzech zestawów danych o różnych rozmiarach [N=110, 1100, 11000], polegający na przekształceniu danych wejściowych na dane binowane. Rysunek ten dostarcza informacji, w jaki sposób binowanie danych wpływa na korelację poprzez obliczenie współczynnika korelacji Persona dla badanych zestawów danych przed i po binowaniu według Kennego[103].

Gdzie:

- ✓ Trzy zestawy danych należy traktować jako losowe próbki o różnych rozmiarach N-liczna danych: N=110, 1100, 11000, które zostały wygenerowane z tej samej populacji danych.
- ✓ Punktem wyjścia dla każdego zestawu danych jest czerwona linia prosta [Y = A+BX] zapewniająca wizualne odniesienie.
- ✓ Punkty danych dla wartości całkowitych X w linii prostej [Y=A+BX] zostały równomiernie rozłożone (dla bin= 0.5).
- ✓ Podstawowe zestawy danych w ustalonych przedziałach (rysunek 27a, 27b, 27c) przetransformowano na reprezentatywne zestawy danych dla danego przedziału - średnia wartość (Y) dla każdego X (rysunek 27a', 27b', 27c').
- ✓ Dla każdego zestawu danych obliczono przed i po binowaniu danych r- współczynnik korelacji Pearsona z 95% przedziałem ufności[103].

Z analiza rysunku 27. wynika, że:

- ✓ Korelacje dla zestawów danych wejściowych (rysunek 27a, 27b i 27c) nie różnią się znacząco bez względu na rozmiar danych, natomiast korelacje dla zestawów danych po binowaniu (rysunek 27a', 27b' i 27c') rosną wraz z rozmiarem danych.
- ✓ Binowanie danych przyczynia się do zwiększenia korelacji między danymi/zmiennymi wraz z rozmiarem danych a wzrost korelacji jest zazwyczaj wynikiem uśrednienia zestawów danych w odpowiednie zakresy.
- ✓ Binowanie danych dobrze sprawdza się dla dużych zbiorów danych ułatwiając przedstawienie w sposób graficzny analizę wielkich danych[103].



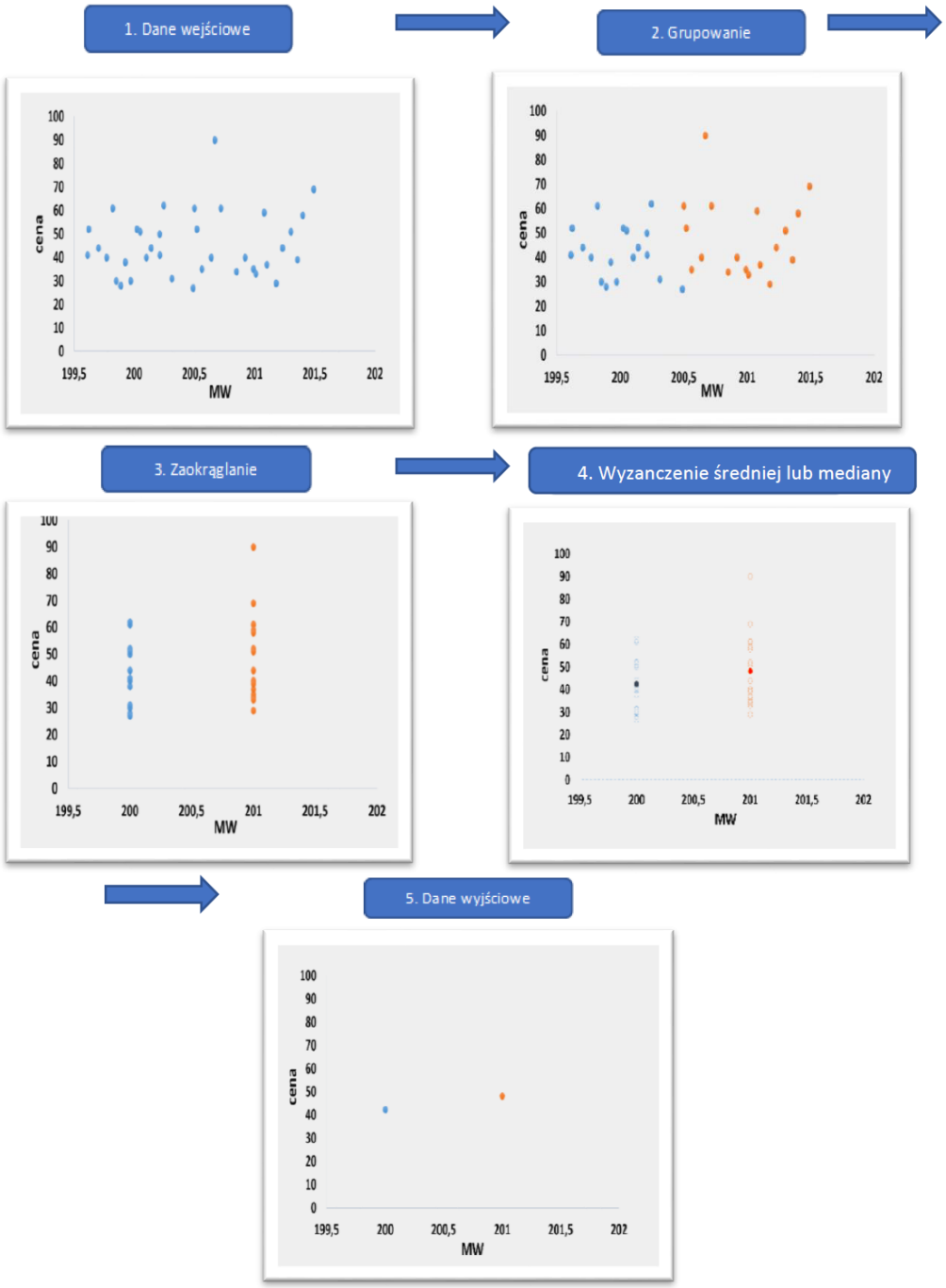
Rysunek 27. Schemat binowania danych dla trzech zestawów danych o różnych rozmiarach [N=110, 1100, 11000] według[103].

Binowanie zwiększając korelacje, tworzy ryzyko obserwowania korelacji losowych. Z punktu widzenia wielkich danych umożliwia jednak obserwację występujących zależności między zmiennymi. W niniejszej pracy binowanie stosowano jako podstawową metodę modelowania statystyk molekularnych.

Na rysunku 28. przedstawiono w sposób graficzny przykładowe binownia danych zastosowane w moich badaniach. Rysunek ten ilustruje zależność ceny od masy cząsteczkowej w dwóch ustalonych przedziałach masowych [pierwszy przedział masowy: 199,5-200,4] oraz [drugi przedział masowy 220,5-201,4]. Wartości danych zostały równomiernie rozłożone dla bin = 1 (liczby całkowite).

Gdzie:

- ✓ Rysunek 1 prezentuje dane wejściowe, przedstawiające zależność ceny od MW - masy cząsteczkowej związku chemicznego.
- ✓ Rysunek 2 przedstawia grupowanie się danych wejściowych, tzn. związków chemicznych z uwzględnieniem ich mas cząsteczkowych. W tym przypadku w pierwszym przedziale masowym: 199,5-200,4 (kolor niebieski) oraz drugim przedziale masowym: 220,5-201,4 (kolor pomarańczowy).
- ✓ Rysunek 3 prezentuje sposób zaokrąglenia mas cząsteczkowych związków chemicznych w dwóch ustalonych przedziałach. Zastosowano binowanie dla bin = liczby całkowite. W tym przypadku dla bin = 200 kolor (niebieski) oraz bin = 201 (kolor pomarańczowy).
- ✓ Rysunek 4 przedstawia wyznaczanie średnich wartości/lub mediany kółka - pełne w zaokrąglonych przedziałach mas związków.
- ✓ Rysunek 5 zawiera dane wyjściowe przedstawiają średnią/lub medianę wartość z danych przedziałów.



Rysunek 28. Schemat binowania danych zastosowany w badaniach szeregu Abamachem.

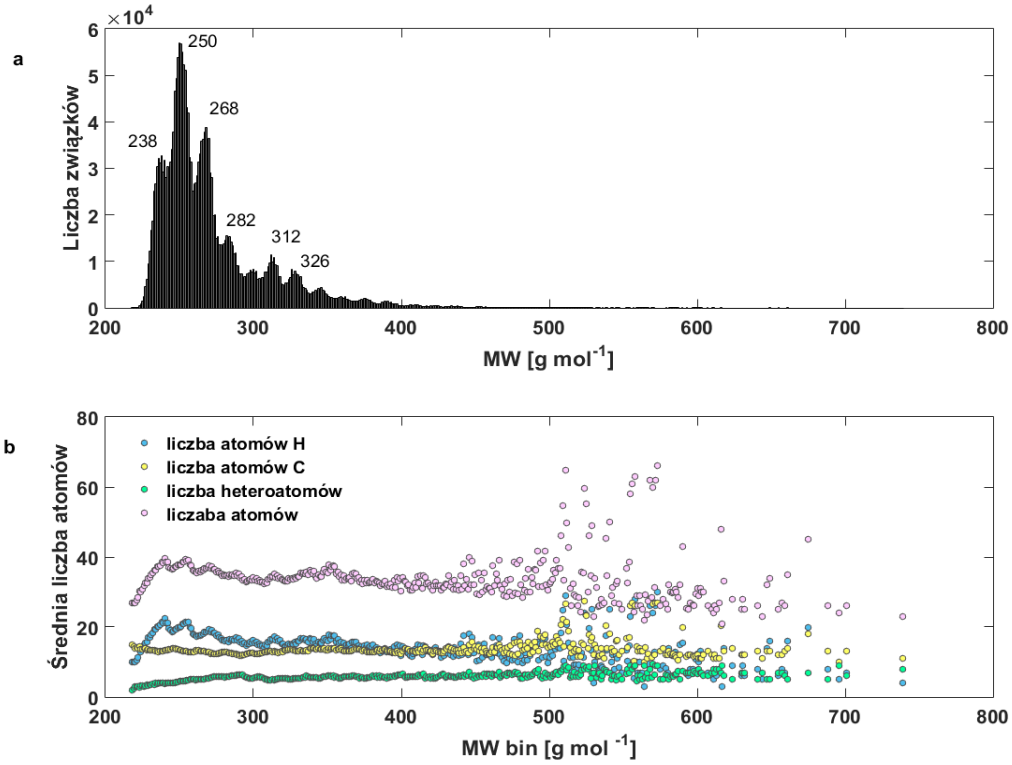
4. Charakterystyka danych Abamachem

Abamachem jest firmą badawczą zajmującą się badaniami w dziedzinie syntetycznej chemii organicznej. Firma została założona w 2008 roku przez zespół chemików organików. Siedziba i laboratoria Abamachem znajdują się w Kijowie. Firma ta oferuję katalog bloków budulcowych zawierający ponad 2 miliony związków chemicznych. Wszystkie próbki są dokładnie testowane za pomocą spektroskopii LC / MS i NMR i muszą uzyskać 95 % czystości. Do jednoznacznej identyfikacji związków firma wykorzystuje FT-IR i analizę elementarną. Na życzenie klienta dla każdego zamówionego związku chemicznego mogą zostać udostępnione wszystkie dane analityczne[4]. W katalogu Abamachem podaje się cenę każdego oferowanego związku. Formaty danych oraz problemy ich konwersji omówiłam w części eksperymentalnej.

5. Statystyka danych Abamachem: zależności struktura-ekonomia dla wielkiej biblioteki związków chemicznych

Na rysunku 29. przedstawiono analizę statystyczną mas cząsteczkowych MW i binowanych mas cząsteczkowych MW bin dla komercyjnej biblioteki firmy Abamachem zawierającej ok. 2.2 miliona związków chemicznych. Z interpretacji wykresu 29a przedstawiającego częstości występowania odpowiednich izobinów MW wynika, że masy cząsteczkowe związków w badanej bibliotece danych mieszczą się w przedziale od 218 do 739 Da. Maksimum występuje dla MW ok. 250 Da. Co ciekawe, porównując analizę MW dla katalogu firmy Abamachem (rysunek 29a) z analizą MW dla bazy Beilstein (rysunek 16.), obserwuje się wyraźne podobieństwo rozkładu mas MW związków chemicznych dla analizowanych zbiorów. W bazie Beilstein najwięcej związków posiada masy cząsteczkowe mieszące się w przedziale od ok. 250 do 300 Da, natomiast w katalogu Abamachem w przedziale od 238 do 268 Da. Na rysunku 29b przedstawiono zależność między średnią liczbą atomów a binowaną masą cząsteczkową MW z uwzględnieniem liczby atomów wodoru i węgla, a także wszystkich atomów i heteroatomów dostępnych w badanym katalogu chemicznym. Z rysunku 29b wynika, że średnia całkowita liczba atomów dobrze koreluje z średnią liczbą atomów wodoru, co na wykresie ilustruje podobny kształt obu krzywych. Porównując rysunek

29a z rysunkiem 29b można zaobserwować wyraźne podobieństwo dla zbiorów, które mieszczą się w podobnym przedziale mas. W przedziale od 200 do 400 MW można zaobserwować ciągłość wskazanej wyżej zależności. Po przekroczeniu wartości 400 Da zależność ta nabiera wyraźnie charakteru chaotycznego. Efekt ten wynika najprawdopodobniej ze statystycznie małej reprezentatywności próbkowanej przestrzeni w tym zakresie MW (mała liczba związków o pojedynczych wartościach MW).



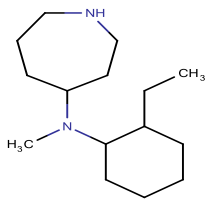
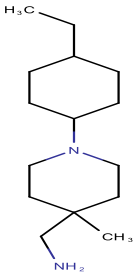
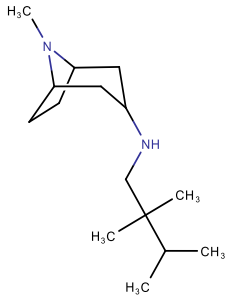
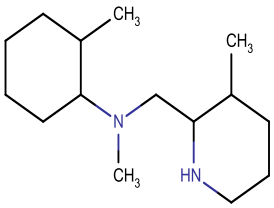
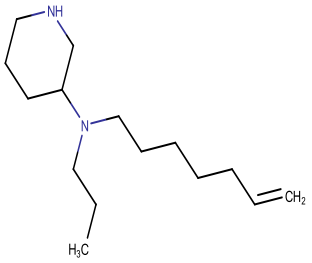
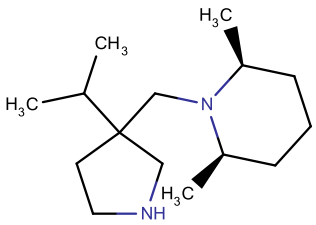
Rysunek 29. Rozkład mas (a) Średnia liczba atomów vs. binowana MW dla danych Abamachem (b).

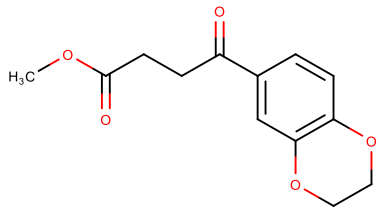
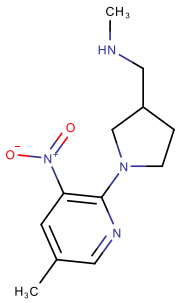
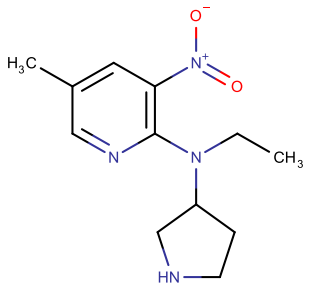
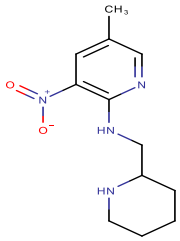
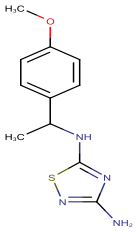
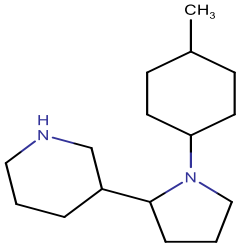
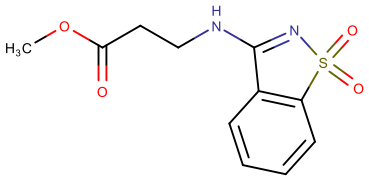
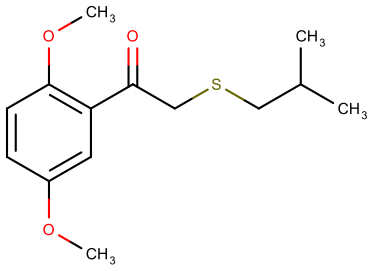
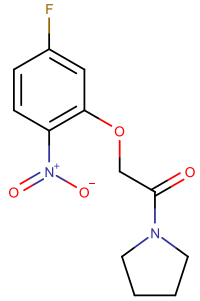
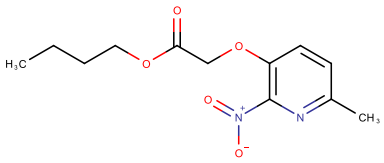
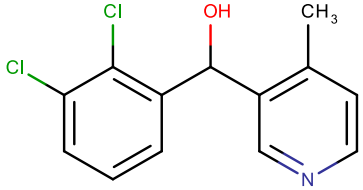
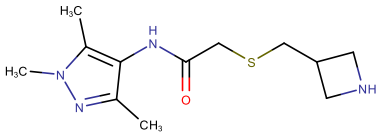
W celu lepszego zrozumienia problemów przetwarzania wielkich danych typu Abamachem w tabeli 3. przedstawiono liczby wybranych związków chemicznych posiadających określone izobiny MW zilustrowane na histogramie (Rysunek 29a) oraz wybrane ich struktury posiadające właśnie takie masy cząsteczkowe. Z analizy tabeli 3. wynika, że w badanym katalogu najwięcej związków chemicznych posiada MW= 250 Da: liczba związków: 56936, natomiast najmniej

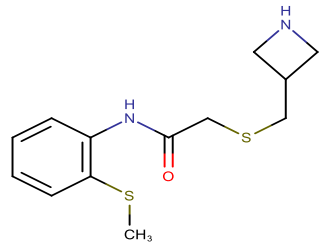
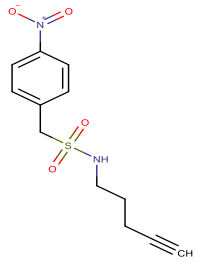
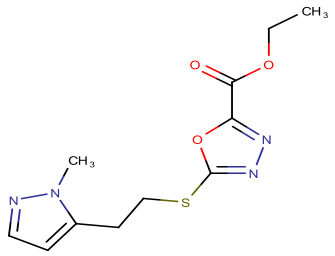
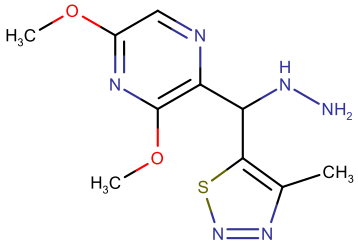
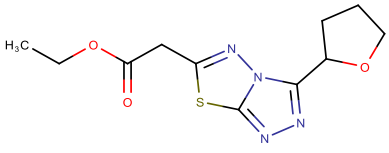
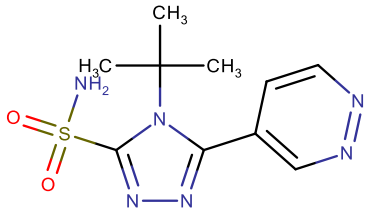
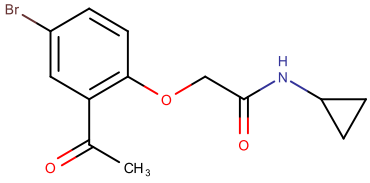
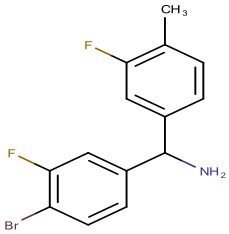
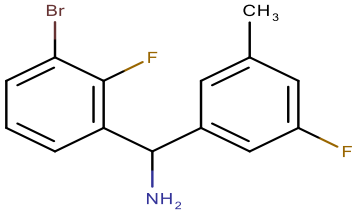
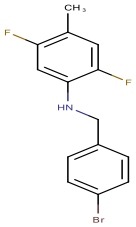
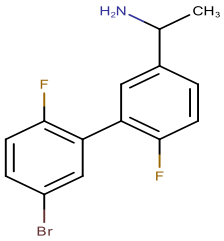
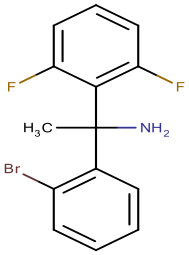
związków chemicznych posiada MW= 326 Da: liczba związków: 8219. Ponadto zaprezentowane tu wybrane struktury z katalogu różnią się masą, liczbą oraz rozmieszczeniem przestrzennym heteroatomów tworzących daną strukturę chemiczną.

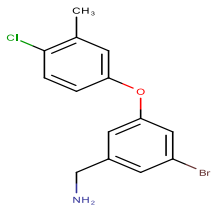
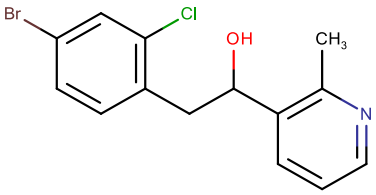
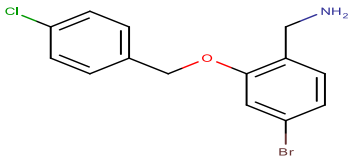
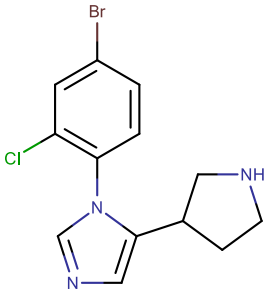
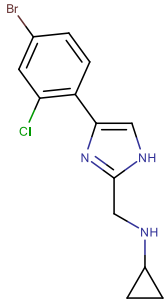
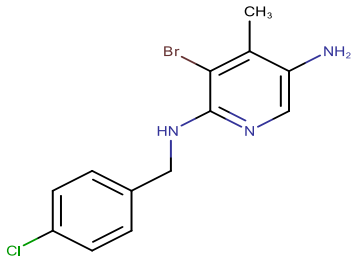
Tabela 3. Populacja związków chemicznych o określonych MW oraz wybrane struktury chemiczne z katalogu Abamachem posiadające właśnie takie MW.

Masa cząsteczkowa [g/mol]	238 [g/mol]	250 [g/mol]	268 [g/mol]	282 [g/mol]	312 [g/mol]	326 [g/mol]
Liczba związków chemicznych	32639	56936	38818	15468	11453	8219

Wybrane związki chemiczne o MW= 238,42 g/mol		
		
		

Wybrane związki chemiczne o MW= 250,39 g/mol		
		
		
Wybrane związki chemiczne o MW= 268,75g/mol		
		
		

Wybrane związki chemiczne o MW= 282,32 g/mol		
		
		
Wybrane związki chemiczne o MW= 312,16g/mol		
		
		

Wybrane związki chemiczne o MW= 326,62 g/mol		
		
		

W tabeli 4. zamieszczono macierz korelacji dla ceny molowej i wagowej oraz wybranych deskryptorów molekularnych dla danych bazy Abamachem. Obliczone współczynniki korelacji są miarą oszacowania zdolności prognozowania uzyskanego modelu. Pamiętać musimy, że wartości te odnoszą się do wielkich danych. Tutaj korelacja oznacza istnienie pewnej relacji w znacznie większym stopniu, niż w przypadku danych klasycznych. Jeżeli jego wartość zawarta jest w przedziale 0.5-1.0, świadczy to o znacznej zdolności prognozowania (przewidywalności) modelu. Korelacje pomiędzy danymi ekonomicznymi a różnymi deskryptorami molekularnymi są stosunkowo niskie. Można wyróżnić tylko jedną wysoką wartość współczynnika korelacji obliczonego dla ceny MBM i masy R= 0.474 (tabela 4.).

Tabela 4. Macierz korelacji pomiędzy wybranymi deskryptorami molekularnymi a ceną wagową- WBM i molową- MBM dla danych Abamachem.

	P1	P2	MW	AC	C	H	N	SAS1
P1	1	0.857	-0.033	0.171	-0.105	0.177	0.216	0.341

P2	0.857	1	0.474	0.045	-0.108	0.005	0.096	0.239
MW	-0.033	**0.474**	1	-0.213	-0.034	-0.302	-0.187	-0.116
AC	0.171	0.045	-0.213	1	0.598	0.983	-0.102	0.380
C	-0.105	-0.108	-0.034	0.598	1	0.542	-0.446	0.081
H	0.177	0.005	-0.302	0.983	0.542	1	-0.093	0.378
N	0.216	0.096	-0.187	-0.102	-0.446	-0.098	1	0.031
SAS1	0.341	0.239	-0.116	0.380	0.081	0.378	0.031	1

gdzie:

P1 - cena wagowa WBM [$/g];

P2 - cena molowa MBM [$/mol];

MW- masa cząsteczkowa;

AC - liczba atomów;

C - liczba atomów C;

H - liczba atomów H;

N - liczba atomów N;

SAS1- dostępność syntetyczna.

Deskryptory molekularne, takie jak: masa cząsteczkowa oraz liczby atomów węgla, wodoru i azotu obliczono na podstawie struktury cząsteczki przy użyciu programu MATLAB, natomiast syntetyczną dostępność związków (ang. synthetic accessibility score) obliczono za pomocą programu SYLVIA-XT 1.4 (Molecular Networks, Erlangen, Germany).

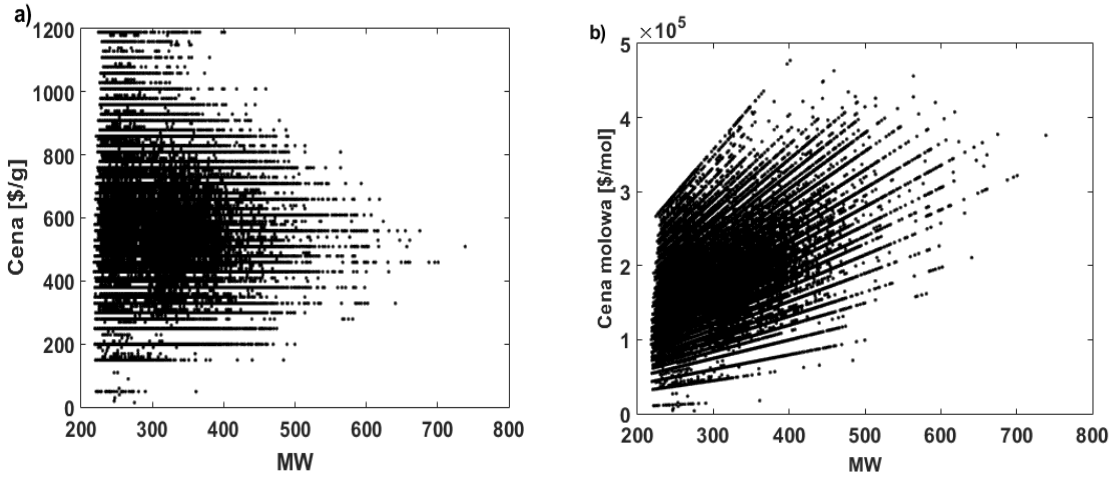
Ważnym elementem bez którego nie byłyby możliwe obliczenie stanowiło przetworzenie bazy (katalogu) Abamachem do postaci rozpoznawanej przez stosowane przeze mnie programy. To bardzo istotny element obliczeń typu big data. Liczba danych decyduje, że nie da się przeprowadzić tego elementu ręcznie. Wszystkie poprawki należy przeprowadzić poprzez odpowiednie procedury softwarowe. Stosuje się tu wiele procedur, np. „data cleansing". W gruncie rzeczy procedury te są operacjami technicznymi i nie będą tutaj przeze mnie

omawiane dalej. Stanowiły one jednak istotny element warunkujący możliwości przetwarzania danych molekularnych o wymiarowości rzędu 10^6.

6. Badanie zależności między wybranymi deskryptorami molekularnymi a ceną

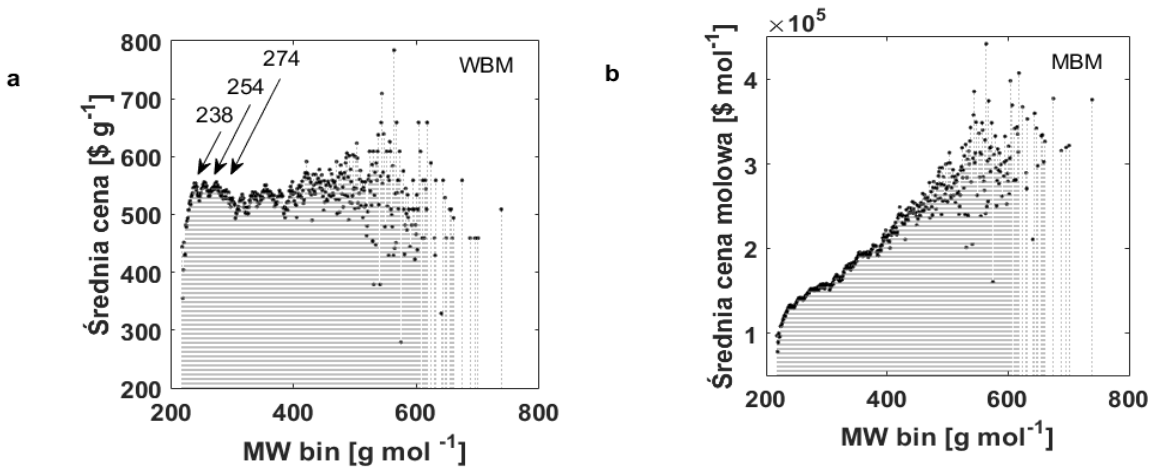
6.1. Wpływ masy cząsteczkowej na cenę cząsteczki

Na rysunku 30. zilustrowano statystykę molekularną zależności ceny wagowej [$/g] oraz ceny molowej [$/mol] w stosunku od masy cząsteczkowej w badanym katalogu Abamachem. Relacje przedstawione na rysunku 30. określają zachowanie się pełnej biblioteki związków chemicznych, w których zamiast pojedynczego modelu można zidentyfikować serię linii. Cena wagowa [$/g] tworzy serię poziomych linii w ramach każdej indywidualnej ceny, co sugeruje, że cena wagowa, która jest właściwością ekonomiczną w określonych seriach związków nie zależy od masy cząsteczkowej (rysunek 30a), podczas gdy cena molowa [$/mol] tworzy serię modeli liniowych, co może wskazywać, że cena molowa zależy liniowo od MW (rysunek 30b)[98].



Rysunek 30. Analiza właściwości ekonomicznych - cen względem MW dla 2.2 mln związków chemicznych: (a) cena wagowa [$/g] vs. MW (b) cena molowa [$/mol] vs. MW.

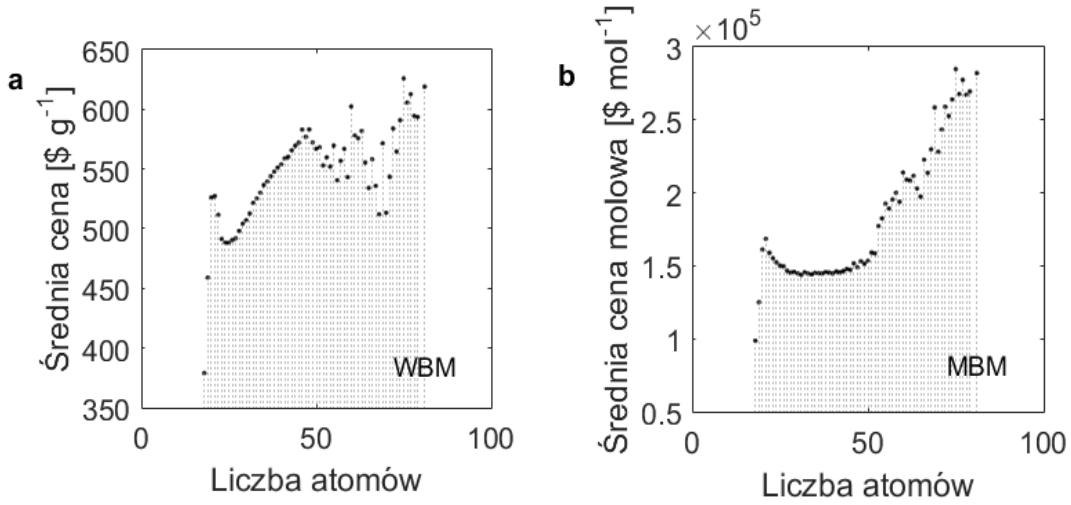
Na rysunkach 31a, zilustrowano statystykę molekularną binowanych mas cząsteczkowych, z której wynika, że w ekonomii średnia cena związku znormalizowana w postaci miary [$/g] jest w dużym stopniu niezmienna w szerokim zakresie MW. Świadczy to o tym, że cena w mierze WBM-$/g jest średnio stała dla stałej masy i formalnie taki typ relacji ceny i MW oznacza brak zależności ceny od MW. Wyjątek stanowią niskie zakresy MW, które wykazują wyraźny trend spadkowy. Natomiast na rysunku 31b przedstawiono analizę średniej ceny molowej [$/mol] w stosunku do binowanej MW, z której wynika, że w badanym katalogu wraz ze wzrostem MW od ok. 220 Da do 600 Da średnia cena molowa rośnie w przedziale od $1*10^5$ do ok. $3.5*10^5$ $/mol. Ponadto z analizy rysunku 31b wynika, że średnia cena związku normalizowanego w postaci miary molowej jest w przybliżeniu liniową funkcją MW. Oznacza to, że cena zależy liniowo od molarności substancji oraz że średnio im większa jest waga sprzedawanej substancji, tym większa jest cena. W ramach zbliżonych produktów, które moglibyśmy w chemii określić jako kongeneryczne wydaje się to dobrze oddawać zasady ekonomii[97]. Zależność liniową ceny molowej vs. MW można także wytłumaczyć w następujący sposób. Wraz ze wzrostem MW rośnie ilość materii (substancji), którą musimy kupić, aby otrzymać 1 mol substancji. Na przykład, kupując 1 mol benzenu kupujemy(otrzymujemy) 78 g, podczas gdy, kupując 1mol annulenu trzy razy tyle - 3*78 g (rysunek 40). Współczynnik korelacji dla średniej ceny molowej a binowaną masą wynosi R_{bin}=0.93. Wysoka wartość korelacji wskazuje, że na rynku cząsteczek płaci się za *ilość materii*. Liniowa zależność ceny [$/mol] oznacza także, że cena w skali wagowej (WBM) nie zależy od MW. Ciekawym efektem jest występowanie niewielkich maksimów średnich cen WBM (Rysunek 31a).



Rysunek 31. Analiza średniej ceny (a) WBM [$/g] (b) MBM [$/mol] w stosunku do wartości binowanych mas cząsteczkowych MW dla Abamachem.

6.2. Wpływ liczby atomów na cenę cząsteczki

Wzory cząsteczkowe przedstawiane są przy pomocy struktur atomowych (węgiel i heteroatomy). Można przypuszczać, że prosty wzór cząsteczkowy i liczba atomów ma wpływ na cenę cząsteczki już poprzez sam fakt, że to właśnie taką strukturę klient ogląda w katalogu. Z ekonomicznego punktu widzenia bardziej atrakcyjne wydają się cząsteczki o wyższej liczbie atomów. Rzeczywiście taki efekt uwidacznia się na rysunku 32., który przedstawia analizę średnich cen z uwzględnieniem miar WBM- $/g i MBM- $/mol w stosunku do całkowitej liczby atomów dla badanej biblioteki związków Abamachem. Z rysunku 32a wynika, że wraz ze wzrostem liczby atomów w przedziale od ok. 30 do 60 cena cząsteczki wzrasta od ok. 500 do 550 $/g. Powyżej 60 atomów obserwuje się dużą różnicę między danymi i nie można w jednoznaczny sposób określić trendów cen. Rysunek 32b prezentujący rozkład średniej ceny molowej MBM w stosunku do całkowitej liczby atomów wskazuje najwięcej cząsteczek posiada liczbę atomów mieszczącą się w przedziale od ok. 30 do 60 atomów. W tym przedziale zależność średniej ceny od liczby atomów AC ma charakter ciągły dla MBM, jest to wręcz wartość zbliżona do stałej, tj. ok. $1.5*10^5$ $/mol. Powyżej 60 atomów obserwuje się wzrost średnich cen molowych w zakresie od ok. $1.7*10^5$ do $2.7*10^5$ $/mol.

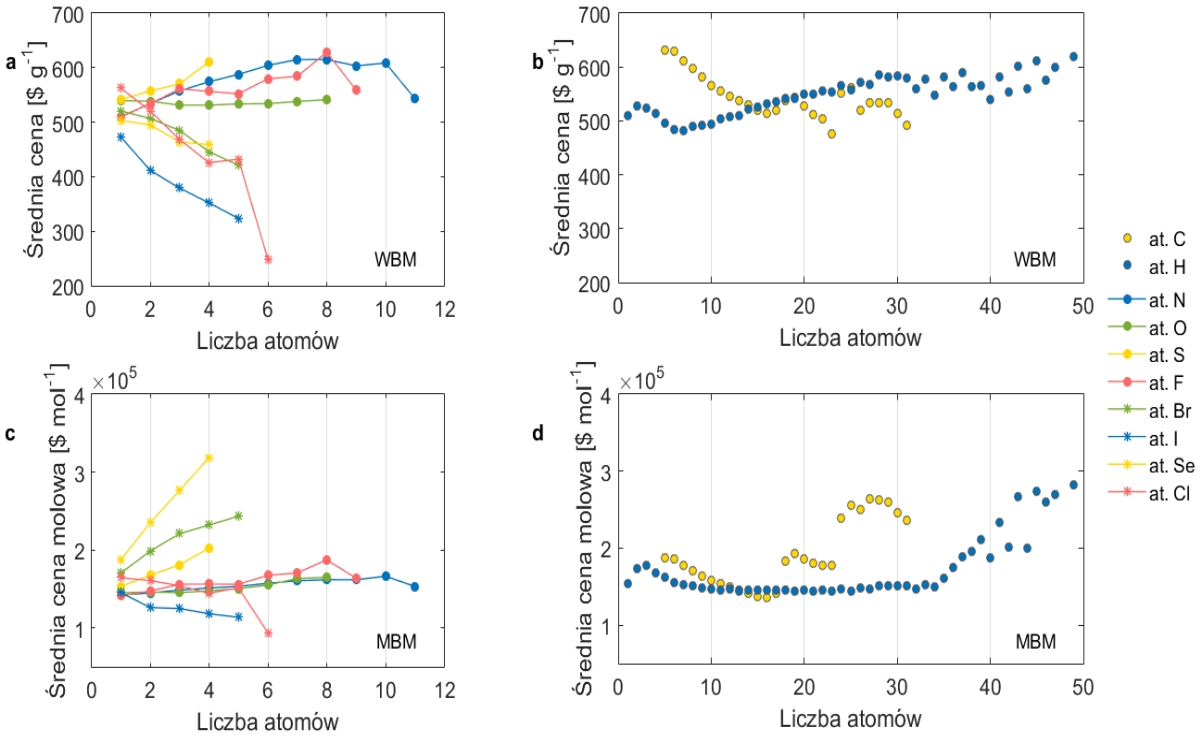


Rysunek 32. Statystyczny rozkład średnich cen związków (a) WBM i (b) MBM w stosunku do całkowitej liczby atomów.

6.3. Wpływ składu pierwiastkowego na cenę cząsteczki

Na rysunku 33. przedstawiono wpływ składu chemicznego, ilości atomów węgla i wodoru oraz heteroatomów na cenę substancji normalizowanej WBM- $/g i MBM- $/mol dla danych z katalogu Abamachem. Interpretując rysunki 33a i 33c opisujące statystyczny rozkład heteroatomów wchodzących w skład związków można dojść do wniosku, że podział heteroatomów dokonuje się na dwa rodzaje. Pierwszy z nich zawiera takie heteroatomy, jak: azot, tlen, siarka czy fluor, których cechą szczególną jest to, że ich obecność prowadzi do wzrostu ceny cząsteczki. Poza tym są one nieco droższe od takich heteroatomów, jak: chlor, jod czy brom. Powyższy podział można wytłumaczyć, na dwa sposoby. Po pierwsze, w trakcie syntezy takie atomy jak siarka czy tlen są najczęściej „wbudowane w syntetyzowane struktury" związków, zaś chlor, jod czy brom są zwykle wykorzystywane jako reaktywne grupy funkcyjne. Mimo że ich obecność jest absolutnie wymagana, to ostatecznie zostają one często usunięte podczas reakcji chemicznej. Innymi słowy *ekonomia atomowa* sprzyja temu pierwszemu rodzajowi atomów, podczas gdy ten drugi stanowi balast syntezy, obniżając jej *ekonomię atomową*. Na rysunkach 33b i 33d przedstawiono statystyczny rozkład ilości atomów węgla i wodoru w badanej bibliotece związków.

Z analizy rysunku 33b i 33d wynika, że wraz ze wzrostem liczby atomów węgla w cząsteczce cena związku wyrażona WBM maleje, a MBM rośnie.



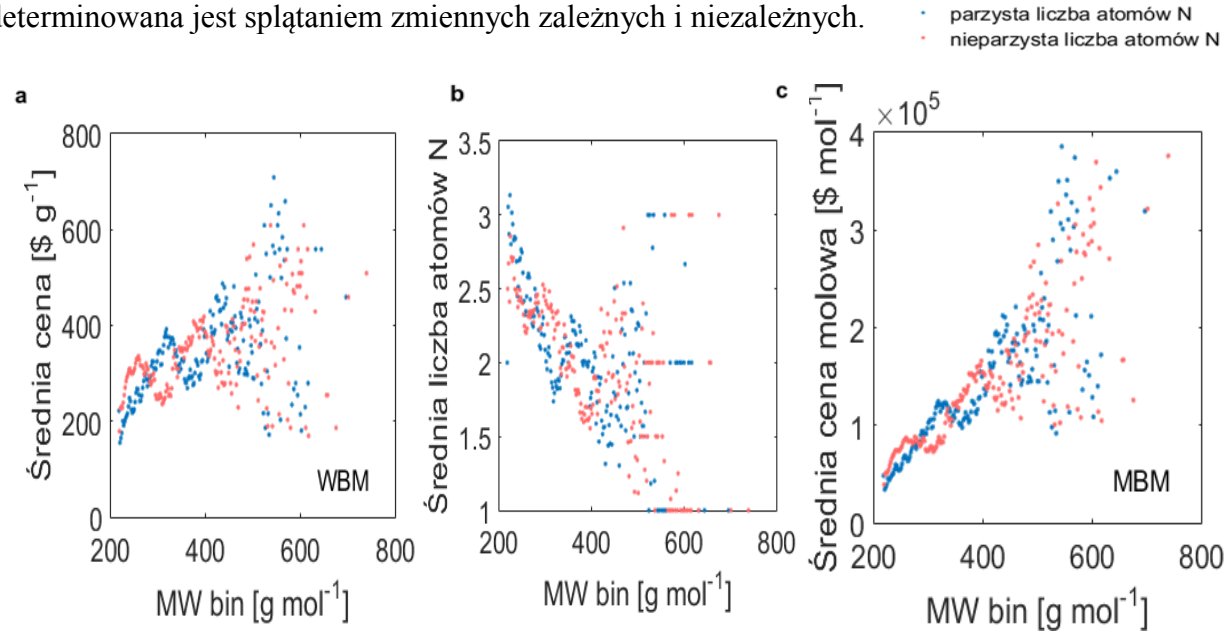
Rysunek 33. Porównywanie średnich cen WBM i MBM związków z uwzględnieniem statystycznego rozkładu atomu węgla, wodoru i heteroatomów w badanym katalogu Abamachem.

7.Reguła azotowa

Reguła azotowa w spektroskopii mas dla jonów nieparzystoelektronowych powstających głównie w EI mówi, że parzysta liczba atomów azotu odpowiada parzystej masie cząsteczkowej, natomiast nieparzysta liczba atomów azotu odpowiada nieparzystej masie cząsteczkowej. Natomiast ta sama reguła dla jonów parzysto-elektronowych powstających głównie w ESI, APCI i MALDI decyduje o tym, że parzysta liczba atomów azotu odpowiada nieparzystej masie cząsteczkowej, a nieparzysta liczba atomów azotu odpowiada parzystej masie cząsteczkowej.

Jony nieparzystoelektronowe to $M^{+\cdot}$, a jony parzystoelektronowe to $M+H^+$, $M-H^+$, $M+Na^+$, $M+Cl^-$ itp.[104].

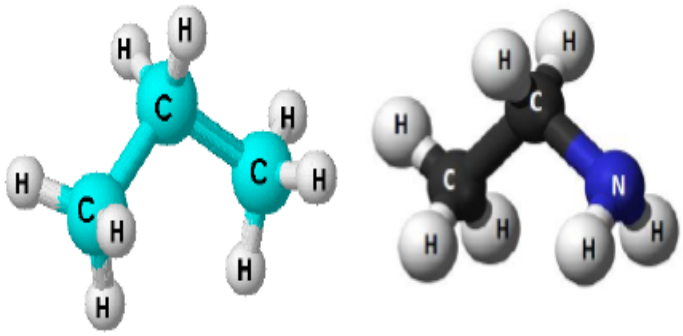
Z molekularnego punktu widzenia reguła azotowa porządkuje MW cząsteczek zawierających atomy azotu. Na rysunku 34. przedstawiono ceny WBM i MBM odpowiednio dla cząsteczek z parzystą i nieparzystą liczbą atomów azotu. Na rysunkach 34a i 34c można zaobserwować interesujący efekt odpowiadający regule azotowej. Parzyste i nieparzyste cząsteczki przedstawiane są przez oddzielne krzywe cenowe. Ponadto średnio droższe są zawsze te cząsteczki, w których jest mniej atomów azotu (rysunek 34b). Taki spadek liczby atomów azotu obserwuje się w przedziale masy od ok. 220 do 400 Da. Powyżej 400 Da zależność nabiera charakteru bardziej chaotycznego. Nie jest jasne, czy obserwowana „reguła azotowa" jest wyrazem jakiejś prawidłowości elementarnej lub chemicznej, czy też, co wydaje się bardziej prawdopodobne, determinowana jest splątaniem zmiennych zależnych i niezależnych.



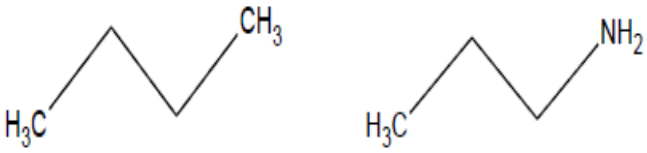
Rysunek 34. Cena jako funkcja liczby atomu azotu (a) WBM vs. MW bin (b) średnia liczba atomów azotu vs. MW bin (c) MBM vs. MW bin.

Jednym z wyjaśnień obserwowanej reguły azotowej może być fakt, że koordynacja atomów węgla i azotu w cząsteczkach chemicznych jest różna. Jeśli cząsteczka zawiera atomy azotu, to zmniejsza się całkowita liczba atomów węgla i rośnie ciężar cząsteczkowy związku (C=12

N=14). Fakt ten decyduje o tym, że analogiczne związki z azotem przesuwają się ku wyższym wartościom MW, co przedstawia rysunek 35. porównujący analogi węglowe z analogami azotowymi.



Liczba atomów (ang. atom count)	AC	11	10
Masa cząsteczkowa (ang. molecular weight)	MW	44,40 g/mol	45,08 g/mol



Liczba atomów (ang. atom count)	AC	14	13
Masa cząsteczkowa (ang. molcular weight)	MW	58,12 g/mol	59,11 g/mol

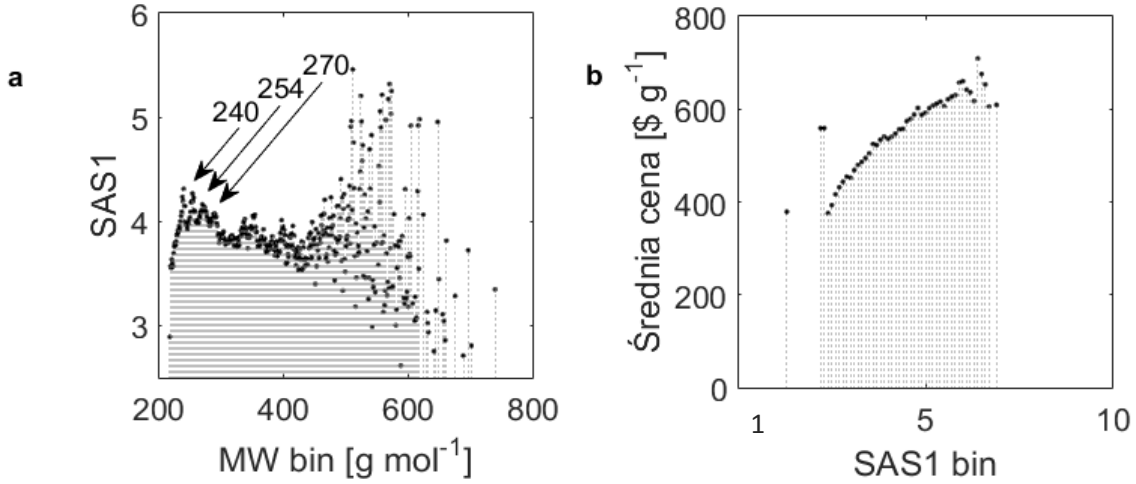
Rysunek 35. Zależność między MW a AC dla grafów związków C→N.

8. Wpływ dostępności syntetycznej na cenę biblioteki Abamachem

Badania i rozwój rynku farmaceutycznego wymagają szybkiej syntezy potencjalnie nowych leków (drug candidates). Systemy projektowania mogą łatwo tworzyć wirtualne biblioteki struktur *de novo*, jednakże to chemicy muszą następnie te struktury syntetyzować, co nie zawsze jest zadaniem łatwym. Dlatego narzędzia obliczeniowe/komputerowe wspomagające projektowanie syntez są wykorzystywane przez chemików, ponieważ przyczyniają się do skrócenia czasu wstępnego etapu projektowania nowych związków oraz do obniżenia jego kosztów. Narzędzia obliczeniowe/komputerowe wspomagające projektowanie syntez to na przykład deskryptory dostępności syntetycznej SAS związku lub wskaźniki złożoności cząsteczek organicznych[99].

W przeprowadzonych badaniach przeanalizowano wpływ syntetycznej dostępności na cenę cząsteczki dla biblioteki Abamachem. Na rysunku 36. przedstawiono wpływ syntetycznej dostępności na cenę cząsteczki dla związków organicznych katalogu Abamachem. Z interpretacji rysunku 36a prezentującego zależność średniej syntetycznej dostępności w stosunku do binowanej masy cząsteczkowej wynika, że wraz ze wzrostem MW w przedziale ok. 200 Da do 400 Da deskryptory SAS1 mieszczą się w zakresie od 3.5 do 4.5. Średnio syntetyczna dostępność związków biblioteki Abamachem przyjmuje spodziewaną średnia wartość, lokując się w pobliżu średniej między 1-10, która wynosi 5. Powyżej 400 Da zależność nabiera charakteru chaotycznego, ze względu na małą reprezentatywność związków w tym zakresie masowym. Powyżej 400 Da wynik binowania SAS1 jest chaotyczny. Oznacza to, że wykres binowanych wartości nie ma charakteru ciągłego. Na rysunku 31b przedstawiono zależność średniej ceny $/g w stosunku do binowanej syntetycznej dostępności SAS1 bin. Z analizy rysunku 36b wynika, że dla badanej populacji związków średnia cena $/g rośnie wraz ze wzrostem SAS1. Ponadto obliczony współczynnik korelacji między średnią ceną $/g a binowaną syntetyczną dostępnością SAS1 bin wynosi R_{bin}=0.925. Wysoka korelacja świadczy o tym, że *chemia* odgrywa ważną rolę w ustalaniu ceny. Wraz ze wzrostem ceny rośnie SAS1. Nie ma natomiast wyraźniej korelacji między SAS1 a MW (obserwujemy niewielki spadek

SAS1 ze wzrostem MW).Oznacza, to że średnio w każdej grupie związków można spotkać takie, których synteza jest trudna oraz takich, która syntez jest łatwiejsza (rysunek 36a).



Rysunek 36. Analiza obliczonego deskryptora (a) dostępności syntetycznej SAS1 vs. MW bin (b) średniej ceny $/g vs. SAS1 bin.

9. Ewaluacja statystyczna uzyskanych modeli struktura-cena

Modelowanie statystyk molekularnych metodą biniowania wymaga statystycznej walidacji. W badaniach zastosowano metodę randomizacji zmiennej zależnej y. Metody weryfikacji statystycznej modeli wielkich danych stanowią na przykład metody y- randomizacji (ang. y-Randomizatiom) lub walidacji krzyżowej (ang. Cross-validation). W przeprowadzonych badaniach zastosowano metodę y- randomizacji. Metody y- randomizacji i walidacji krzyżowej stosowane są na przykład do sprawdzenia poprawności i wiarygodności modeli QSPR/QSAR[105]. Natomiast do modelowania QSPR/QSAR wykorzystuje się częściej metody regresji liniowej i wielokrotnej regresji liniowej (MLR ang. multilinear regression) czy metodę PLS (ang. partial least square)[106]. Następnie przeprowadza się walidację wewnętrzną i zewnętrzną w celu dokonania oceny stabilności modelu i zdolności jego prognozowania.

9.1. Metoda Y-randomizacji

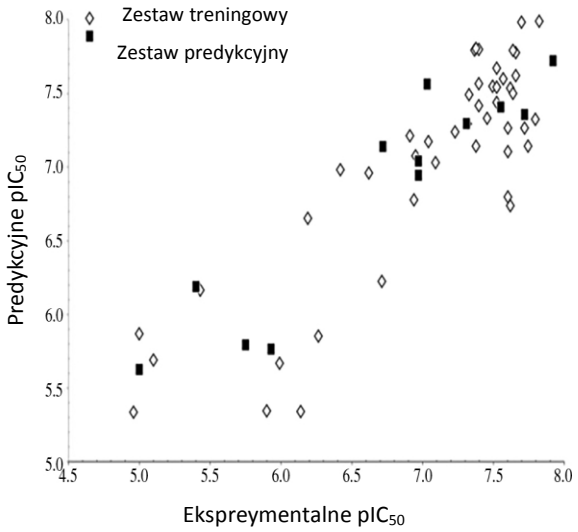
Metoda y-randomizacji polega na badaniu korelacji między zmienną zależną (Y) a zmienną predykcyjną (X), która jest deskryptorem obliczanym na podstawie struktury molekularnej. Walidację statystyczną modelu należy rozpocząć od przygotowania zbioru danych. Kolejnym etapem jest podział zbioru danych na zestaw treningowy używany do budowania modelu rzeczywistego i zbioru, dla którego wartości są prognozowane. Podział na zbiory danych zależy od wielkości całego zbioru oraz celu badania. Y- randomizacja może składać się z kilku etapów, dla których pierwotna macierz deskryptorów X jest stała, a tylko wektor Y jest losowy. Powstaje pytanie, w jaki sposób należy analizować wyniki randomizacji? Odróżnienie modelu rzeczywistego od modeli losowych można ocenić na podstawie odpowiedniego poziomu ufności dla rozkładu normalnego. Innym podejściem do tej kwestii jest obliczenie współczynnika korelacji Persona r między rzeczywistym wektorem y a losowym wektorem y. Analizuje się zatem dwa zbiory korelacji r - Q^2_{yrand} dla modeli losowych i r - R^2_{yrand} dla modeli rzeczywistych[105]. Ich zróżnicowanie pozwala na odróżnienie efektów losowych od rzeczywistych korelacji.

$$Q^2_{yrand} = a_Q + b_Q r \tag{2.1}$$

$$R^2_{yrand} = a_R + b_R r \tag{2.2}$$

Model rzeczywisty charakteryzuje się tym, że jest wolny od korelacji przypadkowej wówczas, gdy punkty przecięcia mają wartości $a_Q < 0{,}05$ i $a_R < 0{,}3$, a korelacja między rzeczywistymi i losowymi wektorami y jest równa zero (r =0).

Zestaw przykładowych analiz wykorzystujących metodę y-randomizację opisano w załączniku **1**.



Rysunek 37. Schemat y- randomizacji dla zestawu danych QSAR składających się z wartości pIC_{50} dla 59 związków według[107]*. Widać, że odróżnienie efektów losowych od rzeczywistych korelacji jest trudne, co świadczy, że w dużym stopniu obserwowane relacje mają charakter losowy.*

gdzie:

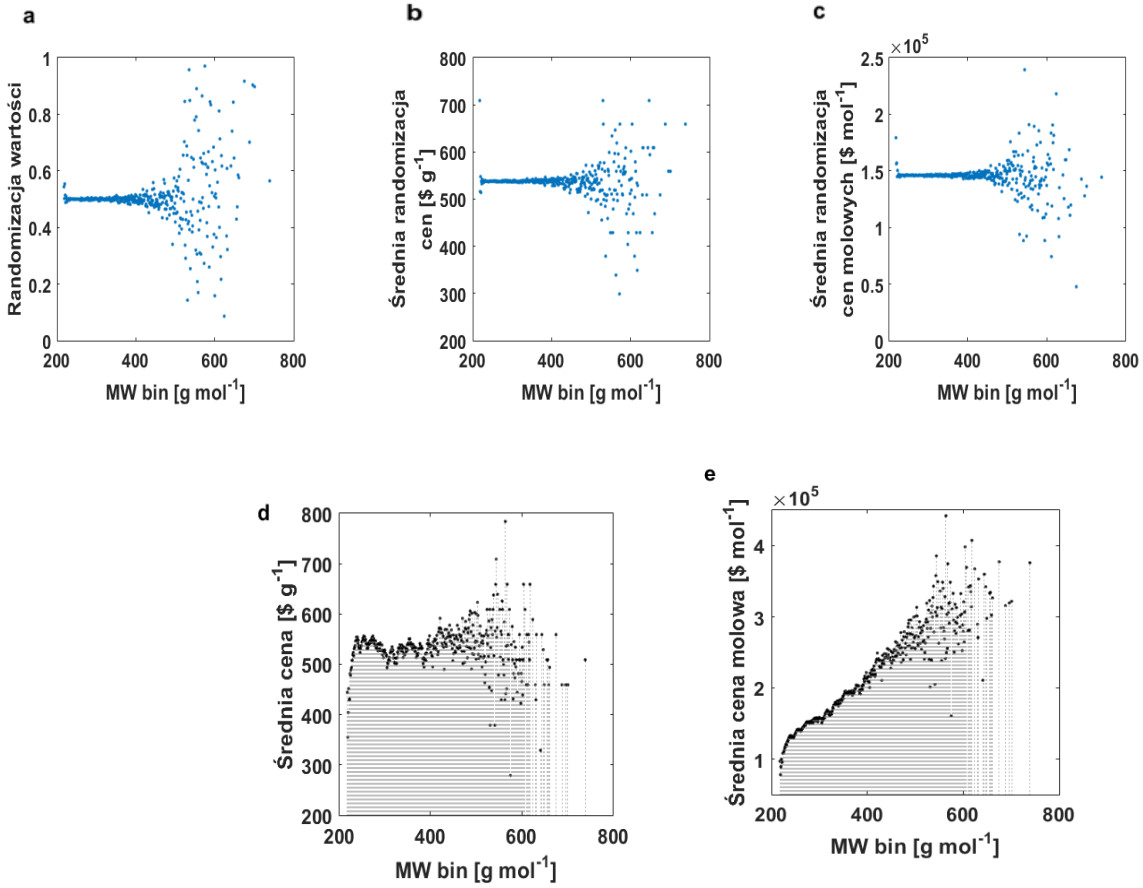
a) porównanie wartości pIC_{50} zestawu treningowego i predykcyjnego;

b) Y- randomizacja, gdzie czarny kwadrat przedstawia model rzeczywisty, a białe kwadraty – model losowy (kodowany).

Na rysunku 38. przedstawiono walidację statystyk molekularnych y- randomizacji dla zestawu danych katalogu Abamachem.

W przeprowadzonych badaniach w celu weryfikacji wiarygodności statystycznej wykonano randomizację wartości oraz cen $/g i $/mol w stosunku do MW bin. Przeprowadzona randomizacja polegała na tym, że dla 10 000 tys. losowych cząsteczek (ze zbioru Abamachem) przypisane zostały randomizowane (losowe) wartości z zakresu 0-1 (rysunek 38a) lub randomizowane (losowo przypisane) ceny $/g i $/mol (rysunek 38b,c). Przeprowadzona analiza wykazuje, że w przypadkach randomizacji wartości cen $/g i $/mol w obszarze, gdzie populacja cząsteczek jest duża [220-400 g/mol] uzyskuje się model, w których średnio ceny $/g i $/mol nie zależą od MW. Tam, gdzie reprezentatywność statystyczna jest mniejsza, model ma

charakter chaotyczny MW>400 g/mol. Porównując rozkład rzeczywistych średnich cen $/g i $/mol (rysunek 3d,e), z modelami randomizowanymi (rysunek 38b,c) obserwuje się, że modele rzeczywistych cen odbiegają od modeli randomizowanych.



Rysunek 38. Analiza statystyczna dla katalogu Abamachem – (a) randomizacja wartości vs. MW bin (b) randomizacja cen $/g vs. MW bin (c) randomizacja cen $/mol vs. MW bin (d) średnia cena $/g vs. MW bin (e) średnia cena $/mol vs. MW bin.

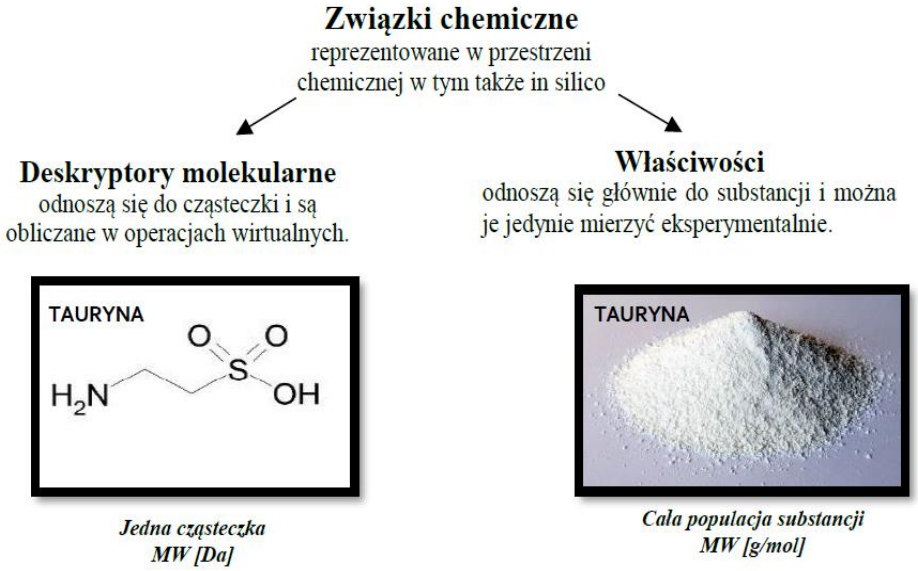
10. Interpretacja uzyskanych wyników

Jednym z pytań, które możemy zadać w aspekcie molekularnych czynników *ekonomii atomowej* jest pytanie, czy płacimy za liczbę cząsteczek, czy ilość materii. Pojęcie związku chemicznego nie jest precyzyjnie definiowane przez IUPAC[2]. W chemii klasycznej chemicy poprzez związek chemiczny intuicyjnie najczęściej rozumieją cząsteczkę (molekułę) składają

się z co najmniej dwóch różnych atomów oraz odpowiadającą jej substancję lub (odmianę chemiczną), złożoną z kilku tych samych cząsteczek (molekuł), które są w ten czy w inny sposób mierzalne. Formalnie jednak związkiem chemicznym określić można zarówno pojedynczą cząsteczkę, jak i substancję, która jest zbiorem takich cząsteczek „in vitro”[1]. Ważnym aspektem w analizowaniu problemu związku chemicznego jest jego opis dokonany przez mierzalne właściwości (substancja) lub obliczeniowe deskryptory (cząsteczka).

10.1. MW deskryptor molekularny czy właściwość

Cząsteczka jest powtarzalnym elementem badań chemicznych. Jej struktura i transformacje są obiektem zainteresowania chemików. Związki chemiczne reprezentowane są w przestrzeni chemicznej, w tym także w postaci in silico, przez dwa typy parametrów. Pierwsze z nich to deskryptory molekularne, drugie zaś to właściwości. Deskryptory molekularne odnoszą się do cząsteczki i są obliczane w operacjach wirtualnych, natomiast właściwości odnoszą się głównie do substancji i można je jedynie mierzyć eksperymentalnie (Rysunek 39.). Często odróżnienie deskryptorów i właściwości nie jest proste. Niektóre parametry, jak na przykład masa cząsteczkowa MW, mogą być zarówno deskryptorem, jak i właściwością. Masa cząsteczkowa MW może być właściwością, jeśli jest mierzona dla jednej cząsteczki, na przykład w spektrometrii mas MS. Natomiast MW może być deskryptorem w wyniku zsumowania udziału mas poszczególnych atomów w łączną MW dla pojedynczej cząsteczki. Ciężarem 1 mola substancji jest jego MW [g/mol], podczas gdy masą pojedynczej cząsteczki - MW [Da]. Korelacja między tymi dwiema zmiennymi wynosi 100% wówczas, gdy odwzorowujemy zbiór substancji w zbiór cząsteczek i odwrotnie. Formalnie transformacja ta wymaga liczby Avogadro (N_A), MW [Da] * N_A = MW [g / mol][98]. W chemii często zapomina się o tej różnicy między miarą Daltona versus g/mol.

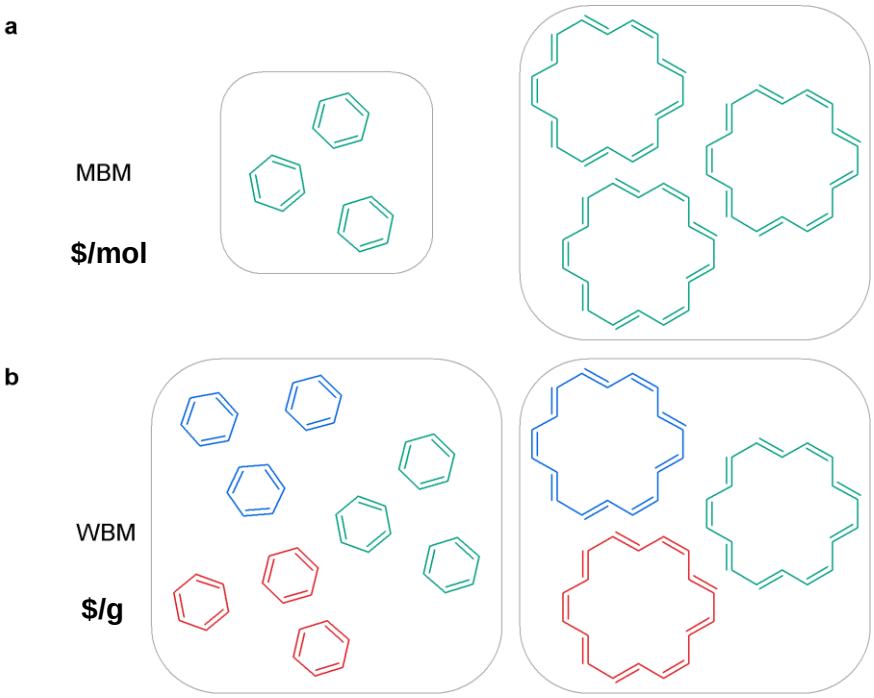


Rysunek 39. Deskryptor molekularny cząsteczki a właściwość substancji.

10.2. Masa cząsteczkowa jako miara złożoności cząsteczki

Obiektem zainteresowania chemików są właściwości cząsteczki chemicznej oraz jej przemiany. Często przyjmuje się, że wraz ze wzrostem cząsteczki rośnie jej złożoność. W takim ujęciu masa cząsteczkowa MW może być postrzegana jako najprostsza, szeroka miara złożoności cząsteczkowej[108].

Na rynku związków chemicznych stosuje się wagową miarę ilości związku WBM- $/g. Taka miara jest typowa dla rynku, na którym większość towarów wycenia się w $/kg. Inną miarą jest molowa miara ceny - MBM- np. $/mol; miary molowe powszechnie stosowane są w chemii. Na rysunku 40. zilustrowano porównanie miary wagowej WBM i molowej MBM na przykładzie cząsteczki benzenu i anulenu, która jest dokładnie trzykrotnością masy benzenu. WBM normuje substancję przez stałą masę, MBM natomiast – przez stałą liczbę cząsteczek.



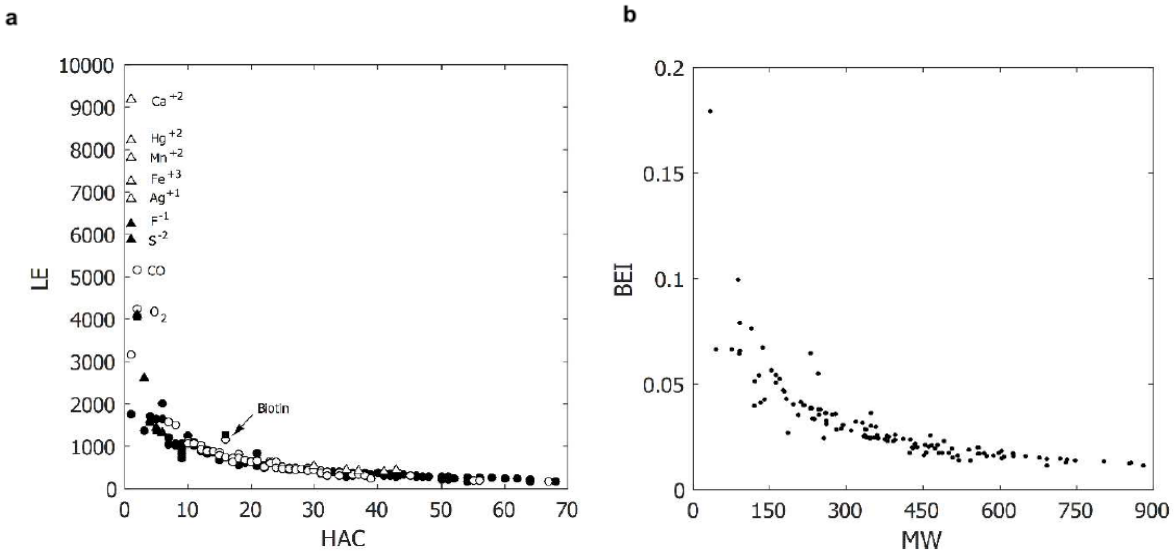
Rysunek 40. Porównanie miar (a) molowych MBM (b) wagowych WBM mas

gdzie :

- MBM- reprezentuje tę samą liczba cząsteczek,
- WBM – reprezentuje liczbę cząsteczek odwrotnie proporcjonalnych do MW.

11. Efekty hiperboliczne

Ciekawym efektem obserwowanym w zależności ceny WBM jest nieliniowość obserwowana dla niskich wartości MW. Efekt ten jest bardzo trudny do zrozumienia i wyjaśnienia. Warto jednak zauważyć, że podobne trudności interpretacyjne występują dla przebiegu wydajności ligandu od HAC lub MW (rysunek 41.)



Rysunek 41. Zależność LE vs. HAC and BEI vs. MW dla szeregu ligandów[84]. Rysunek (b) dane przeliczone na BEI wg.[84], według publikacji własnej[109].

Warto także zwrócić uwagę na analogię między definicją ceny WBM a LE (BEI):

LE = (1.37*p$\mathbf{IC_{50}}$)/HAC

BEI= p$\mathbf{IC_{50}}$/MW

Cena WBM \$/g = (cena dla mola substancji)/MW

LE definiowane jest jako stosunek wiązania ligandu do liczby atomów ciężkich:

$$\mathrm{LE} = \frac{\mathrm{P}}{\mathrm{HAC}} \tag{2.3}$$

gdzie:

P - to właściwość wiążąca (dowolna właściwość mierzona w celu zdefiniowania interakcji pomiędzy ligandem a receptorem),

HAC- to liczba atomów ciężkich lub HC (liczba atomów wodoru), AC (całkowita liczba atomów) itp.

Natomiast BEI definiowane jest jako stosunek właściwości wiążącej (ang. binding property) lub właściwości wiążącej znormalizowanej molowo (ang. molar- normalized property) do masy cząsteczkowej związku:

$$\mathrm{BEI} = \frac{\mathrm{P}}{\mathrm{MW}} \qquad \mathrm{BEI} = \frac{\mathrm{P_{mol}}}{\mathrm{MW}} \tag{2.4}$$

gdzie:

P – to właściwość wiążąca (dowolna właściwość mierzona w celu zdefiniowania interakcji pomiędzy ligandem a receptorem),

Pmol- odnosi się do właściwości znormalizowanej molowo, np. powinowactwa wiązania molowego.

MW- masa cząsteczkową w Daltonach [Da].

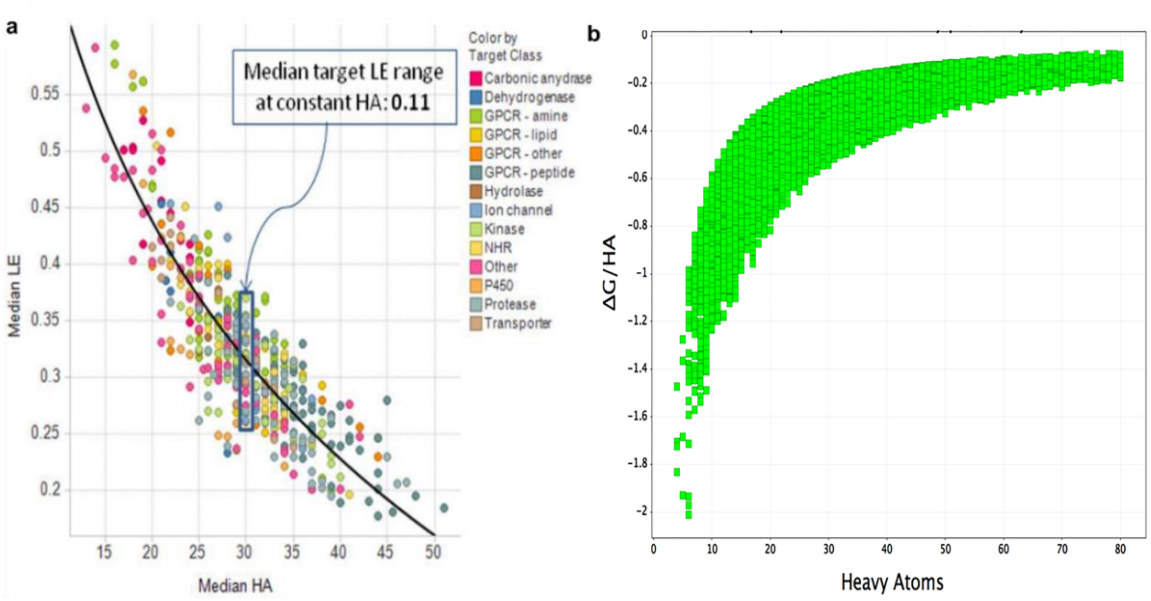
Można poszerzyć definicję LE obejmującą dowolną właściwość znormalizowaną molowo P_{mol} lub deskryptor molekularny (MD). W ten sposób definiujemy parametr efektywności PE (ang. efficiency parametr), dla którego dowolna właściwość lub deskryptor molekularny jest znormalizowany względem: HAC, HC, AC:

$$\mathrm{PE_{HC}} = \frac{\mathrm{Pmol}}{\mathrm{HC}} \tag{2.5}$$

gdzie:

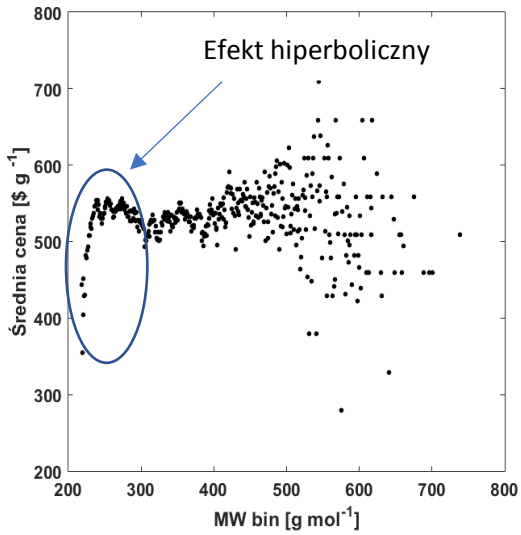
Pmol- odnosi się do właściwości znormalizowanej molowo, np. powinowactwa wiązania molowego.

Obserwowane zależności LE od HAC zawsze wykazują trend krzywoliniowy zbliżony do hiperboli np. (rysunek 42a)[52] oraz (rysunek 42b)[110]. Trend taki przez wiele lat był zagadkowy. Dopiero ostatnio zespół Polańskiego zaproponował wyjasnienie tego efektu efektem przypadkowej zmiany skali molowej do skali wagowej, jaką obserwuje się w przypadku oblicznia LE[98]. Efekty takie, określili oni jako hiperboliczne. Można je powszechnie obserwować w lieraturze na przykład modele pokazane na rysunku 42.



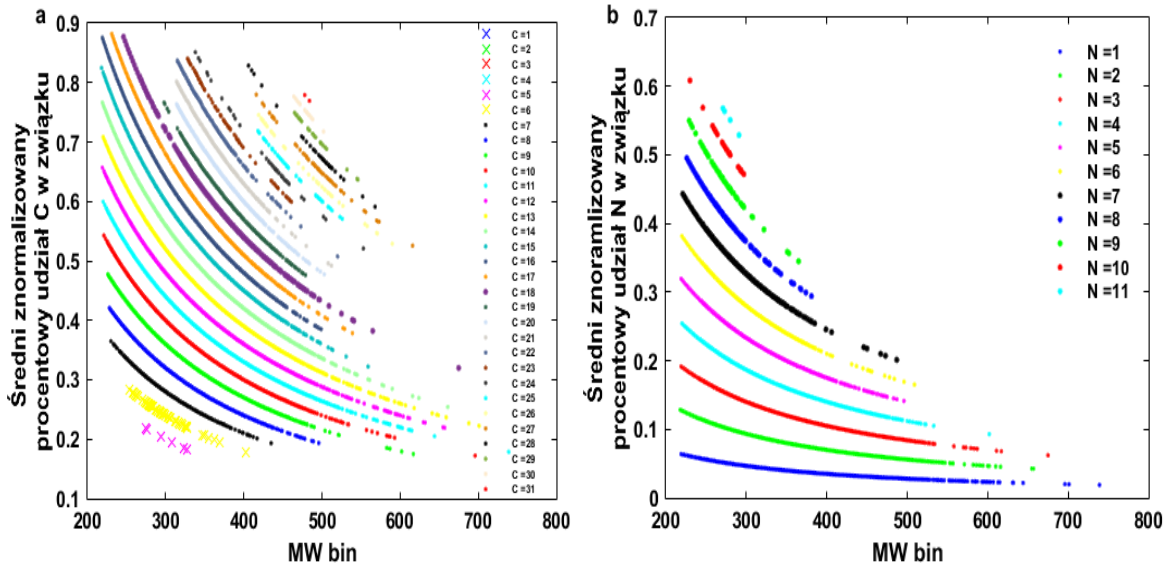
Rysunek 42. Efekt hiperboliczny miary LE, gdzie: LE=IC50/HAC vs. HAC[52,110].

Warto zauważyć, że efekt nieliniowości obserwowany dla niskich wartości MW cena $/g vs. MW (rysunek 43.) przypomina właśnie trend LE (BEI). Efekt ten można zinterpretować jako *trend hiperboliczny*, który może być obserwowany tylko w zakresie najniższych MW. Rodzi się pytanie, z czego wynika taki kształt zależności cena vs. MW?



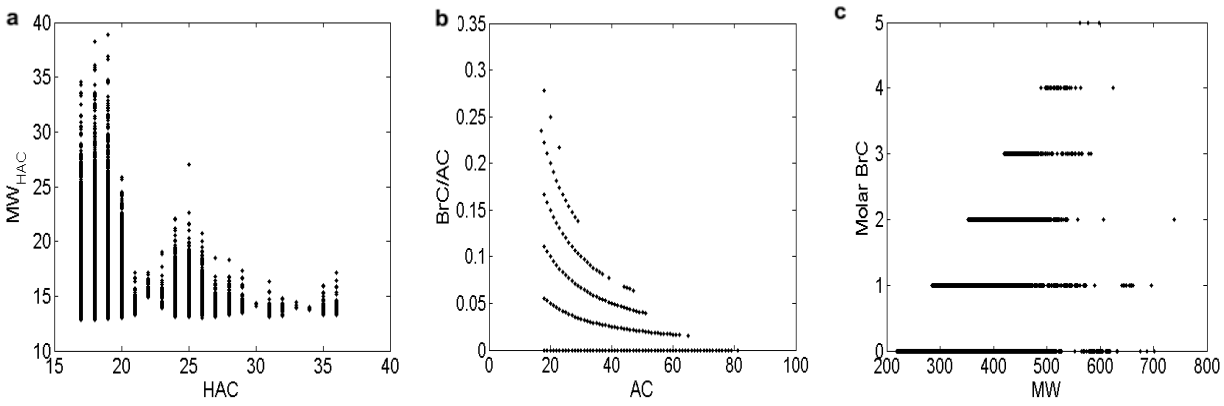
Rysunek 43. Zależność średnich cen [$/g] vs. MW bin dla związków chemicznych Abamachem.

Zależności przedstawione na rysunku 44. przedstawiają efekty hiperboliczne dla średniego znormalizowanego procentowego udział atomów węgla (rysunek 44a) i azotu (rysunek 44b) w związku chemicznym w stosunku do wartości binowanej masy cząsteczkowej w badanym katalogu Abamachem. Interesujące, że także w tym przypadku obserwujemy nieliniowe zależności. Tak więc efekty podobne do trendu LE obserwować można dla podstawowych układów właściwości/deskryptorów cząsteczek i substancji takich jak skład procentowy.



Rysunek 44. Średni znormalizowany procentowy udział (a) atomu węgla (b) atomu azotu w związku chemicznym w stosunku do wartości zbinowanej masy cząsteczkowej.

Na rysunkach 45a i 45b zaprezentowano zależność składu cząsteczkowego względem HAC, AC. Natomiast analiza rysunku 45c pozwala stwierdzić, że efekt hiperboliczny znika wówczas, gdy skład podaje się jako wartości molowe.



Rysunek 45. Zależności wybranych deskryptorów molekularnych (a) MW/HAC vs. HAC (b) Br/AC vs. AC (c) Liczba atomów bromu (molowo) vs. MW.

Porównując strukturę zależności LE, cena, procent masowy:

LE = (1.37*p**IC$_{50}$**)/HAC

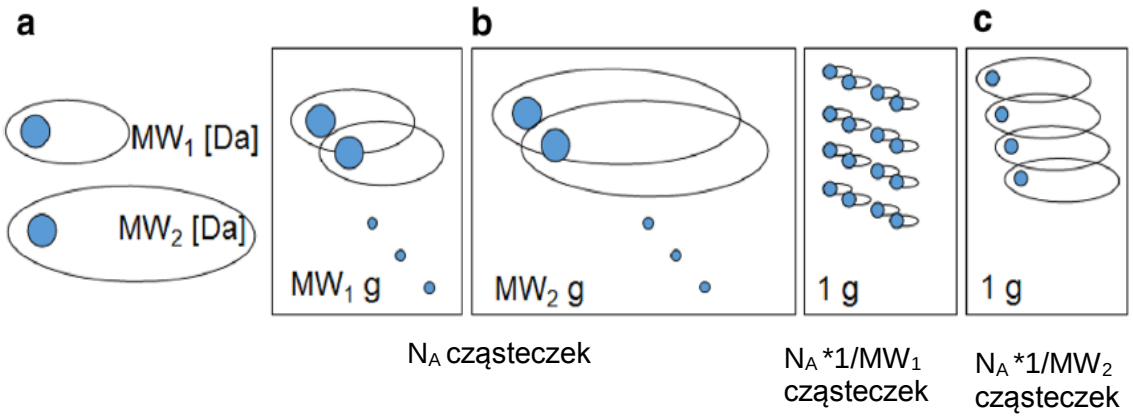
Cena w $/g = (cena mola substancji)/MW

Atomowy skład procentowy= udział **molowy** /MW

można zauważyć istotną analogię, która sprowadza się do tego, że w/w funkcje zawsze przekształcają relacje zależności molowych do zależności w skali wagowej. Wartości IC_{50} podaje się w wartościach stężenia molowego. Wprawdzie formalnie równanie definiujące LE posiada w mianowniku wartość HAC, a nie MW, jednak HAC jest blisko skorelowane z MW[98]. Przekształcenie skali molowej do wagowej wiąże się z występowaniem zniekształcenia trendu (por. np. krzywe dla Rysunek 45b) Br/AC vs. AC c) Molar BrC vs. MW).Taka deformacja trendu długo nie była wytłumaczalna. Wyjaśnienie tego zjawiska dla LE opisano w publikacji Polański et al.,[98]. Poniżej krótko omówiono molekularne uwarunkowania tego efektu.

Aktywność biologiczna IC_{50} (podana jako stężenie molowe) jest zawsze określana w skali molowej, a więc dla stałej liczby cząsteczek. Zmiana skali na wagową powoduje, że w jednostce masy mieści się różna liczba cząsteczek, która koreluje z MW, a ściślej - z 1/MW. Im mniejsze MW, tym większa liczba cząsteczek. Tak więc liczba ligandów, które mogą oddziaływać z receptorem rośnie dla małych MW (HAC) w stosunku 1/MW lub 1/HAC. Tak więc 1/MW

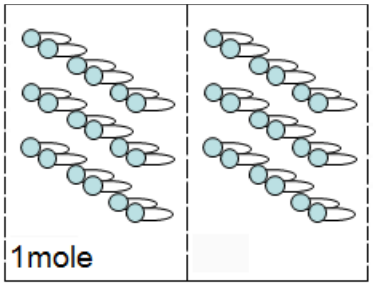
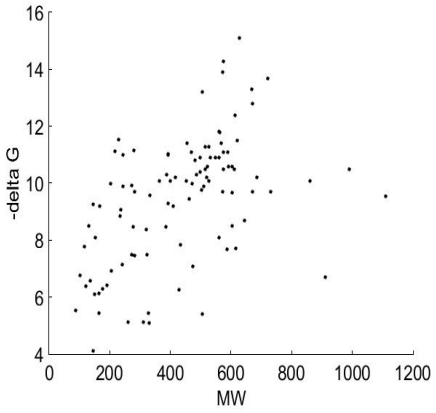
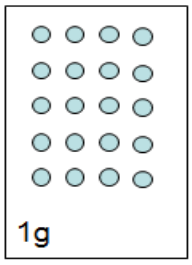
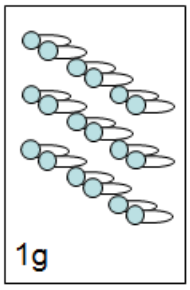
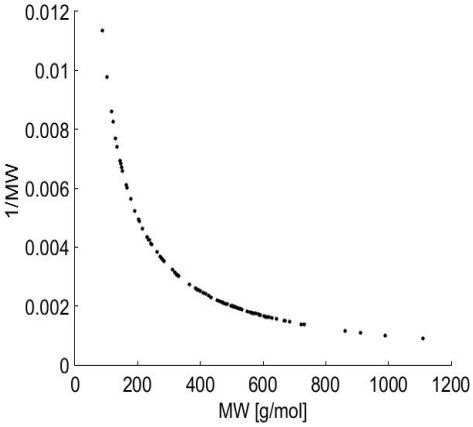
lub 1/HAC są źródłem hiperbolicznego odkształcenia trendu. Innymi słowy statystyka dla skali wagowej nie spełnia wymogów statystyki Avogadry. Schematycznie przedstawiono to na rysunku 46.

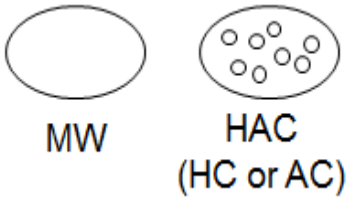
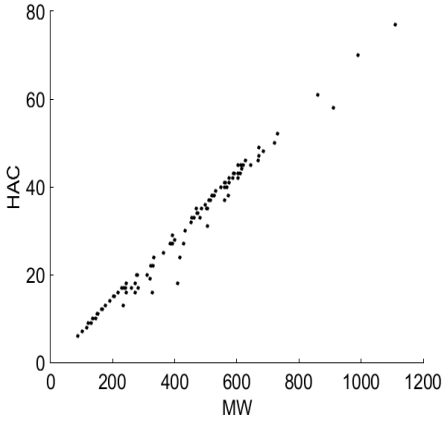


Rysunek 46. Schemat statystyki Avogadry dla skali molowej (b) oraz wagowej (c) dla dwóch różnych cząsteczek (a) według[98,111]*,*

gdzie: dwie cząsteczki MW_1 [Da] i MW_2 [Da] (a) można skalować za pomocą miar molowych do aglomerujących cząsteczek N_A substancji. Ciężar substancji wyniesie odpowiednio MW_1 [g / mol] i MW_2 [g / mol] (b) Alternatywnie, odwzorowanie skali wagowej (c) zachowuje stałą masę, np. 1g, wtedy liczba cząsteczek będzie za każdym razem różna, przyjmując wartość odpowiednio proporcjonalne do $1/MW_1$ i $1/MW_2$. Wirtualny fragment 1 [Da] oznaczono kolorem niebieskim.

Tabela 5. Znaczenie chemiczne i matematyczne LE według[98]

Lp.	Znaczenie chemiczne	Znaczenie matematyczne	Opis
1.	Właściwość molowa MW [g/mol]  Liczba N_A cząsteczek		Właściwość molowa vs. masa cząsteczki
2.	Populacja cząsteczek  Wirtualny fragment 1 [Da]  1 g substancji (1/MW*N_A)		**Deskryptor molekularny** - wirtualny fragment 1 [Da] **Właściwość**- 1g substancji **Deskryptor molekularny** Da $^{-1}$ **Właściwość** 1/MW [mol/g]

3.	Skalowanie przez MW lub HAC  	 	Zależność linowa

Wprawdzie hiperboliczny trend zależności cena vs. MW nie może zostać tak ławo wyjaśniony jak przebieg LE vs. MW lub HAC, lecz występujące w tym przypadku analogie wydają się także sugerować podobne wyjaśnienie.

Tak więc źródłem efektu hiperbolicznego jest matematyczna zależność **1/MW vs. MW** (rów. 2.6), której obrazem jest hiperbola (rysunek 47).

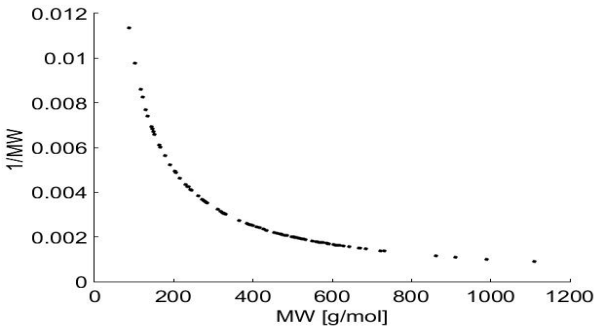
$$\text{ZH} = \text{X} * \frac{\mathbf{1}}{\boldsymbol{Y}} \boldsymbol{vs.Y} \qquad (2.6)$$

gdzie:

ZH – zależność hiperboliczna,

X - dowolna właściwość lub deskryptor (stężenie, aktywność, cena itp.),

Y – taka sama właściwość lub deskryptor (MW, HAC, AC itp.).



Rysunek 47. Zależność hiperboliczna 1/MW vs. MW[98].

Generalnie efekty hiperboliczne wynikają z dualności reprezentacji cząsteczek. Ponieważ w chemii stosujemy dwa typy modeli, pierwszy model dla cząsteczek (single molecule), a drugi model dla molowych reprezentacji cząsteczek (gdzie mol jest substancją, zbiorem cząsteczek). Efekty hiperboliczne wynikają właśnie z koincydencyjnej korelacji tych dwóch modeli. W przypadku aktywności biologicznej (IC_{50}, LE) stosunkowo łatwo wyjaśnić molekularne uwarunkowania takiej nieliniowości. W przypadku relacji ekonomicznych wyjaśnienie jest trudniejsze. Jednak analiza zależności struktura cena (SP: structure - price) na zasadzie analogia matematyczna między LE a ceną WBM sugeruje w tym wypadku podobną etiologię takiego efektu. To ona najprawdopodobniej decyduje o podobnym trendzie nieliniowym. Matematycznie w przypadku ceny idealna hiperbola powstaje w przypadku, kiedy cena molowa jest funkcją stałą MW MBM [$/g] = a (a=stała). W analizowanym przypadku danych Abmachem MBM w dużym obszarze jest jednak funkcją liniową MW: MBM [$/mol] =a *MW. Kiedy MBM [$/mol] = a *MW, to WBM [$/g] jest funkcją stałą WBM [$/g] =a. Relacje takie obserwować można na Rysunku 30.

Obserwowane zależności dobrze ilustrują fakt dominowania efektów ekonomicznym na rynku związków chemicznych, gdzie średnia cena jest funkcją ilości oferowanej *masy*. Efekt chemiczny powodujący występowanie trendu nieliniowego związany jest najprawdopodobniej ze skalą wagową. Efekty tego typu obserwuje się zawsze, w przypadku stosowania skali wagowej. Jego molekularnym uwarunkowaniem jest czynnik związany z odejściem od statystyki Avogadry. Ponieważ statystyka liczny cząsteczek w jednostce masy koreluje z 1/MW efekt taki często przybierać może formę hiperboli.

Podsumowanie i wnioski

Projektowanie in silico to projektowanie przy użyciu systemów obliczeniowych w wirtualnej przestrzeni komputera. Prace obliczeniowe oraz projektowanie in silico są znacznie tańsze w porównaniu od eksperymentów przeprowadzanych w laboratoriach in vitro lub in vivo. Ze względu na koszt poszukiwania nowych farmaceutyków, metody informatyczne in silico są zatem współcześnie powszechnie wykorzystywane przez przemysł farmaceutyczny i chemiczny. Ze względu na rozmiar przetwarzanych danych w projektowaniu leków, zastosowanie metod in silico staje się wręcz niezbędne. Narzędzia chemoinformatyczne służą jako narzędzia do identyfikacji różnic między farmaceutykami a biologicznie aktywnymi cząsteczkami będącymi potencjalnymi kandydatami na leki. Jednak aby w pełni zrozumieć mechanizm wprowadzania leku na rynek, niezbędne jest zrozumienie ekonomicznych uwarunkowań projektowania molekularnego z uwagi na fakt, że w większości przypadków to właśnie rynek decyduje ostatecznie o losach leku.

Niniejsza praca opisuje modelowanie zależności między strukturą a ceną dużej biblioteki związków chemicznych Abamachem. Planując takie badania, stawialiśmy sobie pytanie, czy możliwe jest w ogóle uzależnienie ceny od deskryptorów molekularnych opisujących cechy strukturalne cząsteczki chemicznej. Jak ustala się ceny dla tak dużej biblioteki związków chemicznych. Nie da się tego zrobić w sposób *przesadnie racjonalny*. Pomimo prób nie udało nam się znaleźć odpowiedzi na takie pytania od firmy Abamachem. Nasze modele są pierwszymi modelami typu struktura-ekonomia opisanymi w literaturze. Pokazują także problemy analizy wielkich danych w tym zakresie oraz wskazują jakie metody mogą być użyte w celu modelowania statystyk molekularnych opisujących efekty ekonomiczne. W przygotowanej rozprawie doktorskiej przeprowadzono eksplorację relacji między strukturą a właściwościami QSPR (ang. quantitative structure-property relationship) dla dużej biblioteki danych komercyjnych (ang. building bloks) firmy Abamachem, zawierającej wykaz ponad 2.2 miliona związków chemicznych wraz z ich cenami. Cena jest podstawowym miernikiem efektów ekonomicznych. Wszystkie towary, które dostarcza się na rynek są wyceniane. Tak więc związki chemiczne są również towarem. Ich ceny podawane są w przeliczeniu na jednostkę masy. W badaniach poddano analizie statystyki molekularne wielkiej liczby danych przestrzeni chemicznej poprzez

próbkowanie dużej populacji związków chemicznych. Związki chemiczne reprezentowane są przez dwa typy parametrów reprezentujących deskryptory molekularne lub właściwości. Binowanie okazuje się efektywną metodą modelowania statystyk molekularnych w bibliotekach wielkich danych, pozwalając na detekcję subtelnych efektów molekularnych jak na przykład reguła azotowa lub nieliniowość typowa dla skali wagowej. Wyniki statystyki molekularnej prowadzą ponadto do następujących wniosków:

- Biblioteka bloków budulcowych (Abamachem) oparty jest na średnio stałej cenie za gram substancji, która nie zależy od MW.
- Cena molowa związków (Abamachem) o wyższej liczbie atomów i masie cząsteczkowej jest średnio wyższa.
- Analiza wielkich danych pozwala na detekcję subtelnych efektów molekularnych typu reguły azotowej.
- Dla badanej populacji związków średnia cena $/g (Abamachem) rośnie wraz ze wzrostem kompleksowości cząsteczek opisywanych syntetyczną dostępnością związków SAS1 .
- Skład chemiczny wpływa na cenę - wraz ze wzrostem liczby atomów węgla w cząsteczce cena $/g związku maleje. W przypadku heteroatomów takich jak: tlen, siarka czy fluor ich obecność przyczynia się do wzrostu ceny związku i są one średnio droższe od heteroatomów takich jak: chlor, brom, jod czy selen.
- Zależność ceny w skali wagowej $/g od MW w zakresie niskich wartości MW wykazują efekt nieliniowości. Efekt ten przypomina trend wydajności liganda LE od HAC, którego źródłem jest zmiana skali molowej na wagową.

W wyniku przeprowadzonej analizy stwierdzono korelację między:

- ✓ Średnią ceną molową a binowaną masą, otrzymując wynik R_{bin}=0.93. Wysoką wartość współczynnika korelacji można wytłumaczyć faktem, że w przy wzroście masy wzrasta ilość substancji (materii), którą się kupuje, tak więc powyższa korelacja oznacza, że średnio na rynku cząsteczek płaci się za ilość materii.

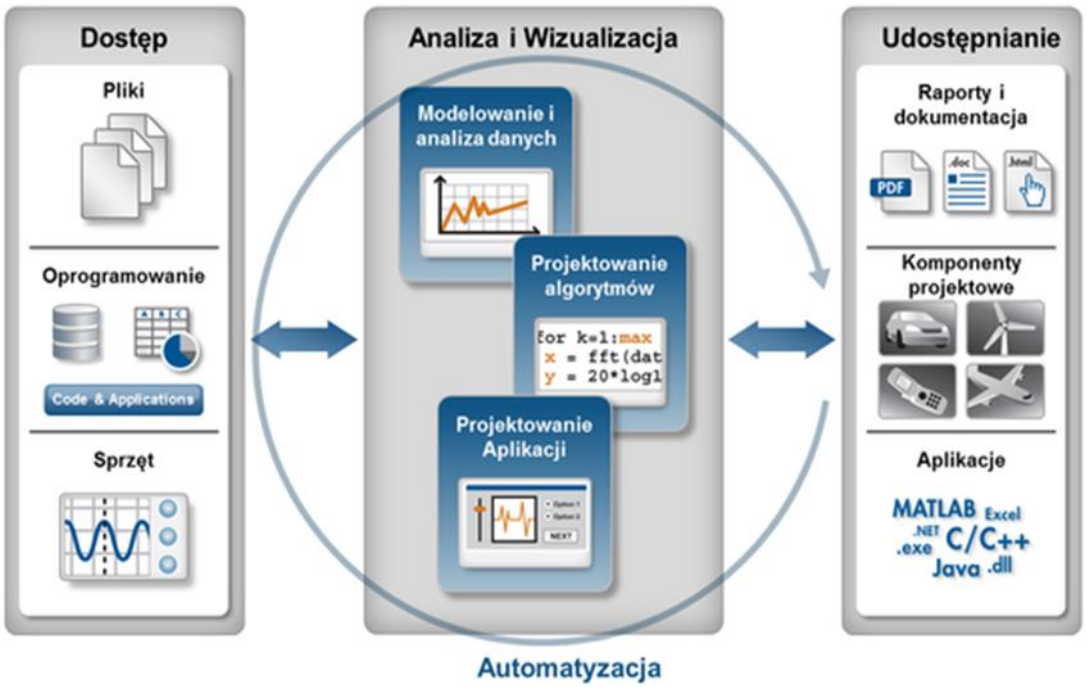
- ✓ Średnią cenę wagową a binowaną syntetyczną dostępnością otrzymując wynik R_{bin}=0.925. Wysoka korelacja świadczy, że chemia odgrywa istotną rolę w kształtowaniu się ceny na rynku związków chemicznych.

CZĘŚĆ EKSPERYMENTALNA

1. Charakterystyka oprogramowania

1.1. Program MATLAB

W literaturze poświęconej chemii organicznej opisano wiele nowych czynności chemoinformatycznych działających w środowisku MATLAB. Jest to środowisko programowania, które rozwija się już od ponad 25 lat. Program ten jest interaktywnym środowiskiem programistycznym umożliwiającym zarówno przeprowadzenie obliczeń, jak i wizualizację i analizę danych. Wysokiego poziomu język programowania zaimplementowany w środowisku MATLABA przeznaczony jest do tworzenia efektywnych algorytmów obliczeniowych[112]. Język ten umożliwia także pracę na macierzach, wektorach i strukturach. Warta podkreślenia jest różnorodność zastosowań umożliwiająca wykorzystanie dostępnych w programie funkcji i narzędzi[113]. Na rysunku 48. zilustrowano schemat działania programu MATLAB.



Rysunek 48. Schemat pracy w programie MATLAB[114].

Praca w środowisku programistycznym MATALB rozpoczyna się od importu danych z plików, baz danych lub sprzętów pomiarowych, takich jak na przykład karta pomiarowa czy oscylokop. Po zaimplementowaniu danych do środowiska programistycznego kolejnym etapem jest analiza i wizualizacja danych. Do tworzenia algorytmów i ich wizualizacji w MATLABIE wykorzystuje się specjalistyczne funkcje i narzędzia, czyli tzw. „Toolboxy" będące modułami rozszerzającymi funkcjonalność środowiska w obrębie danej dziedziny nauki. Trzecim etapem pracy jest udostępnienie wyników poprzez generację raportu w formacie: HTML, Word, LaTEX, PDF, kodu C czy też gotowej aplikacji okienkowej. Omówione wyżej cechy decydują o tym, że środowisko MATLAB jest szczególnie przydatne w procesie przetwarzania danych molekularnych.

1.2. Program Instant JChem

Instant JChem to program pozwalający tworzyć, analizować i zarządzać strukturami chemicznymi i biologicznymi. Środowisko programistyczne posiada narzędzia do zarządzania bazami danych, które zawierają informację na temat milionów struktur, zapewniając wygodne i proste analizy danych chemicznych. Instant JChem dysponuje szeroką funkcjonalnością, w tym umiejętnością konfigurowania baz danych oraz wizualizacji i analizy danych.

Podstawową czynnością dokonywana w Instant JChem jest tworzenie bazy z zaimportowanych danych. Instant JChem może obsługiwać standardowe formaty plików chemicznych (sdf, rdf, nazwa IUPAC, smiles, mrv) podczas importu i eksportu. Zaimportowane struktury chemiczne zostają doprowadzone do postaci standardowej. Oprogramowanie to można wykorzystać dodatkowo do obliczenia różnych deskryptorów, takich jak np. logP czy pKa[115].

1.3. Program SYLVIA

Program SYLVIA (ang. Synthetic Accessibility Score) służy do obliczania oceny syntetycznej dostępności związków organicznych. Został opracowany przez zespół prof. Johana Gasterigera z Friedrich-Alexander-University of Erlangen-Nürnberg. Program oblicza syntetyczny wynik dostępności związków organicznych w skali od 1 do 10, gdzie 1- są

to związki łatwe do zsyntetyzowania, a 10- trudne do zsyntetyzowania. Program obsługuje standardowe formaty plików chemicznych, takie jak SD, RDfile i SMILES, a także posiada opcje filtrowania i sortowania związków oraz graficzny interfejs dla użytkownika. Metoda obliczania dostępności syntetycznej uwzględnia wiele kryteriów, takich jak złożoność struktury molekularnej, złożoność układu pierścieniowego, liczę centrów stereoorganicznych, oraz podobieństwo do związków dostępnych w handlu[99,116].

2. Formaty analizowanych danych

W celu wczytania danych znajdujących się w katalogu Abamachem dostępnych w formacie sdf na potrzeby analizy niezbędne było przekonwertowanie formatu pliku (*.sdf) na format pliku (*.xlsx) , które wykonano w programie Matlab, a następnie zapisanie danych w formacie (*.mat).

Tabela 6. Podział i format przetwarzanych plików w zależności od użytych programów dla katalogu Abamachem

Nazwa pliku i format	Program
Abamachem_BuildingBlocks_150.**sdf**	Instant JChem
economic_data.**xlsx**	MS Excel, MATLAB
liczna_atomow1.**csv** - imput liczba_atomów1_wynik.**csv** - output	CompoundParser.exe
economic_data.**mat**	MATLAB
sylvia_result.**mat**	SYLVIA, MATLAB

Przetwarzanie pojedynczych plików sdf → … nie stanowi istotnego problemu. Problemem jest natomiast automatyczna konwersja danej populacji danych tego typu.

3. Etapy analizy i przetwarzania danych Abamachem

3.1. Pobranie danych Abamachem

Badania rozpoczęto od pobrania ze strony internetowej http://www.abamachem.net katalogu bloków budulcowych firmy Abamachem zawierającej związki chemiczne dostępne na rynku wraz z cenami (nazwa katalogu- Abamachem_BuildingBlocks_1503.sdf.)

3.2. Importowanie danych w programie Instant JChem

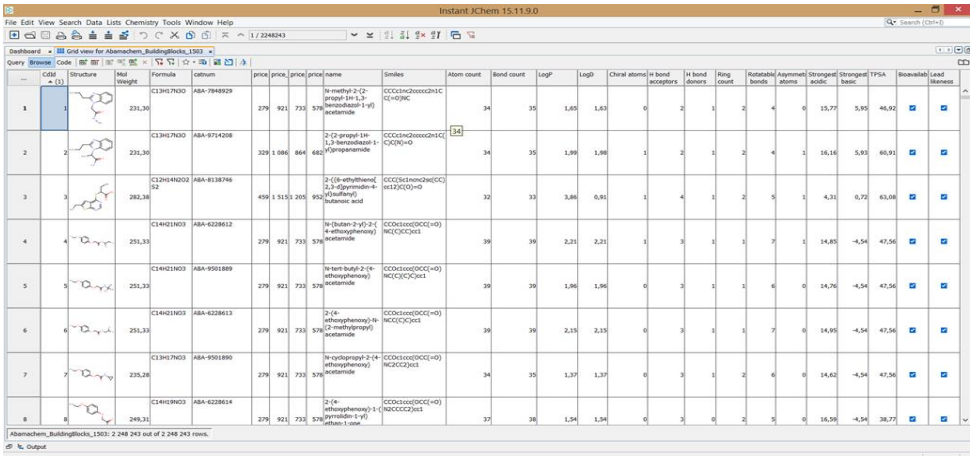
Kolejnym etapem było przetwarzanie surowych danych wejściowych poprzez import katalogu - Abamachem_BuildingBlocks_1503.sdf do programu Instant JChem 15.11.9.0 (Wybierając menu File →Import file→ Next→Finish).

3.3. Generowanie deskryptorów w programie Instant JChem

Po zaimportowaniu katalogu w programie Instant JChem obliczono na podstawie struktury związku wybrane deskryptory molekularne, takie jak:

- ✓ Smiles
- ✓ Atom Count,
- ✓ Bond Count,
- ✓ LogP,
- ✓ LogD,
- ✓ Chiral Atoms,
- ✓ H bond acteptors,
- ✓ H bond donors,
- ✓ Ring Count,
- ✓ Rotable bons,
- ✓ Asymetric atoms,
- ✓ Strondest acid,
- ✓ Strongest basic,
- ✓ TPSA,
- ✓ Bioavailability,
- ✓ Lead lankness.

(Wybierając menu Data→ New Chemical Terms Field →Expression: Smiles →OK).



Rysunek 49. Widok okna programu Instant JChem po zaimportowaniu katalogu Abamachem i obliczeniu wybranych deskryptorów molekularnych.

3.4. Eksportowanie danych w programie Instant JChem

Wygenerowane dane przetworzono w pliku Excela. Eksport wykonano wybierając menu File →Export to file…→economic_data.**xlsx**→ Next→Finish.

Komentarz: po eksporcie plik economic_data.xlsx składa się z trzech arkuszy. Podział na trzy arkusze był niezbędny, ponieważ MS Excel w jednym arkuszu może pomieścić tylko 1048576 pozycji.

- Arkusz 1 zawiera dane od 1-1000000;
- Arkusz 2 zawiera dane od 1000001-2000000;
- Arkusz 3 zawiera dane od 2000001-2248243.

3.5. Obliczanie poszczególnych atomów za pomocą aplikacji CompoundParser.exe

Innym problemem była konwersja danych Abamachem z reprezentacji typu wzór cząsteczkowy do danych. W tym celu zaprogramowano i napisano aplikację CompoundPrser.exe. Aplikacja ta oblicza na podstawie „Formuły C13H15N3O3” wszystkie dostępne atomy, z których składała się dana cząsteczka.

Procedura obliczenia poszczególnych atomów przy pomocy aplikacji CompoundParser.exe:

a) Wstępna obróbka danych

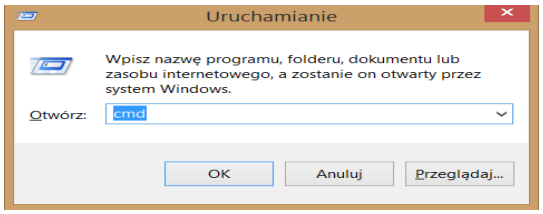
Z pliku economic_data.xlsx skopiowano kolumnę o nazwie Formuła do nowego pustego pliku MS Excel, który następnie zapisano w formacie (*.csv) rozdzielanym przecinkami (nazwa pliku – liczba_atomow1.csv).

Imput	Output
liczba_atomow1.csv (dla związków od 1 -1000000)	liczba_atomow1_wynik.csv
liczba_atomow2.csv (dla związków od 1000001-2000000)	liczba_atomow2_wynik.csv
liczba_atomow3.csv (dla związków od 2000001-2248243)	liczba_atomow3_wynik.csv

Komentarz: ważne jest, aby stworzony plik liczba_atomow1.csv posiadał tę samą ścieżkę dostępu, co aplikacja CompounParser.exe (czyli znajdował się w tym samym katalogu).

b) Uruchomienie aplikacji CompoundParser.exe

Na komputerze należy jednocześnie nacisnąć:  R. W pojawiającym się okienku należy wpisać Otwórz→ cmd→ OK.



W pojawiającym się oknie dialogowym należy wpisać następujące polecenia:

C:\User\Urszula>cd C:\Users\Urszula\Desktop\CompundParser_bin

Enter

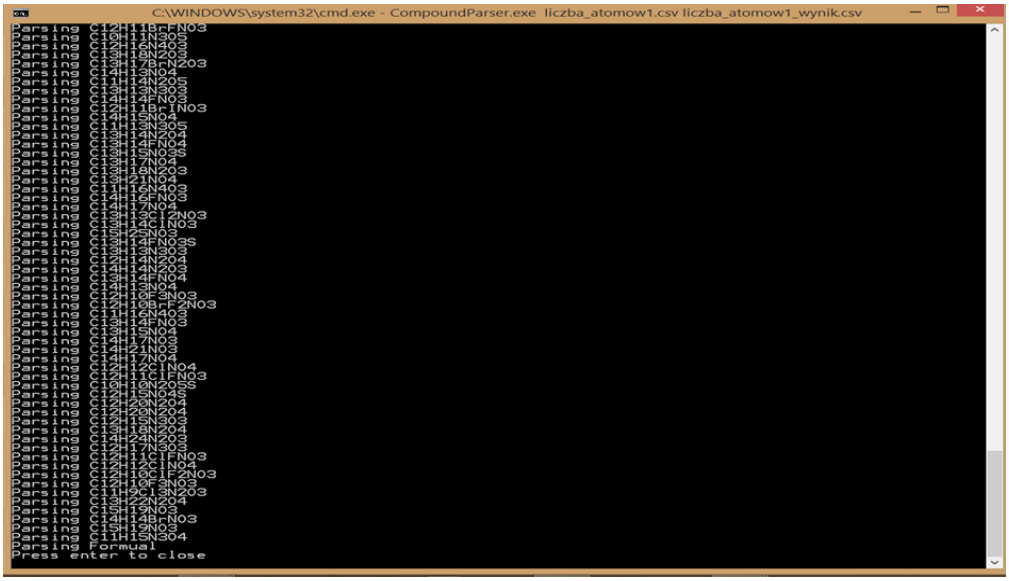
C:\Users\Urszula\Desktop\CompundParser_bin >CompoundParser.exe

Enter

C:\Users\Urszula\Desktop\CompundParser_bin

>CompoundParser.exe liczba_atomow1.csv liczba_atomow1_wynik.csv

Enter



Rysunek 50. Wygląd okna dialogowego po uruchomieniu aplikacji CompoundParser.exe

Komentarz: Obliczenia te zostały wykonane dla wszystkich związków w badanym katalogu Abamachem. Ponadto obliczone liczby atomów zostały następnie otwarte w MS Excel i przekopiowane jako kolejne kolumny do pliku economic_data.xlsx

3.6. Wczytanie danych do programu MATLAB R2015a

Wczytanie danych **econommic_data.xlsx** do programu MatlabaR2015a wykonano wpisując następujące polecenia do okna dialogowego Command Window:

>>economic_data_1=xlsread('C:\Users\Urszula\Desktop\abamachem\economic_data.xlsx',1);

>>economic_data_2=xlsread('C:\Users\Urszula\Desktop\abamachem\economic_data.xlsx',2);

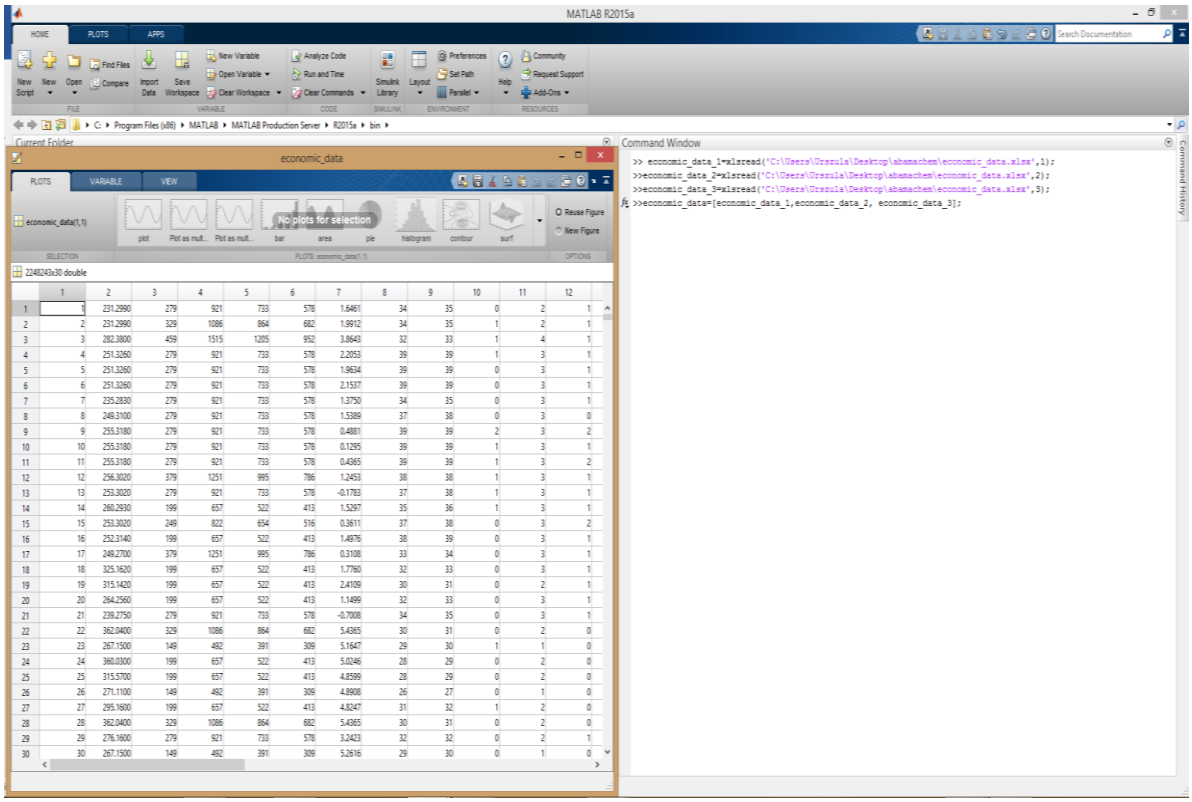
>>economic_data_3=xlsread('C:\Users\Urszula\Desktop\abamachem\economic_data.xlsx',3);

>> economic_data=[economic_data_1,economic_data_2, economic_data_3];

gdzie:

- economic_data_1- nazwa pliku
- xlsread- komenda umożliwiająca wczytanie pliku MS Excela do programu Matlab
- 'C:\Users\Urszula\Desktop\abamachem economic_data.xlsx' - ścieżka dostępu dla wczytywanych danych
- 1-pierwszy arkusz w pliku economic_data.xlsx

Komentarz: ze względu na duży rozmiar analizowanych danych niezbędne było wprowadzenie każdego arkusza Excela z osobna, a następnie złączenie trzech arkuszy w jeden plik o nazwie economic_data.



Rysunek 51. Widok okna programu MATLAB po wczytaniu danych economic_data.xlsx

Tabela 7. Nazwy kolumn wczytanych danych economic_data.xlsx do programu Matlab

Kolomna 1	Kolomna 2	Kolomna 3	Kolomna 4	Kolomna 5	Kolomna 6	Kolomna 7	Kolomna 8	Kolomna 9	Kolomna 10
LP	MW	Cena dla 1 g	Cena dla 10 g	Cena dla 5 g	Cena dla 2.5 g	LogP	AC	BC	Chiralne atomów
Kolomna 11	Kolomna 12	Kolomna 13	Kolomna 14	Kolomna 15	Kolomna 16	Kolomna 17	Kolomna 18	Kolomna 19	Kolomna 20
H bond acteptors	H bond donors	Liczba pierścieni	Rotable bonds	Liczba atomów C	Liczba atomów H	Liczba atomów N	Liczba atomów O	Liczba atomów S	Liczba atomów Br
Kolomna 21	Kolomna 22	Kolomna 23	Kolomna 24	Kolomna 25	Kolomna 26	Kolomna 27	Kolomna 28	Kolomna 29	Kolomna 30
Liczba atomów F	Liczba atomów I	Liczba atomów Se	Liczba atomów Na	Liczba atomów Sn	Liczba atomów P	Liczba atomów Si	Liczba atomów B	Liczba atomów K	Liczba atomów Cl

Ostatnim krokiem jest zapisanie wczytanych danych poprzez naciśniecie menu Save Workplace pliku o nazwie **economic_data.mat**

3.7. Obliczenie syntetycznej dostępności przy użyciu programu SYLVIA

Syntetyczne wyniki dostępności (ang. synthetic accessibility score) obliczono przy użyciu programu SYLVIA wersji 1.4 (Molecular Networks GmbH). Wszystkie z nich, oprócz 180 rekordów, zostały pomyślnie przetworzone w programie SYLVA.

Obliczenia syntetycznej dostępności przeprowadzono dla katalogu Abamachem_BuildingBlocks_1503.sdf dla wszystkich związków chemicznych. Następnie przetworzono plik do formatu (*.mat) i nazwano go sylvia_result.mat, który wykorzystano do dalszych analiz.

3.8. Obliczenie współczynników korelacji

Współczynniki korelacji obliczono pomiędzy wybranymi deskryptorami, a ceną (Tabela 4.) Obliczenia wykonano w programie MATLAB wpisując następujące polecenia do okna dialogowego Command Window:

>>wsp_korelacji=[economic_data, sylvia];

>>correcoef (wsp_korelacji (:,2), (wsp_korelacji (:,3).*wsp_korelacji(:,2)), ‘rows’, ‘pairwise’);

Komentarz: Na potrzeby obliczenia współczynników korelacji połączono dwa pliki economic_data.mat i sylvia.mat w jeden plik.

METODY

Katalog firmy Abamachem z dużą bibliotekę danych 2,248,243 chemicznych oferowanych na rynku zostały pobrane ze strony internetowej (http://www.abamachem.net/~~HEAD=pobj). Dane zostały pobrane w formacie SDF. Następnie zapisane dane zostały przygotowane do dalszego przetwarzania (powielone indeksy zostały usunięte). Użyto wersji programu JChem 14.7.28 wydanej w 2014 do otwierania pliku w formacie sdf. Obliczenia przeprowadzono przy użyciu programu MATLAB wersji R2015b. W programie MATLAB również napisano dodatkowe własne skrypty, które zostały wykorzystane do analizowania badanych struktur. Użyto komputera z systemem operacyjnym Windows 64- bitowym z procesorem Intel Core i5-4210U 1.7 GHz i pamięcią 8.0 GB. Syntetyczne wyniki dostępności Tabela 1 SAS 1 obliczono przy użyciu programu SYLVIA wersji 1.4 (Molecular Networks GmbH). Wszystkie oprócz 180 rekordów zostało pomyślnie przetworzonych w programie SYLVIA.

BIBLOGRAFIA

1. Polanski, J. *Chemoinformatics: From Chemical Art to Chemistry in Silico*. *Elsevier* (2019).

2. Polanski, J. & Gasteiger, J. *Computer representation of chemical compounds*. *Handbook of Computational Chemistry* (2017).

3. Polanski, J., Bak, A. *Podstawy chemoinformatyki leków Wydanie drugie rozszerzone*. *Wydawnictwo Uniwersytety Śląskiego* (2018).

4. Products. Available at: http://www.abamachem.net/?products.

5. Kurczab, R. Projektowanie leków in silico. *Przem. Farm.* **1**, 32–34 (2011).

6. Vistoli, G., Pedretti, A. & Testa, B. Assessing drug-likeness--what are we missing? *Drug Discov. Today* **13**, 285–94 (2008).

7. Polanski, J., Kurczyk, A., Bak, A. & Musiol, R. Privileged structures - dream or reality: preferential organization of azanaphthalene scaffold. *Curr. Med. Chem.* **19**, 1921–45 (2012).

8. Todeschini, R. & Consonni, V. *Handbook of molecular descriptors. WileyVCH, Weinheim*. **11**, (2000).

9. Kubinyi, H. E. *3D QSAR in drug design*. *Theory, Methods and Applications, ESCOM, Science Publishers B.V.* (1993).

10. Bunin, B. A., Siesel, A., Morales, G. A. & Bajorath, J. *Chemoinformatics: Theory, practice, & products*. *Chemoinformatics: Theory, Practice, & Products* (Springer Netherlands, 2007).

11. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**, 493–500 (2003).

12. Bak, A. & Polanski, J. Modeling robust QSAR 3: SOM-4D-QSAR with iterative variable elimination IVE-PLS: application to steroid, azo dye, and benzoic acid series. *J. Chem. Inf.*

Model. **47**, 1469–80

13. Polański, J. Wybrane problemy projektowania substancji biologicznie aktywnych. *Wiadomości Chem.* **[Z] 53**, **1**-, 1–16 (1999).

14. Zyrianov, Y. Distribution-based descriptors of the molecular shape. *J. Chem. Inf. Model.* **45**, 657–672 (2005).

15. Polanski, J. Drug design using comparative molecular surface analysis. *Expert Opin. Drug Discov.* **1**, 693–707 (2006).

16. Martin, R. L., Gardiner, E. J., Senger, S. & Gillet, V. J. Compression of molecular interaction fields using wavelet thumbnails: Application to molecular alignment. *J. Chem. Inf. Model.* **52**, 757–769 (2012).

17. Polanski, J., Gieleciak, R., Magdziarz, T. & Bak, A. GRID formalism for the comparative molecular surface analysis: Application to the CoMFA benchmark steroids, azo dyes, and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **44**, 1423–1435 (2004).

18. Hopfinger, A. J. *et al.* Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **119**, 10509–10524 (1997).

19. Vedani, A., Dobler, M. & Zbinden, P. Quasi-atomistic receptor surface models: A bridge between 3-D QSAR and receptor modeling. *J. Am. Chem. Soc.* **120**, 4471–4477 (1998).

20. Vedani, A. & Dobler, M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.* **45**, 2139–2149 (2002).

21. Vedani, A. & Dobler, M. Multidimensional QSAR: Moving from three- to five-dimensional concepts. *Quant. Struct. Relationships* **21**, 382–390 (2002).

22. Vedani, A. *et al.* Novel ligands for the chemokine receptor-3 (CCR3): A receptor-modeling study based on 5D-QSAR. *J. Med. Chem.* **48**, 1515–1527 (2005).

23. Ivanciuc, O., Ivanciuc, T. & Cabrol-Bass, D. 3D quantitative structure activity relationships with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. *Analusis* **28**, 637–642 (2000).

24. Polanski, J., Bak, A., Gieleciak, R. & Magdziarz, T. Modeling robust QSAR. *J. Chem. Inf. Model.* **46**, 2310–8

25. Wang, T. & Wade, R. C. Comparative Binding Energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. *J. Med. Chem.* **45**, 4828–4837 (2002).

26. Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **59**, 96–103 (1937).

27. Taft, R. W. The General Nature of the Proportionality of Polar Effects of Substituent Groups in Organic Chemistry. *J. Am. Chem. Soc.* **75**, 4231–4238 (1953).

28. Hansch, C. & Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **86**, 1616–1626 (1964).

29. Samula, K. & Cieniecka, A. *Wstep do projektowania leków*. (Państwowy Zakład Wydawn. Lekarskich, 1979).

30. Fujita, T., Iwasa, J. & Hansch, C. A New Substituent Constant, ir, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **86**, 5175–5180 (1964).

31. Marszałł, M, Kupcewicz, B. Statystyczne i chemometryczne metody analizy danych w chemii medycznej i bilogii. 1–72 (2013).

32. Polanski, J., Bogocz, J. & Tkocz, A. Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. *Drug Discov. Today* **20**, 1300–4 (2015).

33. Szlezák, N., Evers, M., Wang, J. & Pérez, L. The role of big data and advanced analytics in

drug discovery, development, and commercialization. *Clin. Pharmacol. Ther.* **95**, 492–495 (2014).

34. Polanski, J. Big Data in Structure-Property Studies—From Definitions to Models. in 529–552 (2017).

35. Polanski, J., Bogocz, J. & Tkocz, A. The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *J. Comput. Aided. Mol. Des.* **30**, 381–389 (2016).

36. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **110**, 5959–5967 (1988).

37. Sippl, W. & Höltje, H. D. Structure-based 3D-QSAR - Merging the accuracy of structure-based alignments with the computational efficiency of ligand-based methods. *J. Mol. Struct. THEOCHEM* **503**, 31–50 (2000).

38. Cho, S. J., Serrano Garsia, M. L., Bier, J. & Tropsha, A. Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *J. Med. Chem.* **39**, 5064–5071 (1996).

39. Plik:CADD-2.png - Centrum Obliczeniowe, ICM Uniwersytet Warszawski. Available at: https://kdm.icm.edu.pl/kdm/Plik:CADD-2.png. (Accessed: 21st September 2019)

40. Kubinyi, H. *Comparative Molecular Field Analysis (CoMFA)*. *Handbook of Chemoinformatics: From Data to Knowledge, (Eds.) Gasteiger,J.* **4**, (Wiley Blackwell, 2008).

41. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: A basic tool of chemometrics. in *Chemometrics and Intelligent Laboratory Systems* **58**, 109–130 (2001).

42. Klebe, G., Abraham, U. & Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem.* **37**, 4130–4146 (1994).

43. Klebe, G. *Comparative Molecular Similarity Indices Analysis: CoMSIA*. **3**, (1998).

44. Polanski, J. & Walczak, B. Comparative molecular surface analysis (COMSA): A novel tool for molecular design. *Comput. Chem.* **24**, 615–625 (2000).

45. Kohonen, T. *Self-Organizing Maps*. **30**, (Springer Berlin Heidelberg, 2001).

46. Polański, J., Gieleciak, R. & Bąk, A. The comparative molecular surface analysis (COMSA) - A nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pKa values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **42**, 184–191 (2002).

47. Polański, J. The non-grid technique for modeling 3D QSAR using self-organizing neural network (SOM) and PLS analysis: application to steroids and colchicinoids. *SAR QSAR Environ. Res.* **11**, 245–261 (2000).

48. Vedani, A., Briem, H., Dobler, M., Dollinger, H. & McMasters, D. R. Multiple-conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. *J. Med. Chem.* **43**, 4416–4427 (2000).

49. Hahn, M. Receptor Surface Models. 1. Definition and Construction. *J. Med. Chem.* **38**, 2080–2090 (1995).

50. Hahn, M. & Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure-Activity Relationships Studies. *J. Med. Chem.* **38**, 2091–2102 (1995).

51. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).

52. Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C. & Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.* **13**, 105–121 (2014).

53. Gleeson, M. P., Hersey, A., Montanari, D. & Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.* **10**, 197–208 (2011).

54. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **10**, 507–519 (2011).

55. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).

56. Goracci, L. *et al.* Lipostar, a Comprehensive Platform-Neutral Cheminformatics Tool for Lipidomics. *Anal. Chem.* **89**, 6257–6264 (2017).

57. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The 'wired' universe of organic chemistry. *Nat. Chem.* **1**, 31–36 (2009).

58. Fuller, P. E., Gothard, C. M., Gothard, N. A., Weckiewicz, A. & Grzybowski, B. A. Chemical network algorithms for the risk assessment and management of chemical threats. *Angew. Chemie - Int. Ed.* **51**, 7933–7937 (2012).

59. Kowalik, M. *et al.* Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chemie - Int. Ed.* **51**, 7928–7932 (2012).

60. Gothard, C. M. *et al.* Rewiring chemistry: Algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angew. Chemie - Int. Ed.* **51**, 7922–7927 (2012).

61. Fialkowski, M., Bishop, K. J. M., Chubukov, V. A., Campbell, C. J. & Grzybowski, B. A. Architecture and evolution of organic chemistry. *Angew. Chemie - Int. Ed.* **44**, 7263–7269 (2005).

62. Trost, B. M. The atom economy - A search for synthetic efficiency. *Science (80-.).* **254**, 1471–1477 (1991).

63. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).

64. Merck KGaA to buy Chematica | Business | Chemistry World. Available at:

https://www.chemistryworld.com/news/merck-kgaa-to-buy-chematica/3007276.article.

65. Suplemental Infrmation Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–701 (2018).

66. Clark, P. G. K. *et al.* LP99: Discovery and synthesis of the first selective BRD7/9 bromodomain inhibitor. *Angew. Chemie - Int. Ed.* **54**, 6217–6221 (2015).

67. *Is it true FDA is approving fewer new drugs lately?* (2006).

68. Elebring, T., Gill, A. & Plowright, A. T. What is the most important approach in current drug discovery: Doing the right things or doing things right? *Drug Discov. Today* **17**, 1166–1169 (2012).

69. Bunnage, M. E. Getting pharmaceutical R&D back on target. *Nature Chemical Biology* **7**, 335–339 (2011).

70. Baillie, T. A. & Rettie, A. E. Role of biotransformation in drug-induced toxicity: Influence of intra- and inter-species differences in drug metabolism. *Drug Metabolism and Pharmacokinetics* **26**, 15–29 (2011).

71. Smith, D. A., Di, L. & Kerns, E. H. The effect of plasma protein binding on in vivo efficacy: Misconceptions in drug discovery. *Nature Reviews Drug Discovery* **9**, 929–939 (2010).

72. Gabrielsson, J. *et al.* Early integration of pharmacokinetic and dynamic reasoning is essential for optimal development of lead compounds: strategic considerations. *Drug Discovery Today* **14**, 358–372 (2009).

73. Elg, M., Gustafsson, D. & Deinum, J. The importance of enzyme inhibition kinetics for the effect of thrombin inhibitors in a rat model of arterial thrombosis. *Thromb. Haemost.* **78**, 1286–92 (1997).

74. Tresadern, G., Bartolome, J. M., Macdonald, G. J. & Langlois, X. Molecular properties

affecting fast dissociation from the D2 receptor. *Bioorg. Med. Chem.* **19**, 2231–41 (2011).

75. Southan, C., Varkonyi, P., Boppana, K., Jagarlapudi, S. A. R. P. & Muresan, S. Tracking 20 years of compound-to-target output from literature and patents. *PLoS One* **8**, e77142 (2013).

76. Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* **10**, 428–38 (2011).

77. Southan, C., Vrkonyi, P. & Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminform.* **1**, (2009).

78. Leeson, P. D. & Springthorpe, B. Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nat. Rev. Drug Discov. 6, 881-890. *Nat. Rev. Drug Discov.* **6**, 881–890 (2007).

79. Lobell, M. *et al.* In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* **1**, 1229–1236 (2006).

80. Wunberg, T. *et al.* Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* **11**, 175–80 (2006).

81. Heifets, A. & Jurisica, I. SCRIPDB: A portal for easy access to syntheses, chemicals and reactions in patents. *Nucleic Acids Res.* **40**, (2012).

82. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).

83. Martin, Y. C. A bioavailability score. *J. Med. Chem.* **48**, 3164–3170 (2005).

84. Kuntz, I. D., Chen, K., Sharp, K. A. & Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9997–10002 (1999).

85. Kenny, P. W. The nature of ligand efficiency. *J. Cheminform.* **11**, (2019).

86. Murray, C. W. *et al.* Validity of ligand efficiency metrics. *ACS Med. Chem. Lett.* **5**, 616–618 (2014).

87. Hopkins, A. L., Groom, C. R. & Alex, A. Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today* **9**, 430–431 (2004).

88. Leeson, P. D. & Davis, A. M. Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* **47**, 6338–6348 (2004).

89. Proudfoot, J. R. The evolution of synthetic oral drug properties. *Bioorganic Med. Chem. Lett.* **15**, 1087–1090 (2005).

90. Leeson, P. D., St-Gallay, S. A. & Wenlock, M. C. Impact of ion class and time on oral drug molecular properties. *Medchemcomm* **2**, 91–105 (2011).

91. Walters, W. P., Green, J., Weiss, J. R. & Murcko, M. A. What do medicinal chemists actually make? A 50-year retrospective. *Journal of Medicinal Chemistry* **54**, 6405–6416 (2011).

92. Leeson, P. D. & St-Gallay, S. A. The influence of the 'organizational factor' on compound quality in drug discovery. *Nat. Rev. Drug Discov.* **10**, 749–765 (2011).

93. Southan, C., Boppana, K., Jagarlapudi, S. A. R. P. & Muresan, S. Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds. *J. Cheminform.* **3**, (2011).

94. Loving, K., Alberts, I. & Sherman, W. Computational Approaches for Fragment-Based and De Novo Design. *Curr. Top. Med. Chem.* **10**, 14–32 (2010).

95. Shultz, M. D. The thermodynamic basis for the use of lipophilic efficiency (LipE) in enthalpic optimizations. *Bioorganic Med. Chem. Lett.* **23**, 5992–6000 (2013).

96. Harikrishnan, L. S. *et al.* Diphenylpyridylethanamine (DPPE) derivatives as cholesteryl ester transfer protein (CETP) inhibitors. *J. Med. Chem.* **55**, 6162–6175 (2012).

97. Polanski, J. *et al*. Molecular descriptor data explain market prices of a large commercial chemical compound library. *Sci. Rep.* **6**, (2016).

98. Polanski, J., Tkocz, A. & Kucia, U. Beware of ligand efficiency (LE): Understanding LE data in modeling structure-activity and structure-economy relationships. *J. Cheminform.* **9**, (2017).

99. Gasteiger, J. Cheminformatics: Computing target complexity. *Nat. Chem.* **7**, 619–620 (2015).

100. Boda, K., Seidel, T. & Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput. Aided. Mol. Des.* **21**, 311–325 (2007).

101. Li, J. & Eastgate, M. D. Current complexity: a tool for assessing the complexity of organic molecules. *Org. Biomol. Chem.* **13**, 7164–7176 (2015).

102. Kurczyk, A. *et al.* Ligand-Based Virtual Screening in a Search for Novel Anti-HIV-1 Chemotypes. *J. Chem. Inf. Model.* **55**, 2168–2177 (2015).

103. Kenny, P. W. & Montanari, C. A. Inflation of correlation in the pursuit of drug-likeness. *J. Comput. Aided. Mol. Des.* **27**, 1–13 (2013).

104. Vetter, W. *F. W. McLafferty, F. Turecek. Interpretation of mass spectra. Fourth edition (1993). University Science Books, Mill Valley, California*. *Biological Mass Spectrometry* **23**, (Wiley, 1994).

105. Kiralj, R. & Ferreira, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application. *J. Braz. Chem. Soc.* **20**, 770–787 (2009).

106. Modelowanie w ochronie środowiska-Ćwiczenia at. https://chemia.ug.edu.pl/sites/default/files/_nodes/strona-chemia/16609/files/mwos_3.pdf.

107. Araujo, S. C., Maltarollo, V. G., Silva, D. C., Gertrudes, J. C. & Honorio, K. M. ALK-5 inhibition: A molecular interpretation of the main physicochemical properties related to

bioactive ligands. *J. Braz. Chem. Soc.* **26**, 1936–1946 (2015).

108. Böttcher, T. An Additive Definition of Molecular Complexity. *J. Chem. Inf. Model.* **56**, 462–70 (2016).

109. Polanski, J., Pedrys, A., Duszkiewicz, R. & Kucia, U. Ligand Potency, Efficiency and Drug-likeness: A Story of Intuition, Misinterpretation and Serendipity. *Curr. Protein Pept. Sci.* **20**, (2019).

110. Reynolds, C. H. & Reynolds, R. C. Group Additivity in Ligand Binding Affinity: An Alternative Approach to Ligand Efficiency. *J. Chem. Inf. Model.* **57**, 3086–3093 (2017).

111. Polanski, J. & Tkocz, A. Between Descriptors and Properties: Understanding the Ligand Efficiency Trends for G Protein-Coupled Receptor and Kinase Structure-Activity Data Sets. *J. Chem. Inf. Model.* **57**, 1321–1329 (2017).

112. MATLAB - MathWorks - MATLAB & Simulink. Available at: https://www.mathworks.com/products/matlab.html.

113. Sradowski, W. *MATAB Praktyczny podręcznik moelowania*. (Helion, 2015).

114. Oprogramowanie Naukowo-Techniczne. Available at: http://www.ont.com.pl/produkty/lista-produktow/matlab.

115. About Instant JChem. Available at: https://docs.chemaxon.com/display/docs/About+Instant+JChem.

116. SYLVIA - Estimation of the Synthetic Accessibility of Organic Compounds | MN-AM. Available at: https://www.mn-am.com/products/sylvia.

117. Akhurst, R. J. & Hata, A. Targeting the TGFβ signalling pathway in disease. *Nature Reviews Drug Discovery* **11**, 790–811 (2012).

118. Walczak, B. & Daszykowski, M. Chemometria w metabolomice i proteomice. in *Proteomika i metabolomika* (Warsaw University Press, 2010).

SPIS RYSUNKÓW

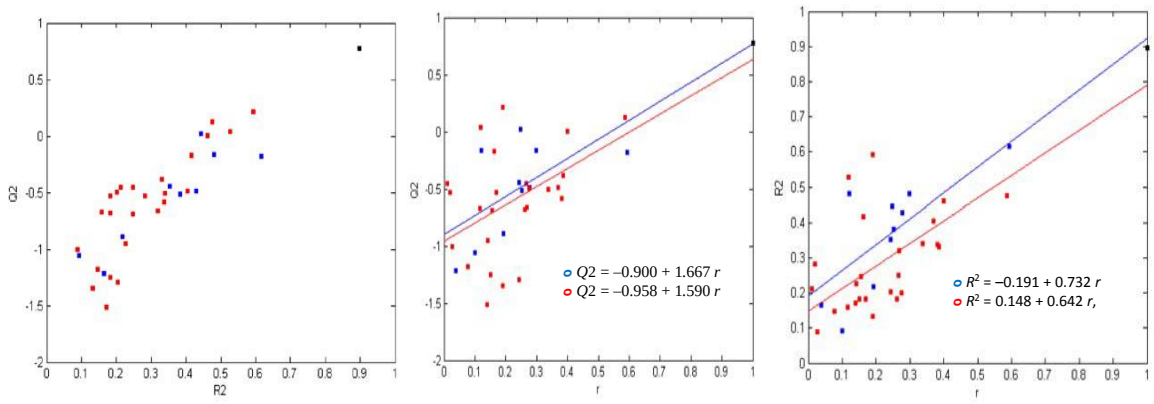
SPIS TABEL

ZAŁĄCZNIKI

Załącznik 1 Metody y-randomizacji i walidacji krzyżowej (przykłady literaturowe)

Metoda y-randomizacji

Na rysunku 52. przedstawiono y-randomizację dla zestawu danych QSPR składającego się z wartości przesunięć chemicznych karbonyl - tlen [δ/ppm] dla 50 próbek (związków benzaldehydów) opisanych przez osiem deskryptorów molekularnych[117]. Dla każdego zestawu danych dokonano 10 i 25 randomizacji losowych aby pokazać wpływ liczby przebiegów na statystykę korelacji. Modele PLS (treningowy i predykcyjny) zostały zbudowany przy użyciu danych wcześniej zrandomizowanych.



Rysunek 52. Schemat prognozowania modelu PLS metodą y-randomizacji[105].

gdzie:

- czarny kwadrat – model rzeczywisty,
- niebieskie kwadraty – 10 modeli losowych,
- czerwone kwadraty – 25 modeli losowych

Tabela 8. Statystyka korelacji dla y – randomizacji modeli PLS[105].

Patamert	10 iteracji	25 iteracji
Maximum Q^2_{yrand}	0,020	0,218
Maximum R^2_{yrand}	0,616	0,592
Odchylenie standardowe (Q^2_{yrand})	0,418	0,467
Odchylenie standardowe (R^2_{yrand})	0,162	0,133
y- randomizacja (r_{yrand} vs. Q^2_{yrand})	-0,900	-0,958

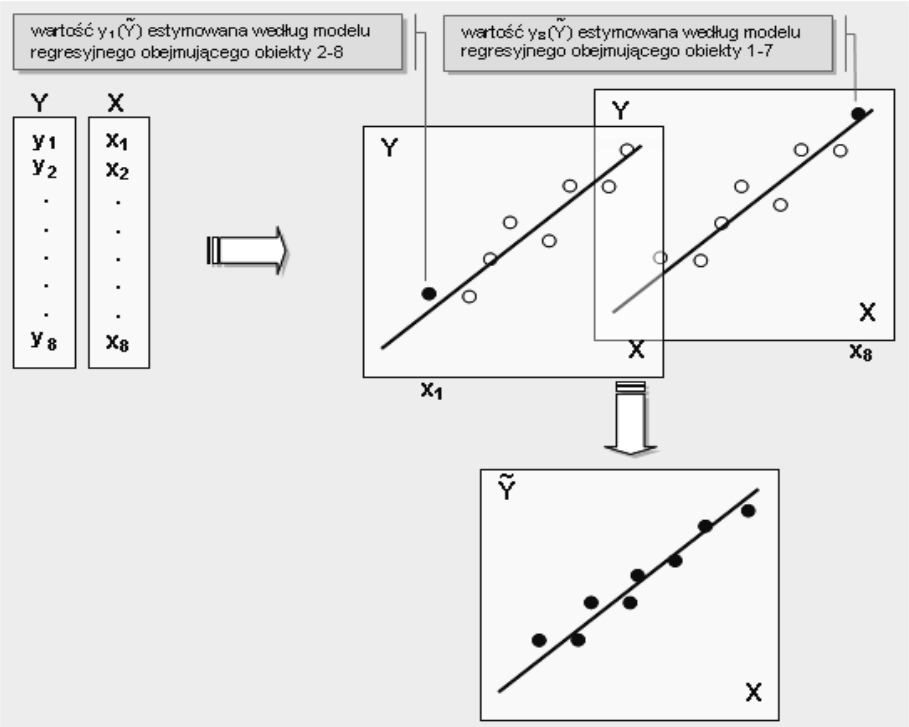
y –randomzacja (r_{yrand} vs. R^2_{yrand})	0,191	0,148

Metoda walidacji krzyżowej

Inna często stosowaną metodą statystyczną jest walidacja krzyżowa. Procedura walidacji krzyżowej jest wykorzystywana do oceny zdolności predykcyjnej modeli. Walidację statystyczną modelu należy rozpocząć od przygotowania zbioru danych. Kolejnym etapem walidacji krzyżowej jest podział danych na dwie grupy (zbiór modelowy i zbiór testowy), w których wartości zmiennych zależnych y i niezależnych x są znane. W zbiorze modelowym optymalizacji podlega model PLS, który w zbiorze testowym wykorzystywany zostaje do przewidywania wartości ỹ. Ponieważ w zbiorze danych wartości y są znane możliwe jest obliczanie wartości między y- ỹ i obliczenie błędu. Kolejny przebieg procedury polega na redukcji ze zbioru danych ilości próbek (jednej próbki LOO ang. leave one out lub kilku próbek LSO ang. leave several out). Natomiast dla pozostałych próbek buduje się model o różnej liczbie czynników na podstawie którego wylicza się wartość zmiennej zależnej y wykluczonego rzędu. Tak więc uśnięte próbki to zbiór testowy a pozostałe próbki to zbiór modelowy. Następnie po eliminacji ze zbioru danych X, Y określonych rzędów tworzone zostają modele regresyjne, które są następnie wykorzystywane do prognozowania wartości zmiennej zależnej y wykluczonej próbki[3]. Moc predykcyjną modeli posiadającą różna liczbę czynników wyznacza się w odniesieniu do liczby usuniętych próbek. Otrzymane średnie wartości wyników dla modeli o różnej liczbie czynników wykorzystywane są do wyboru modelu o optymalnej kompleksowość (złożoności obliczeniowej). W zależności od jego rodzaju dla modeli kalibracyjnych zwykle jest to średni błąd dopasowania modelu[118]. Natomiast podwójna walidacja krzyżowa pozwala ocenić moc predykcyjną zmiennych. Procedura podwójnej walidacji krzyżowej, polega na iteracyjnym redukcji liczby próbek ze zbioru danych i przeprowadzeniu walidacji krzyżowej dla pozostałych próbek oraz oszacowaniu konkretnych parametrów modelu. Do wyznaczenia optymalne kompleksowości nie są rozważane usunięte próbki co umożliwia na obiektywną walidację konstruowanego modelu[118].

Na rysunku 53. przedstawiono schemat prognozowania modelu regresyjnego metodą walidacji krzyżowej. W tym przypadku ze zbioru danych wyjściowych każdorazowo eliminowano jedną próbkę (zbiór testowy) a z pozostałych próbek budowano model (zbiór modelowy).

Po wykluczeniu ze zbioru danych X,Y konkretnych rzędów zbudowano modele regresyjne 1-8 czarne kółka. Rysunek 53. przedstawia tylko dwa wykresy dla modeli regresyjnych 1 oraz 8.

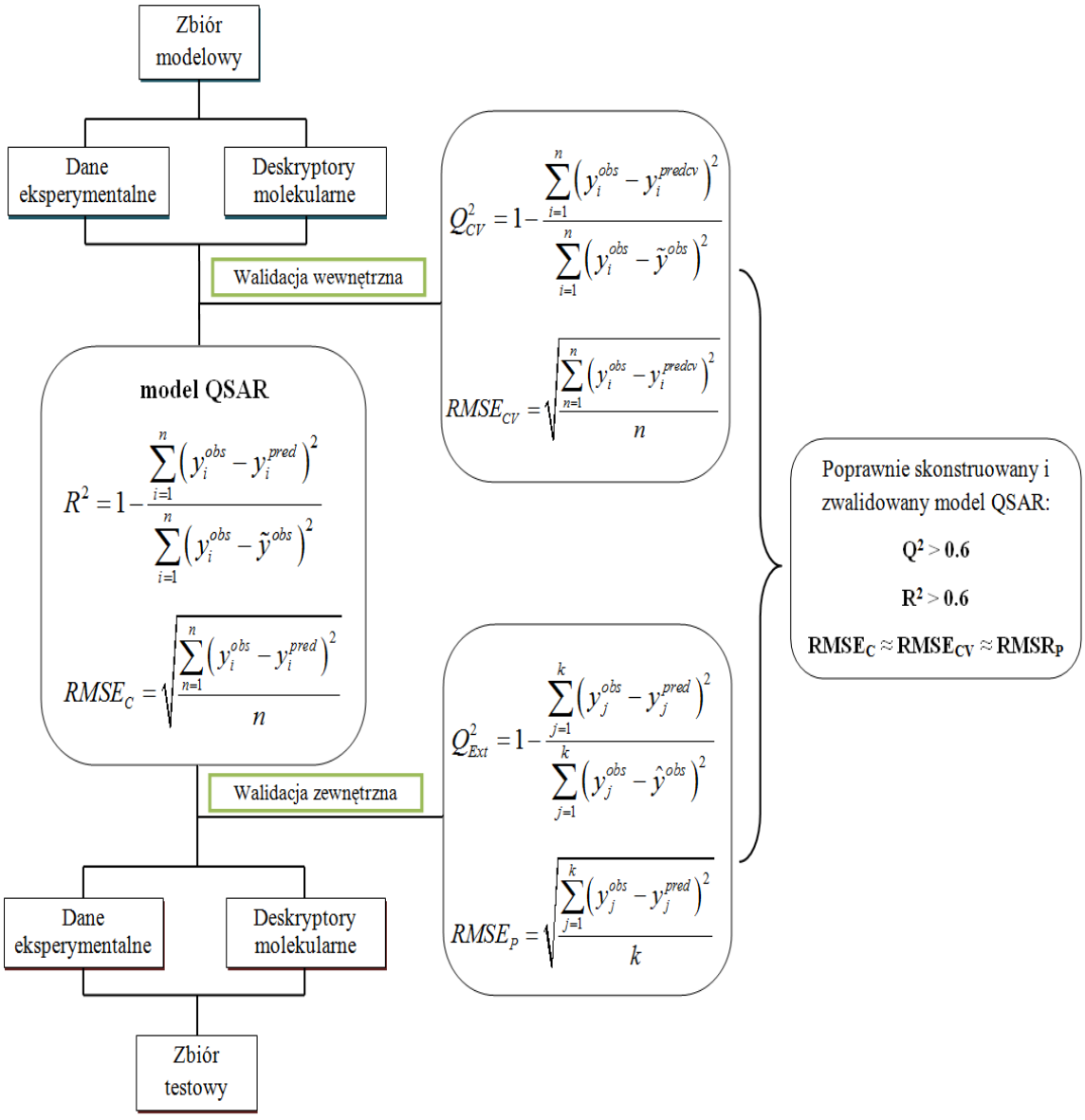


Rysunek 53. Schemat prognozowania modelu regresyjnego metodą walidacji krzyżowej[3].

Na rysunku 54. przestawiono w schemat procedury walidacji krzyżowej: wewnętrznej i zewnętrznej z uwzględnieniem obliczanych parametrów.

gdzie:

- <u>model QSAR:</u> R^2- współczynnik determinacji modelu (miara oceny jakości dopasowania modelu), $RMSE_c$- średni błąd kwadratowy zbioru modelowego
- <u>Walidacja wewnętrzna:</u> Q^2_{CV}- współczynnik kroswalidacji (miara oceny stabilności modelu), $RMSE_{CV}$ średni błąd kwadratowy kroswalidacji.
- <u>Walidacja zewnętrzna:</u> Q^2_{Ext}- współczynnik walidacji (miara oceny zdolności prognostycznych modelu), $RMSE_{Ext}$ średni błąd kwadratowy przewidywania[106].



Rysunek 54. Schemat prognozowania modelu regresyjnego metodą walidacji krzyżowej[106].

gdzie:

- y_i^{obs} –wartość eksperymentalna dla i- tej cząsteczki,
- y_i^{pred}- wartość prognozowana dla i- tej cząsteczki,
- $\tilde{y}^{obs}$ – średnia wartość eksperymentalna dla cząsteczek zbioru modelowego,
- n- liczna cząsteczek zbioru modelowego,
- k –liczba cząsteczek zbioru testowego,
- $y_i^{pred}{}_{CV}$- wartość przewidywana dla i-tej wykluczonej cząsteczki ze zbioru modelowego w walidacji wewnętrznej,
- y_j^{obs}- wartość prognozawna dla j-tej cząsteczki
- $\hat{y}^{obs}$ – średnia wartość eksperymentalna dla cząsteczek zbioru testowego[106].

Załącznik 2 Kserokopia publikacji naukowych

1.Polanski, J., **Kucia, U**., Duszkiewicz, R., Kurczyk, A., Magdziarz, T., Gasteiger, J., Molecular descriptor data explain market prices of a large commercial chemical compound library, *Sci. Rep.* **6**, (2016). **MNiSW=40**

2.Polanski, J., Tkocz, A. & **Kucia, U.** Beware of ligand efficiency (LE): Understanding LE data in modeling structure-activity and structure-economy relationships. *J. Cheminform.* **9**, (2017). **MNiSW=45**

3.Polanski, J., Pedrys, A., Duszkiewicz, R. & **Kucia, U**. Ligand Potency, Efficiency and Drug-likeness: A Story of Intuition, Misinterpretation and Serendipity. *Curr. Protein Pept. Sci.* ***20****,* (2019). **MNiSW=25**