Rozprawa doktorska

OCENA LIGANDA JAKO POTENCJALNEGO LEKU NA PRÓBIE WYBRANYCH BAZ WIELKICH DANYCH

mgr inż. Roksana Duszkiewicz

Promotor pracy

Prof. dr hab. inż. Jarosław Polański

Uniwersytet Śląski

Instytut Chemii

Katowice 2020

Najserdeczniejsze podziękowania składam prof. Jarosławowi Polańskiemu za niezastąpioną pomoc w realizacji niniejszej pracy, inspirację oraz krytyczne uwagi oraz Tomaszowi Zychowi za wsparcie i wyrozumiałość.

SPIS TREŚCI

WYKAZ ZASTOSOWANYCH SYMBOLI I SKRÓTÓW

Skrót	Rozwinięcie
AC_{50}	ang. *active concentration* – połowa maksymalnej aktywności (wg PubChem)
ADMET	ang. *absorption, distribution, metabolism, excretion and toxicity* – wchłanianie, dystrybucja, metabolizm, wydalanie i toksyczność
BEI	ang. *binding efficiency index* – wskaźnik wydajności wiązania
BLOB	ang. *Binary Large OBjects* – duże obiekty binarne
EC_{50}	ang. *half-maximal effective concentration* – połowa maksymalnego stężenia efektywnego
ED	ang. *effective dose* – dawka skuteczna
FDA	ang. *Food and Drug Administration* – Agencja Żywności i Leków
HAC	ang. *heavy atom count* – liczba atomów ciężkich
HAC bin	ang. *binined heavy atom count* – zbinowana liczba atomów ciężkich
HTS	ang. *high throughput screening* – wysokowydajny screening
IC_{50}	ang. *half maximal inhibitory concentration* – połowa maksymalnego stężenia hamującego
Kd	ang. *dissociation constant* – stała dysocjacji
Ki	ang. *inhibition constant* – stała inhibicji
LE	ang. *ligand efficiency* – wydajność liganda
LELP	ang. *lipophilicity dependent ligand efficiency* – wydajność liganda zależna od lipofilowości
LLE	ang. *lipophilic efficiency* – wydajność lipofilowa
MW	ang. *molecular weight* – masa cząsteczkowa
MW bin	ang. *binning molecular weight* – zbinowana masa cząsteczkowa
NOR	ang. *number of records* – liczba wystąpień
SEI	ang. *surface efficiency index* – wskaźnik wydajności powierzchni
SILE	ang. *size independent ligand efficiency* – wydajność liganda niezależna od wielkości

CEL PRACY

Projektowanie leku jest złożonym procesem, który wciąż obarczony jest dużym ryzykiem oraz niepewnością [Sözüdoğru 2020; Hassanzadeh 2010]. Generalnie projektowanie leku można interpretować jako odwzorowanie w obrębie dwóch grup zmiennych reprezentujących związki chemiczne, z których pierwszą grupę stanowią deskryptory molekularne, związane z cząsteczkową reprezentacją związku chemicznego. Druga grupa zmiennych związana jest z biologiczną aktywnością leku, to znaczy właściwościami związku chemicznego, a w zasadzie substancji chemicznej. Deskryptory i właściwości to dwa typy reprezentacji związków chemicznych [Polański 2016A]. O ile formalizm deskryptorów został opisany w wyczerpujący sposób [Todeschini 2008], typologią właściwości zajmowano się w znacznie mniejszym zakresie. Polanski i Gasteiger wyjaśniają to znacznie mniejszą populacją dostępnej puli właściwości [Polański 2016A, Polański 2019]. Nierównowaga w dostępnych populacjach reprezentacji molekularnych jest względnie prosta do wyjaśnienia. Deskryptory można obliczyć w stosunkowo tanich operacjach *in silico*. Ustalenie właściwości głównie substancji wymaga pomiarów i jest kosztowne.

Celem pracy był problem oceny liganda jako potencjalnego leku, w szczególności analiza *estymatora* wydajności liganda (LE: *ligand effciency*) jako jednego z typów reprezentacji właściwości substancji stosowanego w projektowaniu leków. W pracy przeanalizowano również inne typy LE oraz kontrowersje związane ze stosowaniem LE, szeroko dyskutowane w literaturze [Schultes 2010, Shultz 2013, Shultz 2014, Polański 2017B, Polański 2017C, Sheridan 2016]. W zakres pracy wchodziła analiza dużych molekularnych baz danych ChEMBL i PuChem. W ramach pracy utworzono także bazę mniejszych danych literaturowych, które poddano analizom porównawczym: Binding Database (BindingBD, PTaylaorLa, USPaten, 5HT, AChE), Psychoactive Drug Screening Program PDSP oraz dane z publikacji Mortenson 2018, Johnson 2016, Johnson 2019. Wszystkie analizowane dane znajdują się na załączonym dysku pamięci.

Praca łączy się z cyklem badań prowadzonych w Zakładzie Chemii Organicznej, po obecnej reformie w zespole profesora Polańskiego (CHEMO-BIO-INFO), których

celem jest wyjaśnienie kontrowersji związanych ze stosowaniem LE w projektowaniu leków [**D3, D4, D5**, Tkocz 2020, Kucia 2020]. W wyniku badań przeprowadzonych w ramach badań opisanych w niniejszej pracy w szczególności opisano systematykę reprezentacji związku chemicznego jako właściwości substancji [**D1**], która może być przedstawiana w różnych miarach: skali molowej, skali wagowej lub skali pojedynczej molekuły. Pokazano także jak wzajemne koincydencje tych reprezentacji generują niepewność, w szczególności związaną z fragmentacją cząsteczki (potencjalnych) leków [**D1**]. W pracy pokazałam także, że fizycznie niepewność fragmentacji molekularnej połączyć można z paradoksem analogicznym do paradoksu Zenona, który jest wynikiem niepewności podziału czasu i przestrzeni.

WSTĘP

Interesującym przykładem ciągle rozwijającej się branży gospodarki jest przemysł farmaceutyczny, który w związku z wydłużaniem długości życia społeczeństwa stanowi ważny element jego funkcjonowania. Wraz z rozwojem technologicznym rozwija się również nauka, tak jest między innymi w przypadku chemoinformatyki i bioinformatyki – stosunkowo nowych dziedzin, które łączą odpowiednio wiedzę chemiczną i biologiczną z matematyką, statystyką i osiągnięciami współczesnej informatyki. Dziedziny te mają również ogromny wpływ na rozwój medycyny i farmacji, w tym na proces projektowania leków.

Projektowanie leku wymaga ustalenia dwóch grup zmiennych. Pierwszą grupą są tzw. deskryptory molekularne, związane z cząsteczkową reprezentacją związku chemicznego. Druga grupa zmiennych związana jest z biologiczną aktywnością leku, to znaczy właściwościami związku chemicznego, a w zasadzie substancji. Deskryptory i właściwości to dwa typy reprezentacji związku chemicznego [Polański 2016A]. O ile deskryptory zostały opisane w wyczerpujący sposób [Todeschini 2008], właściwościami zajmowano się w znacznie mniejszym zakresie. Polanski i Gasteiger wyjaśniają to np. luką dostępnej populacji właściwości [Polański 2016A, Polański 2019D]. Nierównowaga w dostępnych populacjach reprezentacji molekularnych jest względnie prosta do wyjaśnienia. Deskryptory można obliczyć w stosunkowo tanich operacjach *in silico*. Ustalenie właściwości, głównie substancji, wymaga pomiarów i jest kosztowne.

Ważnym imperatywem uformowania pod koniec lat 90. ubiegłego wieku chemoinformatyki był brak narzędzi umożliwiających efektywne zarządzanie stale powiększającymi się zasobami danych chemicznych. Informatyka tworzy narzędzia i efektywnie buduje wiedzę chemiczną [Gasteiger 2003; Gasteiger 2008].

Wraz z dostępnością metod chemo- i bioinformatycznych, wzrosła ilość danych zdeponowanych w wirtualnych bazach danych. Bazy te zawierają informacje dotyczące wyników przeprowadzonych już badań eksperymentalnych *in vivo* lub *in vitro*. Ich analiza – wirtualne badanie przesiewowe (ang. *virtual screening*) jest często jednym z pierwszych kroków mających na celu poszukiwanie związków chemicznych o właściwościach leczniczych.

W szczególności w niniejszej pracy analizowano duże baz danych molekularnych ChEMBL i PubChem. Wyjaśniono kontrowersje związane ze wskaźnikiem efektywności – wydajnością liganda (LE – ang. *ligand efficiency*) jako jednego z typów reprezentacji molekularnej, a także omówieno przydatność LE w projektowaniu leków.

Niniejsza praca składa się z części literaturowej (rozdział I) części obejmującej badania własne (rozdział II) oraz części eksperymentalnej (rozdział III).

Rozdział I został podzielony na siedem podrozdziałów. W podrozdziale 1. znalazły się informacje dotyczące trzech pojęć: siły działania, powinowactwa oraz skuteczności ligandów. Omówiono ich definicje, wzajemne powiązania oraz znaczenie w procesie projektowania leków.

Podrozdział 2. został poświęcony pojęciom: deskryptora i właściwości, które często w literaturze używane są zamiennie, jednak zrozumienie zróżnicowania tych dwóch pojęć może znacznie ułatwić prowadzenie operacji w trakcie projektowania molekularnego. W tym kontekście omówiono problemy związane z precyzyjnym definiowaniem niektórych estymatorów projektowania molekularnego, czy są deskryptorami, czy właściwościami związków chemicznych.

Podrozdział 3. skupia się na wyjaśnieniu znaczenia lekopodobności jako głównej cechy, którą powinny posiadać związki będące kandydatami na przyszłe leki.

W podrozdziale 4. szczegółowo przedstawiono informacje na temat aktywności biologicznej takie jak: termodynamika wiązania ligand-receptor, zależność parametrów charakteryzujących aktywności biologiczne, rodzaje, sposoby pomiaru oraz znaczenie aktywności biologicznych w projektowaniu leków.

Podrozdział 5. omawia reprezentację właściwości typu LE, stosowaną ostatnio jako estymator w procesie projektowania leku.

W kolejnym, 6. podrozdziale omówiono coraz powszechniej stosowane pojęcie wielkich danych (*big data*) oraz wskazano ich wykorzystanie w procesie projektowania leków. W tej części pracy omówiono również dwie komercyjne bazy danych: ChEMBL i PubChem, będące głównymi źródłami danych, które zostały wykorzystane w niniejszej pracy.

W ostatnim podrozdziale części teoretycznej zwrócono uwagę na matematyczne aspekty związane z projektowaniem leków.

Część zawierająca badania własne przeprowadzone w Instytucie Chemii Uniwersytetu Śląskiego w Katowicach została zamieszczona w rozdziale II. Rozdział ten został w całości poświęcony wynikom obszernych badań na różnego typu zbiorach danych zawierających informacje o lekach i kandydatach na leki.

W rozdziale III zamieszczono informacje dotyczące części eksperymentalnej. W ostatniej części znalazł się m.in.: spis rycin i tabel, bibliografia oraz załączniki dotyczące dorobku naukowego autorki pracy. Do pracy został również dołączony nośnik CD zawierający wszystkie analizowane dane wraz ze skryptami stosowanymi w programie MATLAB.

Publikacje, których jestem współautorką w niniejszej pracy zostały oznaczone symbolami **D1-D5**, a pełne ich wersje zeskanowane i umieszczone na końcu rozprawy.

I. PODSTAWY TEORETYCZNE – ANALIZA LITERATURY PRZEDMIOTU BADAŃ

Opracowanie nowego leku jest złożonym i czasochłonnym procesem, który wymaga znacznych nakładów finansowych. Dlatego firmy farmaceutyczne i ośrodki akademickie pracują nad usprawnieniem szeregu wczesnych procesów projektowania leków. Takie podejście ma na celu identyfikację cząsteczek posiadających odpowiednie właściwości. Kluczowe przedkliniczne etapy procesu poszukiwania leków przedstawione na rycinie 1. obejmować mogą: identyfikację docelową i walidację, opracowywanie testów, wysokowydajny screening (HTS – ang. *high throughput screening*), identyfikację wyników, optymalizację i ostatecznie wybór cząsteczki [Hann 2012, Meyer 2000, Hughes 2011]. Etapy te mogą ulegać powtórzeniu, a każda kolejna faza może składać się z kilku równoległych i/lub kolejnych podetapów.

Rycina 1. Przedkliniczne etapy projektowania leków

Jednym z głównych celów podczas projektowania leków jest ustalenie zależności między dawką leku a jego efektem terapeutycznym. Znaczny udział w wyjaśnieniu tej zależności miał rozwój farmakokinetyki (dziedzina farmakologii zajmująca się badaniem wpływu organizmu na lek) i farmakodynamiki (wpływ leku na organizm). Matematyczne modele farmakodynamiki były szeroko stosowane przez farmakologów w celu opisywania działania leków [Holford 1981].

Termin „projektowanie leków” odnosi się do szukania nowych farmaceutyków, jednak w rzeczywistości bezpośrednimi celami są tzw. potencjalne leki (ang. *drug candidates*). W projektowaniu leków optymalizujemy zatem interakcje molekularne między potencjalnymi lekami a ich receptorami. Interakcje te budzą duże zainteresowanie, stanowiąc często temat badań sam w sobie. Związki chemiczne wiążące się z receptorami (makrocząsteczkami) noszą nazwę ligandów (łac. *ligare* – wiązać). Opis interakcji ligand-receptor wymagał opracowania szeregu koncepcji – najważniejsza z nich, siła działania (ang. *potency*), opisuje aktywność związaną z liczbą (ilością) ligandów potrzebnych do uzyskania efektu o danej intensywności. Zarówno w przypadku ligandów egzo-, jak i endogennych siła działania jest podstawową cechą determinującą interakcje z receptorem. Precyzyjne znaczenia pojęć związanego z oddziaływaniem ligandów wciąż są dalekie od precyzyjnego zrozumienia poniżej przedstawiono znacznie terminów *affinity*, *potency* oraz *efficacy* stosowanych w literaturze angielskiej wg [Salahudeen 2017].

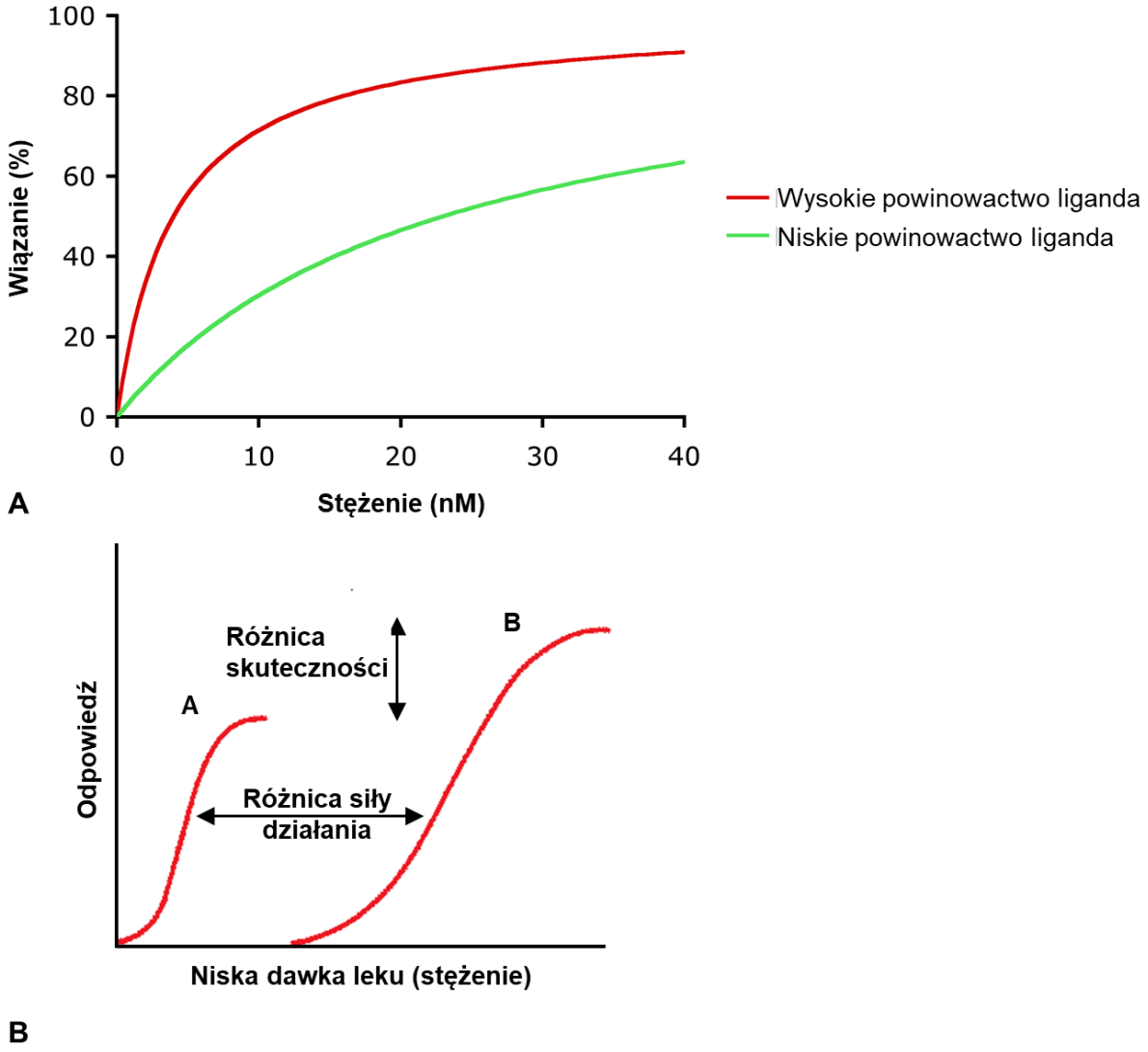
Affinity (powinowactwo), można zdefiniować jako parametr opisujący siłę działania liganda (leku) z receptorem przy danym stężeniu leku. Wskaźnik ten opisuje także trwałość z jaką lek wiąże się z receptorem. Matematyczny model powinowactwa leku do receptora został po raz pierwszy opisany przez Irvinga Langmuira [Kenakina (2004)]. *Affinity* jest jednym z czynników decydujących o wartości siły działania (*potency*).

Innym wskaźnikiem wiązania jest stała dysocjacji kompleksu lek-receptor Kd. Siłę wiązania (interakcji) liganda i jego receptora można opisać poprzez powinowactwo. Im wyższa wartość Kd, tym słabsze wiązanie i mniejsze powinowactwo odwrotna sytuacja ma miejsce gdy wartość Kd jest niska.

Siła działania leku (*potency)* jest z kolei miarą ilości leku niezbędnego do wywołania efektu o określonej wartości. Na ogół *potency* jest oznaczana jako mediana skutecznego stężenia/dawki (IC50/EC50/ED50/Kd), przy czym warto zauważyć, że wielkości stosowane do wyrażania siła działania formalnie nie są identyczne.

Kolejnym parametrem opisującym wiązanie lek-receptor jest efficacy (skuteczność) tzn. miara maksymalnej odpowiedzi farmakologicznej leku (fizjologicznej), gdy zachodzi interakcja z receptorem (związek między odpowiedzią a zajętością

receptora). Skuteczność zależy od wydajności aktywacji receptora na odpowiedzi komórkowe i tworzenia wielu kompleksów lek-receptor (ryc.2).



Rycina 2. Rózne reprezentacje właściwości opisujące oddziaływanie leku i receptora: powinowactwo % ligandów związanych przez receptor przy stałym stężeniu liganda (A); porównanie siły (potency) i suktecznośći (efficacy) działania (B) (opis w tekście). Często nie rozróżnia się tych różnych typów aktywności biologicznej ligandów zmodyfikowene wg [1,2]

1. SIŁA DZIAŁANIA, POWINOWACTWO ORAZ SKUTECZNOŚĆ

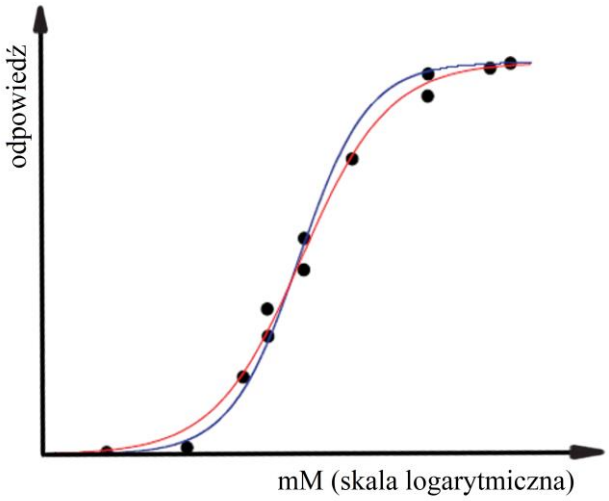
Pod względem ilościowym powinowactwo liganda do jego receptora jest związane z populacją jaką wśród dostępnych receptorów zajmuje ligand przy określonym stężeniu, co jest niekiedy określane mianem zajętości receptora (ang. *receptor*

occupancy) [Borgert 2013]. Innym ważnym aspektem działania liganda podczas wiązania receptorów jest jego skuteczność.

Mianem skuteczności określa się wydajność sygnalizacyjną interakcji receptor-ligand lub, innymi słowy, siłę liganda do wydajnej aktywacji receptora w celu wytworzenia silnej odpowiedzi *in vivo*. Skuteczność mierzy wielkość i profil odpowiedzi fizjologicznej po podaniu leku. Zatem skuteczność określa ilościowo korzyści terapeutyczne wynikające z całej kaskady skomplikowanych interakcji gość-gospodarz odpowiednio na etapach farmaceutycznym, farmakokinetycznym i farmakodynamicznym. Z drugiej strony, przy braku mierzalnego efektu funkcjonalnego, ligand może nadal wykazywać silne powinowactwo wiązania [Filimonov 2008]. Siła działania (*potency*) jest zatem rodzajem superpozycji powinowactwa i skuteczności liganda, jednak parametry te nie mogą być ze sobą ściśle powiązane. Siła działania nie koreluje także w znaczący sposób z dawką terapeutyczną [Walkman 2002].

W badaniach związków chemicznych przeprowadzanych *in vitro* stężenie poszczególnych cząsteczek jest z góry ustalone. Nie dochodzi w nich do niekontrolowanego wprowadzania lub utraty poszczególnych związków, jak to ma miejsce w badaniach przeprowadzanych *in vivo*, w których siła działania oznacza również zależność między reakcją fizjologiczną a poziomem ekspozycji na lek. Taka równowaga pomiędzy celem/receptorem a stężeniem leku może być opisana zależnością Hilla–Langmuira [Gabrielsson 2018]. Idea siły działania leku została sformułowana przez angielskiego fizjologa, laureata nagrody Nobla, Archibalda Hilla, który w 1909 roku wyprowadził równanie opisujące wiązanie tlenu przez hemoglobinę [Hill 1910]. W literaturze model ten określany jest modelem *typu Langmuira*, pomimo, iż oryginalny model Hilla został opracowany około dziesięć lat przed opracowaniem równania Langmuira. Hill opracował swój model, aby opisać sygnalizację aktywowaną przez układ ligand-receptor. Amerykański fizykochemik Irving Langmuir badał adsorpcję gazów. Zjawisko to dla ówczesnych chemików było znacznie łatwiejsze do wyobrażenia niż efekty receptorowe leków. Na początku XX wieku opisy białek i receptorów makrocząsteczkowych były całkowicie nieznane. Hill doszedł do błędnego wniosku, opierając się na zależności od temperatury, że bieg czasu jest ograniczony raczej przez oddziaływanie receptorów niż dyfuzję. Ta jego błędna interpretacja trwała przez dziesięciolecia wyraźnie wskazując na

przypadkową genezę modelu Hilla. Jednak ten intuicyjny model okazał się niezwykle przydatny dla chemii i biologii [Colquhoun 2006, D4]. Rycina 3. porównuje matematyczne modele efektów Hilla i Langmuira. Sigmoidopodobna krzywa stężenia hamującego (pIC_{50}), która wykazuje efekt biologiczny, jest reprezentacją powszechnie znaną we współczesnej chemii i biologii, jednak można ją wykorzystać w o wiele bardziej złożony sposób. Aby to wyjaśnić, porównajmy reprezentację odpowiedzi biologicznej taką jak siłę działania *(potency*) z tak zwanym modelem jednocząsteczkowym [Knight 2009, Leake 2013], w którym opis zależności ligand-receptor modeluje się poprzez dane mikroskopowe dostępne przy użyciu obecnych zaawansowanych metod biofizycznych. Oba podejścia stawiają w centrum modelu sygnalizację systemu ligand-receptor. Jednak model jednocząsteczkowy nie prowadzi do pełnego zrozumienia reprezentacji Hilla, która transformuje aktywność w skalę stężenia [mol/L]. Różnica między modelem jednocząsteczkowym a siłą działania przypomina dychotomię między deskryptorami molekularnymi odnoszącymi się do cząsteczek (zwykle, ale nie zawsze, obliczanymi) a właściwościami, które są najczęściej mierzone dla zespołów molekularnych układających reprezentacje molowe (mole). Pojęcia deskryptorów i właściwości zostały szerzej omówione w podrozdziale 2.



Rycina 3. Sigmoidalna krzywa stężenie-odpowiedź modelowana za pomocą równań Hilla (czerwony) i Langmuira (niebieski), zmodyfikowane wg [Colquhoun 2006]

Siła działania liganda jest przykładem intuicyjnej, innowacyjnej idei, która ukształtowała rozumienie terminu biologicznej aktywności ligandów. Zarówno intuicja, jak i odkrycia przez szczęśliwy traf (ang. *serendipity*) mają znaczącą wartość w rozwoju nauki, gdyż przełomowe, innowacyjne teorie wymagają nie tylko wzmożonych wysiłków i lat badań, ale także często zmiany dominujących przekonań i ogólnie przyjętych zasad w nauce. Rutynowe podejście do problemów naukowych jest elementem zachowawczym kontrolującym ludzką aktywność, co ogranicza innowacyjne pomysły, które są niezbędne do opracowywania nowych idei i pobudzania kreatywności, z kolei nowe pomysły często są nieprecyzyjne.

W szczególności rozwój leków opiera się na nieoczekiwanych odkryciach. Zasadniczo to, co obecnie rozumiemy przez projektowanie molekularne, to poszukiwanie cząsteczek, które mogą potencjalnie oddziaływać z biologicznymi receptorami makrocząsteczkowymi. Interakcje te powinny wpływać na prawidłowe działanie szlaków sygnałowych, których zadaniem jest prawidłowe funkcjonowanie organizmów. Organizmy są jednak złożonymi systemami, a znajomość struktury i funkcji receptorów białkowych do niedawna była dość ograniczona [Polański 2019].

W ostatnich latach świadomość procesów determinujących aktywność biologiczną ligandów znacznie wzrosła, opanowanie matematycznego i fizycznego formalizmu siły działania wciąż stanowi wyzwanie. Opisano wiele mechanizmów, które determinują funkcje i interakcje białek wiele rzeczy pozostaje jednak wciąż niewyjaśnionych.

Tabela 1. Słownik terminów związanych z wiązaniem lek-receptor

Termin	Polskie tłumaczenie
Affinity	Powinowactwo
Efficacy	Skuteczność
Potency	Siła działania
Single molecule model	Model jednocząsteczkowy

Receptor occupancy	Zajętość receptora (obsadzenie receptorów przez ligandy)

2. DESKRYPTORY A WŁAŚCIWOŚCI

Związki chemiczne, zarówno cząsteczki, jak i substancje, mogą być przedstawiane w postaci deskryptorów molekularnych (ang. *molecular descriptors*), bądź właściwości molekularnych (ang. *property*). Deskryptory są to wskaźniki najczęściej numeryczne odnoszące się do cząsteczek lub struktur cząsteczkowych, które można obliczyć na cząsteczkowej reprezentacji związku cheicznego. Z kolei właściwości, jeśli dostępne są odpowiednie substancje mierzone są eksperymentalnie, jeśli nie – wymagają prognozowania przez projektowanie molekularne. Jednak nie zawsze możliwe jest precyzyjne określenie jaką kategorię reprezentuje dana wartość, czy jest ona deskryptorem molekularnym, czy właściwością [**D2**, Polański 2016A]. Takim przykładem może być np. masa cząsteczkowa – uzyskana w wyniku sumowania poszczególnych atomów będzie deskryptorem molekularnym, natomiast mierzona za pomocą spektrometrii mas (lub ważąc jeden mol substancji) będzie właściwością związku chemicznego [Polański 2016A]. Wówczas masa cząsteczkowa jednego mola substancji ma jednostkę gram na mol, a masa cząsteczkowa pojedynczej cząsteczki – jednostkę Dalton. Do opisu relacji opisu pomiędzy tymi dwoma reprezentacjami konieczna jest liczba Avogadro (NA). Relację opisać można wówczas postać [Polański 2017]:

$$MW\left[\frac{g}{mol}\right] = MW[Da] \cdot NA$$

gdzie MW [g/mol] i NA to właściwości, natomiast MW [Da] jest deskryptorem [Polański 2017A].

Analiza podobieństwa reprezentacji cząsteczkowych może być wykorzystana w projektowaniu leków np. wirtualnych metodach przesiewowych opartych na strukturze liganda (ang. *ligand-based virtual screening methods*), m.in. do identyfikowania nowych związków o pożądanym działaniu, np. farmakologicznym.

Deskryptory molekularne odgrywają istotną rolę w ustalaniu korelacji pomiędzy daną cząsteczką a jej działaniem, toksycznością czy właściwością, ponieważ dają możliwość prostego przedstawienia struktury cząsteczek w sposób numeryczny [Willett 2014, Schneider 2010].

W literaturze można odnaleźć wiele przykładów prac, w których nie odróżnia się deskryptorów i właściwości molekularnych. Pojęcia te traktuje się równoważnie, definiując deskryptory jako numeryczny sposób przedstawienia cząsteczki, który ma za zadanie kodować strukturę i jednocześnie właściwości molekularne analizowanych związków [Grisoni 2017, Schneider 2005].

Przykład masy cząsteczkowej wydaje się być banalny, jednakże problem rozróżniania właściwości i deskryptorów jest sprawą bardziej skomplikowaną [Polański 2018]. W zasadzie masa cząsteczkowa obliczona dla związku chemicznego na podstawie sumy mas atomowych jest prognozą MW. Ten fakt można zrozumieć, jeśli uświadomimy sobie, że chociaż taka suma jest bardzo zbliżona do rzeczywistej wartości MW, formalnie obliczenia nie uwzględniają udziału relatywistycznego. Kolejnym problemem jest to, że pojedynczy akronim MW odnosi się zarówno do pojedynczej cząsteczki, jak i do mola substancji określonej przez liczbę Avogadro, wyrażoną odpowiednio w różnych jednostkach, odpowiednio Daltonach i kmol/kg.

Zazwyczaj właściwości odnoszą się do substancji, natomiast deskryptory z założenia obliczane są na podstawie reprezentacji molekularnej. Nie jest to jednak zawsze tak oczywiste, zwłaszcza w kontekście projektowania leków. Należy tutaj wskazać, że wraz z rozwojem technologicznym pojawiły się tak zwane metody tzw. pojedynczej cząsteczki, w szczególności biologia lub biofizyka pojedynczej cząsteczki, które dostarczyły eksperymentalnych sposobów pomiaru właściwości pojedynczej cząsteczki [Knight 2009]. W związku z powyższym w pełni spójne rozróżnienie między właściwościami a deskryptorami może stanowić problem, co można wyraźnie zaobserwować na podstawie prostego porównania nomenklatury IUPAC i tej opracowanej w chemoinformatyce. Terminologia cheminformatyki zwykle używa pojecia deskryptory, podczas gdy IUPAC stosuje termin właściwości zarówno dla deskryptorów, jak i właściwości. Stworzenie jasnej i spójnej taksonomii dla

rozróżnienia deskryptorów od właściwości mogłoby zatem rozwiązać wiele problemów pojawiających się w chemii [**D2**, Polański 2017].

Innym ważnym przykładem jest współczynnik podziału (LogP), na podstawie którego pomiarów (ok. 30 tys. danych) stworzono model regresji w celu prognozowania właściwości rzeczywistych lub wirtualnych cząsteczek [Martel 2013]. W rzeczywistości po zsyntetyzowaniu związku chemicznego jego rzeczywisty pomiar logP jest rzadko wykonywany. Zatem w znacznym stopniu są to deskryptory molekularne obliczane dla reprezentacji molekularnych w celu przewidywania pożądanych właściwości odpowiednich substancji. Rzutuje to z kolei na takie metody jak ADMET, reguła Lipińskiego czy podobieństwa do leków (ang. *druglikness*), gdyż są to metody stosowane do wstępnego filtrowania związków, jednak filtry te zwykle nie mogą opierać się na rzeczywiście mierzonych właściwościach, ponieważ cząsteczki, które są projektowane na tym etapie, nie są poddawane badaniom eksperymentalnym w tym zakresie [Polański 2018]. Inną przyczyną większego znaczenia deskryptorów – oprócz faktu, że nie można zmierzyć właściwości związków, które są w fazie projektowania, ponieważ nie są one dostępne – jest to, że nawet po zsyntetyzowaniu tych związków właściwości inne niż siła (ang. *potency*) nie byłyby mierzone, ponieważ pomiary właściwości są znacznie droższe. Braki te mogłyby zostać zredukowane poprzez rozwój nowych, niedrogich technologii pomiaru właściwości [Polański 2018].

Projektowanie leków w oparciu o spektrum właściwości nie jest rzadkością. Zrobotyzowane metody pomiaru siły działania [Michael 2008] lub nowe metody skutecznego, szybkiego i niedrogiego pomiaru profili lipidów z udziałem kilku tysięcy lipidów (ang. *lipidomic fingerprints*) [Gorraci 2017] to udane przykłady takich prac.

3. LEKOPODOBIEŃSTWO POTENCJALNYCH LEKÓW (*DRUG CANDIDATES*)

Jedną z metod oceny liganda jako potencjalnego leku jest badanie jego podobieństwa do zbioru znanych leków. Na przestrzeni ostatnich 25 lat stworzono szereg zasad empirycznych, których celem jest uproszczenie i usprawnienie reguł

projektowania leków. Reguły te sprowadzają się do badania statystyk szeregów związków, które okazały się bardziej przydatne w praktyce farmakologicznej od tych, które odrzucone zostały na różnych etapach projektowania leków. Zasady te to między innymi: reguła pięciu, ADMET, kryteria lekopodobienstwa, struktury uprzywilejowane. Reguła Lipinskiego inaczej nazywana regułą pięciu (Ro5) umożliwia identyfikację struktur chemicznych o potencjalnym zastosowaniu jako leki. Kryteria te odnoszą się do masy cząsteczkowej (MW), współczynnika podziału pomiędzy dwie niemieszające się wzajemnie fazy oktanol/woda (log P), liczby donorów wiązania wodorowego (HBD) oraz liczby akceptorów wiązania wodorowego (HBA). Reguła Lipinskiego powstała w wyniku analizy podobieństwa zbiorów wszystkich leków. Okazuje się, że znane leki wykazują podobne wartości właściwości ADMET (ang. *Absorption, Distribution, Metabolism, Excretion and Toxicity*). Reguła Lipinskiego opisuje właściwości ważne ze względu na farmakokinetykę przyszłego leku, czyli absorpcję, dystrybucję, metabolizm, wydalanie oraz toksyczność [Shultz 2013, Hansch 1962, Cherkasov 2014]. Z drugiej strony reguła Lipińskiego wyznacza po prostu „klasę leku” (ang. *drug-like*), zbiór związków chemicznych, które spełniają określone kryteria lekopodobieństwa (ang. *druglikeness*) [Guziałowska-Tic 2016; Alves 2017]. Projektowanie potencjalnego leku (*drug candidate*) w zbiorze, który mieści się w klasie leku (*druglike*) znacznie zwiększa prawdopodobieństwo uzyskania nowego leku. Istnieje wiele metod oceny lekopodobieństwa. Polegają one na badaniu zależności aktywności biologicznej od właściwości fizykochemicznych. HTS umożliwia szybkie badanie aktywności dużej liczby związków chemicznych. Jednak nie wszystkie cząsteczki, które wykazują obiecującą aktywność biologiczną podczas badań przesiewowych, będą odpowiednie do użytku klinicznego. W farmacji wykorzystuje się koncepcję lekopodobieństwa, aby zmniejszyć duże ryzyko projektowania leków uwarunkowaną między innymi niepewnością co do tego, czy biologicznie aktywna cząsteczka będzie odpowiednia do użytku farmakologicznego [Sözüdoğru 2020].

Innym parametrem wykorzystywanym do oceny potencjału ligandów jest tzw. wydajność liganda (LE – ang. *ligand efficiency*) lub pokrewne jej wielkości wydajność lipofilowa (LipE – ang. *lipophilic efficiency*) [Manallack 2013]. Z matematycznego punktu widzenia LE jest transformatą siły działania (*potency*). W zasadzie LE można

także interpretować w kontekscie lekopodobieństwa. Więcej informacji na temat wydajności liganda zawarto w podrozdziale 5.

Powyższe reguły powstały w wyniku badań statystyki struktur molekularnych leków lub potencjalnych leków. Zaobserwowano między innymi, że znaczącemu wzrostowi masy cząsteczkowej (MW) potencjalnych leków (kandydatów na leki) w ostatnich dziesięcioleciu towarzyszył niższy wzrost MW dla leków. Statystycznie potencjalne leki o mniejszej masie cząsteczkowej charakteryzują się lepszymi właściwościami fizykochemicznymi [Hopkins 2014].

4. AKTYWNOŚĆ BIOLOGICZNA

Znalezienie ligandów, które są selektywne oraz generowanie pożądanego profilu farmakologicznego leku jest trudnym zadaniem ze względu na złożony charakter różnorodnych interakcji lek-receptor [Davis 2013, Sharp 2012]. Wskazuje się zatem na konieczność ilościowego porównania zdolności hamowania różnych związków, które działają na ten sam cel.

4.1. TERMODYNAMIKA WIĄZANIA LIGAND-RECEPTOR

Kinetyka interakcji z receptorem, czyli innymi słowy szybkość, z jaką lek wiąże się z receptorem i oddziela się od niego oraz termodynamika wiązania to dwa najważniejsze modele służące do fizykochemicznego opisu wiązania ligand-receptor (LR). Powstaje pytanie, czy wiązanie lek-receptor można traktować zerojedynkowo. W rzeczywistości bowiem część cząsteczek leku utworzy kompleks lek-receptor, podczas gdy inna jego część pozostanie niezwiązana [Pan 2013, Barril 2015, Keseru 2015]. Uważa się, że zrozumienie oraz pełny opis kinetyki wiązania LR a także zrozumienie molekularnych uwarunkowań tego dopasowania stanowić będzie ważny element projektowania leków w przyszłości. Ważnymi problemami w tym kontekście są kwestiach związane ze stabilizacją interakcji lek-receptor, takie jak [Pan 2013]:

- dostępność miejsca wiązania;
- wielkość leku;
- fluktuacje konformacyjne;
- efekt elektrostatyczny;
- efekt hydrofobowy;
- udział cząsteczek wody.

Reakcję liganda z receptorem można zapisać modelowo jako reakcję dwóch cząsteczek:

$$\boldsymbol{L + R = LR}$$

Reakcji tej towarzyszy zmiana energii swobodnej zdefiniowanej w 1873 roku przez Josiaha Gibbsa jako zależnej od czynnika entalpicznego (ΔH) i entropicznego (ΔS) [Bohm 2003]:

$$\boldsymbol{\Delta G = \Delta H - T\Delta S}$$

W odniesieniu do reakcji LR, czynnik entalpiczny można odnieść do tworzenia się i/lub rozerwania wiązań wodorowych, oddziaływań jonowych/dyspersyjnych/ polarnych/hydrofobowych/kationowo-π lub kompleksowania metali, z kolei czynnik entropiczny związany jest ze zmianami stopni swobody po utworzeniu kompleksu.

Zatem zapis energii swobodnej dla wiązania LR wygląda następująco:

$$\boldsymbol{\Delta G_{wiązania} = \Delta G_{LR} - (G_L + G_R)}$$

Siłę interakcji w dwumolekularnym układzie LR można określić ilościowo, mierząc stałą równowagi procesu asocjacji (K_r) lub stałą dysocjacji (K_d) [Krumrine 2003]. Stała równowagi K_r jest powiązana ze stałą szybkości reakcji na jej końcu (k_2) do stałej szybkości reakcji na jej początku (k_1):

$$\boldsymbol{K_r = \frac{k_2}{k_1} = \frac{[LR]}{[L][R]}}$$

gdzie: [LR] stężenie kompleksu ligand-receptor, [L] stężenie wolnego liganda, [R] stężenie wolnego receptora.

Odwrotnością stałej równowagi procesu asocjacji jest stała dysocjacji K_d:

$$K_r = \frac{1}{K_d}$$

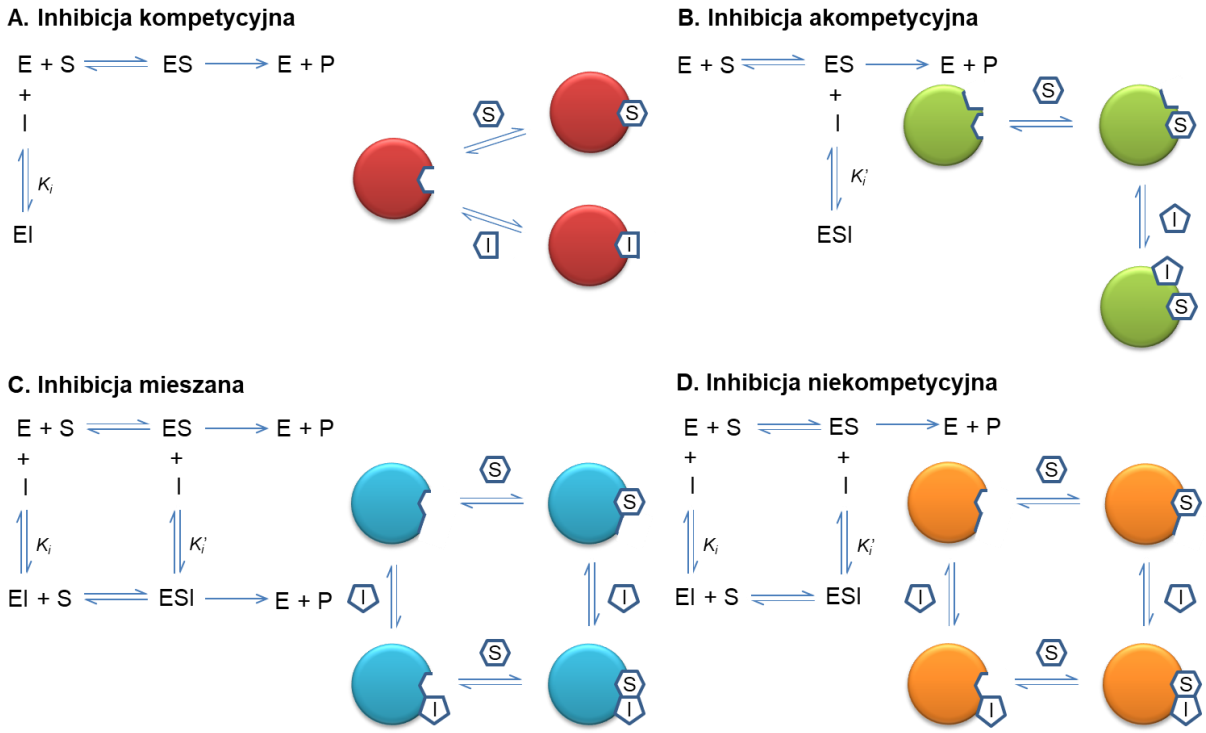
Na poziomie molekularnym wartość K_r osiąga się, gdy stężenie wolnego liganda przyjmuje wartość K_d, co oznacza również, że 50% miejsca wiązania receptora jest zajęte przez ligand [Cheng 2001].

$$K_d = \frac{[L][R]}{[LR]}$$

4.2. ZALEŻNOŚĆ PARAMETRÓW K_D ORAZ K_I

W literaturze często stała hamowania (K_i), która utożsamiana jest z K_d. Należy jednak pamiętać, że precyzyjnie rzecz biorąc nie są to terminy, które można stosować zamiennie. Oba są używane do opisania powinowactwa wiązania cząsteczki lub makrocząsteczki do enzymu lub receptora. Różnica polega na tym, że K_d jest pojęciem bardziej ogólnym, natomiast stała K_i reprezentuje również stałą dysocjacji, ale w węższym zakresie, gdyż dotyczy przypadku wiązania inhibitora z enzymem. Równowaga wiązania opisana wartością K_i zależy od kinetycznego mechanizmu hamowania. W odniesieniu do enzymatycznych mechanizmów działania można wyróżnić dwie ich kategorie: odwracalne i nieodwracalne, oparte odpowiednio na wiązaniu niekowalencyjnym lub kowalencyjnym. Odwracalne inhibitory enzymów mogą być reprezentowane przez cztery modele kinetyczne: kompetycyjny, akompetycyjny, niekompetycyjny oraz mieszany. Schematy poszczególnych typów zamieszczono na rycinie 4. W hamowaniu kompetycyjnym (ryc. 4A.) inhibitor wiąże się tylko z wolnym enzymem (E), a nie z kompleksem enzym-substrat (ES). W przypadku akompetycyjnym (ryc. 4B.) inhibitor wiąże się tylko z kompleksem enzym-substrat. Hamowanie mieszane (ryc. 4C.) obejmuje wiązanie inhibitora zarówno z wolnym enzymem, jak i kompleksem enzym-substrat

o różnych stałych wiązania (K_i i K_i'). Z kolei, hamowanie niekompetycyjne (ryc. 4D.) jest szczególnym przypadkiem hamowania mieszanego, w którym wiązanie substratu nie ma wpływu na wiązanie inhibitora [Cleland 1963].



Rycina 4. Hamowanie odwracalne enzymów: A. niekompetycyjne inhibitory wiążą się z aktywnym miejscem enzymu. B. akompetycyjne inhibitory wiążą się w oddzielnym miejscu, ale wiążą się tylko z kompleksem ES. C. Mieszane inhibitory wiążą się w oddzielnym miejscu, ale mogą wiązać się z E lub ES. D. Niekompetycyjne inhibitory wiążą się w oddzielnym miejscu, ale mogą wiązać się z E lub ES z identycznym powinowactwem. K_i jest stałą równowagi wiązania inhibitora z E; K_i' jest stałą równowagi wiązania inhibitora z ES. Opracowanie własne na podstawie [Cleland 1963].

Co ważne, wartości K_i dokładnie definiują stałą wiązania tylko wtedy, gdy mechanizm kinetyczny jest poprawnie zidentyfikowany. Określenia „K_i" używa się za każdym razem, gdy tę stałą wiązania mierzy się za pomocą kinetyki hamowania, natomiast „K_d" jest preferowane, gdy wiązanie jest mierzone bardziej bezpośrednio (np. za pomocą wygaszania fluorescencji, izotermicznej kalorymetrii miareczkowej lub powierzchniowego rezonansu plazmonowego) [Gerlza 2014].

4.3. ZALEŻNOŚĆ MIĘDZY K_i ORAZ IC_{50}

W optymalizacji leków powszechnie stosuje się nie tylko omówiony powyżej metryczny wskaźnik wiązania molowego K_i, używany do scharakteryzowania aktywności biologicznej, ale również parametr IC_{50} – połowę maksymalnego stężenia hamującego (ang. *half maximal inhibitory concentration*) [Sheridan 2016]. IC_{50} i K_i są najczęściej stosowanymi narzędziami pomiaru skuteczności leku [Aykul 2016].

IC_{50} jest miarą skuteczności substancji w hamowaniu określonego procesu biologicznego lub biochemicznego. W przypadku leków, IC_{50} oznacza stężenie leku, które jest wymagane do 50% zahamowania aktywności enzymu docelowego [Chen 1973]. Często IC_{50} jest przedstawiane w skali logarytmicznej jako pIC_{50}, co znacznie ułatwia obliczenia [Dagliyan 2009]:

$$pIC_{50} = -log(IC_{50})$$

Z kolei K_i (lub K_{i50}) wskazuje jak silny jest inhibitor, jest to stężenie wymagane do uzyskania połowicznego zahamowania aktywności liganda. Wskaźnik K_i leku jest zasadniczo taką samą wartością liczbową, co wartość liczbowa IC_{50}, podczas gdy w przypadku inhibicji kompetycyjnej (inhibicji konkurencyjnej) i niekompetycyjnej (niekonkurencyjnej) wartość K_i wynosi około połowy wartości IC_{50}. Dlatego im mniejsza wartość K_i, tym mniejsza ilość leków jest potrzebna w celu zahamowania aktywności enzymu. K_i odnosi się do powinowactwa wiązania (ang. *binding affinity*), natomiast IC_{50} odzwierciedla wytrzymałość inhibitora leku [Bachmann 2005].

Wartości K_i i IC_{50} mogą być powiązane ze sobą, na przykład dla konkurencyjnego inhibitora w reakcji enzymatycznej monosubstratu przy użyciu wzoru Chenga-Prusoffa:

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}}$$

gdzie [S] jest stężeniem badanego substratu, a K_m jest stałą Michaelisa substratu dla enzymu [Cheng 1973]. Należy jednak pamiętać, że obliczona K_i może nie być równoznaczna z mierzalnymi K_i badanych związków oraz że K_i jest stała dla danego

związku, a IC_{50} jest wartością względną. Z biologicznego punktu widzenia zależność K_i od IC_{50} jest bardziej złożona niż zależność Chenga-Prusoffa. IC_{50} zmienia się w zależności od schematu eksperymentalnego. W związku z tym nie należy go traktować jako bezpośredni wskaźnik powinowactwa wiązania ze względu na jego zależność od stężenia substratu, chyba że zastosowane zostaną identyczne warunki testu. Oprócz zależności od stężenia substratu wartość IC_{50} zależy również od sposobu działania inhibitora [Kuntz 1999]. Wśród wymienionych powyżej rodzajów inhibicji związek między K_i a IC_{50} jest inny [Burlingham 2003]. Gdy badane są akompetycyjne, niekompetycyjne i mieszane inhibitory, równanie Chenga-Prusoffa zmienia się w sposób przedstawiony w tabeli 2. W przypadku mechanizmu mieszanego inhibitory wiążą się z wolnym enzymem (K_{iE}), a także kompleksem enzym-substrat (K_{iES}).

Tabela 2. Powiązania pomiędzy IC_{50} i K_i w zależności od mechanizmu inhibicji

Mechanizm	Równanie IC_{50}
Kompetycyjny	$IC_{50} = K_i\left(1 + \frac{[S]}{K_m}\right)$
Akompetycyjny	$IC_{50} = K_i\left(1 + \frac{K_m}{[S]}\right)$
Mieszany	$IC_{50} = \frac{(K_m + S)}{\left(\frac{S}{K_{iES}} + \frac{K_m}{K_{iE}}\right)}$

Wartości IC_{50} i K_i mogą być stosowane wymiennie gdy:

- związki mają ten sam sposób reakcji, gdy wywierają działanie hamujące;
- warunki testu są takie same.

W związku z tym jeżeli wartość [S] jest znacznie niższa niż wartość K_m ([S] << K_m), to dla inhibitorów kompetycyjnych IC_{50} zbliża się do K_i z kolei gdy [S] >> K_m,

to relacja jest uproszczona do IC_{50} = K_i dla akonkurencyjnych inhibitorów. W przypadku inhibicji mieszanej takie uproszczenia nie mają zastosowania. Zasadniczo każda podana wartość IC_{50} dla (nie)konkurencyjnego inhibitora jest górną granicą dla K_i tego związku. Porównanie średnich zmierzonych wartości pK_i z odpowiadającymi wartościami pIC_{50} zawartymi w repozytorium ChEMBL wskazuje na to, że są one wyższe o 0,335 jednostki log [Kalliokoski 2013]. Testy przeprowadzane są w tych samych warunkach, relację między dwoma inhibitorami o identycznym mechanizmie działania można określić jako:

$$\frac{K_{i,1}}{K_{i,2}} = \frac{IC_{50,1}}{IC_{50,2}}$$

Możliwość wspólnego analizowania tych danych powinna być uzależniona od celu badania i dokładności z jaką ma zostać przygotowane. Statystyczna analiza repozytorium ChEMBL wykazała, że zmienność heterogenicznych wartości IC_{50} (wyrażonych przez odchylenie standardowe) jest o około 25% niższa w porównaniu z odpowiednikami K_i. W związku z tym dodanie IC_{50} do danych K_i nie jest zalecane ze względu na pogorszenie jakości danych. Z kolei jeśli głównym celem jest analiza danych IC_{50}, dodanie danych K_i będzie wzbogaceniem zbioru IC_{50} [Kalliokoski 2013].

4.4. INNE WŁAŚCIWOŚCI CHARAKTERYZUJĄCE AKTYWNOŚĆ BIOLOGICZNĄ

Oprócz omówionych już parametrów używanych do charakteryzowania aktywności biologicznej wyróżnia się również inne rodzaje przedstawione w tabeli 3.

Tabela 3. Parametry stosowane w celu scharakteryzowania aktywności biologicznych

Nazwa	Skrót	Definicja
Stała inhibicji	**K_i**	Stała równowagi dysocjacji kompleksu enzymatycznego związanego z inhibitorem.

Nazwa	Skrót	Definicja
Połowa maksymalnego stężenia hamującego	**IC_{50}**	Stężenie inhibitora wymagane do zahamowania danej funkcji biologicznej o połowę.
Połowa maksymalnego stężenia efektywnego	**EC_{50}**	Stężenie leku w osoczu wymagane do uzyskania 50% maksymalnego efektu biologicznego.
Dawka skuteczna	**ED**	Dawka lub ilość leku, która wywołuje odpowiedź terapeutyczną lub pożądany efekt w określonych warunkach.
Dawka śmiertelna	**LD**	Dawka (zwykle rejestrowana jako dawka na kilogram masy ciała pacjenta), przy której następuje śmierć chorego.
Mediana dawki śmiertelnej	**LD_{50}**	Dawka powodująca śmierć połowy badanej populacji po określonym czasie trwania.
Minimalna dawka śmiertelna	**LDlo**	Najmniejsza ilość leku, która może spowodować śmierć w danej populacji w kontrolowanych warunkach.
Dawka toksyczna	**TD**	Ilość, przy której stała, nierozpuszczalna w wodzie substancja wywołuje szkodliwe skutki dla określonej populacji w określonym czasie.
Najniższa dawka toksyczna	**TDlo**	Najniższa dawka substancji, która może wywoływać efekty toksyczne w danej populacji w kontrolowanych warunkach.

Nazwa	Skrót	Definicja
Dawka wzorcowa	**BMD**	Dawka, która odpowiada określonej zmianie w reakcji niepożądanej w porównaniu z odpowiedzią u nienarażonych osób.
Dawka referencyjna	**RfD**	Dawka jest szacunkową codzienną doustną ekspozycją na populację ludzi (w tym wrażliwe podgrupy), która prawdopodobnie nie będzie miała znaczącego ryzyka szkodliwych skutków w ciągu życia.

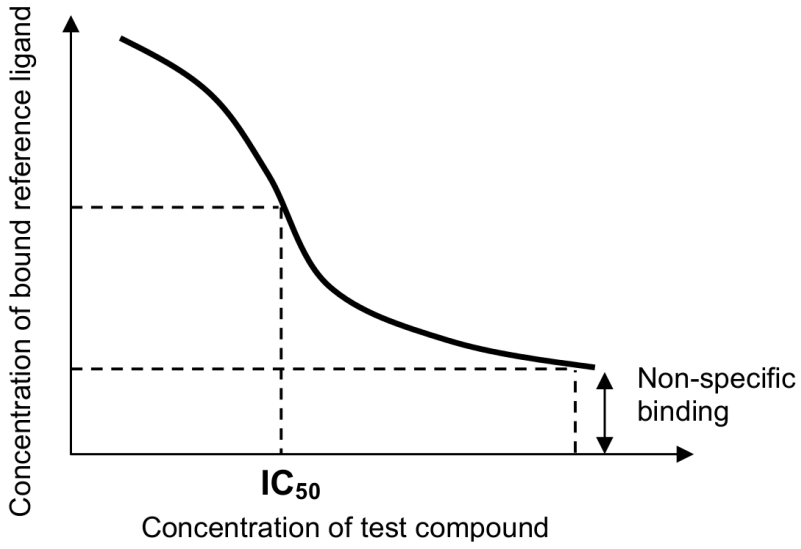
Ciekawym typem właściwości mierzonej dla leku jest minimalne stężenie hamujące (MIC – ang. *minimum inhibitory concentration*). Interesującym faktem jest w tym wypadku odniesienie nie do mola lecz do wagi działającego leku. MIC jest oczywistą miarą właściwości związaną z aktywnością związku. Miarę MIC stosuje się np. w przypadku leku, który jest mieszaniną np. produktów naturalnych, nie ma więc możliwości stosowania skali molowej. Skala MIC stosowana jest na przykład w przypadku antybiotyków i uważana jest za miarę siły działania chemioterapeutyku.

4.5. SPOSOBY POMIARU AKTYWNOŚCI BIOLOGICZNEJ

Określenie interakcji ligand-receptor stanowi istotny problem w praktyce laboratoryjnej. Udział entalpii w tworzeniu kompleksu LR można określić bezpośrednio, stosując technikę izotermicznej kalorymetrii miareczkowej, w której wykonuje się serię automatycznych etapów miareczkowania, aż miejsca wiązania do receptora zostaną całkowicie zapełnione [Klebe 2000, Gohlke 2002]. Podczas miareczkowania mierzy się stale temperaturę i odpowiednio systematycznie koryguje się jej wartość. Wynikiem jest termogram przedstawiający kolejne kroki miareczkowania w postaci pików. Odpowiednie przetworzenie otrzymanych wyników

pozwala na określenie stechiometrii, parametrów K_d i ΔH, podczas gdy inne (ΔS lub ΔG) można łatwo obliczyć przy użyciu standardowych zależności.

Stałą dysocjacji (K_d) dla danego liganda często określa się za pomocą testów konkurencyjnych, w których mierzy się różnicę w stosunku do znanego receptora [Kenny 2017, Dinse 2011]. Im większa jest różnica, tym silniejsze jest wiązanie badanej cząsteczki, dlatego stosuje się również termin: stała hamowania (K_i). Obecnie zwykle wykonuje się pomiary siły działania w celu scharakteryzowania interakcji ligand-receptor [Polański 2019]. Innymi słowy, określamy stężenie (C), które prowadzi do określonego profilu farmakologicznego. Zazwyczaj efekt ten jest określony wg miary 50%, a wynik oznaczany jest jako IC_{50}. W celu określenia stężenia konkurującego liganda, który wypiera 50% swoistego wiązania liganda odniesienia (IC_{50}), przeprowadza się „współzawodnictwo" w zakresie stężeń badanego związku. Schemat przedstawiono na rycinie 5.



Rycina 5. Krzywa współzawodnictwa dla badanego związku w teście wiązania receptora

W publiczne dostępnych katalogach danych jak np. ChEMBL zdeponowanych jest prawie trzy razy więcej informacji dotyczących IC_{50} niż K_i, a różnica ta stale się powiększa [Kalliokoski 2013]. Jedna z przyczyn różnic w dostępnych danych dotyczących IC_{50} i K_i wiąże się z trudnością w przeprowadzania badań eksperymentalnych wymaganych do oznaczenia K_i w porównaniu do IC_{50}. Według

różnych źródeł IC_{50} można wygenerować przy od ok. 25% do ok. 50% mniejszym nakładzie prac eksperymentalnych w porównaniu do wartości K_i [Perola 2010, Nowicki 2008].

Wynika to z faktu, że w celu ustalenia wartości K_i należy zbadać szybkość reakcji katalizowanej enzymem, poprzez niezależne badanie stężenia substratu i inhibitora. Określenie szybkości reakcji katalizowanej enzymem wymaga szeregu pomiarów dla zakresu stężeń substratu w stosunku do jednego stężenia inhibitora. W rzeczywistości protokół należy powtórzyć iteracyjnie dla różnych stężeń inhibitorów [Caldwell 2012]. Następnie do oszacowania parametru K_i stosuje się metody statystyczne. W praktyce potrzeba około 75-100 indywidualnych pomiarów wykonywanych zwykle w trzech powtórzeniach. Z drugiej strony, IC_{50} jest określone tylko przy jednym stężeniu substratu w zakresie stężeń inhibitora – zwykle sześć do ośmiu razy, co daje 18-24 wartości mierzonych w trzech kolejnych powtórzeniach [Hu 2012]. Wartość IC_{50} dla określonego inhibitora wyprowadza się na podstawie krzywej inhibitor-odpowiedź przy użyciu różnych procedur dopasowania z różnymi wartościami IC_{50} dla tych samych nieprzetworzonych danych [Riera 2016].

4.6. AKTYWNOŚĆ BIOLOGICZNA I JEJ TRANSFORMATY W PROJEKTOWANIU LEKÓW

W celu usprawnienia optymalizacji cząsteczek lekopodobnych (ang. *drug-like molecules*) wykorzystuje się nie tylko proste wartości aktywnownosci biologicznej lecz także jej matematyczne transformaty. Zaproponowano wiele takich transformat jako wskaźników efektywności (parametrów złożonych). Wskaźniki efektywności to proste funkcje matematyczne aktywności biologicznej (np. IC_{50}) zwykle odwzorowane względem prostego deskryptora molekularnego (HAC, MW). Parametry takie tworzone są z myślą o prostej kategoryzacji populacji związków chemicznych na takie dla których prawdopodobieństwo znalezienia leku lub potencjalnego (kandydata) leku. W praktyce zamiast podejmować decyzję w oparciu o samą wartość IC_{50} angażujemy także inne zmienne, np. MW [Shultz 2013]. Najczęściej stosowane wskaźniki efektywności to: wydajność liganda (LE – *ligand efficiency*), wskaźnik wydajności wiązania (BEI – *binding efficiency index*), wydajność lipofilowa

(LipE – *lipophilic efficiency*), wydajność liganda zależna od lipofilowości (LELP – *lipophilicity dependent ligand efficiency*), wydajność liganda zależna od wielkości (SILE – *size independent ligand efficiency*), maksymalna wydajność liganda (%LE – *maximal ligand efficiency*) oraz wskaźnik wydajności powierzchni (SEI – ang. *surface efficiency index*) (Tab. 4) [Bembene 2009; Reynolds 2008; Abad-Zapatero 2011].

W przypadku wszystkich wskaźników efektywności wartości pIC_{50} używane są zamiennie z wartościami K_i, podobnie w przypadku log P, który może być zastępowany wartością log D [Shultz 2013]. Taka praktyka jest szczególnie przydatna, podczas badania statystyk dużych zbiorów danych, gdzie często nie wszystkie parametry są dostępne.

Tabela 4. Wybrane wskaźniki efektywności i sposoby ich wyznaczania

Wskaźnik	Skrót	Sposób obliczania
Wydajność liganda	LE	$1,37\left(\frac{pIC_{50}}{HAC}\right)$
Wskaźnik wydajności wiązania	BEI	$\frac{pIC_{50}}{MW}$
Wydajność lipofilowa	LipE	$pIC_{50} - logP$
Wydajność liganda zależna od lipofilowości	LELP	$LogP/LE$
Wydajność liganda zależna od wielkości	SILE	$\frac{pIC_{50}}{HAC^{0,3}}$
Maksymalna wydajność liganda	%LE	$\frac{LE}{maxLE} \cdot 100$
Wskaźnik wydajności powierzchni	SEI	$pIC_{50} \cdot \frac{100Å^2}{PSA}$

5. PODSTAWY ZASTOSOWAŃ PARAMETRÓW TYPU LE W PROJEKTOWANIU LEKÓW

W projektowaniu leków optymalizuje się zdolność wiązania ligandów w zależności od ich struktury chemicznej. Jest wiele metod, które koncentrują się na tym problemie. W szczególności analizowano także wpływ wielkości cząsteczek (ang. *molecular size*) na efektywność opracowywania nowych leków [Hopkins 2014, Williams 2017]. Zasadniczo wzrost wielkości (MW) cząsteczki zwiększa również jej złożoność. Interesujące jest zatem to, czy wzrost MW wpływa na prawdopodobieństwo zidentyfikowania nowych leków, w szczególności – czy szanse na znalezienie mniejszych i mniej złożonych ligandów są wyższe niż w przypadku większych ligandów. Z kolei wzrost złożoności molekularnej może zwiększyć siłę wiązania ligand-cel, która dla małych cząsteczek lub fragmentów może spadać poniżej mierzalnego poziomu [Hann 2001, Zartler 2005].

Oznacza to, że rozmiar liganda odgrywa ważną rolę w dopasowywaniu i wiązaniu ligand-cel. Ocena tej roli jest ważnym narzędziem w poszukiwaniu bardziej skutecznego sposobu projektowania leków. Liczba atomów ciężkich (HAC – ang. *heavy atom count*) to prosty deskryptor opisujący rozmiar cząsteczki. HAC używany jest do definiowania wydajności liganda (LE), który bywa stosowany do oceny zdolności wiązania ligandów. LE jest zdefiniowany jako stosunek energii wiązania do HAC [Hopkins 2014, Kuntz 1999, Reynolds 2013, Reynolds 2014, Reynolds 2017].

5.1. KONTROWERSJE WOKÓŁ ZASTOSOWANIA LE W PROJEKTOWANIU LEKÓW

Obecnie najczęściej opisywanym wskaźnikiem efektywności jest wydajność liganda LE (ang. *ligand efficiency*). LE często stosuje się także w przemyśle farmaceutycznym do wstępnego szacowania potencjału optymalizacji projektowanych leków lub struktur wiodących (*drug candidate, lead structures*) poprzez ocenę zdolności wiązania ligandów [Hopkins 2004, Hopkins 2014, Kuntz 1999, Reynolds 2007, Reynolds 2008, Reynolds 2017].

LE opracowano w celu oceny średniej wartości energii swobodnej wiązania Gibbsa (ΔG^o) w stosunku do liczby atomów ciężkich (HAC), czyli atomów nie będących atomami wodoru [Kenny 2014]. Zależność tę można zapisać w postaci [Polański 2017B]:

$$LE = \frac{\Delta G^\circ}{HAC}$$

LE jest powszechnie obliczany jako funkcja stężenia hamującego pIC_{50}, mianowicie:

$$LE = \frac{-RTln(IC_{50})}{HAC} \approx 1,37\left(\frac{pIC_{50}}{HAC}\right)$$

gdzie: IC_{50} – połowa maksymalnego stężenia hamującego (miara aktywności biologicznej), R – stała gazowa, T – temperatura bezwzględna.

W ostatnich latach pojawiły się pytania dotyczące trafności definicji wydajności liganda. Zakwestionowano nie tylko matematyczną poprawność LE z jego wysoką preferencją dla małych ligandów. Wielu autorów wskazało również, że przydatność parametru LE można kwestionowac [Schultes 2010, Shultz 2013, Shultz 2014, Polański 2017B, Polański 2017C, Sheridan 2016]. Konkretnie jeżeli HAC dwóch potencjalnych ligandów różnią się znacznie to delta-HAC (w zasadzie 1/HAC) decyduje o różnych wartościach LE. Z kolei, kiedy porównujemy dwa ligandy o takim samym HAC informacja o samym HAC nie wnosi żadnej nowej informacji. Tak więc sama wartość LE jest w tym wypadku mało znacząca [Scott 2018]. Jednym z paradoksów związanych z zastosowaniem LE jest wiec problem, który polega na tym, że nie powinno się porównywać związków o bardzo różnych wartościach HAC, co powoduje wątpliwość, gdyż wg definicji LE zależy właśnie od wielkości cząsteczki [Scott 2018]. Pomimo tych problemów w literaturze dostępnych jest wiele różnych analiz opartych na LE [Hopkins 2014]. Obserwowany trend LE wielu autorom wydawał się paradoksalny i nie był w pełni zrozumiany (Murray 2014, Zhou 2009, Shultz 2013, Shultz 2014, Polański 2017B, Polański 2017C, Sheridan 2016]. LE odnosić się może zarówno do modelu jednocząsteczkowego jak i do modelu wielocząsteczkowego (molowego), przy czym obserwujemy tutaj ciekawy paradoks, który wynika z faktu, że fragmenty (1 HAC lub odpowiadający mu 1 Dalton) nie mają

rzeczywistej reprezentacji molowej, chociaż formalnie 1 mol Daltonów stanowi 1 g [Polański 2017B, Polański 2017C].

W kontekście projektowania leków zwrócono uwagę na wysoką preferencję LE dla małych ligandów, która wpisuje się w nowsze trendy w farmacji, faworyzujące małe ligandy molekularne (tak zwana koncepcja *slim pharma*) [Hann 2011], mające korzystniejsze profile lekopodobieństwa [Shultz 2019]. Dlatego też LE działa nieoczekiwanie dobrze, jeśli zostanie zastosowana jako wiodący estymator optymalizacji [Mignani 2018, Meanwell 2016, Cavalluzzi 2007, Schultes 2010], pomimo niepewności co do jej fizycznego znaczenia [Polański 2017B, Polański 2017C]. Wydaje się zatem, że pełniejsze zrozumienie LE jest istotne dla racjonalizacji projektowania molekularnego.

5.2. EKONOMICZNE WSKAŹNIKI EFEKTYWNOŚCI LEKU

Związek między strukturą a właściwościami danego związku chemicznego (QSPR – ang. *quantitative structure-property relationship*) stanowi teoretyczną podstawę projektowania leków. Oprócz właściwości molekularnych proces tworzenia nowego leku uzależniony jest jednak w równym stopniu od parametrów ekonomicznych [Manallack 2013]. Całościowe zrozumienie losu nowych leków wymaga zatem spojrzenia na ten proces również ze strony efektywności ekonomicznej. Relacje ekonomiczne w chemii przedstawiane mogą być za pomocą modeli ilościowych zależności struktura-ekonomia (QSER – ang. *quantitative structure-economy relationships*). Parametrem ekonomicznym może być m.in. cena jednego grama danego związku [D2].

Badanie dużych ilości danych molekularnych różni się od klasycznego podejścia do modelowania QSAR (obejmującego dużo mniejsze zestawy danych) tym, że obserwacje charakteryzujące poszczególne związki są często zastępowane średnimi, maksymalnymi lub minimalnymi wartościami opisującymi większe serie związków lub ich klasy [D2].

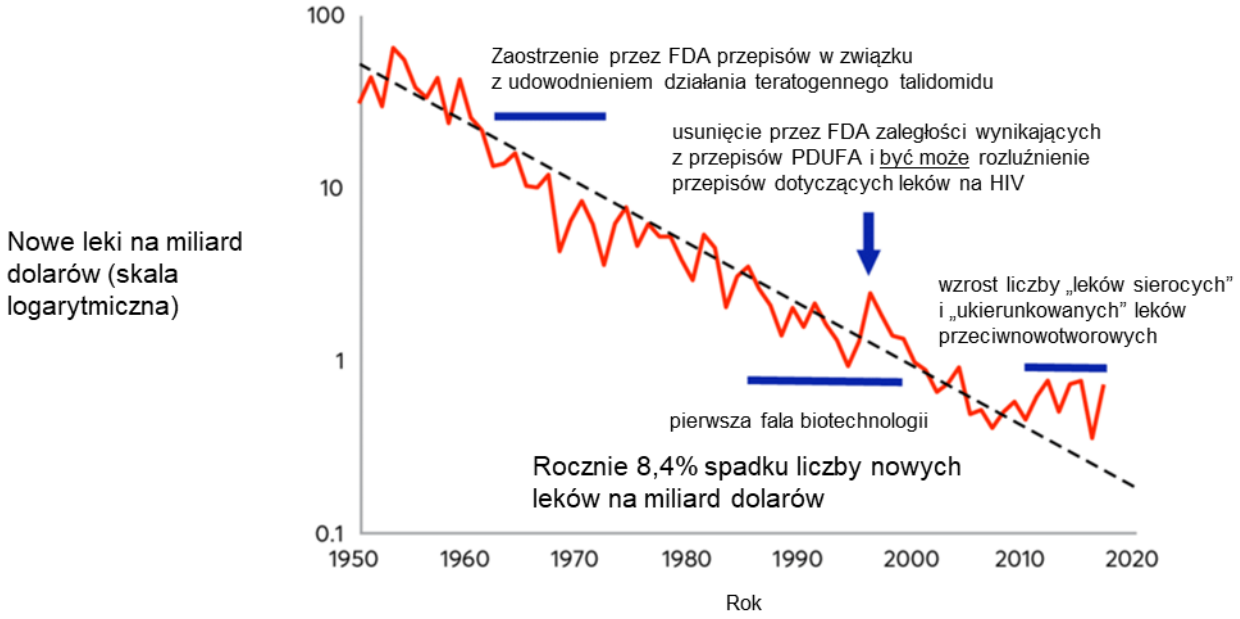
Chociaż projektowanie potencjalnych leków jest trudne, w rzeczywistości ich ostateczna konwersja do farmaceutyków okazuje się znacznie trudniejsze.

Efektywność projektowania leków jest niska i wciąż może zostać zilustrowana przy pomocy danych ekonomicznych prawem Erooma [Scannell 2012]. Prawo Erooma mówi
o tym, że pomimo rozwoju technologii odkrywanie nowych leków staje się coraz wolniejsze i kosztowniejsze [Schulthess 2014]. Statystycznie co dziesięć lat podwajają się koszty związane z poszukiwaniem, projektowaniem, analizą i rozwojem nowych farmaceutyków. Prawo to wynika z kilku reguł ekonomicznych, które mogą odnosić się do rynku farmaceutycznego m.in. w wyniku [Scannell 2012]:

- Podniesienia norm bezpieczeństwa i wprowadzenia nowych regulacji prawnych, w związku z czym wydłużony zostaje czas prowadzenia badania danego potencjalnego leku, a tym samym wprowadzenia go na rynek.

- Dużej konkurencyjności na rynku farmaceutyków. Duży sukces jednego leku powoduje, że sukces innego leku na tym samym rynku wydaje się być mało prawdopodobny.

Prawo Erooma w przemyśle farmaceutycznym można zilustrować na rycinie 6., która przedstawia stosunek liczby nowo rejestrowanych leków (FDA) do inwestycji w R&D firm farmaceutycznych od lat 50. XX wieku. W latach 60. XX wieku, będących „złotą epoką" dla przemysłu farmaceutycznego, opracowanie udanego leku kosztowało średnio 100 milionów dolarów. Od tego czasu liczba nowych leków na miliard dolarów (skorygowanych o inflację) zmniejsza się o połowę co dziewięć lat. Z kolei ok. 2000 r. koszt nowego leku przekroczył 1 miliard dolarów, a postęp prac badawczo rozwojowych spadał od kolejnej dekady. Nieznaczna stabilizacja rynku ma miejsce od roku 2012, gdy wzrosło zainteresowanie lekami sierocymi (lekami chorób rzadkich) i celowanymi (ukierunkowanymi na określone działanie) [Jones 2018].



Rycina 6. Prawo Erooma: liczba nowych cząsteczek zatwierdzonych przez amerykańską Agencję ds. Żywności i Leków (farmacja i biotechnologia) na miliard $ ogólnoświatowych wydatków na badania i rozwój, opracowanie na podstawie [Jones 2018]

Badania wskazują na istnienie dwóch głównych przyczyn spowolnienia wydajności badań i rozwoju. Po pierwsze uważa się, że klasyczny model rozwoju wiedzy w zakresie biologii komórki i genomiki, a następnie przełożenie tej wiedzy na odkrycie nowych leków, są wadliwe. Innym problemem jest zgodność między modelami chorób *in vitro* i *in vivo.* Taka zgodność jest podstawowym założeniem wielu badań biomedycznych, ale rzadko poddaje się je krytycznej ocenie. Taka ocena wydaje się istotna w kontekście wzrostu klinicznej produktywności badań farmaceutycznych. [Horrobin 2003]. Po drugie proces odkrywania nowych leków napotyka coraz większe problemy wynikające z tego, że wszystko co było łatwe do wykorzystania zostało już wykorzystane (wg powiedzenia „zebraliśmy wszystkie nisko wiszące owoce"), a dalszy rozwój wymaga większych nakładów pracy. Do rozważenia pozostają jeszcze kwestie związane z tym, że nowy lek musi być lepszy niż istniejące. Rosnąca oferta istniejących leków zmniejsza potencjalną wartość tych nieodkrytych, aż do momentu, w którym nie warto wydawać pieniędzy na ich opracowanie. Przyczynia się to na przykład do niskiego poziomu inwestycji w produkty lecznicze na nadciśnienie, pomimo jego ciągłego znaczenia klinicznego.

Nowe leki musiałyby zastąpić istniejące leki generyczne, które są zazwyczaj tanie i ogólnie skuteczne.

Ciekawe badania dotyczące efektywności ekonomicznej leku, np. zależność między siłą działania a wartością średniej sprzedaży bestsellerów leków (lista TOP100) znaleźć można w publikacjach [Polański 2016B, Polański 2015].

6. OD BIG DATA DO NOWEJ WIEDZY CHEMICZNEJ

Najszersza definicja interpretuje *dane* jako wszystko, co można zapisać. Dotyczy to również metadanych, tj. danych, które odnoszą się do innych danych. Oznacza to, że mogą one stanowić zarówno uporządkowane, jak i nieuporządkowane zbiory wartości. Ponadto wartości mogą być zarówno nominalne, jak i liczbowe, podczas gdy te ostatnie mogą być liczbami dyskretnymi, przedziałami lub stosunkami. Inny typ danych służy do opisu plików audio, wideo i graficznych (duże obiekty binarne, BLOB – ang. *Binary Large Objects*), a do ich eksploracji wykorzystywany jest specjalny typ analizy [Maheshwari 2014]. Ważnym problemem w przemyśle farmaceutycznym jest dostępność i wymiana danych, tzw. *data sharing*.

Analiza *Wielkich Danych* (ang. *Big Data*) może być wykorzystywana w różnych dziedzinach i ma specjalne zastosowanie w chemii, biologii i medycynie. Zwiększenie dostępu do komputerów w ciągu ostatnich 25 lat przyczyniło się do powstania dużych wirtualnych repozytoriów danych molekularnych, a tym samym do wzrostu zainteresowania wielkimi zbiorami danych [Seiler 2007]. Projektowanie molekularne, w szczególności w przypadku odkrywania nowych leków, wymaga analizy dużej liczby modeli molekularnych. Możliwość modyfikacji struktur chemicznych w kontrolowany i zautomatyzowany sposób pozwala na obiektywne, przejrzyste i powtarzalne wykorzystanie wyników analiz dużych liczb struktur w rozwijaniu nowych idei [Kenny 2005, Barnard 2011].

Istnieje wiele definicji *Big Data*, jednakże wyróżnia się trzy zasadnicze czynniki decydujące o różnicy pomiędzy konwencjonalnymi a dużymi zbiorami danych. Czynniki te przedstawione zostały na rycinie 7., są to: objętość, szybkość i różnorodność. W tym przypadku objętość odnosi się do ogromnej ilości zawartych

w zbiorach danych, szybkość – do tempa przyrostu informacji w nich zawartych, natomiast różnorodność – do różnych form zdeponowanych danych [Laney 2016 Polański 2017A].

Rycina 7. Czynniki decydujące o interpretacji danych jako Big Data [Polański 2017A]

Big Data są niekiedy również definiowane przez wysoki stopień złożoności informacji jakie zawierają [Laney 2001]. Opracowanie statystyk poprzez analizę dużych ilości danych pozwala m.in. na ujawnienie szczegółów, których nie można wykryć przy użyciu analizy mniejszych grup. Przykładem mogą być aktywności małych cząsteczek i bioterapeutyków, na temat których w literaturze znajduje się wiele informacji, a dostęp do nich poprzez zdeponowanie we wspólnych zbiorach może przyspieszyć rozwój i zastosowanie w projektowaniu nowych leków [Gaulton 2012, Keiser 2009, Paolini 2006, Mestres 2009]. Uważa się również, że analizy dużych zbiorów danych mogą przynieść wymierne korzyści oraz wpłynąć na innowacje w gospodarce. Przykładem są badania przeprowadzone w Stanach Zjednoczonych, które sugerują, że potencjalne wykorzystanie dużych zbiorów danych w służbie zdrowia może obniżyć koszty o 300 miliardów dolarów rocznie [Ślezak 2014]. Analiza dużych danych pozwoli na lepsze opracowywanie statystyk, gdyż większa liczba danych oznacza dokładniejszą analizę informacji dając możliwość zrozumienia

tendencji w nich występujących. Oznacza to, że możliwości wyjaśnienia, modelowania lub zrozumienia skutków medycznych, farmakologicznych lub chemicznych reprezentowanych przez duże zbiory danych wciąż mają duży potencjał wzrostowy [Polański 2017B].

Należy jednak pamiętać, że analiza tego typu danych wiąże się z różnego rodzaju problemami tj.: niewystarczające ich zdefiniowanie w zbiorach i/lub ich nieuporządkowanie.

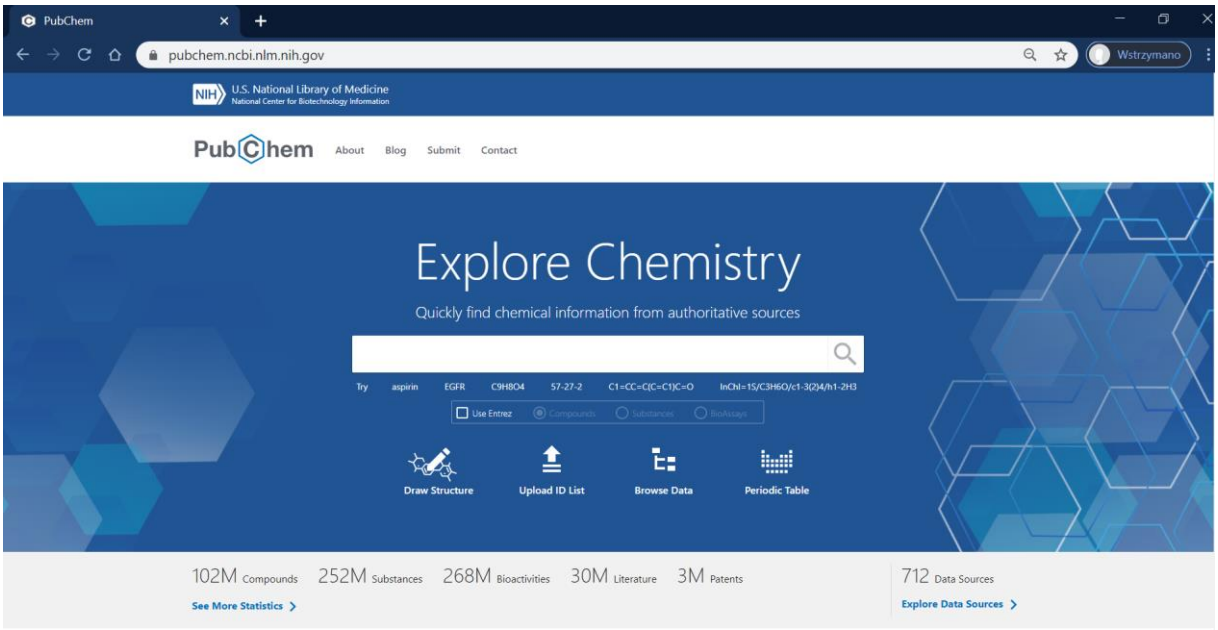
Należy zwrócić również uwagę na różnice metodyki oraz interpretacji analizy wielkich danych. Problem ten omówiono bliżej w publikacji [Polański 2017A]. Dobrym wczesnym przykładem efektywności przetwarzania wielkich danych może być symulacja epidemii grypy Google'a [Ginsberg 2009].

6.1. WIRTUALNE BAZY DANYCH ŹRÓDŁEM INFORMACJI O ZWIĄZKACH CHEMICZNYCH

W odpowiedzi na rosnące zapotrzebowanie na dostęp do danych dotyczących bioaktywności związków chemicznych stworzono szereg wirtualnych katalogów. Przykładami takich katalogów są bazy: PubChem, ChemBank, BindingDB, ChEMBL, DrugBank, ZINC oraz wiele innych, które zawierają informacje dotyczące bioaktywności pochodzące z danych literaturowych. Każdy z tych katalogów jest unikalny, ponieważ zawiera różne zbiory informacji dotyczące wybranych obszarów tematycznych [Seiler 2007]. DrugBank zawiera szczegółowe informacje dotyczące właściwości i mechanizmów działania zatwierdzonych leków [Knox 2010]. W BindingDB zawarte są informacje o powinowactwie wiązania małych cząsteczek do celów białkowych [Liu 2007]. ZINC to z kolei baza trójwymiarowych złożonych struktur przeznaczonych do ich komputerowej analizy (VS – ang. *virtual screening*) pod względem potencjalnego zastosowania farmakologicznego [Irwin 2005]. Poniżej szerzej omówiono repozytoria wykorzystywane w niniejszej pracy.

6.2. PUBCHEM

Baza danych PubChem (dostępna pod adresem internetowym https://pubchem.ncbi.nlm.nih.gov) jest publicznym, uruchomionym w 2004 roku repozytorium cyfrowym zawierającym informacje na temat substancji chemicznych oraz ich aktywności biologicznych. Zarządzana jest przez firmę National Center for Biotechnology Information (NCBI), stanowiącą część National Library of Medicine – instytucji podległej United States National Institutes of Health (NIH). NCBI jest również twórcą m.in. internetowej wyszukiwarki PubMed zawierającej artykuły z dziedziny medycyny i nauk biologicznych.



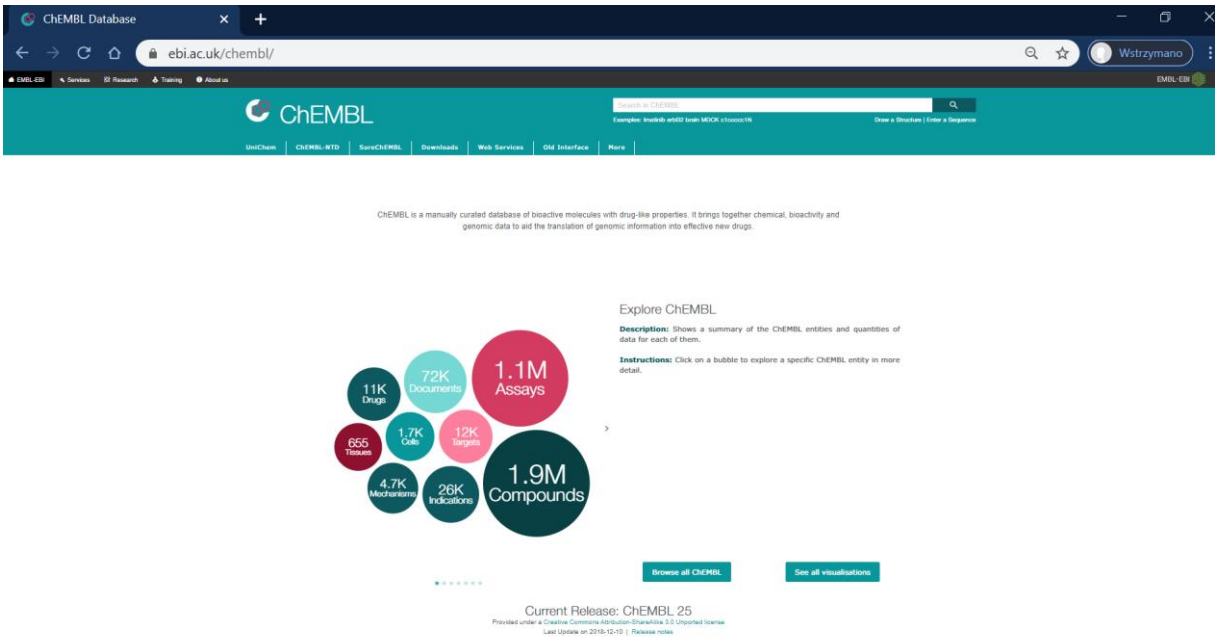
Rycina 8. Strona internetowa bazy danych PubChem [https://pubchem.ncbi.nlm.nih.gov/, data: 01.09.2020]

Baza PubChem od 16 lat stanowi cenne źródło informacji chemicznych wykorzystywanych w takich obszarach badań jak bioinformatyka, chemoinformatyka, chemia medyczna czy projektowanie leków [Kim 2016]. Cały system PubChem składa się z trzech połączonych ze sobą wewnętrznych baz: *Substance*, *Compound* i *BioAssay*. *Substance* zawiera opisy substancji chemicznych (241 545 248 rekordów) dostarczane przez użytkowników bazy PubChem. *Compound* zawiera unikalne struktury chemiczne (94 524 858 rekordów) wyodrębnione z bazy

Substance. W *BioAssay* znajdują się dane dotyczące aktywności biologicznej (1 252 826 rekordów). Dostęp do dużej ilości danych, rozbudowany interfejs graficzny i zaawansowane narzędzia wyszukiwania powodują, że PubChem jest jedną z najszerzej publicznie dostępnych i jednocześnie najczęściej stosowanych cyfrowych bibliotek informacji chemicznych. Jest ona systematycznie aktualizowana i rozwijana dzięki współpracy z innymi serwerami i firmami partnerskimi.

6.3. ChEMBL

Baza danych ChEMBL (dostępna pod adresem internetowym https://www.ebi.ac.uk/chembl) jest otwartą bazą zawierającą informacje dotyczące właściwości wiązania, funkcjonalności oraz właściwości ADMET dużej liczby bioaktywnych związków lekopodobnych. Dane te są często pobierane z pierwotnie opublikowanej literatury, a następnie dalej opracowywane i standaryzowane w taki sposób, aby zmaksymalizować ich użyteczność. ChEMBL zawiera również struktury i adnotacje dotyczące leków zaaprobowanych przez amerykańską Agencję Żywności i Leków (FDA – ang. *Food and Drug Administration*). W bazie tej zawarte są wszystkie informacje o każdym zatwierdzonym przez FDA produkcie, w tym nazwa handlowa, drogi podawania, informacje o dawkowaniu i data zatwierdzenia.

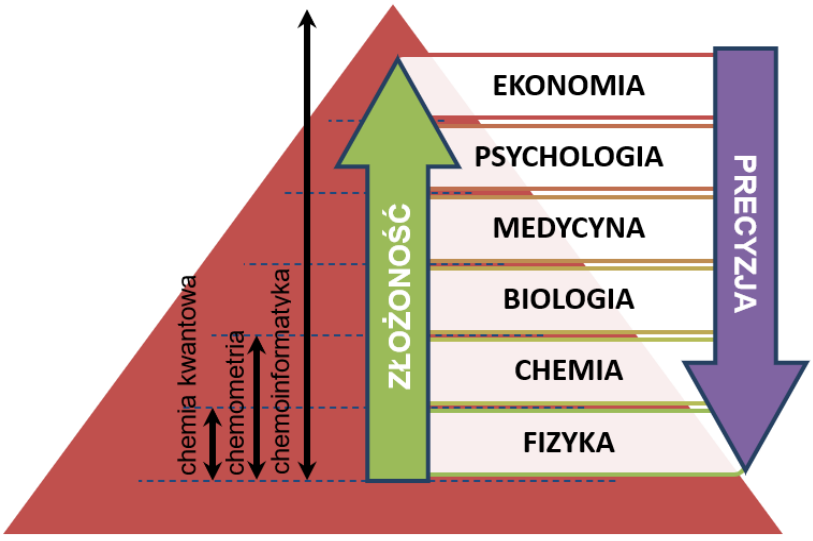


Rycina 9. Strona internetowa bazy danych ChEMBL [https://www.ebi.ac.uk/chembl/, data:01.09.2020]

Baza ChEMBL również obejmuje trzy części: *Compounds*, *Assays* i *Targets*. *Compounds* dotyczy substancji, które zostały przetestowane pod kątem ich bioaktywności (2 101 843 rekordów). *Assays* zawiera wiadomości o indywidualnych eksperymentach, które przeprowadzono w celu oceny aktywności biologicznej (14 675 320 rekordów). W *Targets* znajdują się informacje o białkach docelowych (11 538).

7. STATYSTYKI MOLEKULARNE

Analiza wielkich danych molekularnych różni się od analiz danych klasycznych. Dlatego zaproponowaliśmy by tego typu analizy określać jako statystyki molekularne. Ich przykładem mogą być analizy typu struktura-ceny związku chemicznego [**D2**]. Ciekawym problemem jest złożoność problemów i danych w różnych naukach. (ryc. 10). Według tej interpretacji fizyka jest mniej złożoną reprezentację obiektów chemicznych. Podobnie chemia redukuje złożoność nauk biologicznych [Polański 2019, Polański 2016A, Rosenblum 2006, Polański 2017A]. Warto zwrócić uwagę na to, że chemicy zsyntetyzowali nie więcej niż 200 milionów cząsteczek chemicznych, podczas gdy na świecie żyje ponad siedem miliardów ludzi. Zarówno chemia, jak i biologia nie znajdują się jednak na szczycie piramidy, która zajmowana jest przez medycynę czy psychologię. Na samym szczycie piramidy znajdują się ekonomia.



Rycina 10. Złożoność a precyzja w badaniach naukowych, zmodyfikowano wg [Polański 2019]

II. BADANIA WŁASNE

Praca doktorska została zrealizowana w Zakładzie Chemii Organicznej Uniwersytetu Śląskiego w Katowicach, w którym prowadzone są badania teoretycznych aspektów doskonalenia metod projektowania molekularnego, eksploracji danych chemicznych oraz baz danych o lekach. W niniejszej pracy badano w szczególności statystyki molekularne modelowane w dwóch bazach wielkich danych: ChEMBL oraz PubChem, które są największymi dostępnymi zbiorami informacji o wartościach aktywności biologicznych związków chemicznych. Innymi analizowanymi zbiorami danych były Binding Database (BindingBD, PTaylaorLa, USPaten, 5HT, AChE) oraz Psychoactive Drug Screening Program (PDSP).

1. ZBIORY ANALIZOWANYCH DANYCH

W tabeli 5. przedstawiłam liczbę rekordów związków chemicznych zarejestrowanych w bazach danych chemicznych [**D4**]. Wszystkie rekordy analizowane w niniejszej pracy zostały pobrane ze stron internetowych baz: ChEMBL i PubChem. Liczba rekordów pobranych z bazy ChEMBL (sierpień 2018 ChEMBL wersja 24, www.ebi.ac.uk) wynosiła 779 714, natomiast z bazy PubChem (sierpień 2017, pubchem.ncbi.nlm.nih.gov) – 2 435 467. Zapisy powtarzające się dla poszczególnych związków traktowano jako niezależne wpisy. Dodatkowo przeanalizowano dane dla wybranej serii leków zebrane w Binding Database. Obejmuje to: BindingBD – 570 927 rekordów, PTaylaorLab – 180 rekordów, USPatent – 210 254 rekordów, 5HT – 830 rekordów lub AChE – 726 rekordów (pobrano: grudzień 2018 r., www.bindingdb.org) oraz bazę danych Psychoactive Drug Screening Program PDSP – 22 273 rekordów (pobrano: grudzień 2018, pdsp.unc.edu/databases). Przeanalizowano również dane pochodzące z publikacji zamieszczonych w bibliografii w Mortenson 2018, Johnson 2016, Johnson 2019, Hopkins 2014, Schultes 2010 oraz dotyczące punktów wrzenia w Gharagheizi 2013 – 17 768 związków chemicznych.

Tabela 5. Liczebność dużych zbiorów danych dotyczących aktywności biologicznej

Źródło	Liczebność danych/ dane unikalne [mln]*	Dostępna liczba aktywności
PubChem	97,1/35,8	2 435 467
ChEMBL	2,0/0,06	779 714
Mcule	32,9/23.8	
Molport	22,4/0,14	
SureChEMBL	18,6/2,2	
ZINC	16,9/1/1	
IBM	7,9/0,31	
Emolecules	5,2/0,07	
tpharma	3,8/0,1	
nikkaji	3,2/0,3	

*na podstawie [Southan 2018]

na podstawie [D3**]

2. LE, INTERAKCJA IC_{50} I 1/MW JAKO REPREZENTACJA ZWIĄZKU CHEMICZNEGO

Celem niniejszej pracy była analiza wielkich baz danych molekularnych ChEMBL i PubChem oraz innych wybranych zbiorów danych szczegółowo opisanych w podrozdziale 1 oraz próba wyjaśnienia kontrowersji związanych ze stosowaniem LE, które zostały szeroko opisane w literaturze [Schultes 2010, Shultz 2013, Shultz 2014, Polański 2017B, Polański 2017C, Sheridan 2016]. Z jednej strony w wielu pracach LE uważa się za interesującą reprezentację [Hopkins 2004, Hopkins 2014, Kuntz 1999, Reynolds 2007, Reynolds 2008, Reynolds 2017]. Z kolei inni autorzy dyskwalifikują taką formę prezentacji właściwości związku chemicznego [Schultes 2010, Shultz 2013, Shultz 2014, Polański 2017B, Polański 2017C, Sheridan 2016].

W najszerszym znaczeniu statystycznym LE jest transformacją pAC_{50} i HAC:

$$\mathbf{1,37}\left(\frac{\boldsymbol{pIC_{50}}}{\boldsymbol{HAC}}\right)$$

Interesujące jest to, że taką transformację bada się w dziedzinie HAC. Mamy w takim wypadku efekt „splątania zmiennych", gdzie argument funkcji 1/HAC staje się swego rodzaju zmienną uwikłaną o silnym bezpośrednim wpływie na funkcję.

Z LE związana jest też wydajność liganda niezależna od wielkości, SILE (ang. *Size independent ligand efficiency*) definiowana jako:

$$\boldsymbol{SILE} = \frac{\boldsymbol{AC_{50}}}{\boldsymbol{HAC^{0,3}}}$$

LE wykorzystywana jest podczas wczesnych procesów projektowania leków [Shoultes 2010, Mignani 2018, Meanwell 2016, Cavalluzzi 2017].

3. PLE, INTERAKCJA IC_{50} I MW JAKO REPREZENTACJA ZWIĄZKU CHEMICZNEGO

W pracy [**D2**] zaproponowaliśmy modyfikacje LE do formy iloczynowej – PLE (ang. *product ligand efficiency*) definiowaną jako iloczyn siły działania i liczby atomów ciężkich:

$$\boldsymbol{PLE} = \boldsymbol{AC50} \cdot \boldsymbol{HAC}$$

i odpowiednio w postaci logarytmicznej:

$$\boldsymbol{pPLE} = \boldsymbol{pAC50} \cdot \boldsymbol{HAC}$$

Poniżej znaczenie PLE dla reprezentacji liganda molowego jako MW (kg/mol)/MW (Da) = 1, PLE można zdefiniować przez:

$$PLE = AC50 \cdot HAC \cdot \left(\frac{MW\left(\frac{kg}{mol}\right)}{MW\ (Da)} \right)$$

Ponieważ AC_{50} jest parametrem opartym na stężeniu, które ma wymiar (mol/L) wymiar PLE (jednostka) wynosi: (kg/L) · HAC/MW (Da). Oznacza to, że fizyczne znaczenie PLE to minimalne stężenie hamujące (MIC) skalowane do HAC. Ponieważ AC_{50} zwykle odnosi się do 50% stężenia hamującego, multiplikatywna LE odnosi się również do MIC_{50}, który jest podawany w kg/L. W niniejszej pracy przeanalizowano fizyczne znaczenie tego estymatora oraz przetestowanie jego zależności na danych PubChem i ChEMBL, a także na wybranej serii leków i kandydatów na leki [Hopkins 2014, Schultes 2010].

4. FUNKCJA OCENIAJĄCA (*SCORING FUNCTION*) SCORE JAKO REPREZENTACJA ZWIĄZKU CHEMICZNEGO

W niniejszej pracy [**D5**] zdefiniowałam również nowy elastyczny predyktor: *SCORE:*

$$SCORE = a \cdot pAC50 + b \cdot pHAC$$

Gdy a i b = 1, wówczas SCORE = pPLE

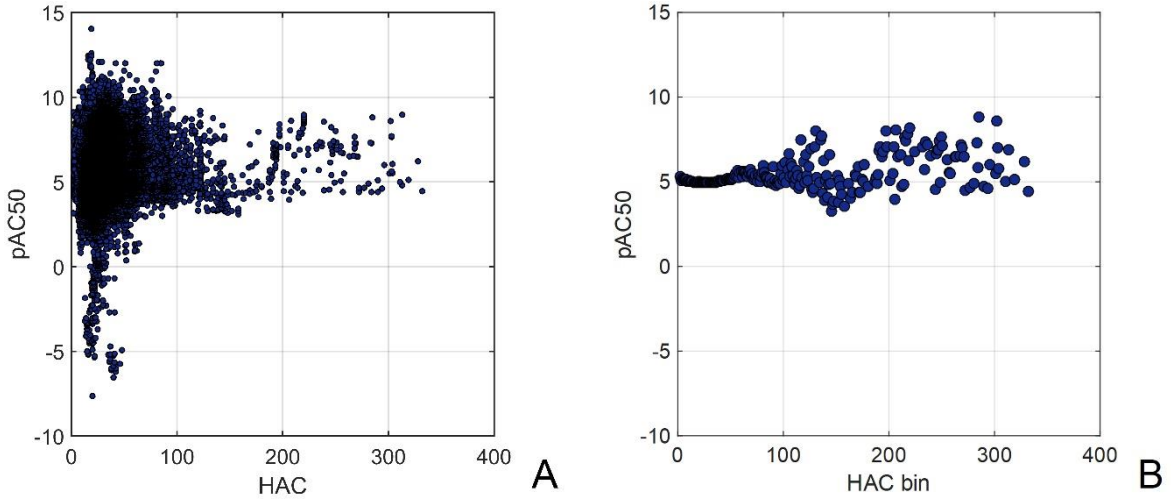
SCORE można interpretować jako transformatę pAC_{50} oraz HAC. Wartość SCORE może być regulowana w zależności od potrzeb konkretnego projektu przez zmianę wartości a i b (zmienne typu idem).

5. STATYSTYKI LE I PLE W DOMENIE WIELKICH DANYCH

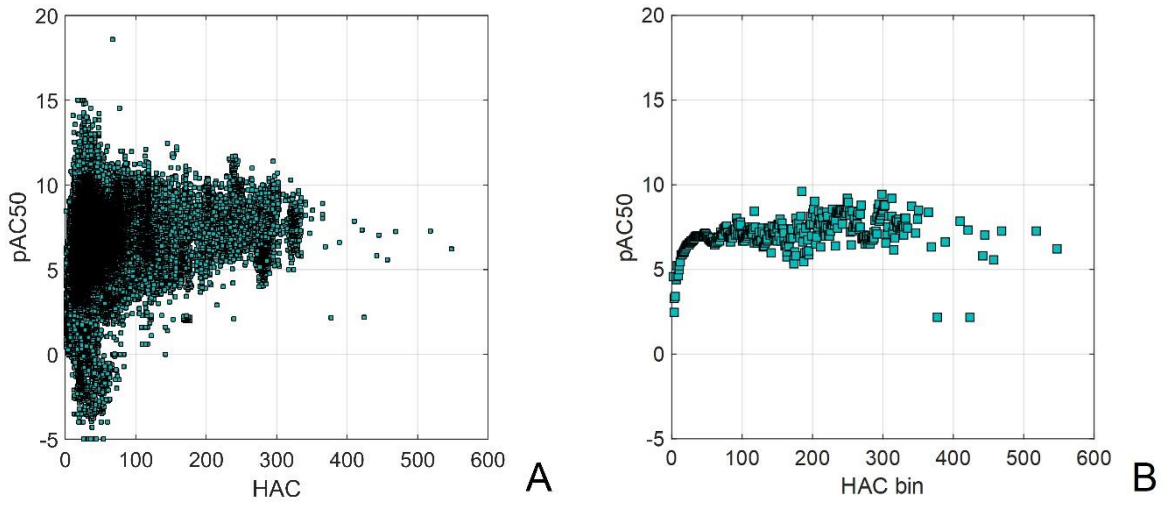
W niniejszej pracy badano statystyki różnych zbiorów wielkich danych (*Big data)*, w tym największych powszechnie dostępnych baz danych: PubChem oraz ChEMBL. Przeanalizowane zostały zależności pAC_{50} oraz LE względem tzw. zliczeń ciężkich atomów (HAC). W szczególności pobrane i przygotowane do analizy dane zostały

poddane procesowi binowania w celu określenia tendencji w nich występujących, przy czym pojedyncze biny definiują średnią wartość funkcji dla kolejnych liczb HAC.

Na rycinie 11A. i rycinie 12A. przedstawiono wyniki odpowiednio bez binowania, a następnie zestawiono je z danymi poddanymi binowaniu – rycina 11B. i 12B. Zależność pokazana na rycinach 11A. i 12A wykazuje brak uporządkowania. Z kolei na rycinie 11B. i 12B. dane wydają się znacznie bardziej uporządkowane. Do ok. 70. HAC średnia wartość pAC_{50} utrzymuje się na poziomie ok. 5. Powyżej 70. HAC zależność pAC_{50} vs. HAC przybiera formę bardziej chaotyczną. Podobnie sytuacja wygląda na rycinie 12A. (ChEMBL), gdzie przedstawiono ponad 3-krotnie mniejszą ilość danych w porównaniu do bazy PubChem. Na rycinie 12B. do pewnego momentu widoczny jest wzrost pAC_{50} wraz ze wzrostem HAC. Próba wyjaśnienia przebiegu profili na rycinie 11. i 12. musi obejmować także liczbę dostępnych rekordów vs. HAC, określonych jako liczba wystąpień znajdujących się w kolejnych binach (NOR).

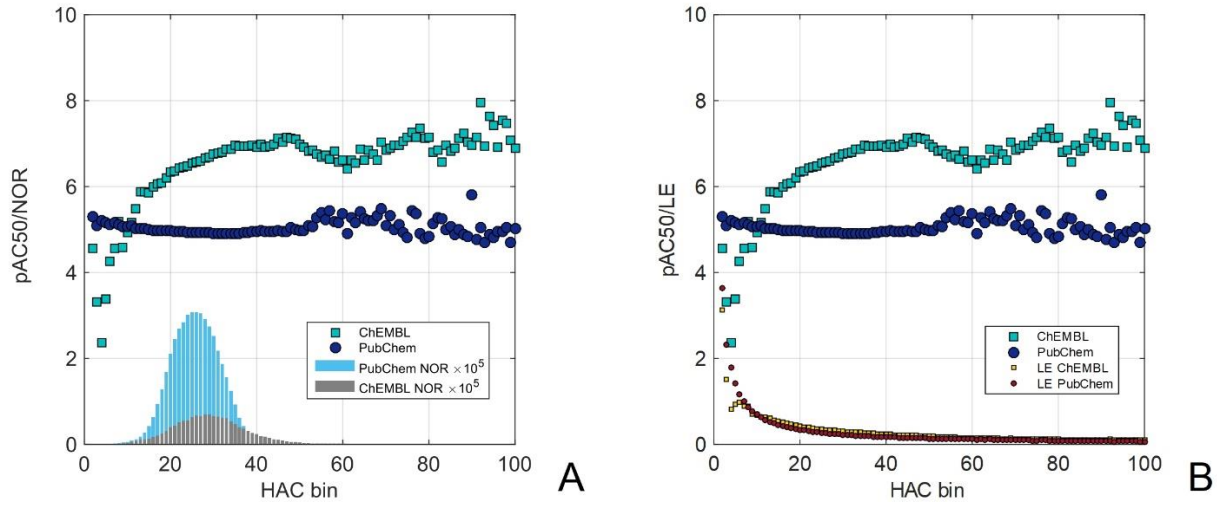


Rycina 11. Wartości aktywności niebinowane **(A)** i binowane **(B)**, względem liczby atomów ciężkich HAC w populacji PubChem



Rycina 12. Wartości aktywności niebinowane **(A)** i binowane **(B)**, względem liczby atomów ciężkich HAC w populacji ChEMBL

Rycina 13.A. przedstawia porównanie rozkładów wartości pAC_{50} i ich liczby wystąpień względem liczby HAC. Zauważalne jest wyraźne „wygładzenie" w przedziale 2.-50. HAC zarówno dla ChEMBL, jak i dla PubChem. ChEMBL, w przeciwieństwie do PubChem, wykazuje wzrost średniego stężenia pAC_{50} wraz ze zwiększaniem liczby atomów ciężkich do wartości HAC = 47, w którym pAC_{50} = 7,14. Po osiągnięciu wartości maksymalnej przy HAC = 47 spada przy HAC = 61, w którym pAC_{50} osiąga wartość 6,41, co intepretować można jako spadek precyzji dopasowania (projektowania) liganda w tym zakresie wraz ze wzrostem HAC. Ciekawe, że zależność pAC_{50} vs. HAC istotnie różni się dla zbioru PubChem i ChEMBL (ryc. 13) dla niskich wartości HAC. O ile dla zbioru ChEMBL pAC_{50} rośnie ze wzrostem HAC dla danych PubChem pAC_{50} nie zależy od HAC. Statystycznie ten ostatni przypadek oznacza, że pAC_{50} nie zależy od HAC. A więc średnia wartość siły działania leku nie zależy w tym wypadku od HAC dla danych PubChem. Inaczej w przypadku danych Pubchem, które rejestrują szerszą bibliotekę ligandów średnie pAC_{50} praktycznie nie zależy od HAC. W przypadku danych ChEMBL, który rejestruje dane projektów farmaceutycznych o wyższej jakości pAC_{50} rośnie ze wzrostem HAC.



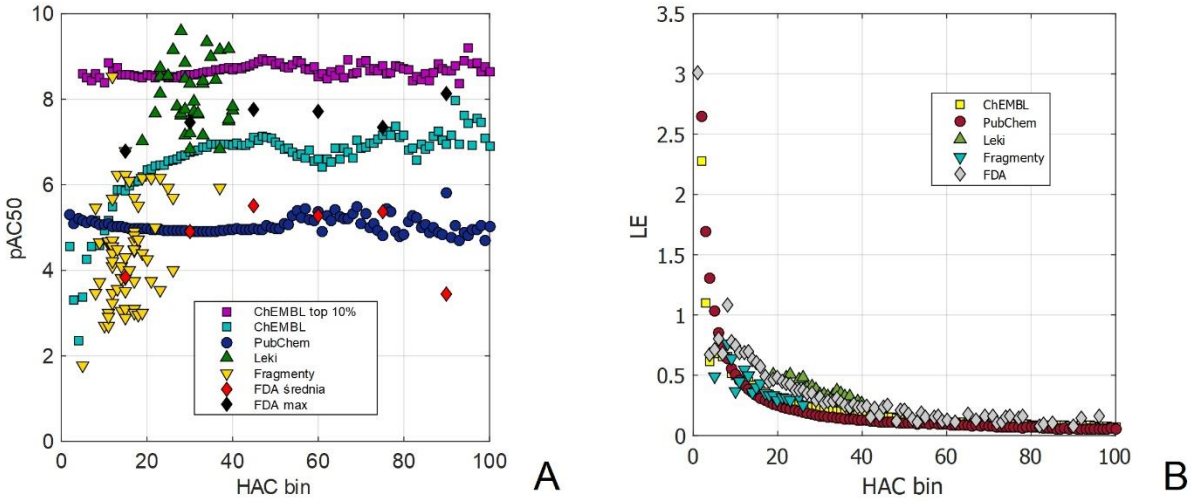
Rycina 13. Średnie wartości aktywności (*potency*) binowane, względem liczby atomów ciężkich HAC populacji ChEMBL (zielone punkty) w porównaniu z populacją PubChem (granatowe punkty), z uwzględnieniem liczby rekordów w danym binie dla obu baz **(A)** oraz wartości LE względem liczby atomów ciężkich **(B)**

Inny interesujący efekt można zaobserwować w zakresie poniżej 10. HAC, gdzie średni poziom aktywności pAC_{50} jest wyższy dla większej populacji PubChem (ale mniej wiarygodnych danych), niż dla bardziej wiarygodnej, ale mniejszej populacji ChEMBL. Średnia aktywność w tym regionie jest więc zależna od wielkości populacji. Wskazuje to na osiągnięcie stosunkowo łatwego dopasowania ligand-target dla niskich wartości HAC. Dla HAC wzrastającego powyżej 10., wyższą aktywność wykazuje z kolei populacja ChEMBL.

Według analiz przedstawionych na rycinie 13. dla danych PubChem dopasowanie ligand-target jest trudniejsze dla HAC powyżej 10 (pAC_{50} maleje ze wzrostem HAC), ponieważ dla większych ligandów brak dopasowania ligand-target może łatwiej doprowadzić do całkowitego braku aktywności. W związku z tym znalezienie optymalnych ligandów dla wyższych HAC wśród mniej systematycznych danych jest mniej prawdopodobne, zatem region poniżej 10. HAC wydaje się być obiecującym obszarem do dalszych badań. Rycina 13B. przedstawia wyniki analizy rozkładu energii wiązania lub tzw. wydajności liganda (LE) dla poszczególnych liczb atomów ciężkich (HAC). Na rycinie widoczny jest spadek LE wraz ze wzrostem HAC. Warto zauważyć, że zarówno dla ChEMBL, jak i PubChem, które mają różne populacje

i wartości aktywności, wykresy są niemal identyczne – LE nie korelują z aktywnością, tylko z zależnością 1/HAC według modelu zbliżonego do hiperboli.

Rycina 14. przedstawia analizę możliwości wykorzystywania statystyk aktywności PubChem oraz CHEMBL jako tła do oceny jakości leków zatwierdzonych przez FDA w latach 1939-2014, serii wybranych 37 leków (Imitanib, Vandetanib, Vorinostat, Vemurafenib, Sunitinib, Crizotinib (ALK), Nilotinib, Bosentan, Sitagliptin, Raltegravir, Sorafenib, Ambrisentan, Vildagliptin, Gefitinib, Pazopanib, Lapatanib (EGFR), Erlotinib, Vismodegib, Lapatanib (erbB2), Elvitegravir, Tofacitinib, Sildenafil, Crizotinib (MET), Da-satinib, Bosutinib, Tadalafil, Ruxolotinib, Alogliptin, Sax-agliptin, Sitaxentan, Linagliptin, Maraviroc, Roflumilast, Aliskiren, Vardenafil, Axitinib, Aprepitant) [Hopkins 2014] oraz fragmentów leków [Schultes 2010] w odniesieniu do ich wielkości HAC na podstawie aktywności molowej.

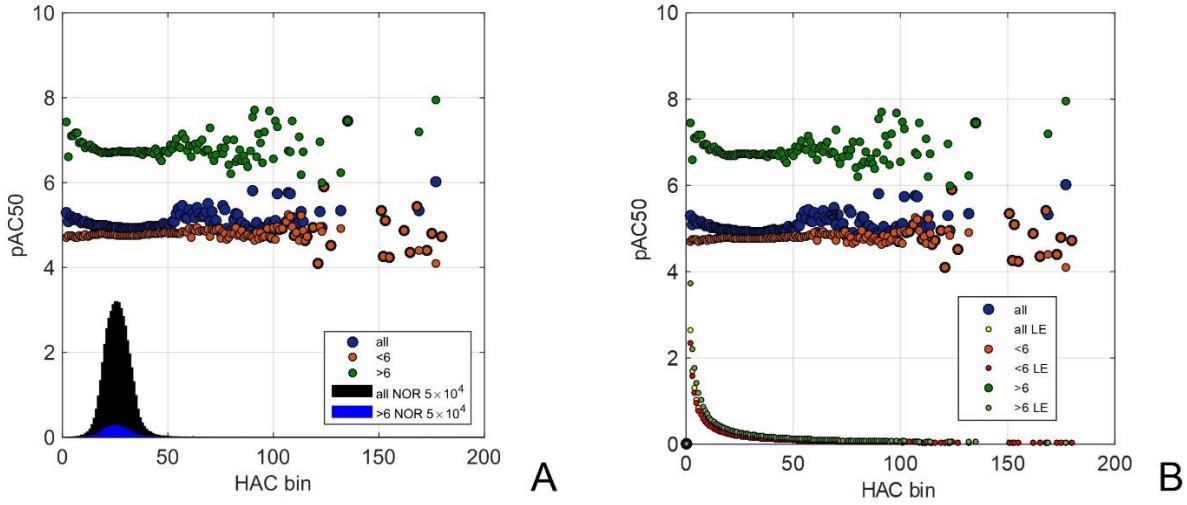


Rycina 14. Średnie wartości pAC_{50} względem HAC **(A)** oraz LE względem HAC **(B)** dla baz: PubChem, ChEMBL, FDA, Leków i Fragmentów

W tym kontekście regiony powyżej i poniżej wykresu średniej aktywności PubChem mogą kategoryzować dwie klasy. Większość z wybranych leków cechuje się aktywnością powyżej średniej ChEMBL, fragmenty można znaleźć w prawie wszystkich regionach, 74% wartości aktywności znajduje się poniżej średniej PubChem, 13% – pomiędzy wartością PubChem i średnią ChEMBL, 13% – powyżej

średniej ChEMBL. Średnia aktywność FDA zbliżona jest do średniej aktywności PubChem.

Wpływ rosnącej złożoności interakcji ligand-target zilustrowano za pomocą odrębnych wykresów dla przedziałów aktywności (poniżej i powyżej 6 HAC) względem HAC. Dla populacji powyżej 6 HAC obserwujemy wyraźny spadek średniej pAC_{50}, aż do 21 HAC, gdzie pAC_{50} = 6,69. Dla lepszego zrozumienia różnicy pomiędzy aktywnością pAC_{50} a wydajnością LE, na rycinie 15. porównano wszystkie analizowane populacje. Hiperboliczny kształt wykresu LE jest niemal jednakowy dla wszystkich danych.

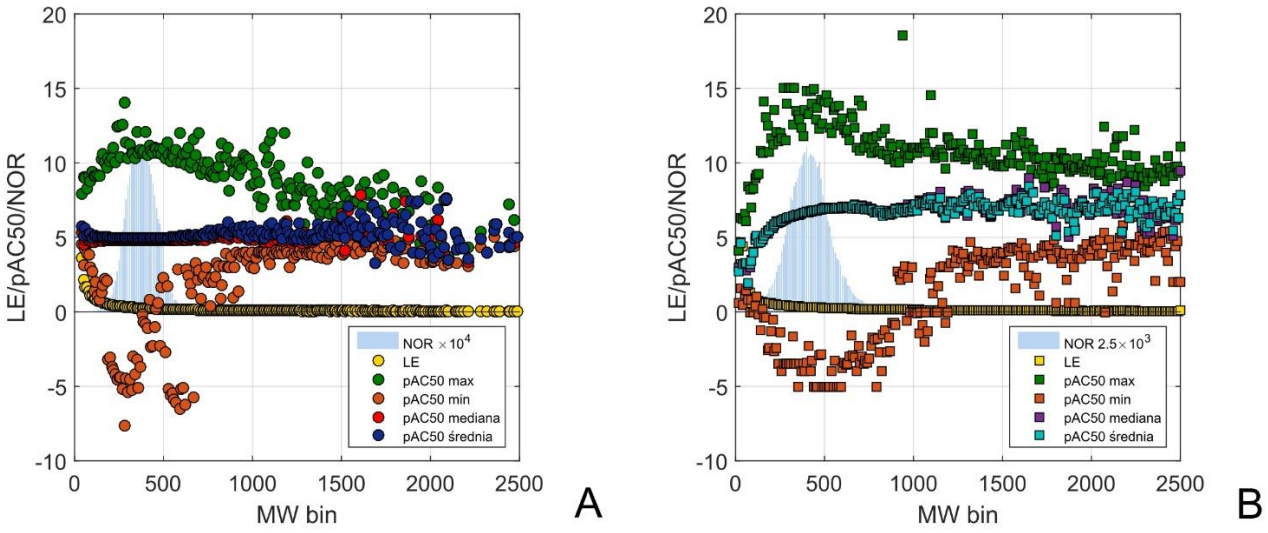


Rycina 15. Średnie wartości pAC_{50} i NOR (dla wszystkich pAC_{50}, <6. pAC_{50}, >6. pAC_{50}) względem HAC **(A)** oraz średnie wartości pAC_{50} i LE (dla wszystkich pAC_{50}, <6. pAC_{50}, >6. pAC_{50}) względem HAC **(B)** dla bazy PubChem

Średnia wartość aktywności pAC_{50} w przedziale od 2. do 5. HAC nieznacznie spada od 5,3 do 4,9, przyjmując wartość ok. 5.0 w całym przebiegu. Niewielki spadek wyjaśniać można poprzez wzrost złożoności interakcji ligand-target wraz ze wzrostem wielkości cząsteczki, co powoduje także wzrost trudności uzyskania optymalnego dopasowania ligand-receptor. W przypadku danych z PubChem wpływ HAC na średnią aktywność jest niewielki.

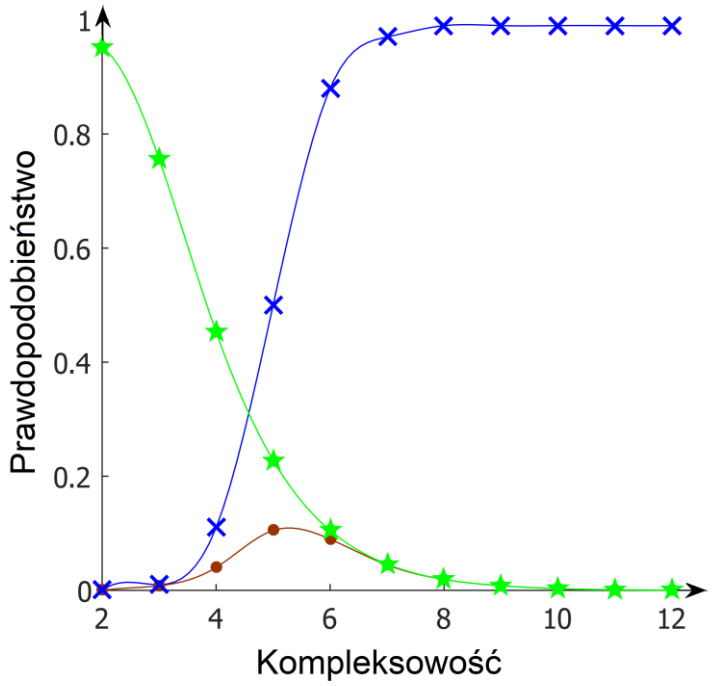
Podobne wykresy jak w dyskutowane powyżej można przedstawić dla prostych zależności pAC_{50} czy LE względem masy cząsteczkowej. Rycina 16. przedstawia

statystyki molekularne zarówno dla bazy ChEMBL, jak i PubChem. Na rycinie zostały również ukazane maksymalne i minimalne wartości pAC_{50} jakie pojawiają się w poszczególnych binach. Zarówno dla bazy ChEMBL, jak i dla Pubchem najniższe i najwyższe wartości pAC_{50} pojawiają się w miejscach gdzie populacja dostępnych danych jest najwyższa.



Rycina 16. Zależność LE, pAC_{50} i HAC względem MW dla danych z PubChem **(A)**, ChEMBL **(B)**

Co ciekawe, po porównaniu do rozkładu wartości pAC_{50} z prawdopodobieństwem wykrycia, pomiaru i dopasowania modelu ligand-receptor można zauważyć interesującą korelację między poszczególnymi wykresami rozkładu prawdopodobieństwa zaprezentowano na rycinie 17., a aktywnością pAC_{50}.



Rycina 17. Wykres prawdopodobieństwa wzajemnych oddziaływań lek-receptor. kolory kodują: prawdopodobieństwo dopasowania lek-receptor w dowolny sposób (kolor zielony), prawdopodobieństwo wykrycia aktywności liganda (kolor niebieski), mierzoną wartość aktywności (kolor czerwony) zmodyfikowany wg [Zartler 2005]

Pojawia się zatem pytanie: jak wysoka powinna być optymalna siła działania (*potency*) leku? Często uznaje się projektowanie leków za proces, którego celem jest maksymalizacja wartości siły działania, co nie zawsze znajdzie odzwierciedlenie w praktyce. W rzeczywistości o potencjale tzw. kandydatów na leki decyduje szereg tzw. właściwości lekopodobnych, np. ADMET i właściwości lipofilowe. Co ciekawe, podczas projektowania leków zawsze mierzy się siłę działania, z kolei właściwości ADMET czy parametry lekopodobieństwa są zwykle prognozowane *in silico* i rzadko mierzone. Formalnie wartości ADMET czy logP nie są więc właściwościami *sensu stricte*, lecz deskryptorami molekularnymi, które wyliczane są na podstawie struktur odpowiednich cząsteczek tak, by symulować prognozowane wartości właściwości nieznanych projektowanych struktur.

Nie można zweryfikować projektowania żadnego leku bez pomiaru jego aktywności biologicznej. Natomiast ADMET i parametry lekopodobieństwa, które są również bardzo ważne nie są jednak absolutnie konieczne do projektowania. Powód, dla

którego pomiary są ograniczone do jednej cechy jest jasny: pomiary są drogie. Efekt ten został już opisany jako tzw. deficyt właściwości [Polański 2019,Polański 2017A]. Okazuje się zatem, że nie tylko chemia czy projektowanie leków, ale również aspekty ekonomiczne ostatecznie determinują obecny kształt farmakologii [Polański 2015].

5.1. Odwzorowania ΔpAC_{50}, ΔPLE i ΔLE względem HAC dla danych opisujących projekt od fragmentu do celu (dane F2L)

Jako estymator LE zawsze wykazuje preferencję względem małych ligandów, przybierając w tym zakresie najwyższą wartość. Pojawia się zatem pytanie: czy można zaprojektować estymator, który równoważyłby bardziej równomiernie interakcję między siłą działania a HAC. W praktyce niskie wartości HAC i niskie wartości AC_{50} (wysoka siła działania) wskazują atrakcyjne potencjalne leki (kandydatów na leki). Próbą zdefiniowania takiego parametru jest iloczynowa interakcja siły działania i HAC (Product LE: PLE) [**D5**].

Ciekawą biblioteką AC_{50} są dane opisane w publikacjach [Mortenson 2018, Johnson 2016, Johnson 2019] obejmujące molekuły wiodące (ang. *lead compounds*) oraz fragmenty określane razem jako F2L (ang. *fragment to lead*). Autorzy tych publikacji zauważyli, że w procesie F2L (tzn. zamiany fragmentu w rzeczywisty cel) LE może zarówno zwiększać się, jak i zmniejszać (delta-LE może być dodatnia jak i ujemna). Obserwację tę określamy poniżej jako efekt F2L. Jest to ciekawy efekt ponieważ przejście F2L powinno z założenia łączyć się ze wzrostem pAC_{50}. W tym kontekście Autorzy zalecają jednak w stosunku do zebranych danych ostrożność, ze względu na to, że pochodzą one z publikacji a nie przemysłu farmaceutycznego. Ich zdaniem względy publikacyjne mogą decydować o „przekłamaniach".

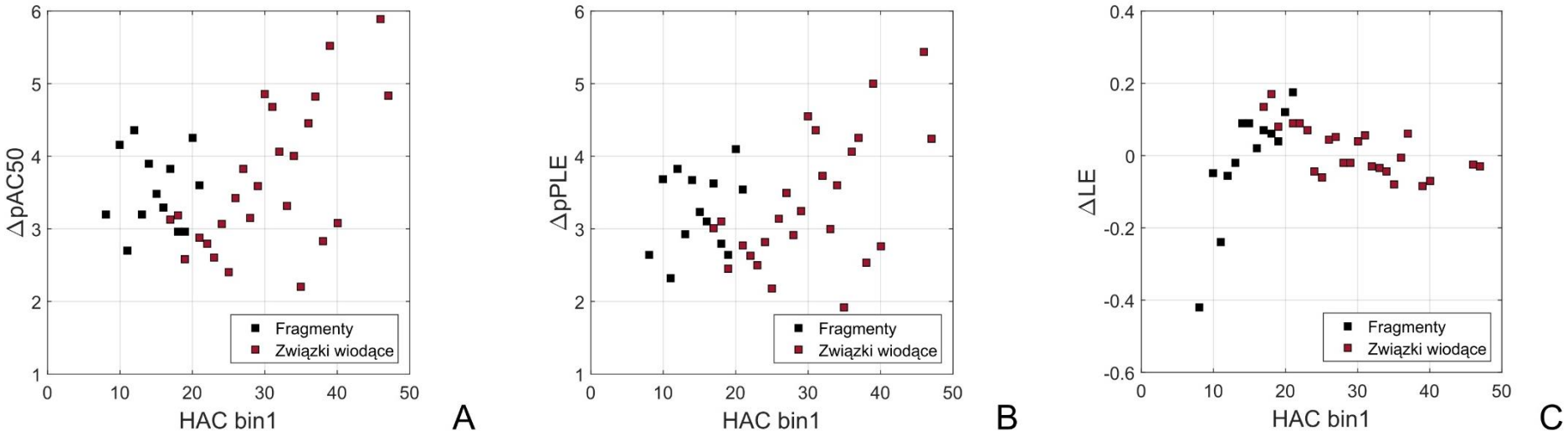
W kontekście pokazanych analiz pojawiają się dwa pytania:

- Czy można wskazać kluczowe czynniki decydujące o efekcie F2L?

- Czy należy oczekiwać jakiejkolwiek regularności dla stosunkowo małej populacji ligandów opisanych w publikacjach [Mortenson 2018, Johnson 2016, Johnson 2019]?

Na rycinie 18. przeanalizowałam dane F2L [Mortenson 2018, Johnson 2016, Johnson 2019], gdzie pokazałam zmiany wartości ΔpAC_{50}, ΔLE lub ΔpPLE w funkcji HAC. Inaczej niż w oryginalnej pracy pokazano wartości ΔpAC_{50}, ΔLE lub ΔpPLE w funkcji HAC nie tylko dla fragmentów, lecz także dla ligandów. Zgodnie z oczekiwaniami nie zaobserwowano korelacji między ΔpAC_{50}, ΔpPLE i HAC dla fragmentów. Z kolei dla fragmentów ΔLE vs. HAC wskazuje trend rosnący. Im niższy HAC fragmentu, tym niższy jest również delta-LE dla związanych z nimi związków wiodących. W przypadku fragmentu o niskich HAC wzrost ΔAC_{50} przez związek wiodący nie może zrównoważyć wkładu HAC decydującego o wartości LE fragmentu poprzez hiperboliczny składnik 1/HAC. Te zależności wyraźnie ilustrują dominujący wpływ HAC na ΔLE. Innymi słowy to wielkość liganda decyduje o ΔLE, a przecież proces F2L ma za zadanie doskonalenie fragmentu, tak by stał się udanym projektem leku (liganda). Jeżeli ΔLE maleje to wydawało by się, że projekt jest nieudany. Ciekawe też, że przejście od fragmentów do związków wiodących na osi X w zasadzie nie zaburza ciągłości funkcji delta-LE (ryc. 18C). To także zastanawia, sugerując dominujące znaczenie HAC niezależnie, czy analizujemy fragment, czy związek.

Podsumowując efekt F2L pokazuje, że LE jest estymatorem, którego wartość w dużym stopniu jest zdominowana przez wartość HAC.



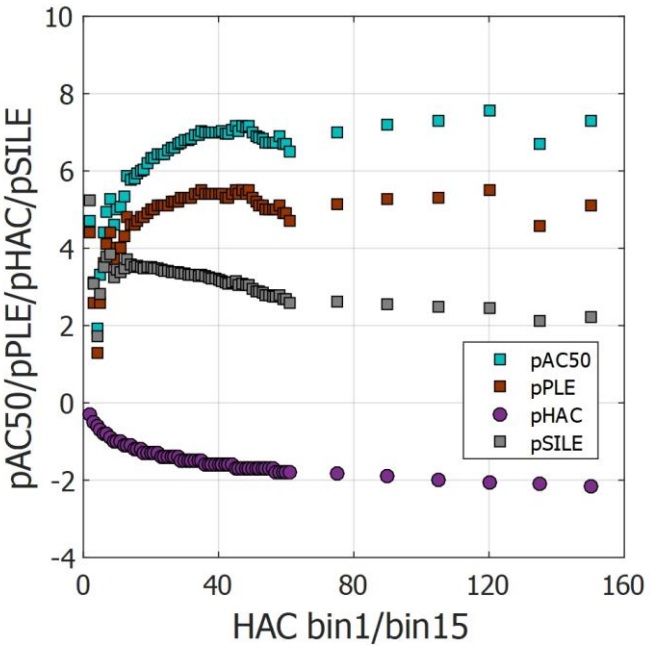
Rycina 18. Zależność różnicy siły działania fragmentów i związanych z nimi związków wiodących mierzona jako ΔpAC_{50} (A), ΔpPLE (B) lub ΔLE (C) jako funkcja HAC, dane zmodyfikowane wg [Mortenson 2018, Johnson 2016, Johnson 2019].

5.2. ODWZOROWANIA pAC_{50}, PLE, pPLE I LE WZGLĘDEM HAC DLA DANYCH PUBCHEM CHEMBL

Mała populacja F2L niekoniecznie musi odzwierciedlać trendy w większych zbiorach danych. Dlatego poniżej przedstawiono statystyki pAC_{50}, PLE i LE dla dużych baz aktywności PubChem oraz CHEMBL.

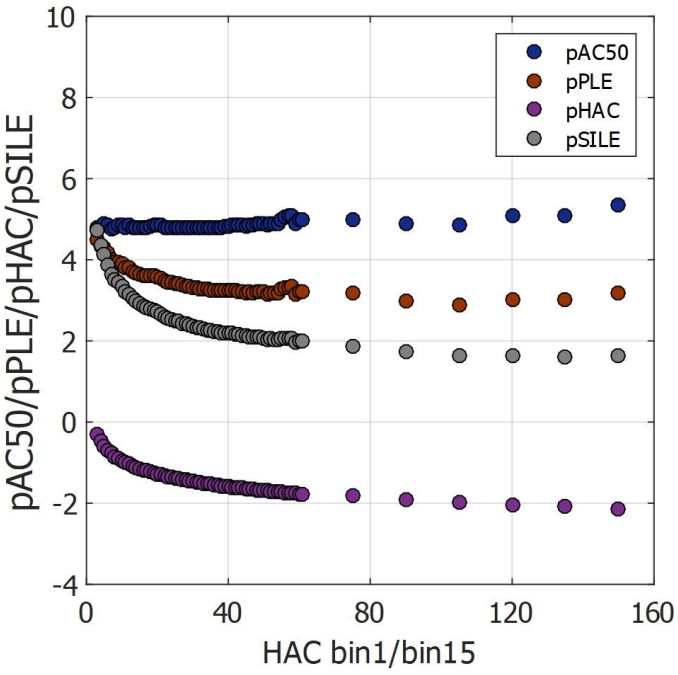
Powszechnie stosowaną reprezentacją AC_{50} jest ujemna skala logarytmiczna (pAC_{50}), podobna do skali pH, w której wyższe wartości pAC_{50} wskazują większą siłę działania w skali wykładniczej. Miarę PLE można łatwo przekształcić do miary pPLE, która jest analogiczna do pAC_{50} poprzez wykorzystanie funkcji logarytmicznej. Poniżej przedstawiłam analizy przeprowadzone z wykorzystaniem miary pPLE (wyższa pPLE oznacza lepszą jakość). Ponieważ logarytm iloczynu jest sumą logarytmów, pPLE można defragmentować do jego składników logHAC (pHAC) i pAC_{50}, jak pokazano na rycinie 19.

Na rycinie 19. dane ChEMBL pokazują wzrost pAC_{50} aż do wartości około 50 dla HAC. Jednocześnie interakcja między pAC_{50} i pHAC jest widoczna w pPLE, którego optimum przesuwa się nieznacznie w kierunku niższych wartości HAC w porównaniu do maksimum wykresu pAC_{50}. W szczególności dla wykresu pPLE maksimum przy ok. 30-50 HAC jest szersze, podczas gdy obniżenie wysokiego HAC (160) jest wyższe.

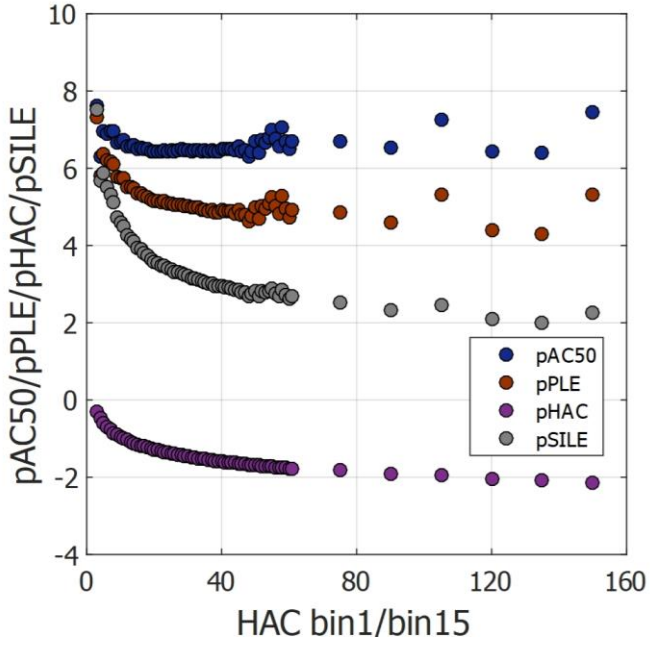


Rycina 19. Zależność pAC_{50}, pPLE, pHAC i pSILE od HAC dla danych z ChEMBL

Na rycinach 20. i 21. zilustrowano dane PubChem, które wskazują, że pAC_{50} jest w zasadzie stałą funkcją HAC, potwierdzając w ten sposób, że siła działania pAC_{50} nie jest generalnie funkcją wielkości cząsteczki, jeśli sondowana populacja substancji czynnych jest wystarczająco duża a ich projektowanie nie jest precyzyjne. Z kolei dla najbardziej aktywnych związków (dane) PubChem (ligandy pAC_{50} >6) pAC_{50} zmniejszał się wraz ze wzrostem HAC (ryc. 21.), co wskazuje, że prawdopodobieństwo prawidłowego dopasowania ligand-cel maleje wraz ze wzrostem wielkości cząsteczki dla ligandów o wysokiej aktywności. Wg funkcji pPLE w obu przypadkach (ryc. 20.-21.) najniższy zakres HAC wskazuje optymalne potencjalne leki (*drug candidates*), Odpowiednio, w danych ChEMBL, PubChem, w różnych zakresach HAC, zaobserwować można wszystkie możliwe scenariusze interakcji pAC_{50} vs. HAC przewidziane przez funkcję pPLE.



Rycina 20. Zależność pAC_{50}, pPLE, pHAC i pSILE od HAC dla danych z PubChem

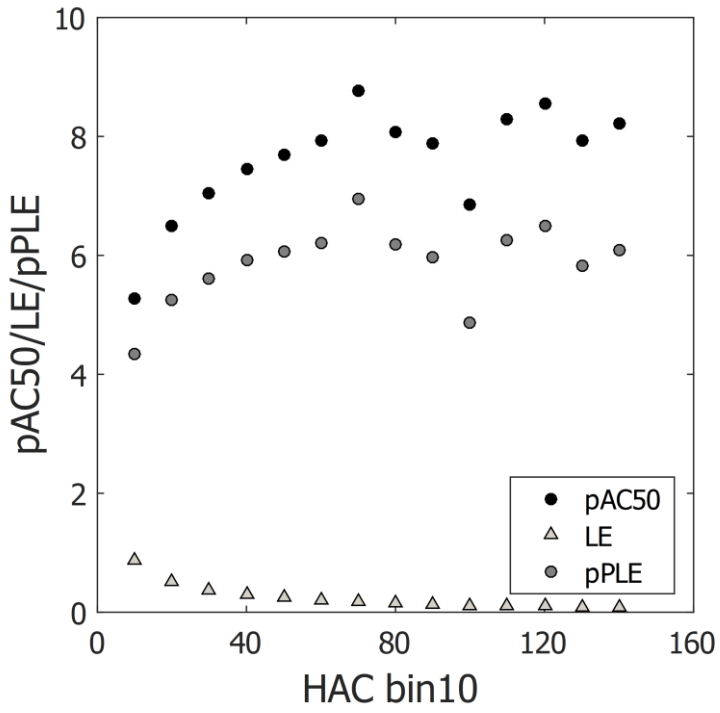


Rycina 21. Zależność pAC_{50}, pPLE, pHAC i pSILE od HAC dla danych z subpopulacji PubChem z pAC_{50} >6.

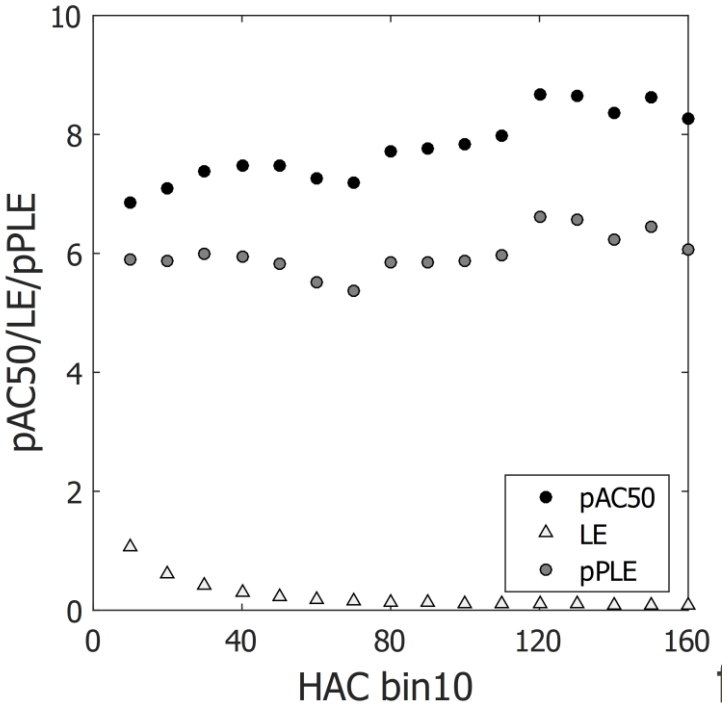
Dodatkowo dla porównania na rycinach 20.-21. zilustrowano wykresy SILE, który zajmuje miejsce pomiędzy wykresami pPLE i pHAC, gdy:

$SILE = AC_{50}/HAC^{0,3}$

Podobne zależności przedstawiam dla danych dotyczących opatentowanych w Stanach Zjednoczonych potencjalnych leków (*drug candidates*) (ryc. 22.) oraz dla psychoaktywnych leków PDSP (ang. *Psychoactive Drug Screening Program*) (ryc. 23.). W przypadku pierwszych danych typowy trend wzrostu AC_{50} i pPLE vs. HAC można zaobserwować dla HAC poniżej 50, z kolei dla psychoaktywnych leków PDSP wzrost wartości pAC_{50} w stosunku do z HAC nie był wystarczająco silny, aby obserwować wzrost trendu pPLE vs. HAC.



Rycina 22. Zależność pAC_{50}, LE i pHAC od HAC dla *drug candidates* z bazy Binding DB

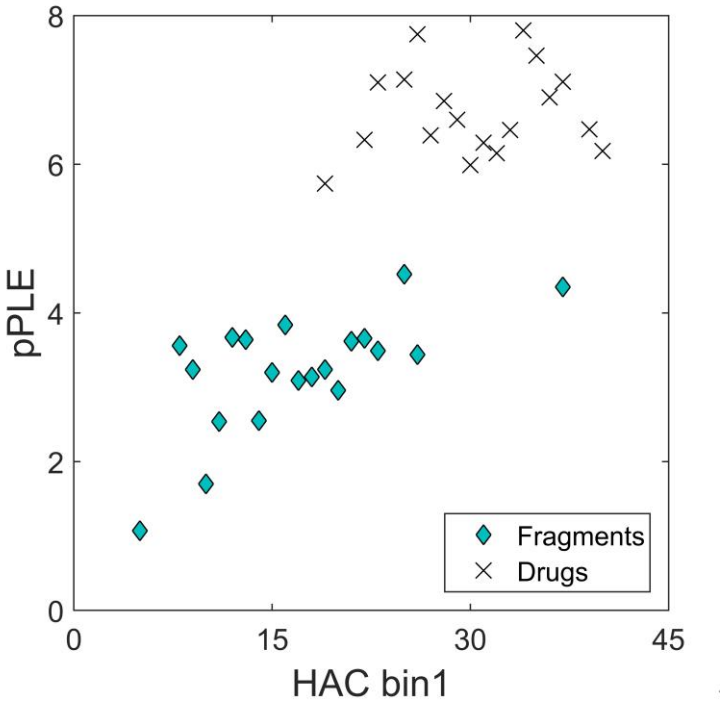


Rycina 23. Zależność pAC_{50}, LE i pHAC od HAC leków PDSP (*Psychoactive Drug Screening Program*)

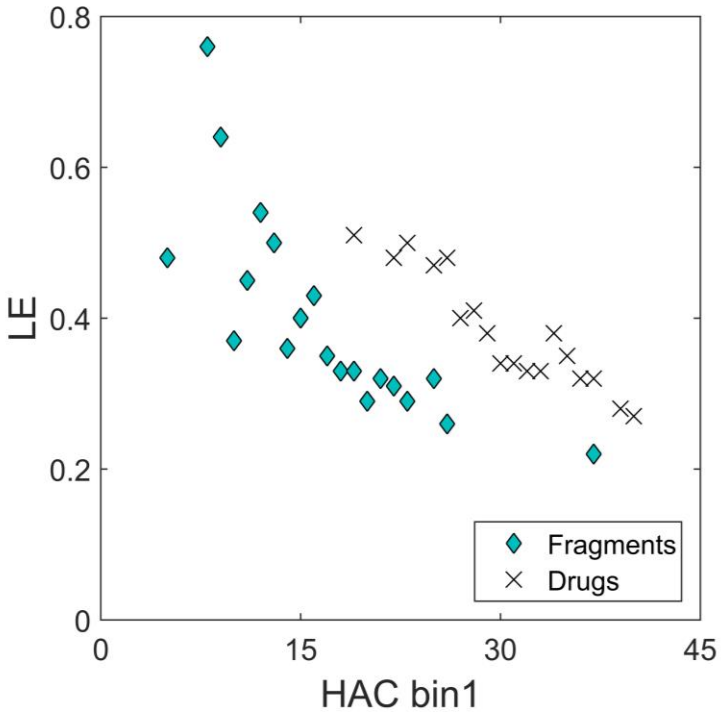
5.3. STATYSTYKI PPLE DLA LEKÓW I FRAGMENTÓW

Na rycinie 24. pokazano zastosowanie pPLE do oceny serii leków [Hopkins 2014] oraz fragmentów leków [Schultes 2010]. Niezależnie od zakresu HAC wartość pPLE była zawsze wyższa dla leków niż dla fragmentów. Nastąpiło wyraźne oddzielenie klasy leków od klasy fragmentów. Ponadto wszystkie leki miały wyższą wartość pPLE niż fragmenty. W związku z tym pPLE może być dobrze zrównoważonym predyktorem, który może wyraźnie wskazywać rozwój od fragmentów do leków.

Z kolei rycina 25. ilustruje statystyki LE dla tych samych danych. Analiza wykresu LE vs. HAC pokazuje, że dla danego HAC powyżej 20. leki miały nieco wyższą wartość LE niż fragmenty. Jednak dla wartości HAC poniżej 20. LE dla fragmentów jest wyższy. Wartości takiej nie może osiągnąć jakikolwiek lek.



Rycina 24. Zależność pPLE od HAC dla fragmentów leków i leków



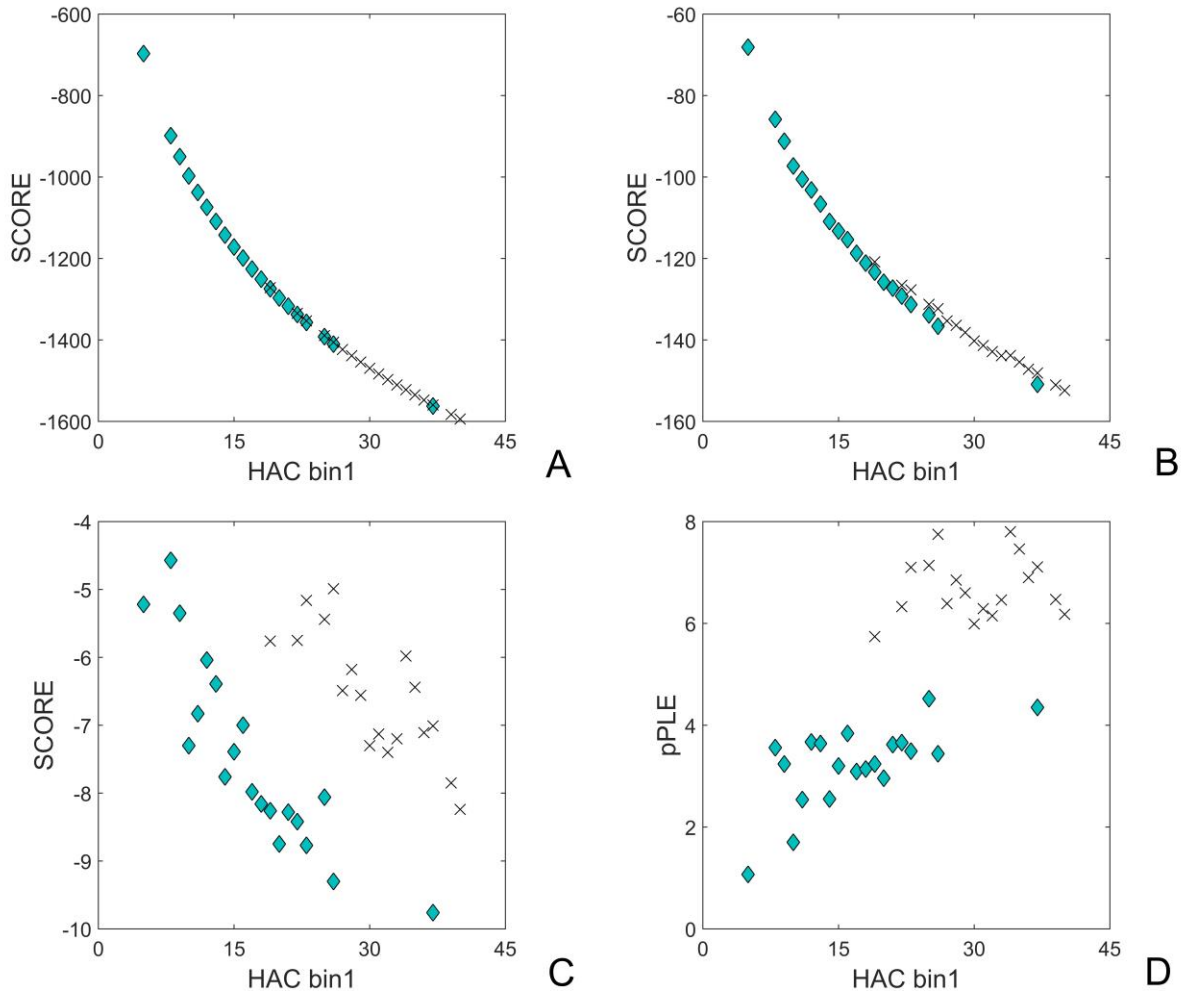
Rycina 25. Zależność LE od HAC dla fragmentów leków i leków

Powstaje wobec tego problem czy oznacza to, wyższość fragmentu nad zbudowanym z niego gotowym lekiem. Oczywiście ponieważ odpowiedź jest negatywna powstaje problem, jaki estymator jest właściwy dla szacowania jakości optymalizacji F2L. Poniżej przedstawiłam jedno z rozwiązań, którym jest funkcja scoring (*scoring function*) *SCORE* **[D5]**. Matematyczną podstawą jest fakt, że efekt składników pHAC i pAC_{50} można dostosować za pomocą dodatkowych parametrów a i b, których wartości można zmieniać (a,b = *idem*):

$$\mathbf{SCORE = a \cdot pAC_{50} + b \cdot pHAC}$$

jeśli a i b = 1, to SCORE = pPLE.

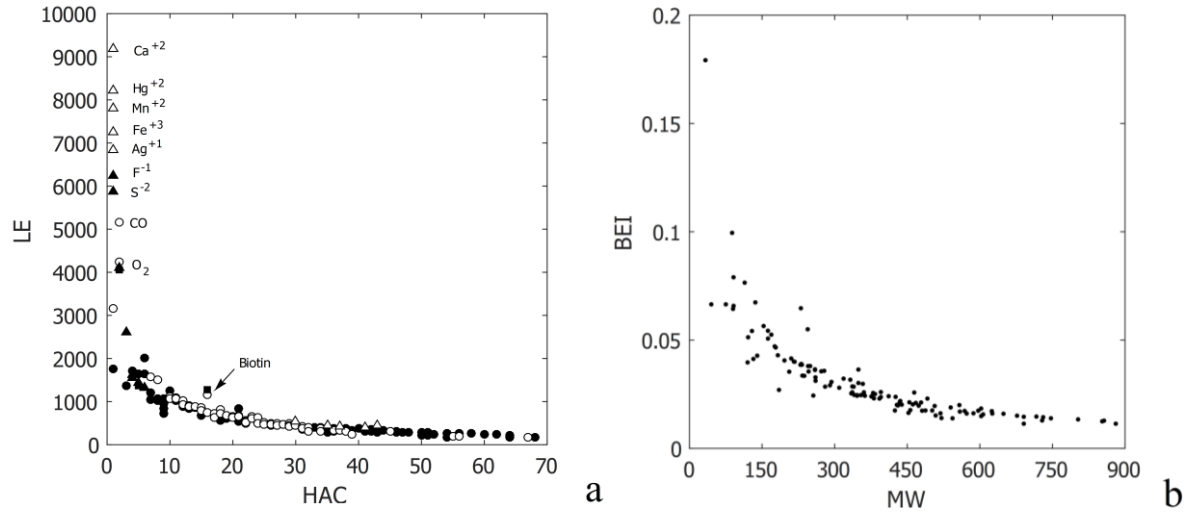
Matematycznie wartości liczbowe HAC wynoszą zwykle od 1 do 200, podczas gdy pAC_{50} zwykle mieści się w zakresie od 1 do 7; zatem podczas obliczania wartości LE udział 1/HAC wyraźnie dominuje w LE. Innymi słowy, LE jako predyktor kładzie większy nacisk na niski HAC niż na wysoką siłę wiązania pAC_{50}. Z kolei pAC_{50} (w zakresie HAC od 10^1 do 10^7) dominuje w pPLE (ryc 26). SCORE został zaprojektowany jako predykator, w którym wpływ pAC_{50} i HAC można dostrajać do potrzeb przez zmianę wartości parametrów a i b, dla bardziej zrównoważonej interakcji między AC_{50} a HAC. Funkcję SCORE można elastycznie dostosować, aby zmienić relacje SCORE np. fragmentów i leków, a także łącznego zbioru fragmentów i leków. W szczególności zmiana parametrów a i b może wspomagać różne strategie opracowywania leków. W związku z tym, w zależności od preferencji, możemy precyzyjnie dostroić wartości a i b w funkcji SCORE w celu przyjęcia strategii preferowanej w konkretnym projekcie rozwoju leku.



Rycina 26. Wartość SCORE jako funkcja HAC dla fragmentów leków i leków dla różnych wartości a i b: a = 1, b = 1000 **(A)**; a = 1, b = 100 **(B)**; a = 1, b = 10 **(C)**; a = 1 b = 1 **(D)**

5.4. EFEKTY HIPERBOLICZNE DLA PROFILI LE VS. HAC I CENY ZWIĄZKU CHEMICZNEGO VS. MW

LE stosowana jako predyktor preferuje ligandy o niskiej masie cząsteczkowej lub niskiej liczbie atomów ciężkich (HAC). Przez długi czas efekt tak dużego wpływu HAC pozostawał zagadką (ryc. 27).



Rycina 27. Zależność LE vs. HAC **(a)** i BEI vs. MW **(b)** dla serii ligandów zmodyfikowany wg [Kuntz 1999] według pracy [**D4**]

Obserwowane zależności LE od HAC zawsze wykazują trend krzywoliniowy zbliżony do hiperboli. LE został zaprojektowany jako parametr, który ma odpowiedzieć na pytanie jaka jest najwyższa możliwa wartość siły oddziaływania (*potency)* liganda dla pojedynczego HAC. Statystycznie LE (BEI) jest interakcją siły oddziaływania (pIC_{50} lub pAC_{50}) i 1/HAC (1/MW). Jednak ze względu na dualizm reprezentacji związku chemicznego reprezentowanego przez pojedynczą cząsteczkę lub mol substancji dla reprezentacji molowej interpretacja czynnika 1/MW (1/HAC) jest zupełnie inna. Działanie 1/MW sprowadza się bowiem do zamiany efektu molowego na efekt wyrażony w skali wagowej. Np. stężenie molowe pomnożone przez 1/MW daje wartość stężenia wagowego ([mol/l] x [1/g/mol] = [g/l]). Na rycinie 28. przedstawiłam interpretację fizycznego znaczenia LE (BEI) jako transformaty właściwości molowej, która odpowiada skali wagowej. Ponieważ skala wagowa nie zachowuje stałej liczby cząsteczek w pojedynczej jednostce (g, kg) liczba ligandów dla substancji małych cząsteczek (niskie HAC, MW) dostępna dla receptora jest znacznie większa niż dla cząsteczek dużych (wysokie HAC, MW). Tak więc użyteczność LE w farmacji sprowadza się do promowania niskich wartości HAC, MW na skutek ww. efektu, co jest zgodne z koncepcją *slim pharma* czy *molecular obesity*. Warto także podkreślić, że LE w literaturze bada się wyłącznie dla rzeczywistych leków. Nigdy zaś nie prowadzi się prognozowania wartości tej wielkości. Prognozowane są

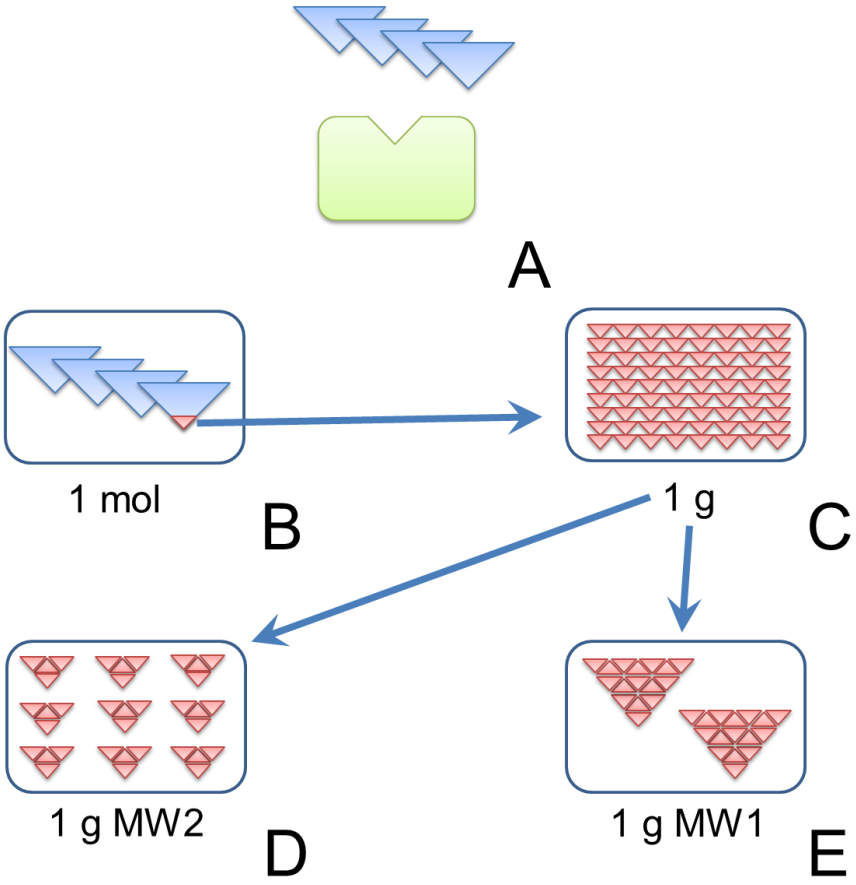
wyłącznie wartości pAC_{50}. W wyniku tego LE jest najbardziej wiarygodnym obrazem lekopodobieństwa. Jeżeli związek chemiczny okaże się chybionym projektem leku jego wartość LE nigdy nie pojawi się w literaturze.

Matematycznie, skala LE może być interpretowana jako interakcja między właściwościami molowymi a operatorem 1/MW. Hiperboliczne działanie operatora 1/MW istotnie wpływa na funkcje LE. Jeśli hipotetyczny proces projektowania leków chcielibyśmy oprzeć wyłącznie o maksymalizację wartości pAC_{50}, wówczas wartości pAC_{50} jako samodzielna zmienna dają większą precyzję odwzorowania vs. HAC, pokazując, gdzie szanse na znalezienie ligandów o wysokim pAC_{50} są największe.

Jednak w obecnym projektowaniu leków należy kierować się nie tylko wysokim pAC_{50}, ale także biodostępnością, właściwościami ADMET czy też regułą Lipińskiego, w których niska wartość MW ma kluczowe znaczenie, a *Slim pharma* jest uznawany za preferowany trend w projektowaniu leków [Hann 2011]. Niespodziewanie okazuje się, że LE spełnia potrzeby estymatora (w praktyce jednoparametrycznego) do wczesnej optymalizacji projektowania leków, poprzez nacisk na minimalizację MW. Taką interpretację LE i MW omówiono dokładniej w pracach [**D4,** Polański 2017A, Polański 2017B].

Na rycinie 28. przedstawiono rzeczywiste znaczenie LE i wskaźnika efektywności wiązania (BEI), który jest miarą związaną z LE. Siła działania jest miarą opartą na termodynamice, która jest zgodna ze statystyką Avogadro, która dla jednego mola zawsze zachowuje stałą liczbę cząsteczek niezależnie od rozmiaru cząsteczki. Działanie operatora 1/HAC (1/MW) decyduje o tym, że po transformacji siły działania (np. pIC_{50}) do miary typu LE interpretowanej jako właściwość substancji nie jest zachowana statystyka Avogadry.

W kontekście statystyki Avogadry obowiązującej dla reprezentacji molowej ww. efekt opisać można także następującym paradoksem. Odpowiednio 1g jest jednym molem Daltonów. W rzeczywistości nie istnieje jednak mol Daltonów [**D3, D4**].



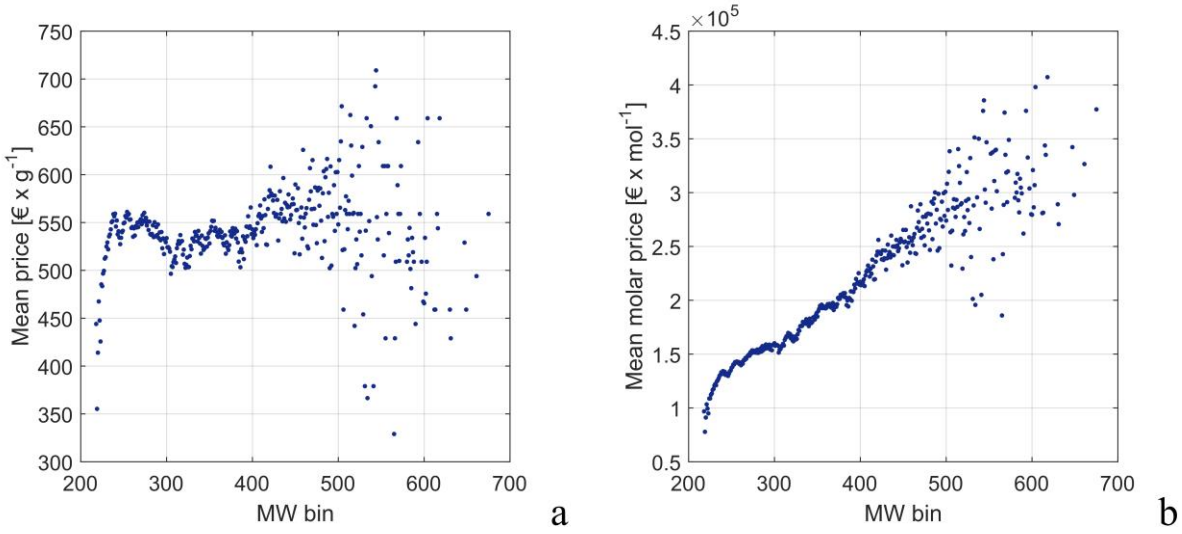
Rycina 28. Fizyczne znaczenie BEI (LE)

Z matematycznego punktu widzenia interakcja IC_{50} i 1/MW (1/HAC) odwzorowywana vs. MW (HAC) prowadzi do efektu splątania danych, gdzie argument funkcji 1/HAC staje się swego rodzaju zmienną uwikłaną o silnym bezpośrednim wpływie na funkcję.

Z teoretycznego punktu widzenia interesujące analogie pojawiają się podczas porównania statystyk LE z tymi, które są używane w ekonomii. W ekonomii powszechnie stosowaną miarą wyrażania ceny jest skala wagowa ($/g), a nie skala molowa ($/mol). Dane ekonomiczne są rzadko dostępne, jednakże Polański wraz z zespołem podjęli próbę analizy 2 mln cen związków chemicznych dostępnych w katalogu Abamachem [**D2**].

Analiza danych ekonomicznych jest zdecydowanie bardziej złożona niż danych chemicznych. Generalnie analiza biblioteki ABAMACHEM prowadzi do wniosku,

że płacimy za ilość materii, która jest skorelowana z MW. Ciekawym efektem jest „efekt hiperboliczny obserwowany dla „ceny w skali wagowej" przy niskich wartościach MW (ryc. 29), gdzie obserwujemy nieliniowe zniekształcenia trendu. Należy pamiętać, że cena nie jest wartością niezmienną, a użycie jej w tym miejscu miało na celu porównanie skal molowych i wagowych oraz wyjaśnienie, że nawet tak „nieostra" zmienna (formalnie jest to właściwość substancji) może wykazywać efekt skali wagowej.



Rycina 29. Binowane ceny biblioteki 2 mln. zw. chemicznych ABAMACHEM [**D2**]

5.5. BADANIE WSPÓŁCZYNNIKÓW DETERMINACJI (KORELACJI) INTERAKCJI 1/HAC (1/MW) DLA RÓŻNYCH BIBLIOTEK MOLEKULARNYCH [D1]

Statystycznie siłę poszczególnych czynników w iloczynowej reprezentacji LE można określić poprzez korelację indywidualnych czynników z mierzoną wartością funkcji LE. W tabeli 6 przedstawiono wyniki takiej analizy, podając korelację (współczynniki determinacji) dla różnych zbiorów danych (z binowaniem lub bez binowania danych). Szersza analiza otrzymanych wyników omówiona została w pracy **D1.**

Tabela 6. Współczynnik korelacji R dla danych aktywności biologicznych i temperatury wrzenia [**D1**]

Pozycja	Dane[1]	Współczynnik korelacji, R
1	MW vs. HAC	0.995
2	AC_{50} vs. pAC_{50}	-0.019
3	AC_{50} vs. pAC_{50}	-0.439 (binowane)
4	BP vs. MW	0.719
5	pAC_{50} vs. HAC	0.138
6	LE vs. 1/MW	0.752
7	LE vs. 1/HAC	0.759
8	LE vs. IC_{50}	-0.001
9	LE vs. pIC_{50}	-0.287
10	BP/MW vs. 1/MW	0.896
11	BP/MW vs. BP	-0.409
12	AC_{50}/MW vs. 1/MW	0.183
13	AC_{50}/MW vs. 1/MW	0.993 (binowane)
14	molar P/MW vs. molar P	0.857
15	molar P/MW vs. 1/MW	0,033

[1] niebinowane dane dla: 1,3-12; binowane dane dla 2, 13; serie danych: ChEMBL: 1-3. 5-9, 12-13; BP 4, 10-11

6. LE A FRAGMNENTACJA MOLEKULARNA [D1], CHEMICZNY PARADOKS TYPU ZENONA

Nauki chemiczne zajmują się materią (związkami chemicznymi) i jej (ich) transformacjami, Związek chemiczny reprezentuje przy tym zarówno cząsteczkę, jak i substancję [Bensaude-Vincent 2012]. Zatem używając terminu związki chemiczne, mamy na myśli zarówno cząsteczki, jak i substancje. Niepewność tego terminu polega na tym, że rzeczywiste znaczenie terminu związek chemiczny musi być intuicyjnie rozpoznane przez chemika w zależności od kontekstu informacji [Polański 2015, Polański 2019]. Do ilościowej reprezentacji materii używamy daltonów (Da), które odnoszą się do masy cząsteczkowej (MW) pojedynczych cząsteczek, podczas gdy masa cząsteczkowa substancji (MW), jest również określona w skali molowej w g/mol. Obie wartości MW (substancji i cząsteczki) są reprezentowane przez tę samą liczbę. W związku z tym interpretacja znaczenia MW może być podwójna.

W chemii mnożenie wielkości molowej przez operator 1/MW prowadzi do przekształcenia skali molowej na skalę wagową. Ponieważ operator 1/HAC jest silnie skorelowany z operatorem 1/MW jego rola jest podobna do 1/MW, a operacja mnożenia przez 1/HAC pełni podobną rolę, prowadząc do skali zbliżonej do wagowej, gdzie 1 liczba Avogadry HAC daje 1 mol HAC. Ciekawe, że dla układów jednoatomowych układ 1 mola HAC istnieje w rzeczywistości. Korelacja między MW i HAC jest wysoka i osiąga R = 0,995 dla dużych danych ChEMBL (tab. 6, pozycja 1).

W skali pojedynczej cząsteczki operatory 1/MW (1/HAC) można również interpretować jako operatory molekularne fragmentujące cząsteczki na pojedyncze Daltony lub atomy HAC.

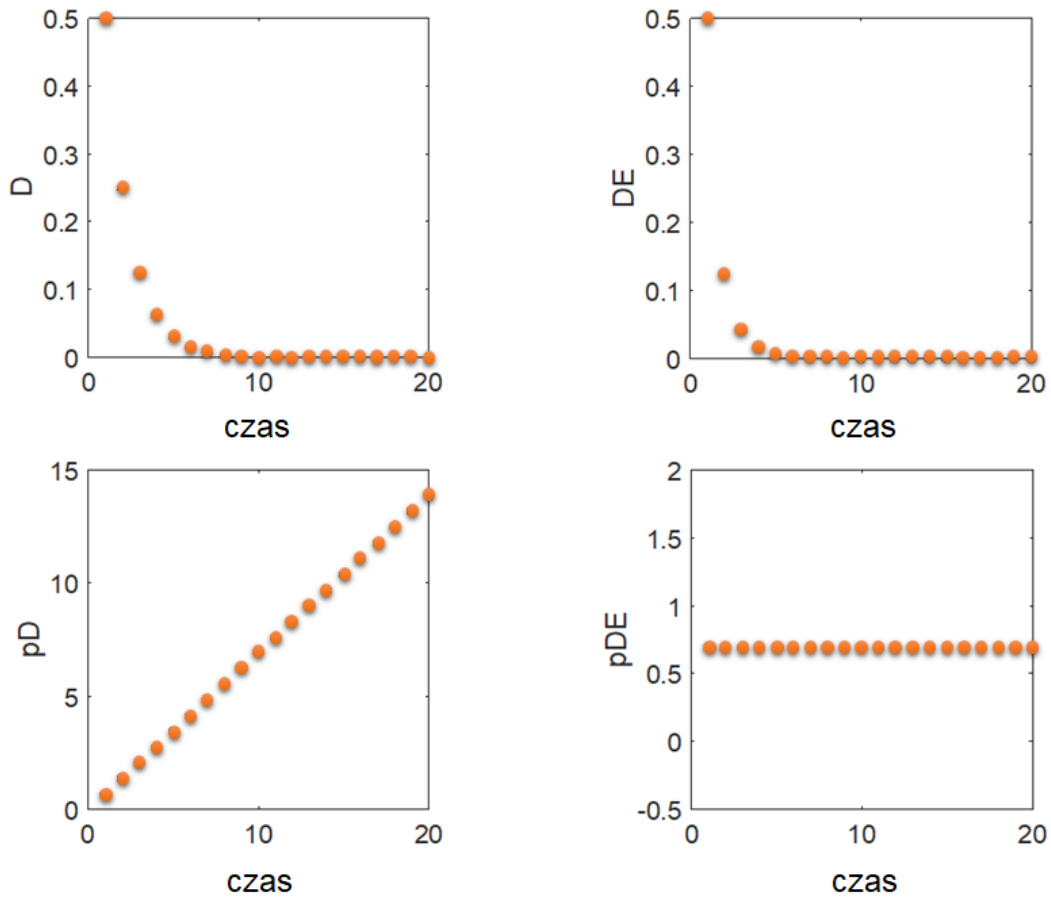
W naszych pracach zauważyliśmy, że wyżej dyskutowany problem dualizmu podziału materii tworzy paradoks analogiczny do paradoksu Zenona **[D1]**. Na rycinie 30. przeanalizowałam matematyczne efekty związane z modelem Zenona, który wynika z niepewności podziału czasu i przestrzeni [Stanford Encyclopedia of Philosophy, Available online: plato.stanford.edu (accessed on 29 August 2019)]. Zastosowane zmienne pozwalają na dostrzeżenie analogii do niepewności podziału (fragmentacji) związku chemicznego**.** Zmienne występujące na rycinie 30. zdefiniowano jako:

DE (*distance efficiency*)= distance/time, przy czym “distance” modelowano jako funkcję **D** = $(1/2)^{\mathbf{time}}$

Z kolei, pD oraz pDE definiowano analogicznie do wielkości chemicznych pIC_{50} oraz LE.

pD =-log**D**

pDE = **pD/time**



Rycina 30. Klasyczny efekt Zenona w ruchu zilustrowany poprzez zmienne distance (D) distance efficiency (DE) oraz ich odpowiednie skale logarytmiczne pD oraz pDE. Definicje w tekście.

Zasadniczo skale molowe lub wagowe są typowymi reprezentacjami właściwości mierzonych substancji. Kwestia maksymalnej siły działania ligandów była inspiracją do opracowania i zastosowania estymatorów wydajności ligandów (LE) w projektowaniu leków. LE można interpretować jako typ reprezentacji właściwości.

Ponieważ LE został pomyślany przede wszystkim jako miara powinowactwa wiązania pojedynczego fragmentu, w literaturze odnoszony jest do skali pojedynczej cząsteczki. W poprzednich rozdziałach wykazano, że trend między liczbą LE względem liczby atomów ciężkich (HAC) można w pełni zrozumieć, jeśli uświadomimy sobie związek LE ze skalą wagową. Ponieważ operator 1/MW (1/HAC) może być interpretowany jako operator łączący wagę molową i wagową. 1/MW (1/HAC) fragmentuje pojedynczą cząsteczkę na fragmenty – Dalton lub pojedynczy HAC. Podwójne znaczenie 1/MW (1/HAC) jest przyczyną niepewności między fragmentacją cząsteczek a konwersją skal. Ponieważ fragmentacja cząsteczek ma zasadnicze znaczenie dla projektowania i/lub optymalizacji ligandów zrozumienie efektów niepewności tworzonej przez dualizm cząsteczki – substancji jako reprezentacji związku chemicznego ma istotne znaczenie dla doskonalenia projektowania molekularnego.

Interesujący jest fakt, że analiza cen związków chemicznych w skali molowej i wagowej pozwala na obserwacje trendów zmienności podobnych do tych obserwowanych dla aktywności biologicznej.

Świadomość niepewności związanej z fragmentacją molekularną oraz jej koincydencje fragmentacji z różnymi typami reprezentacji związków chemicznych może w istotny sposób przyczynić się do uniknięcia wielu nieporozumień projektowania molekularnego oraz QSAR. Jest to istotne, zwłaszcza gdy pojawiają się nowe reprezentacje związku chemicznego jako właściwości związane z pojedynczą molekułą. Szerzej problem ten omówiłam w pracy [**D1**].

7. LE A TEMPERATURA – WIELKOŚĆ FIZYCZNA MODELU POJEDYNCZEJ CZĄSTECZKI [D4]

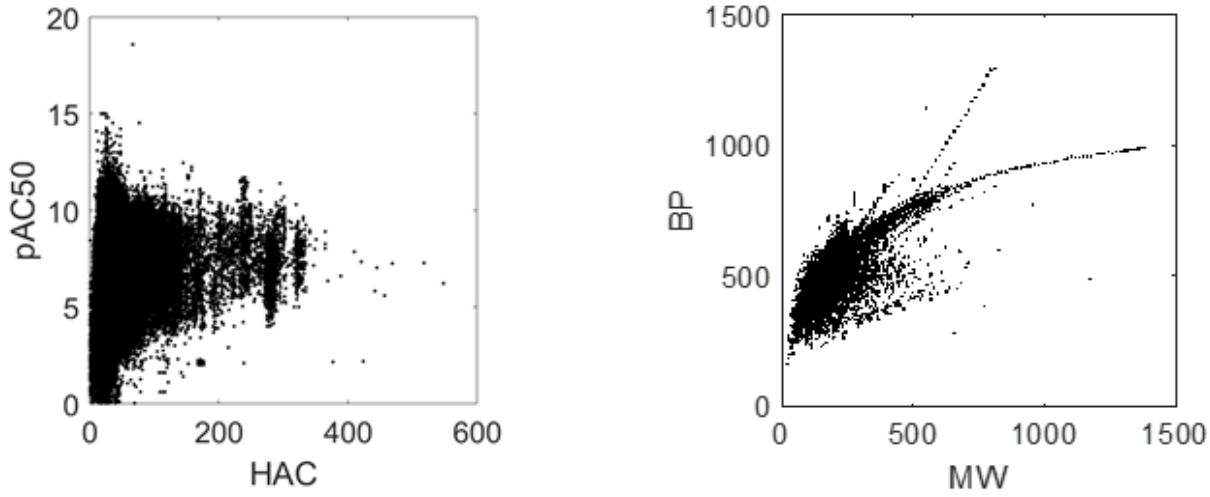
Przedstawione w poprzednich rozdziałach analizy LE pozwalają na lepsze zrozumienie paradoksów związanych z tą wielkością. Poniżej przedstawiłam i porównałam analizę danych siły działania leku do temperatury, a właściwie temperatury wrzenia (BP).

Głębsza analiza znaczenia zmiennej pAC_{50} w kontekście wielkości LE pozwala na ujawnienie pewnych wątpliwości formalnych. I tak, LE jest obliczane jako funkcja $logIC_{50}$ (pIC_{50}), a nie IC_{50}, które jest bezpośrednią reprezentacją stężenia molowego. W tym kontekście pIC_{50} jako logarytm jest wielkością bezwymiarową. pIC_{50} jest jednak proporcjonalne do energii swobodnej wiązania liganda, czyli własności molowej LE =1.4 pIC_{50}/HAC [Shultz 2013]. Z kolei operator 1/HAC używany do obliczania LE jest skorelowany ze skalą wagową. Wcześniej wykazano, że ten dychotomiczny charakter wyjaśnia kontrowersje dotyczące zastosowania funkcji LE w projektowaniu leków i QSAR. W niniejszym rozdziale pokazanie zostaną implikacje wynikające z rozważań dotyczących LE dla systematycznej typologii reprezentacji właściwości w QSAR. W tym kontekście LE porównano z temperaturą wrzenia (BP). BP podawany jest w skali temperatury (BP – ang. *boiling point*). Formalnie temperatura nie odnosi się ani do skali molowej, ani wagowej. Temperatura nie zależy od liczby cząsteczek w układzie. Jest natomiast związana ze średnią energią cząsteczek układu. Oznacza to również, że temperatura jest związana ze średnią energią pojedynczej cząsteczki. W celu szerszego obrazu w przedstawionych poniżej analizach LE i BP porównano z właściwościami, które wyraźnie odnoszą się do skali molowej lub wagowej, takich jak zawartość atomów lub procent wagowy (Cl: procent wagowy atomu chloru) lub ceny związków chemicznych oparte na masie (P – ang. *price*).

Przedstawione analizy dotyczą dużych zbiorów danych (LE, BP, Cl, P). Profil zależności BP vs. MW (R=0,719) jest podobny do profilu pIC_{50} względem HAC. Dla obu wielkości obserwujemy podobne korelacje. BP/MW silnie koreluje z 1/MW (R = 0,896). Podobny efekt jest również wyraźny w przypadku danych „Cl". Z kolei

w danych „P” efekt hiperboliczny 1/MW nie jest tak wyraźny. Dane dotyczące korelacji (wsp. determinacji R) podano w tabeli 6.

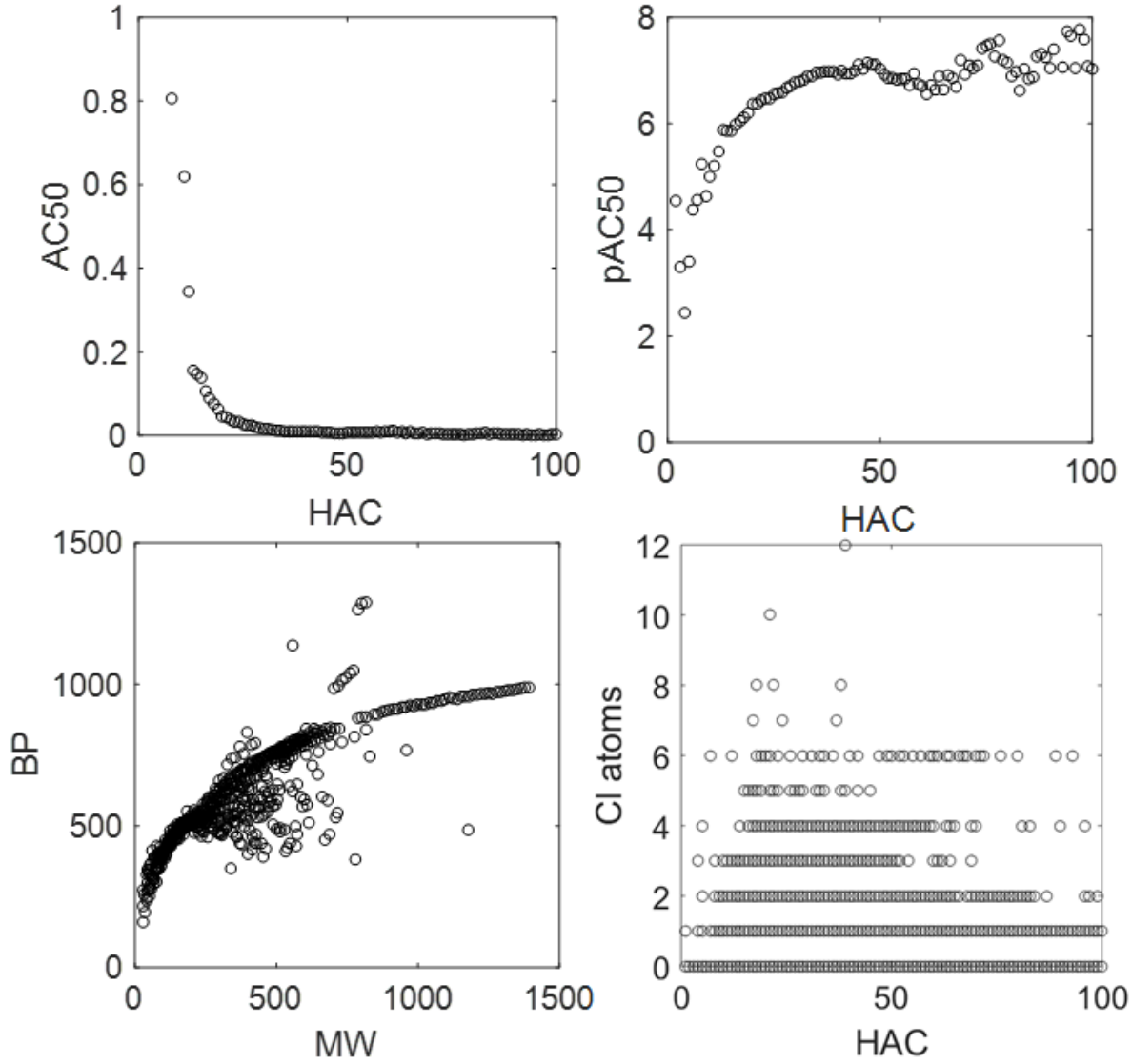
Na rycinie 31. pokazano zależności niebinowanych pAC_{50} i BP dla dużej serii związków chemicznych. Zwłaszcza w przypadku pomiarów biologicznych (pAC_{50}) nie obserwuje się żadnego uporządkowania, Nie uwidacznia się żaden związek między pAC_{50} a HAC, co można zilustrować za pomocą współczynnika korelacji pAC_{50} względem HAC, który wynosi około R = 0 dla danych bez binowania i R = -0,439 po binowaniu. Inaczej zachowuje się BP, gdzie obserwujemy pewne prawidłowości, a R wynosi ok. 0,7. (tab. 6, pozycja 4).



Rycina 31. Wykres zależności pAC_{50} vs. HAC i BP vs. MW dla danych niepoddanych binowaniu.

Właściwości związków chemicznych są często badane w funkcji MW lub HAC, które są prostymi miarami złożoności molekularnej. Na rycinie 32. zilustrowano zmienność AC_{50} i pAC_{50}, oraz BP vs. HAC lub MW dla różnych serii danych modelowanych przy użyciu binowania o wysokiej rozdzielczości. Co ciekawe, BP i pAC_{50} rosną wraz ze wzrostem MW do pewnej wartości granicznej, po której funkcje osiągają wartość zbliżoną do stałej (zależność typu plateau). Temperatura BP oraz siła działania pIC_{50} wykazują wobec tego istotne analogie. Uwidacznia to fakt, że temperatura wrzenia związana jest z molową reprezentacją związku chemicznego. Można to wyjaśnić faktem, że temperatura wrzenia (*boiling point* – punkt wrzenia), chociaż formalnie

wyrażona w skali temperatury, zależy od szeregu własności związku chemicznego, jak molowe ciepło wrzenia itp.



Rycina 32. Wykresy właściwości molowych dla danych dużych zestawów związków chemicznych: AC_{50} (binowane), pAC_{50} (binowane), liczba atomów chloru (niebinowane) w cząsteczkach (co jest równoważne liczbie moli w substancji) lub temperatury wrzenia (BP) (binowane); BP nie zależała od ilości substancji użytej w eksperymencie.

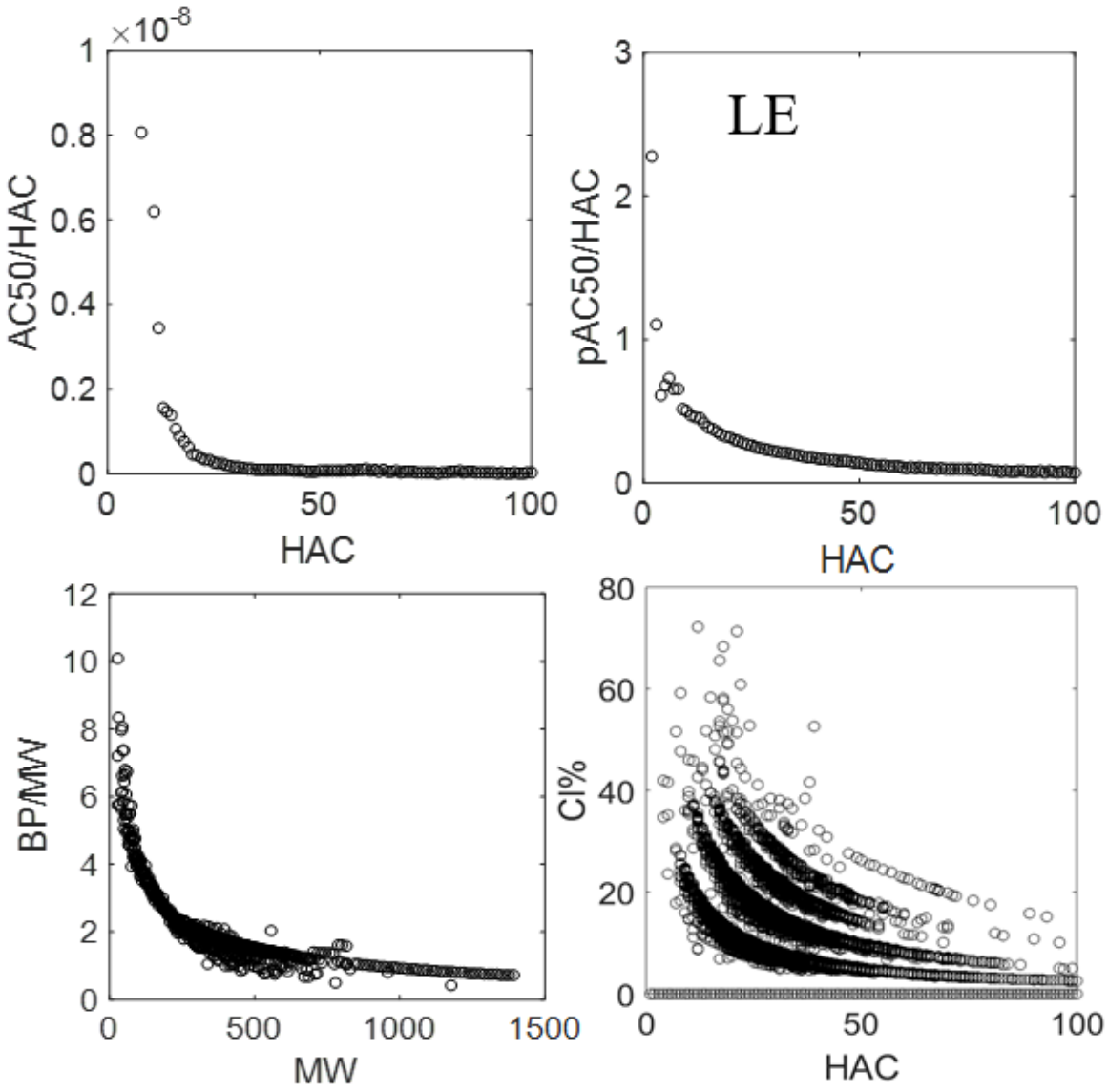
Na rycinach 32 i 33 pokazane zostały zależności szeregu innych wielkości charakteryzujących właściwości związków chemicznych (substancji lub cząsteczek) w funkcji MW (HAC). W szczególności na obrazie zależności liczby poszczególnych rodzajów atomów (dane niebinowane) vs. HAC obserwujemy szereg stałych funkcji

dla HAC = 1; 2; 3 itp. dla kongenerów, które mają taką samą liczbę atomów danego typu (liczbę atomów bromu lub chloru).

Na rycinie 33. przedstawiono wykresy BP, AC_{50}, pAC_{50}, Cl, które zostały obliczone jako stosunek danej właściwości (BP, AC_{50}, pAC_{50}, Cl) do MW lub HAC, względem MW lub HAC. Uzyskane zależności można interpretować jako interakcję indywidualnej właściwości i terminu hiperbolicznego 1/MW (1/HAC).

Wykresy pAC_{50}/HAC i BP/MW vs. HAC lub MW (ryc. 33), a wykresami pAC_{50} i BP vs. HAC lub MW (ryc. 32) są podobne, ale różnią się dla zależności AC_{50} vs. HAC i AC_{50}/HAC vs. HAC (ryc. 32 vs. 33). Oczywiście, jako przyczynę możemy wskazać matematykę, która przekształca podobne zależności pAC_{50} i BP vs. HAC (MW) do podobnych wykresów pAC_{50}/HAC i BP/MW vs. MW. Ważne są jednak chemiczne implikacje obserwowanych analogii profili ww. transformat. Skala pAC_{50} jest transformacją AC_{50}, która wymyślona została przede wszystkim ze względów praktycznych, ponieważ pozwala na bardziej intuicyjną interpretację danych biologicznych.

Substancje o większej sile działania mają również wyższe wartości pAC_{50}. Jednak pAC_{50} jest również proporcjonalne do energii swobodnej wiązania ligand-receptor. Oczywiście istotna różnica na zależnościach AC_{50}/HAC w porównaniu z pAC_{50}/HAC wynika z matematyki funkcji logarytmicznej, koduje to również ważne informacje chemiczne – energia wiązania (pAC_{50}) wzrasta wraz ze wzrostem HAC. Trendy AC_{50} lub pAC_{50} względem HAC określają wpływ złożoności molekularnej na wartości tych parametrów, które można było obserwować w rzeczywistych eksperymentach. Z kolei po fragmentacji molekularnej do HAC trend ten ulega zmianie dla pAC_{50}/HAC (LE) vs. HAC, który zmniejsza się wraz ze wzrostem HAC (ryc. 32 vs. 33).



Rycina 33. Wykresy transformacji 1/MW (1/HAC) właściwości molowych dla zestawów dużych danych związków chemicznych: AC_{50}*1/HAC (binowane), pAC_{50}*1/HAC (LE) (binowane), liczba atomów chloru*1/HAC (niebinowane) w cząsteczkach lub temperatura wrzenia (BP)*1/MW (binowane). BP nie zależało od ilości substancji użytej w eksperymencie.

Temperatura jest związana ze średnią energią cząsteczek w układzie. Oznacza to również, że temperatura jest związana ze średnią energią pojedynczej cząsteczki, co z kolei oznacza, że temperatura jest związana ze skalą pojedynczej cząsteczki. Jednak temperatura jest tylko skalą kodującą temperaturę wrzenia, złożoną właściwość cieczy, która jest skorelowana z ciepłem parowania lub energią potrzebną do przekształcenia substancji z cieczy w fazę pary. Angielski termin *boiling*

point BP dobrze ilustruje tę odmienność punktu wrzenia od typowej skali temperaturowej. Generalnie BP zależy od polarności i wielkości cząsteczek, a także od interakcji molekularnych między cząsteczkami w substancji.

Formalnie BP nie jest reprezentacją ani molową, ani wagą; jednak; BP zależy od MW. .Kiedy cząsteczki są mniejsze (mniejsze MW), rodzaj oddziaływań międzycząsteczkowych jest mniej istotny w kontekście energii potrzebnej do przejścia z fazy ciekłej do gazowej. To interakcji miedzy cząsteczkami decyduje o sile oddziaływania. Podobnie jak w przypadku interakcji ligand-receptor, to nie specyfika interakcji między cząsteczkami, ale liczba oddziałujących cząsteczek determinuje BP przy niskich MW. Rosnąca wartość MW wyznacza granicę, przy której rodzaj interakcji między cząsteczkami zaczyna odgrywać istotną rolę. Z kolei w przypadku dużych MW interakcje międzycząsteczkowe są zwykle mniej ważne. Liczba tych interakcji maleje bowiem ze wzrostem MW ponieważ maleje liczba cząsteczek. To masa cząsteczki decyduje o energii koniecznej do przejścia międzyfazowego. Co ciekawe, widoczne jest, że w obszarze średnich wartości MW 250-900 g/mol można zaobserwować największe odchylenie od tego trendu BP vs. MW (ryc. 20). Intuicyjnie powinien to być również obszar, w którym uwidaczniają się różne omawiane wyżej efekty zależności BP vs. MW.

Profil danych BP vs. MW jest podobny do profilu pIC_{50} względem HAC. Ponadto pIC_{50}/HAC (LE) i BP/MW są silnie skorelowane z operatorem 1/MW (1/HAC). Analogie BP i pAC_{50} wyraźnie wskazują na powiązanie wielkości BP ze skalą molową, pomimo formalnego powiązania tego parametru z miarą temperatury, która łączy się modelem pojedynczej cząsteczki.

8. PODSUMOWANIE I WNIOSKI

1. W niniejszej pracy analizowałam duże populacje biologicznie aktywnych cząsteczek w celu poszukiwania domen przestrzeni chemicznej, które są specyficzne dla pewnych klas związków chemicznych oraz reguł, które kontrolują te domeny.

2. W analizach wykorzystałam dostępne bazy danych ChEMBL oraz PubChem, które poddałam przetwarzaniu (*data curation*) tak by mogły być używane do analiz w programie MATLAB.

3. W wyniku kwerend danych literaturowych przygotowałam biblioteki danych innych bibliotek związków chemicznych:

   - Dane z bazy Binding Database (BindingBD, PTaylaorLa, USPaten, 5HT, AChE) [https://www.bindingdb.org/];

   - Dane z bazy Psychoactive Drug Screening Program (PDSP) [https://pdsp.unc.edu/pdspweb/];

   - Dane dotyczące leków wg publikacji [Hopkins 2014] oraz fragmentów leków wg publikacji [Schultes 2010];

   - Dane F2L wg publikacji [Mortenson 2018, Johnson 2016, Johnson 2019];

   - Temperatury wrzenia (BP) wg publikacji [Gharagheizi 2013].

4. Analizując ww. dane w szczególności wykazałam, że:

   a. Aktywności biologiczne wyrażone w skali pAC_{50} po poddaniu transformacji do LE analizowane jako funkcja 1/HAC w domenach największych dostępnych danych wykazują zawsze typowy efekt hiperboliczny **[D1, D3]**,

   b. Ponieważ LE jest interakcją pAC_{50} oraz czynnika 1/HAC, występowanie efektu hiperbolicznego dowodzi, że wartość LE jest w dużej mierze determinowana przez 1/HAC, którego obrazem w funkcji HAC jest hiperbola **[D4, D5]**,

   c. Występowanie efektu hiperbolicznego wyjaśnić można faktem, że czynnik 1/HAC (1/MW) jest także operatorem zamieniającym wielkości wyrażone

w skali molowej na skalę wagową. Statystyka Avogadry porządkuje związki chemiczne w skali molowej według stałej liczby cząsteczek w jednostce (1 mol). W skali wagowej liczba cząsteczek w jednostce (1 g, 1 kg) nie jest stała i zależy od 1/MW (1/HAC). W skali wagowej liczba cząsteczek w jednostce masy jest różna dla różnych ligandów, a zatem liczba cząsteczek, które oddziałują z receptorami nie jest normalizowana w zależności od liczby ligandów, które są dostępne dla oddziaływań z receptorem. Ponieważ liczba ligandów w jednostce masy rośnie wraz ze wzrostem wartości 1/MW, ligandy o niższej MW (i niższych HAC) wygrywają konkurencję LE, która kształtuje trend LE. Analiza dużych repozytoriów danych IC_{50} wskazała, że funkcja LE nie jest kontrolowana przez statystykę Avogadry typową dla efektów skali molekularnej **[D4].**

d. Skale molowe lub wagowe są typowymi reprezentacjami właściwości mierzonych dla substancji. Pojęcia te zostały w ostatnich latach rozszerzone o tak zwaną skalę pojedynczej cząsteczki (biologia pojedynczych cząsteczek – *single molecule biology*). W pracy porównano duże zbiory danych dotyczące właściwości, analizując ich związek z wyżej wymienionymi reprezentacjami. LE, która powstała by odpowiedzieć na pytanie jaka może być maksymalna siła działania pojedynczego HAC liganda była inspiracją do opracowania szeregu nowych estymatorów wydajności ligandów PLE, pPLE oraz SCORE.

e. LE może być także interpretowana jako deskryptor w skali pojedynczej cząsteczki charakteryzujący aktywność (siłę działania) pojedynczego HAC. Jest to więc metoda fragmentacji cząsteczki. Dualizm związków chemicznych (substancja, molekuła) decyduje o niejednoznaczności takiego podziału. Niejednoznaczność tę opisać możemy paradoksem podobnym do paradoksu Zenona opisującym niejednoznaczność ciągłego i dyskretnego podziału czasu i przestrzeni **[D4].**

f. Ciekawą analogię obserwujemy porównując pAC_{50}, BP oraz ich transformaty pAC_{50}/HAC (LE) i BP/MW. BP mimo formalniej miary temperaturowej (skala pojedynczej cząsteczki) wykazuje podobieństwo do pAC_{50} **[D1]**.

g. Przeprowadzone analizy pozwalają na opisanie pełnej systematyki właściwości jako reprezentacji związku chemicznego, która może być związana ze skalą molową, wagową lub skalą pojedynczej cząsteczki **[D1]**.

h. Wykazałam także, że tzw. *binowanie*, w szczególności binowanie z wysoką rozdzielczością, może być skuteczną metodą poszukiwania relacji struktura-aktywność w ramach dużych zbiorów molekularnych. Podczas kategoryzacji dane, które należały do danego przedziału były zastępowane wartością reprezentatywną dla tego przedziału. Poszczególne pojedyncze numery HAC lub wartości MW określały wielkość przedziałów, natomiast wartości właściwości, np. wartości AC_{50} były reprezentowane przez ich średnią wartość. Formalnie klasy molekularne zdefiniowane tą metodą to: izo-MW lub izo-HAC, które mogą w rzeczywistości zawierać bardzo różne struktury chemiczne. Binowanie w wysokiej rozdzielczości może zatem efektywnie skanować przestrzeń chemiczną z pojedynczą rozdzielczością dla MW i/lub HAC. Uzyskane w ten sposób modele noszą nazwę statystyk molekularnych **[D1-D5]**.

i. Analizowane problemy są szczególnie interesujące ze względu na rosnące zainteresowanie projektowaniem molekularnym opartym na fragmentach molekularnych, biofizyką pojedynczych cząsteczek i biologią pojedynczych cząsteczek, które są obecnie opracowywane i będą wymagały poprawnej interpretacji w cheminformatyce i modelowaniu QSAR.

III. CZĘŚĆ EKSPERYMENTALNA

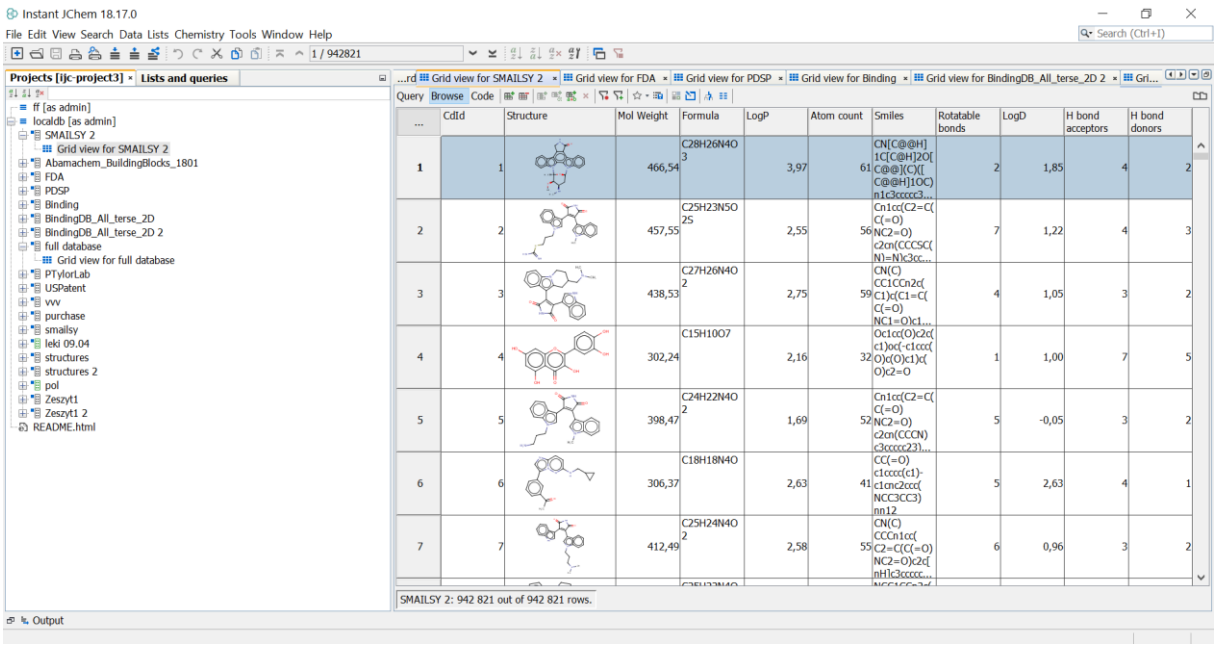
W poniższej części pracy zawarto informacje dotyczące wykorzystanych danych, sposobu ich przetworzenia oraz oprogramowania użytego w tym celu. Zamieszczono również skrypty, które stosowano do wizualizacji otrzymanych wyników.

1. CHARAKTERYSTYKA OPROGRAMOWANIA

Pobrane dane poddano analizie statystycznej korzystając z komputera osobistego o parametrach Intel® Core™ i5-5257U CPU 2.70 GHz, RAM 8 GB, SSD, 128 GB, system operacyjny Microsoft Windows 10. Programy wykorzystywane w analizie danych to InstantJChem 18.17.0, Matlab R2015a, Compound.Parser oraz pakiet Microsoft Office.

1.1. OPROGRAMOWANE INSTANT JCHEM

Instant JChem umożliwia tworzenie, zarządzanie i analizowanie dużej liczbą struktur chemicznych w środowisku umożliwiającym integrowanie różnych baz danych. Podstawowy wygląd interfejsu z zaimplementowanymi danymi przedstawia rycina 34.

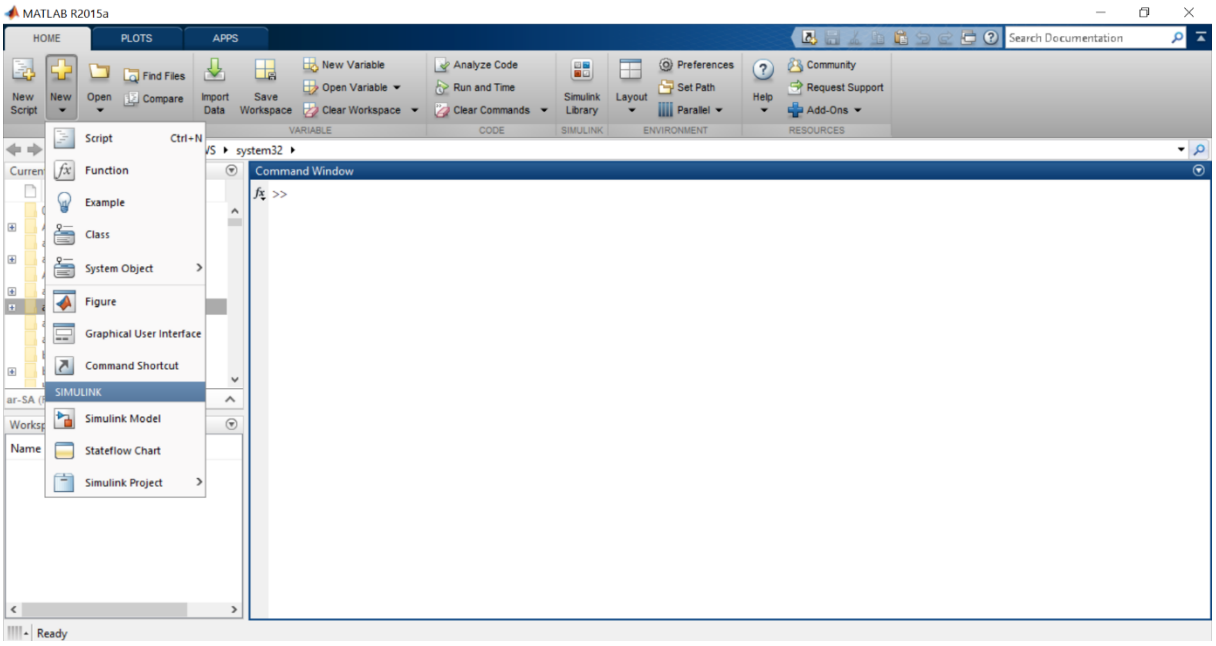


Rycina 34. Interfejs graficzny oprogramowania Instant JChem

Instant JChem to środowisko przeznaczone do przygotowywania baz danych, które obsługuje duże ilości danych (setki tysięcy struktur), zarówno w lokalnych, jak i zdalnych bazach danych. Instant JChem ma szeroką funkcjonalność; możliwość konfigurowania baz danych, automatyczne tworzenie bibliotek, szybkie analizy właściwości molekularnych oraz przewidywanych deskryptorów.

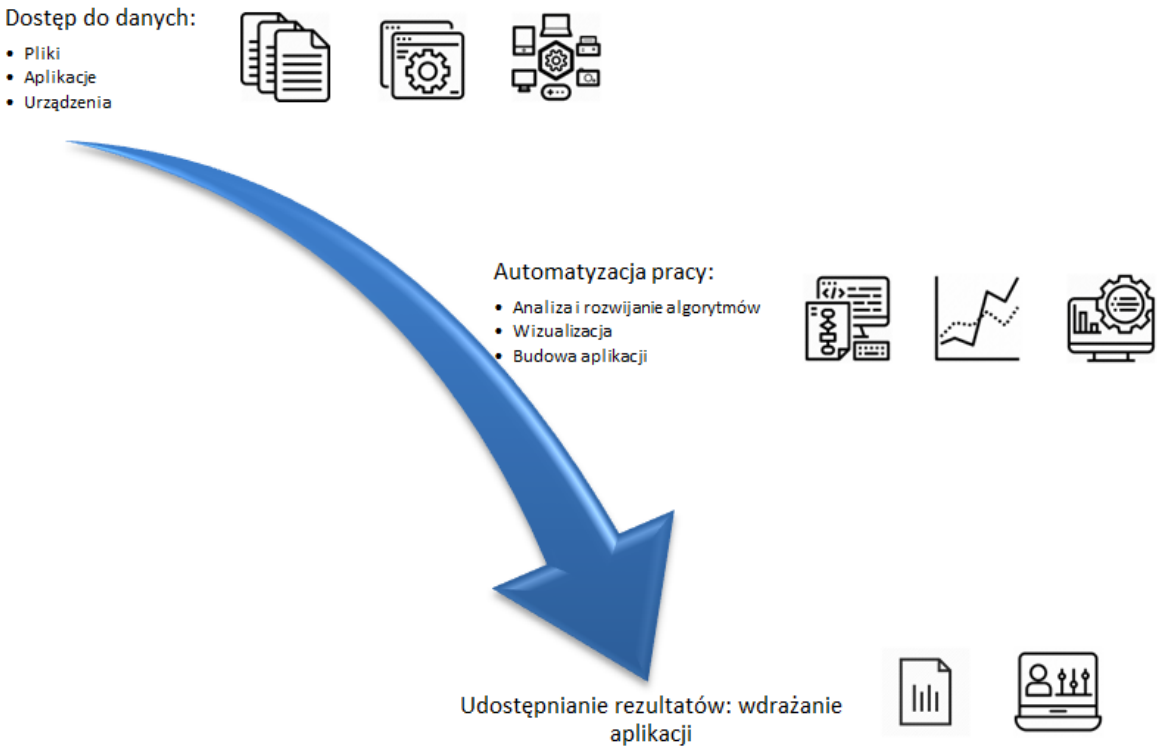
1.2. OPROGRAMOWANE MATLAB

MATLAB jest interaktywnym środowiskiem programistycznym pozwalającym zarówno na przeprowadzenie obliczeń, jak i wizualizację oraz analizę danych. Rycina 35. przestawia interfejs graficzny programu.



Rycina 35. Interfejs graficzny oprogramowania MATLAB

Środowisko MATLAB stanowi jedną z typowych opcji dla współczesnych obliczeniach naukowo-technicznych. Pozwala na przyspieszenie rozwiązywania różnorodnych problemów badawczych poprzez automatyzację rutynowych czynności obejmujących takie etapy jak: zbieranie danych, analiza, rozwijanie algorytmów, prezentacja wyników i wdrażanie aplikacji (ryc. 36).



Rycina 36. Ogólny schemat środowiska MATLAB

1.3. FORMATY DANYCH

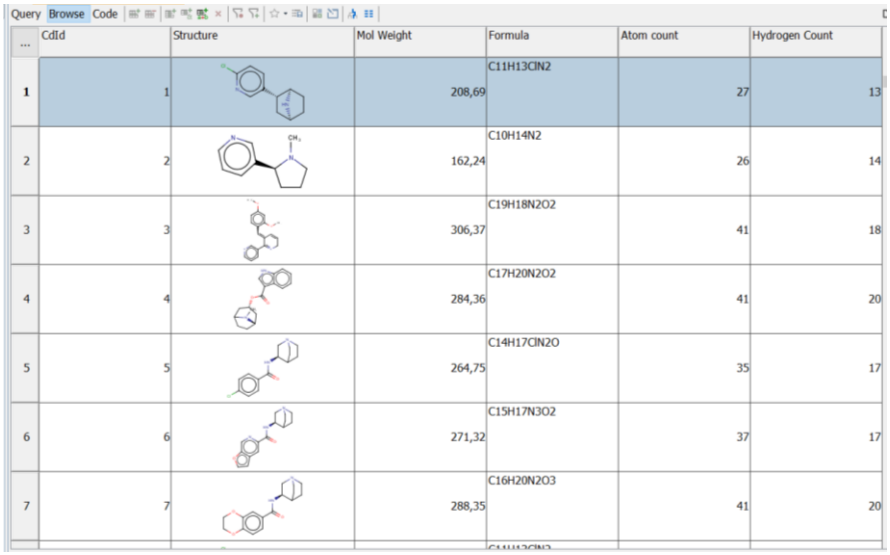
W celu analizy, a następnie wizualizacji danych znajdujących się w katalogach PubChem i ChEMBL dostępnych w formacie *.sdf niezbędne było przekonwertowanie formatu pliku (*.sdf) na format pliku (*.xlsx), które wykonano w programie Matlab, a następnie zapisanie danych w formacie (*.mat).

2. ETAPY ANALIZY I PRZETWARZANIA DANYCH

Pierwszym etapem badań było uzyskanie dostępu do danych z ChEMBL i PubChem, które następnie zostały pobrane w formie kodów SMILES dostępnych w plikach w formacie *.sdf. W celu przypisania kodów SMILES do konkretnych struktur chemicznych wykorzystano program Instant JChem, który pozwala również na policzenie liczby poszczególnych atomów, atomów ciężkich, heteroatomów czy też LogP. Schemat postepowania przedstawiono na rycinie 37. Po zaimportowaniu

katalogu w programie Instant JChem obliczono na podstawie struktury związku wybrane deskryptory molekularne co pokazano na rycinie 38.

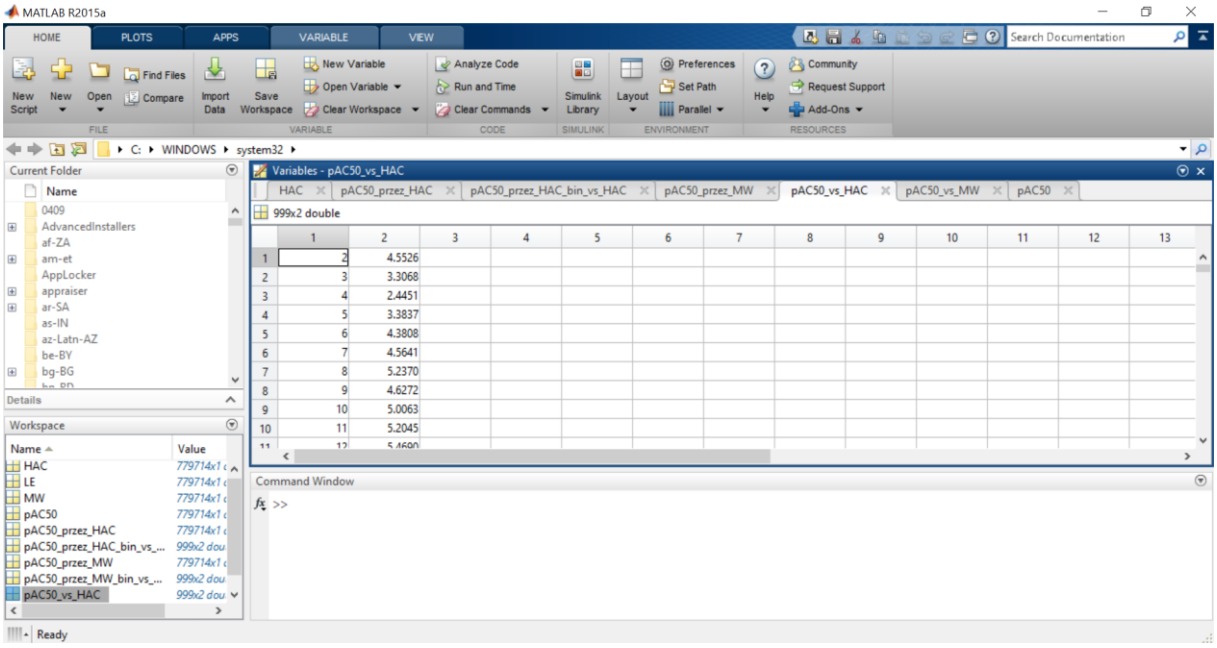
Rycina 37. Schemat procesu importowania danych



Rycina 38. Widok okna programu Instant JChem po zaimportowaniu katalogu PubChem i obliczeniu wybranych deskryptorów molekularnych

Otrzymane wyniki wyeksportowano do pliku Excel. Niezbędny był podział na szereg arkuszy gdyż program ten w jednym arkuszu może pomieścić jedynie 1 048 576 wierszy. Tak przygotowane i pogrupowane dane zostały poddane analizie w dalszym etapie pracy w programie Matlab.

Wczytanie danych do programu MATLAB wykonano, wpisując następujące polecenie „xlsread” do okna dialogowego Command Window. Ze względu na duży rozmiar analizowanych danych niezbędne było wprowadzenie każdego arkusza MS Excel pojedynczo, każdorazowo do osobnego obszaru roboczego (ang. *workspace*). Widok okna dialogowego przedstawiony został na rycinie 39.



Rycina 39. Widok okna programu MATLAB po wczytaniu danych ChEMBL

BIBLIOGRAFIA

Referencja	Opis bibliograficzny
Abad-Zapatero 2011	Abad-Zapatero, C., Blasi, D. (2011). Ligand efficiency indices (LEIs): more than a simple efficiency yardstick. *Molecular informatics*, *30*(2-3), 122-132.
Alves 2018	Alves, V. M., Muratov, E. N., Zakharov, A., Muratov, N. N., Andrade, C. H., Tropsha, A. (2018). Chemical toxicity prediction for major classes of industrial chemicals: Is it possible to develop universal models covering cosmetics, drugs, and pesticides?. *Food and Chemical Toxicology*, *112*, 526-534.
Aykul 2016	Aykul, S., Martinez-Hackert, E. (2016). Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical biochemistry*, *508*, 97-103.
Bachmann 2005	Bachmann, K. A., Lewis, J. D. (2005). Predicting inhibitory drug—drug interactions and evaluating drug interaction reports using inhibition constants. *Annals of Pharmacotherapy*, *39*(6), 1064-1072.
Barril 2015	Barril, X., Danielsson, H. (2015). Binding kinetics in drug discovery. *Drug Discovery Today: Technologies*, (17), 35-36.
Beghes 2011	Hughes, J. P., Rees, S., Kalindjian, S. B., Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, *162*(6), 1239-1249.
Bembenek 2009	Bembenek, S. D., Tounge, B. A., Reynolds, C. H. (2009). Ligand efficiency and fragment-based drug discovery. *Drug Discovery Today*, *14*(5-6), 278-283.
Bernard 2011	Barnard, J. M., Kenny, P. W., Wallace, P. N. (2011). Representing chemical structures in databases for drug design. In *Drug Design Strategies* (pp. 164-191).
Bohm 2003	Bohm, H. J. (2003). Prediction of non-bonded interactions in drug design. *METHODS AND PRINCIPLES IN MEDICINAL CHEMISTRY*, *19*, 3-3.
Borgert 2013	Borgert, C. J., Baker, S. P., Matthews, J. C. (2013). Potency matters: thresholds govern endocrine activity. *Regulatory Toxicology and Pharmacology*, *67*(1), 83-88.
Burlingham 2003	Burlingham, B. T., Widlanski, T. S. (2003). An intuitive look at the relationship of Ki and IC_{50}: a more general use for the Dixon plot. *Journal of chemical education*, *80*(2), 214.
Caldwell 2012	W Caldwell, G., Yan, Z., Lang, W., A Masucci, J. (2012). The IC_{50} concept revisited. *Current topics in medicinal chemistry*, *12*(11), 1282-1290.
Cavalluzzi 2017	Cavalluzzi, M. M., Mangiatordi, G. F., Nicolotti, O., Lentini, G. (2017). Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective. *Expert opinion on drug*

Referencja	Opis bibliograficzny
	discovery, *12*(11), 1087-1104.
Cheng 1973	Chen, Y., Prusoff, W. (1973). Relationship between the inhibition constant and the concentration of an inhibitor that cause a 50% inhibition of an enzyme reaction. *Biochem Pharmacol*, *22*, 3099-3108.
Cheng 2001	Cheng, H. C. (2001). The power issue: Determination of KB or Ki from IC_{50}: A closer look at the Cheng–Prusoff equation, the Schild plot and related power equations. *Journal of pharmacological and toxicological methods*, *46*(2), 61-71.
Cherkasov 2014	Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Consonni, V. (2014). QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, *57*(12), 4977-5010.
Cleland 1963	Cleland, W. W. (1963). The kinetics of enzyme-catalyzed reactions with two or more substrates or products: II. Inhibition: Nomenclature and theory. *Biochimica et Biophysica Acta (BBA)-Specialized Section on Enzymological Subjects*, *67*, 173-187.
Colquhoun 2006	Colquhoun, D. (2006). The quantitative analysis of drug–receptor interactions: a short history. *Trends in pharmacological sciences*, *27*(3), 149-157.
Dagliyan 2009	Dagliyan, O., Kavakli, I. H., Turkay, M. (2009). Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC_{50} using common molecular descriptors. *Journal of chemical information and modeling*, *49*(10), 2403-2411.
Davis 2013	Davis, B. J., Erlanson, D. A. (2013). Learning from our mistakes: the 'unknown knowns' in fragment screening. *Bioorganic medicinal chemistry letters*, *23*(10), 2844-2852.
Dinse 2011	Dinse, G. E., Umbach, D. M. (2011). Characterizing non-constant relative potency. *Regulatory Toxicology and Pharmacology*, *60*(3), 342-353.
Filimonov 2008	Filimonov, D., Poroikov, V. (2008). *Probabilistic approaches in activity prediction* (pp. 182-216). Royal Society of Chemistry, Cambridge, UK.
Gabrielsson 2018	Gabrielsson, J., Peletier, L. A., Hjorth, S. (2018). In vivo potency revisited–keep the target in sight. *Pharmacology Therapeutics*, *184*, 177-188.
Gaulton 2012	Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, *40*(D1), D1100-D1107.
Gerlza 2014	Gerlza, T., Hecher, B., Jeremic, D., Fuchs, T., Gschwandtner, M., Falsone, A., Kungl, A. J. (2014). A combinatorial approach to biophysically characterise chemokine-glycan binding affinities for

Referencja	Opis bibliograficzny
	drug development. *Molecules*, *19*(7), 10618-10634.
Gharagheizi 2013	Gharagheizi, F., Mirkhani, S. A., Ilani-Kashkouli, P., Mohammadi, A. H., Ramjugernath, D., Richon, D. (2013). Determination of the normal boiling point of chemical compounds using a quantitative structure–property relationship strategy: Application to a very large dataset. Fluid Phase Equilibria, 354, 250-258.
Ginsberg 2009	Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.
Gohlke 2002	Gohlke, H., Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition*, *41*(15), 2644-2676.
Goracci 2017	Goracci, L., Tortorella, S., Tiberi, P., Pellegrino, R. M., Di Veroli, A., Valeri, A., Cruciani, G. (2017). Lipostar, a comprehensive platform-neutral cheminformatics tool for lipidomics. *Analytical chemistry*, *89*(11), 6257-6264.
Guziałowska 2016	Guziałowska-Tic, J., Tic, W. J. (2016). Metody QSAR jako alternatywa dla badań na zwierzętach w ocenie toksycznych i ekotoksycznych zagrożeń ze strony substancji chemicznych. *Przemysł Chemiczny*, *95*.
Hann 2001	Hann, M. M., Leach, A. R., Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, *41*(3), 856-864.
Hann 2011	Hann, M. M. (2011). Molecular obesity, potency and other addictions in drug discovery. *MedChemComm*, *2*(5), 349-355.
Hann 2012	Hann, M. M., Keserü, G. M. (2012). Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nature reviews Drug discovery*, *11*(5), 355-365.
Hassanzadeh 2010	Hassanzadeh, H. R., Rouhani, M. (2010, July). A multi-objective gravitational search algorithm. In *2010 2nd international conference on computational intelligence, communication systems and networks* (pp. 7-12). IEEE.
Hill 1910	Hill, A. V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *j. physiol.*, *40*, 4-7.
Holford 1981	Holford, N. H., Sheiner, L. B. (1981). Understanding the dose-effect relationship. *Clinical pharmacokinetics*, *6*(6), 429-453.
Hopkins 2004	Hopkins, A. L., Groom, C. R., Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug discovery today*, *9*(10), 430.
Hopkins 2014	Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C., Reynolds, C. H. (2014). The role of ligand efficiency metrics in drug

Referencja	Opis bibliograficzny
	discovery. *Nature reviews Drug discovery*, *13*(2), 105-121.
Horrobin 2003	Horrobin, D. F. (2003). Modern biomedical research: an internally self-consistent universe with little contact with medical reality?. *Nature Reviews Drug Discovery*, *2*(2), 151-154.
Hu 2012	Hu, Y., Bajorath, J. (2012). Growth of ligand–target interaction data in ChEMBL is associated with increasing and activity measurement-dependent compound promiscuity. *Journal of chemical information and modeling*, *52*(10), 2550-2558.
Irwin 2005	Irwin, J. J., Shoichet, B. K. (2005). ZINC− a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, *45*(1), 177-182.
Johnson 2017	Johnson, C. N., Erlanson, D. A., Jahnke, W., Mortenson, P. N., Rees, D. C. (2017). Fragment-to-Lead Medicinal Chemistry Publications in 2016: Miniperspective. *Journal of medicinal chemistry*, *61*(5), 1774-1784.
Jones 2018	Jones, R., Wilsdon, J. R. (2018). The biomedical bubble: Why UK research and innovation needs a greater diversity of priorities, politics, places and people.
Kalliokoski 2013	Kalliokoski, T., Kramer, C., Vulpetti, A., Gedeck, P. (2013). Comparability of mixed IC 50 data–a statistical analysis. *PloS one*, *8*(4), e61007.
Keiser 2009	Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., ... Whaley, R. (2009). Predicting new molecular targets for known drugs. *Nature*, *462*(7270), 175-181.
Kenny 2005	Kenny, P. W., Sadowski, J. (2005). Structure modification in chemical databases. *Chemoinformatics in drug discovery*, *23*, 271-285.
Kenny 2014	Kenny, P. W., Leitao, A., Montanari, C. A. (2014). Ligand efficiency metrics considered harmful. *Journal of computer-aided molecular design*, *28*(7), 699-710.
Kenny 2017	Kenny, P. W. (2017). Comment on the ecstasy and agony of assay interference compounds. *Journal of Chemical Information and Modeling*, *57*(11), 2640-2645.
Keseru 2015	Keserü, G. M., Swinney, D. C., Mannhold, R., Kubinyi, H., Folkers, G. (Eds.). (2015). *Thermodynamics and kinetics of drug binding*. Weinheim, Germany: Wiley-VCH.
Kim 2016	Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., ... Wang, J. (2016). PubChem substance and compound databases. *Nucleic acids research*, *44*(D1), D1202-D1213.
Klebe 2000	Klebe, G., Grädler, U., Grüneberg, S., Krämer, O., Gohlke, H. (2000). 10–Understanding receptorligand interactions as a prerequisite for virtual screening. *Methods and Principles in*

Referencja	Opis bibliograficzny
	Medicinal Chemistry, 207.
Knight 2009	Knight, A. E. (Ed.). (2009). *Single molecule biology*. Academic Press.
Knox 2010	Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Djoumbou, Y. (2010). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, *39*(suppl_1), D1035-D1041.
Krumrine 2003	Krumrine, J., Raubacher, F., Brooijmans, N., Kuntz, I. (2003). Principles and methods of docking and ligand design. *Methods of Biochemical Analysis*, *44*, 443-476.
Kucia 2020	Praca doktorska Urszuli Kuci, 2020, promotor pracy prof. Jarosław Polański
Kuntz 1999	Kuntz, I. D., Chen, K., Sharp, K. A., Kollman, P. A. (1999). The maximal affinity of ligands. *Proceedings of the National Academy of Sciences*, *96*(18), 9997-10002.
Laney 2001	Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.
Leake 2013	Leake, M. C. (2013). The physics of life: one molecule at a time.
Liu 2007	Liu, T., Lin, Y., Wen, X., Jorissen, R. N., Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, *35*(suppl_1), D198-D201.
Manallack 2013	Manallack, D. T., Prankerd, R. J., Yuriev, E., Oprea, T. I., Chalmers, D. K. (2013). The significance of acid/base properties in drug discovery. *Chemical Society Reviews*, *42*(2), 485-496.
Martel 2013	Martel, S., Gillerat, F., Carosati, E., Maiarelli, D., Tetko, I. V., Mannhold, R., Carrupt, P. A. (2013). Large, chemically diverse dataset of log P measurements for benchmarking studies. *European Journal of Pharmaceutical Sciences*, *48*(1-2), 21-29.
Meanwell 2016	Meanwell, N. A. (2016). Improving drug design: an update on recent applications of efficiency metrics, strategies for replacing problematic elements, and compounds in nontraditional drug space. *Chemical Research in Toxicology*, *29*(4), 564-616.
Mestres 2009	Mestres, J., Gregori-Puigjané, E., Valverde, S., Solé, R. V. (2009). The topology of drug–target interaction networks: implicit dependence on drug properties and target families. *Molecular BioSystems*, *5*(9), 1051-1057.
Meyer 2000	Meyer, E. F., Swanson, S. M., Williams, J. A. (2000). Molecular modelling and drug design. *Pharmacology Therapeutics*, *85*(3), 113-121.
Michael 2008	Michael, S., Auld, D., Klumpp, C., Jadhav, A., Zheng, W., Thorne, N., ... Simeonov, A. (2008). A robotic platform for quantitative high-

Referencja	Opis bibliograficzny
	throughput screening. *Assay and drug development technologies*, *6*(5), 637-657.
Mignani 2018	Mignani, S., Rodrigues, J., Tomas, H., Jalal, R., Singh, P. P., Majoral, J. P., Vishwakarma, R. A. (2018). Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified?. *Drug discovery today*, *23*(3), 605-615.
Mortenson 2018	Mortenson, P. N., Erlanson, D. A., De Esch, I. J., Jahnke, W., Johnson, C. N. (2018). Fragment-to-Lead Medicinal Chemistry Publications in 2017: Miniperspective. *Journal of medicinal chemistry*, *62*(8), 3857-3872.
Murray 2014	Murray, C. W., Erlanson, D. A., Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C., ... Richmond, N. J. (2014). Validity of ligand efficiency metrics.
Nowicki 2008	Nowicki, J. P., Scatton, B. (2008). Measurement and expression of drug effects. In *The practice of medicinal chemistry* (pp. 73-84). Academic Press.
Pan 2013	Pan, A. C., Borhani, D. W., Dror, R. O., Shaw, D. E. (2013). Molecular determinants of drug–receptor binding kinetics. *Drug discovery today*, *18*(13-14), 667-673.
Paolini 2006	Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S., Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature biotechnology*, *24*(7), 805-815.
Perola 2010	Perola, E. (2010). An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *Journal of medicinal chemistry*, *53*(7), 2986-2997.
Polański 2015	Polanski, J., Bogocz, J., Tkocz, A. (2015). Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. *Drug discovery today*, *20*(11), 1300-1304.
Polański 2016A	Polanski, J., Gasteiger, J. (2016). Computer representation of chemical compounds. *Handbook of Computational Chemistry; Leszczynski, J., Puzyn, T., Eds*, 1-43.
Polański 2016B	Polanski, J., Bogocz, J., Tkocz, A. (2016). The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *Journal of computer-aided molecular design*, *30*(5), 381-389.
Polański 2017A	Polanski, J. (2017). Big data in structure-property studies—From definitions to models. In *Advances in QSAR Modeling* (pp. 529-552). Springer, Cham.
Polański 2017B	Polanski, J., Tkocz, A. (2017). Between descriptors and properties: Understanding the ligand efficiency trends for G protein-coupled receptor and kinase structure–activity data sets. *Journal of Chemical Information and Modeling*, *57*(6), 1321-1329.

Referencja	Opis bibliograficzny
Polański 2017C	Polanski, J., Tkocz, A., Kucia, U. (2017). Beware of ligand efficiency (LE): understanding LE data in modeling structure-activity and structure-economy relationships. *Journal of Cheminformatics*, *9*(1), 1-8.
Polański 2018	Polanski, J., Bak, A. (2019). Ligand potency–an essential estimator for drug design: between intuition, misinterpretation and serendipity. *Future medicinal chemistry*, 11(14), 1827-1843.
Polański 2019	Polanski, J. (2019). Chemoinformatics: From Chemical Art to Chemistry in Silico. in Silico Encyclopedia of Bioinformatics and Computational Biology, Ranganathan S., Gribskov M., Nakai H., Schonbach Ch., Eds, Vol. 2, pp. 601-618, Elsevier.
Reynolds 2007	Reynolds, C. H., Bembenek, S. D., Tounge, B. A. (2007). The role of molecular size in ligand efficiency. *Bioorganic medicinal chemistry letters*, *17*(15), 4258-4261.
Reynolds 2008	Reynolds, C. H., Tounge, B. A., Bembenek, S. D. (2008). Ligand binding efficiency: trends, physical basis, and implications. *Journal of medicinal chemistry*, *51*(8), 2432-2438.
Reynolds 2017	Reynolds, C. H., Reynolds, R. C. (2017). Group additivity in ligand binding affinity: an alternative approach to ligand efficiency. *Journal of chemical information and modeling*, *57*(12), 3086-3093.
Riera 2016	Riera, T. V., Wigle, T. J., Copeland, R. A. (2016). Characterization of inhibitor binding through multiple inhibitor analysis: A novel local fitting method. In *High Throughput Screening* (pp. 33-45). Humana Press, New York, NY.
Rosenblum 2006	Rosenblum, B., Kuttner, F. (2006). *Quantum enigma: Physics encounters consciousness*. Oxford University Press.
Sailer 2007	Seiler, K. P., George, G. A., Happ, M. P., Bodycombe, N. E., Carrinski, H. A., Norton, S., ... Ferraiolo, P. (2007). ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic acids research*, *36*(suppl_1), D351-D359.
Salahudee 2017	Salahudeen, M. S., Nishtala, P. S. (2017). An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. Saudi pharmaceutical journal, 25(2), 165-175.
Scannell 2012	Scannell, J. W., Blanckley, A., Boldon, H., Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery*, *11*(3), 191.
Schultes 2010	Schultes, S., de Graaf, C., Haaksma, E. E., de Esch, I. J., Leurs, R., Krämer, O. (2010). Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discovery Today: Technologies*, *7*(3), e157-e162.
Schulthess 2014	Schulthess, D., Chlebus, M., Bergström, R., Baelen, K. V. (2014). Medicine adaptive pathways to patients (MAPPs): using regulatory

Referencja	Opis bibliograficzny
	innovation to defeat Eroom's law. *Chinese clinical oncology*, *3*(2), 21.
Scott 2018	Scott, J. S., Waring, M. J. (2018). Practical application of ligand efficiency metrics in lead optimisation. *Bioorganic medicinal chemistry*, *26*(11), 3006-3015.
Sharp 2012	Sharp, K. A. (2012). Statistical thermodynamics of binding and molecular recognition models. *Protein-Ligand Interactions*, *53*, 3.
Sheridan 2016	Sheridan, R. P. (2016). Debunking the idea that ligand efficiency indices are superior to pIC_{50} as QSAR activities. *Journal of Chemical Information and Modeling*, *56*(11), 2253-2262.
Shultz 2013A	Shultz, M. D. (2013). Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters. *Bioorganic medicinal chemistry letters*, *23*(21), 5980-5991.
Shultz 2013B	Shultz, M. D. (2013). The thermodynamic basis for the use of lipophilic efficiency (LipE) in enthalpic optimizations. *Bioorganic medicinal chemistry letters*, *23*(21), 5992-6000.
Shultz 2014	Shultz, M. D. (2014). Improving the plausibility of success with inefficient metrics.
Shulz 2018	Shultz, M. D. (2018). Two decades under the influence of the rule of five and the changing properties of approved oral drugs: miniperspective. *Journal of medicinal chemistry*, *62*(4), 1701-1714.
Sözüdoğru 2020	Sözüdoğru, E., Clarke, B. (2020). Uncertainty in Drug Discovery: Strategies, Heuristics and Technologies. In *Uncertainty in Pharmacology* (pp. 153-171). Springer, Cham.
Tkocz 2020	Praca doktorska Aleksandry Tkocz, 2020, promotor pracy prof. Jarosław Polański
Todeschini 2008	Todeschini, R., Consonni, V. (2008). *Handbook of molecular descriptors* (Vol. 11). John Wiley Sons.
Walkman 2002	Waldman, S. A. (2002). Does potency predict clinical efficacy? Illustration through an antihistamine model. *Annals of Allergy, Asthma Immunology*, *89*(1), 7-12.
Williams 2017	Williams, G., Ferenczy, G. G., Ulander, J., Keserű, G. M. (2017). Binding thermodynamics discriminates fragments from druglike compounds: a thermodynamic description of fragment-based drug discovery. *Drug discovery today*, *22*(4), 681-689.
Zartler 2005	Zartler, E. R., Shapiro, M. J. (2005). Fragonomics: fragment-based drug discovery. *Current opinion in chemical biology*, *9*(4), 366-370.
Zhou 2009	Zhou, H. X., Gilson, M. K. (2009). Theory of free energy and entropy in noncovalent binding. *Chemical reviews*, *109*(9), 4092-4107.

Referencja	Opis bibliograficzny
1	https://step1.medbullets.com/pharmacology/107007/efficacy-vs-potency;
2	https://en.wikipedia.org/wiki/Ligand_(biochemistry)

SPIS RYCIN

SPIS TABEL

ZAŁĄCZNIKI

Do treści rozprawy doktorskiej dołączono cztery załączniki: życiorys naukowy autora (**Załącznik A**), najważniejsze publikacje w dotychczasowym dorobku naukowym autora (**Załącznik B**), oświadczenie współautorów publikacji (**Załącznik C**) oraz kserokopie pięciu publikacji z dorobku naukowego autora (**Załącznik D**).

ZAŁĄCZNIK A

Życiorys naukowy autora

DANE OSOBOWE

Imię i nazwisko	Roksana Duszkiewicz
Data urodzenia	25.09.1989
Miejsce urodzenia	Miastko

WYKSZTAŁCENIE

2020	Studia doktoranckie w zakresie nauk chemicznych, Instytut Chemii, Wydział Ścisłych
2018	Neurobiologia, studia magisterskie, Wydział Lekarski, Śląski Uniwersytet Medyczny w Katowicach
2015	Technologia chemiczna, studia inżynierskie, Wydział Matematyki Fizyki i Chemii,

	Uniwersytet Śląski w Katowicach
2013	Chemia leków, studia magisterskie, Wydział Matematyki Fizyki i Chemii, Uniwersytet Śląski w Katowicach
2011	Biologia, studia licencjackie, Wydział Biologii i Ochrony Środowiska, , Uniwersytet Śląski w Katowicach

DZIAŁALNOŚĆ NAUKOWA

Asystent w projekcie SONATA BIS	Uniwersytet SWPS
Wolontariat	Centrum Onkologii w Gliwicach
Staż	Centrum Materiałów Polimerowych i Węglowych PAN w Zabrzu

DZIAŁALNOŚĆ DYDAKTYCZNA

Farmakologia; Neurobiologia;	Śląski Uniwersytet Medyczny w Katowicach
Laboratorium chemii organicznej; Informacja naukowa;	Uniwersytet Śląski w Katowicach

UDZIAŁ W KONFERENCJACH

Kraków 2019	wystąpienie: Towards social responsibility of institutions: education, public health and design, 28 listopada 2019, Kraków, Autor prezentacji: *Social Responsibility in Public Health Institutions: problems of Polish paramedics*;
Rzym 2019	wystąpienie: 7th Biennial ESTD Congress, 24-26 października 2019, Rzym, Włochy, Autor prezentacji: *Depersonalisation and derealisation in different groups: Depersonalization in healthy people – theoretical models*;
Katowice 2019	poster: 8th EDITION OF THE CONFERENCE FOR YOUNG SCIENTISTS, 19-20 września 2019, Chorzów, Autor 2 posterów: *Top 200 Farmakoekonomii, czyli analiza najlepiej sprzedających się leków na świecie w latach 2015-2018* i *Zmiany metabolizmu kwasów tłuszczowych w przebiegu wybranych patologii ciąży*;
Katowice 2019	poster: 14th International Congress of the Polish Neuroscience Society, 28-30 sierpnia 2019, Katowice, Autor posteru: *Identification of potential Keap1 inhibitors through database analysis based on the designated pharmacophore and QSAR models*;

Katowice 2019	wystąpienie: Wirtualna konferencja młodych przyrodników, 24-29 czerwca 2019, Autor prezentacji: *Wykorzystanie danych Big data i modelowania molekularnego w procesie projektowania leków*;
Zabrze 2019	wystąpienie: IX Międzynarodowa konferencja naukowa: „Zrównoważony rozwój – Sustainable development 2018”, 10 maja 2019, Zabrze, Autor prezentacji: *Narzędzia chemo- i bioinformatyczne w projektowaniu leków a zrównoważony rozwój*;
Sosnowiec 2019	poster: III Seminarium Ogólnoakademickie „Metody fizykochemiczne w badaniach naukowych”, 17 kwietnia 2019, Sosnowiec, Autor posteru: *Wielowymiarowa analiza baz danych leków i kandydatów na leki – wpływ na proces projektowania nowych substancji leczniczych*
Toruń 2018	wystąpienie: 1st International conference Chemistry For Beauty And Health, 13-16 czerwca 2018, Toruń, Autor prezentacji: Statistic of chemical Big Data in drug design;

ZAŁĄCZNIK B

Polanski, J., Duszkiewicz, R. (2020). *Property representations and molecular fragmentation of chemical compounds in QSAR modeling*. Chemometrics and Intelligent Laboratory Systems, 104-146.

Polanski, J., Kucia, U., Duszkiewicz, R., Kurczyk, A., Magdziarz, T., Gasteiger, J. (2016). *Molecular descriptor data explain market prices of a large commercial chemical compound library*. Scientific reports, 6, 28521.

Polanski, J., Pedrys, A., Duszkiewicz, R., Gasteiger, J. (2019). *Scoring ligand efficiency: potency, ligand efficiency and product ligand efficiency within big data landscape*. Letters in Drug Design Discovery, 16(11), 1258-1263.

Polanski, J., Pedrys, A., Duszkiewicz, R., Kucia, U. (2019). *Ligand Potency, Efficiency and Drug-likeness: A Story of Intuition, Misinterpretation and Serendipity*. Current Protein and Peptide Science, 20(11), 1069-1076.

Polanski, J., Duszkiewicz, R., Pedrys, U., Gasteiger, J. (2019). *Scoring Ligand Efficiency*. Acta Pol Pharm, 76(4), 761-768.