

Rozprawa doktorska



UNIWERSYTET ŚLĄSKI
W KATOWICACH

**Fragmentacja wybranych szeregów związków chemicznych jako
metoda optymalizacji projektowania leków i materiałów**

mgr Anna Pędrys

promotor pracy:

prof. zw. dr hab inż. Jarosław Polański

Instytut Chemii

Katowice

2023

Składam serdeczne podziękowania
Panu prof. dr hab. Jarosławowi Polańskiemu
za inspirację naukową i nieocenione wsparcie, które
przyczyniły się do powstania tej pracy naukowej.

Równocześnie pragnę podziękować moim bliskim za
nieustanną pomoc i motywację, bez których ta praca nigdy
nie zostałaby ukończona.

1. STRESZCZENIE PRACY.....	3
2. WSTĘP.....	5
3. CEL BADAŃ.....	8
4. CZĘŚĆ LITERATUROWA ORAZ PODSTAWY TEORETYCZNE	11
4.1 Optymalizacja.....	11
4.2 Optymalizacja w chemii.....	13
4.2.1 Optymalizacja w chemii - metody <i>in vivo</i>	14
4.2.2 Optymalizacja w chemii - metody <i>in silico</i>	17
4.3 Bazy danych	26
4.3.1 Chemiczne i farmaceutyczne bazy danych.....	26
4.3.1.1 PubChem	28
4.3.1.2 ChEMBL	29
4.3.1.3 Reaxys	30
4.3.1.4 DrugBank	30
4.3.1.5 FDA approvals.....	31
4.3.1.6 ZINC.....	33
4.3.1.7 PharmaCompass	34
4.3.2 Materiałowe bazy danych.....	34
4.4 Chemoinformatyka jako narzędzie optymalizacyjne	36
4.4.1 Fragonomika.....	39
4.4.2 Projektowanie materiałów	40
5. BADANIA WŁASNE	41
5.1 <i>Ligand efficiency</i> jako szczególny przypadek fragonomiki leków	41
5.1.1 Wstęp teoretyczny	41
5.1.2 Metodologia.....	46
5.1.3 Wyniki	47
5.2 Badanie możliwości eksploracji wskaźników innowacyjności na podstawie listy najlepiej sprzedających się leków i <i>FDA approvals</i>	53
5.2.1 Wstęp teoretyczny	53
5.2.2 Metodologia.....	54
5.2.3 Wyniki	55
5.3 Fragmentacja TOP100.....	71
5.3.1 Wstęp teoretyczny	71
5.3.2 Metodologia.....	73
5.3.3 Wyniki	74
5.4 Fragonomika fotoreagentów.....	100
5.4.1 Wstęp teoretyczny	100
5.4.2 Metodologia.....	105
5.4.3 Wyniki	109

6. CZĘŚĆ EKSPERYMENTALNA.....	110
6.1 Charakterystyka oprogramowania	110
6.1.1 Oprogramowanie MATLAB	110
6.1.2 Język programowania Python	113
6.1.3 Oprogramowanie RDKit (biblioteki podstawowe oraz moduł RDKit. Chem.QED)	113
6.1.4 ChemPlot.....	113
7. WNIOSKI	115
8. LITERATURA.....	118
9. SPIS ILUSTRACJI I TABEL.....	128
10. ZAŁĄCZNIKI	132

1. STRESZCZENIE PRACY

Celem pracy była próba wykorzystania deskryptorów molekularnych oraz właściwości farmakologicznych i ekonomicznych wybranych szeregów leków oraz materiałów, w tym listy bestsellerów farmakologicznych (TOP100 lata 2000 do 2019) oraz listy FDA *approvals* (lata 1985 do 2019) do analizy efektywności projektowania molekularnego. Formalnie podstawową metodą analizy była fragmentacja analizowanych związków. Jedną z metod analizy fragmentów w projektowaniu leków jest metoda efektywności liganda „*Ligand Efficiency*” LE. Przeanalizowano zmiany LE dla szeregu leków zarejestrowanych przez FDA i zaproponowano alternatywne narzędzie, *Product Ligand Efficiency*, PLE.

Ekonomia ma decydujące znaczenie w projektowaniu leków i materiałów. Niestety dane ekonomiczne są trudno dostępne. Lista bestsellerów leków jest tutaj wyjątkiem. Porównano najczęściej występujące fragmenty TOP100 kontra wszystkie leki zatwierdzone przez FDA od roku 1985. Czy podobnie jak w przypadku tzw. fragmentów uprzywilejowanych farmakologicznie można wyodrębnić podobne struktury istotne dla ekonomii?

Jest pewnym paradoksem, że w projektowaniu materiałów istotnym mankamentem jest brak baz danych zawierających właściwości syntezowanych układów molekularnych. Przygotowano zestawienie danych dostępnych literaturowo dla wybranych szeregów fotoreagentów. Zintegrowane zostaną one z bazą *Catalytic Material Database*, zawierającą dane o heterogenicznych katalizatorach metanowania (<http://cmd.us.edu.pl/catalog/>).

The aim of the work was an attempt to utilize molecular descriptors, pharmacological properties, and economic data of selected series of drugs and materials, including the list of pharmacological bestsellers (TOP100 of the 2010 to 2019) and FDA approvals (from 1985 to 2019), for the analysis of the efficiency of molecular design. The primary method of analysis was the fragmentation of the analyzed compounds. An example of fragmental analysis in drug design is the ligand efficiency (LE) method. We analyzed the changes in LE for a series of drugs registered by the FDA and proposed an alternative tool, Product Ligand Efficiency, PLE.

Economics plays a crucial role in the design of drugs and materials. Unfortunately, economic data is difficult to access. The list of bestselling drugs is an exception here. The most frequently occurring fragments in the TOP100 were compared to all FDA-approved drugs. Similar to the case of pharmacologically privileged fragments, can we identify similar structures that are significant for economics?

It is paradoxical that in materials design, a significant drawback is the lack of databases of the properties of synthesized molecular systems. We prepared a compilation of literature data for selected series of photoreagents, which will be integrated with a database currently encompassing heterogeneous methanation catalysts, *Catalytic Material Database* available at <http://cmd.us.edu.pl/catalog/>.

2. WSTĘP

Optymalizacja jest uniwersalnym narzędziem umożliwiającym osiągnięcie wyznaczonego celu przy jednoczesnym ograniczaniu towarzyszących efektów niepożądanych. Matematycznie optymalizację wyznacza się jako ekstremum funkcji pod kątem wybranego kryterium jakości.

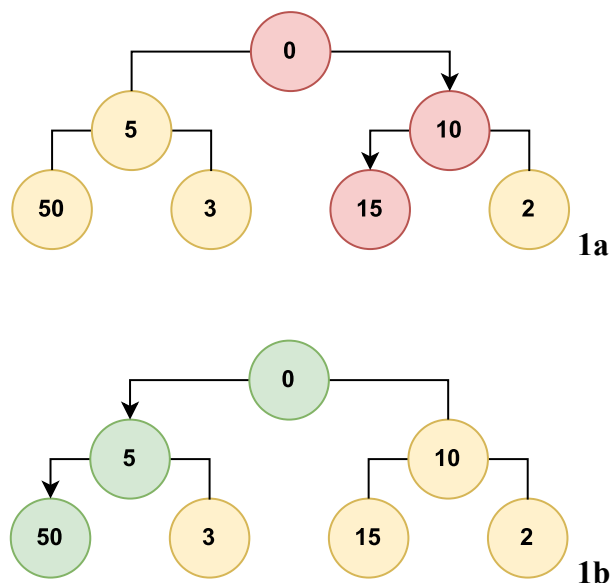
Metody optymalizacyjne stosuje się w niemal każdym sektorze, w przedsiębiorstwie produkcyjnym dążąc do zwiększenia zysków, w programowaniu, aby uzyskać poprawę wydajności kodu czy w centrum logistycznym, planując trasy dostaw. Szczególnym przykładem wykorzystania optymalizacji jest modelowanie ryzyka jako strategia walki z pandemią, powstała w czasie szczytowych zarażeń COVID-19. Szereg rozwiązań technologicznych już obecnych w telefonach komórkowych czy komputerach umożliwia śledzenie kontaktów oraz schematów zachowań danej jednostki. Na podstawie informacji gromadzonych przez dedykowane aplikacje zdrowotne będzie możliwe wyznaczenie kierunków rozprzestrzeniania się pandemii, a co za tym idzie, umożliwienie podjęcia skoncentrowanych działań mających na celu zapobieganie powstawaniu nowych ognisk choroby.

Pomimo wysokiego priorytetu optymalizacji, pozostają obszary, w których trudność sformalizowania metod ich implementacji stanowi poważną przeszkodę [1]. Przykładem może być problem doboru architektury sieci neuronowych, w przypadku którego nie ma jasnych przesłanek sugerujących rekomendowaną metodę rozwiązania tego zagadnienia. Jedną z prób zaadresowania problemu jest twierdzenie Kołmogorowa. Zgodnie z nim, aby przeprowadzić skuteczną aproksymację N -wymiarowego zbioru wejściowego x w M -wymiarowy zbiór wyjściowy y wystarczy jedna warstwa ukryta, z $2N+1$ neuronami [2,3]. Obserwacja Kołmogorowa nie stanowi jednak metody optymalizacyjnej, ponieważ istnieją zbiory, dla których wystarczająca jest znacznie niższa liczba neuronów, \sqrt{NM} lub mniej [4].

Innym przykładem wyzwania związanego z optymalizacją są algorytmy zachłanne. W każdym kroku podejmowana jest decyzja lokalnie najlepsza (optymalna), bez oceny dalszych

kroków [5]. Podjęte w ten sposób decyzje mają poważny wpływ na końcowy rezultat optymalizacji i mogą prowadzić do błędnego wyznaczenia optimum.

Rys. 1a – b Grafy przedstawiające problem podejmowania przez algorytm zachłanny decyzji na podstawie lokalnego optimum. Założeniem eksperymentu jest wyznaczenie ścieżki od góry grafu do dołu, tak by osiągnąć maksymalną sumę. Sposób rozwiązania problemu z wykorzystaniem algorytmu zachłannego, suma wynosi 25 (a) Optymalne rozwiązanie, suma wynosi 55 (b), inspirowane [5].



Istnieją jednak problemy obliczeniowe takie jak poszukiwanie najkrótszych ścieżek (algorytm Dijkstry) czy poszukiwanie minimalnego drzewa rozpinającego (algorytm Kruskala), dla których algorytmy zachłanne dają gwarancję rozwiązania optymalnego [6-8]. Ponadto niewątpliwą zaletą jest ich szybkość oraz niski stopień skomplikowania.

Pojęcie optymalizacji w dziedzinie chemii obejmuje wiele zróżnicowanych problemów na przykład intuicyjne modelowanie warunków reakcji w celu uzyskania możliwie najwyższej wydajności, czystości produktu czy nadmiaru enancjomerycznego. W tym wypadku ścieżkę optymalizacyjną wyznacza się na podstawie obserwacji i doświadczenia wspomaganego wiedzą teoretyczną z zakresu analizowanych procesów. Zwykle potrzeba wielu lat pracy w laboratorium w celu wypracowania tzw. intuicji chemicznej, tak by móc sprawnie optymalizować reakcje.

Wielokierunkową alternatywą klasycznej optymalizacji w chemii jest wprowadzenie technik opartych na algorytmach. Szerokie zastosowanie znajdują one obecnie w procesie projektowania leków, prowadząc do znacznego zwiększenia efektywności pod kątem czasu i materiałów. W rezultacie, uzyskuje się redukcję kosztów oraz zmniejszenie negatywnego wpływu procesu na środowisko.

Chemoinformatyka skupia się na projektowaniu leków. Po raz pierwszy termin pojawia się w artykule Franka K. Browna jako narzędzie do “przekształcenia danych w informacje, a informacji w wiedzę celem szybszego podejmowania lepszych decyzji na polu identyfikacji i optymalizacji leków” [9, 10]. Ćwierć wieku po pierwszym zdefiniowaniu, zastosowanie chemoinformatyki znacznie wykracza poza obszar projektowania leków. Ponadto, obecnie najciekawsze wyniki uzyskuje się zestawiając dane chemiczne (m.in strukturę chemiczną, aktywność czy parametry farmakokinetyczne) z danymi pochodzącymi z pozornie nieskorelowanych z chemią obszarów, w tym ekonomii.

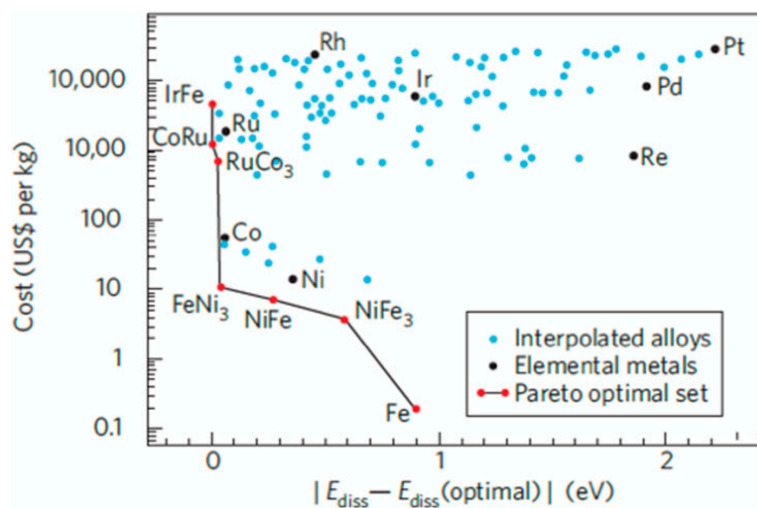
3. CEL BADAŃ

Głównym celem badań było wykorzystanie metod fragmentacyjnych jako narzędzia optymalizującego leki oraz materiały. Obecnie ich optymalizację prowadzi się głównie z wykorzystaniem deskryptorów molekularnych oraz właściwości chemicznych i farmakologicznych (leki) lub funkcjonalnych (materiały) [9-11].

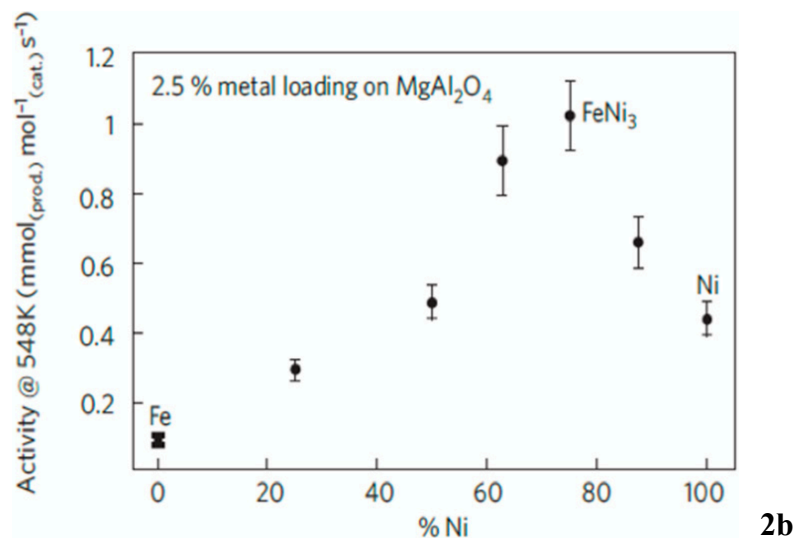
Przeanalizowano zmiany efektywności liganda (ang. *ligand efficiency*, LE) dla szeregu leków oraz ich fragmentów, zarejestrowanych przez FDA. Dokonano analizy znaczenia LE, na podstawie którego podjęto próbę skonstruowania alternatywnego narzędzia do ewaluacji kandydatów na lek (ang. *drug candidates*), *Product Ligand Efficiency* (PLE).

Pomimo tego, że projektowanie w zakresie leków i materiałów (optymalizacja) powinno uwzględniać jako podstawowy parametr aspekt ekonomiczny, bardzo rzadko wykorzystuje się tego typu dane [12]. Do niewielu wyjątków należą prace opisujące dobór katalizatorów w oparciu o ich rynkowe ceny czy ekonomiczne analizy zależności struktura – aktywność (ang. QSAR, *Quantitative Structure–Activity Relationship*) [13-15].

Rys. 2a – b Przykład wykorzystania danych ekonomicznych do wyznaczenia optymalnej ścieżki projektowania katalizatorów metanizacji. Porównanie stosunku ceny do wydajności katalitycznej jedno- i wieloskładnikowej układu Fe|Ni (a) do wykresu aktywności czystych katalizatorów (b) umożliwi wytypowanie optymalnego katalizatora pod względem ekonomicznym i wydajnościowym [14, 15].

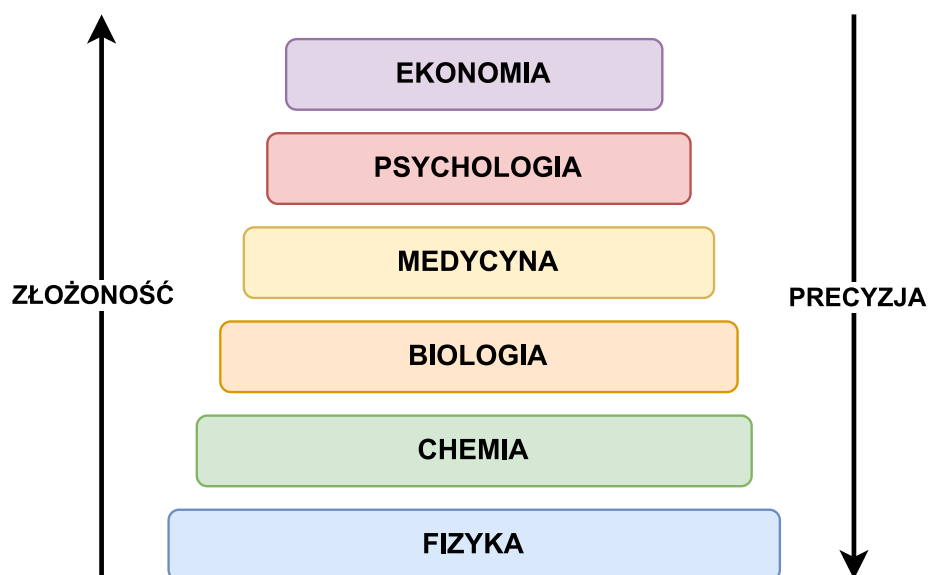


2a



Przyczynę stanowi duża trudność modelowania zależności ekonomicznych. Wyjaśniliśmy to piramidą kompleksowości nauk (rysunek poniżej). Ponadto, poprawne manipulowanie i analizowanie danych ekonomicznych wymagają umiejętności nierozwijanych w toku studiów na kierunkach ścisłych i technicznych, a współpraca pomiędzy zespołami ekonomistów i chemików należy do rzadkości.

Rys. 3 Piramida złożoności i precyzji nauk. Im niższy poziom złożoności danej dziedziny, tym wyższa precyzja jej opisu. Ekonomia zajmuje najwyższe miejsce diagramu. Jest najbardziej złożona i najtrudniejsza do modelowania oraz predykcji [9].



Problem ten zaadresowano wyznaczając drugi cel badań, jakim była próba wykorzystania deskryptorów molekularnych oraz właściwości farmakologicznych w połączeniu z danymi ekonomicznymi najlepiej sprzedających się leków, tzw. bestsellerów (TOP100, TOP), w latach 2010 - 2019. TOP100 to w najprostszym ujęciu model idealnego leku. Analiza tej korelacji umożliwiła więc ocenę innowacyjności leku, parametru, który może zostać wykorzystany do oceny szansy awansu nowych FDA *approvals* na listę najlepiej sprzedających się leków w nadchodzących latach.

Utworzone biblioteki TOP oraz FDA *approvals* poddano fragmentacji oraz wytypowano na ich podstawie struktury uprzywilejowane. Następnie przeprowadzono analizę zbieżności strukturalnej zbiorów drogą wizualizacji przestrzeni chemicznej wraz z redukcją jej wymiaru (t-SNE, SOM), co stanowi trzeci cel badań niniejszej rozprawy doktorskiej.

W ostatnim etapie podjęto próby przeniesienia metod fragmentacyjnych stosowanych z powodzeniem w projektowaniu leków do optymalizacji materiałów. Do analizy wybrano szereg fotokwasów (ang. *photoacids*, PAHs) oraz generatorów fotokwasów (ang. *photoacid generators*, PAGs). Podstawę każdej analizy fragonomicznej stanowi odpowiednio przygotowana baza danych, zawierająca struktury wybranej grupy związków w postaci czytelnej dla komputera, np. w notacji SMILES [9]. Gromadzenie danych dotyczących fotoreagentów doprowadziło do zaobserwowania braku ogólnodostępnych baz danych dedykowanych materiałom, zawierających właściwości syntezowanych układów molekularnych. Popularne bazy chemiczne takie jak PubChem, ChEMBL czy Reaxys, dostarczające setek zróżnicowanych danych o cząsteczkach (zarówno eksperymentalnych jak i dostarczonych metodami obliczeniowymi, *in silico*) od mas molowych, przez struktury przestrzenne, widma spektroskopowe po aktywność, nie są przystosowane do przechowywania danych dotyczących materiałów, jakimi są fotokwasy i ich generatory.

Wobec powyższego, ostatni cel badawczy polegał na przygotowaniu bazy danych dedykowanej wybranej grupie fotoreagentów, która następnie zostanie umieszczona w ogólnodostępnej bazie *Catalytic Material Database* (<http://cmd.us.edu.pl/catalog/>), w sekcji poświęconej fotokwasom i generatorom fotokwasów.

4. CZĘŚĆ LITERATUROWA ORAZ PODSTAWY TEORETYCZNE

4.1 Optymalizacja

Ogólnie optymalizację określa się jako “znalezienie najlepszego rozwiązania spośród wielu możliwych” [16]. Jej dokładna definicja zależy od obszaru zastosowania oraz parametru będącego celem. Potwierdzeniem częstości wykorzystania i uniwersalności optymalizacji może być „Encyklopedia Optymalizacji” wydawnictwa Springer [17], stworzona we współpracy z ponad 400 specjalistami z każdego sektora gospodarki.

Optymalizacja jest obecna w niemal każdym aspekcie ludzkiego życia. Stanowi podstawę dla dziedzin nauk takich jak inżynieria procesów, robotyka, informatyka, matematyka stosowana, medycyna czy ekonomia. Wkracza również w życie prywatne stanowiąc kurs dla zakupów spożywczych, projektowania ubrań czy wyznaczania tras. W tym aspekcie optymalizacją nazywa się metody pozwalające określić kierunek udoskonalenia procesu lub produktu, w celu uzyskania najlepszego wyniku, pod kątem wybranego parametru.

Wielu naukowców stawia hipotezę, iż optymalizacja jest naturalną potrzebą ludzkości. Obserwacja dotychczasowych rozwiązań i poszukiwanie ulepszeń doprowadziły do przełomowych odkryć takich jak wynalezienie koła, silnika parowego czy druku. Jako proces towarzyszyła nam od zawsze, a ujęcie w matematyczne prawa umożliwiło sprawne przełożenie jej na obszary dotychczas nieobjęte optymalizacją [16].

Wbrew powszechnemu przekonaniu, głównym motorem napędowym optymalizacji nie jest jedynie aspekt finansowy, ale świadomość niedoskonałości znanego i dotychczas stosowanego rozwiązania. To ona kieruje technologię ku minimalizacji urządzeń przy jednoczesnym podnoszeniu ich wydajności czy inspiruje przemysł farmaceutyczny do ciągłego udoskonalania dostępnych terapii.

Inną kategorią aspektu optymalizacji, niezwiązanego z ekonomią, może być kryzysowa sytuacja zagrażająca życiu lub prowadząca do nieodwracalnych zmian w środowisku naturalnym. Przykładem mogą być pandemia COVID-19 lub akcja związana z wyciekiem

ropy naftowej w Zatoce Meksykańskiej. Dzięki optymalizacji podejmowanych działań w czasie ekspansji koronawirusa, możliwe było określenie kierunków rozwoju zachorowań i podejmowanie w wyznaczonych strefach wzmoczonych akcji zapobiegawczych. Podczas neutralizacji ropy z wycieku z platformy Deepwater Horizon, głównym celem było ocalenie przebywających na niej pracowników, przy równoczesnej minimalizacji obszaru skażenia.

Czy dążenie do ideału może mieć negatywny skutek na proces? Optymalizacja z definicji posiada wyłącznie pozytywne następstwa, lecz przeprowadzona zbyt wcześnie prowadzi do błędnego wyznaczenia celu. Zjawisko przedwczesnej optymalizacji jest często zauważalne w obszarze IT, jak chociażby podczas premiery rządowej strony www.healthcare.gov w Stanach Zjednoczonych [18, 19].

Portal powstał w celu umożliwienia i ułatwienia wyszukiwania korzystnej oferty ubezpieczenia i zarządzania swoją polisą obywatelom USA, nieobjętych ubezpieczeniem zdrowotnym z ramienia swojego pracodawcy. Strona internetowa na którą przeznaczono z budżetu państwa niecałe 94 miliony dolarów, uległa awarii już po dwóch godzinach od jej opublikowania, umożliwiając pomyślną rejestrację zaledwie sześciu obywatelom.

Rys. 4 Grafika ukazująca awarię rządowego portalu HealthCare, tuż po opublikowaniu [18].



Natychmiast wdrożono rozwiązanie mające umożliwić ponowne działanie strony, dodając do jej obsługi kilka dodatkowych serwerów, bez rzetelnego wyznaczenia przyczyny awarii. Próba optymalizacji nie powiodła się, ponieważ problem w funkcjonowaniu portalu nie leżał w przepustowości obsługujących go serwerów, lecz był spowodowany błędami w kodzie i brakiem przeprowadzenia szczegółowych testów e2e.

Dopiero dwa miesiące później udało się ponownie opublikować stronę i umożliwić pomyślną rejestrację u 80% z odwiedzających. Przygotowano również szczegółowy raport na podstawie wywiadów z 86 pracownikami zaangażowanymi w pracę nad www.healthcare.gov, kilku tysięcy maili, notatek, prezentacji oraz dokumentacji powstałej podczas realizacji projektu. Sprawozdanie ukazuje źródła niepowodzenia premiery portalu oraz stanowi przestrożę dla środowiska IT przed powtórzeniem wskazanych błędów podczas wdrażania projektów.

4.2 Optymalizacja w chemii

Optymalizację w chemii można podzielić analogicznie jak projektowanie leków, na przeprowadzoną w oparciu o metody *in vivo* oraz *in silico*. Pierwsze z nich stanowią rezultat obserwacji na przykład przebiegu reakcji oraz jej produktów, wsparte wiedzą teoretyczną i doświadczeniem zdobytym podczas pracy w laboratorium. Drugie są odpowiedzią na wciąż rosnący udział metod komputerowych oraz obliczeniowych w symulacjach procesów chemicznych. Obie metody czerpią inspirację z innych obszarów nauki i przemysłu, takich jak statystyka, ekonomia czy inżynieria procesów.

W laboratorium badawczym optymalizację wspomaga doświadczenie oraz intuicja wsparta samodzielnie rozszerzoną wiedzą teoretyczną z obszaru optymalizacji. W większości ramowych programów studiów na kierunku chemii, brakuje przedmiotu skupiającego się na metodach optymalizacji w chemii doświadczalnej [20]. W laboratoriach akademickich celem optymalizacji najczęściej jest otrzymanie nowej pochodnej, pożądanego podziału mieszaniny enancjomerów lub produktu o wysokiej czystości.

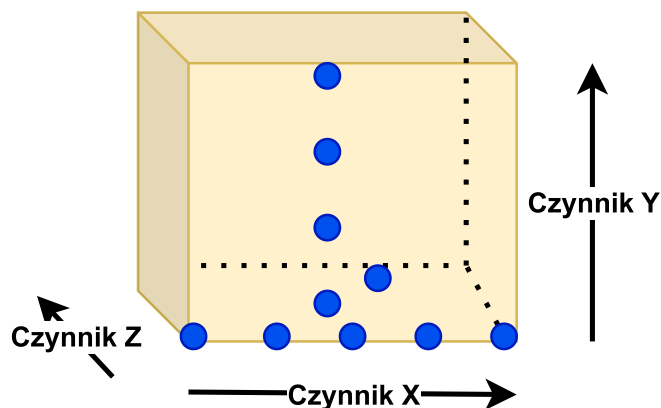
Laboratoria przemysłowe stanowią jeden z elementów przedsiębiorstwa i są pod stałym nadzorem działu finansów oraz R&D (ang. *Research and Discovery*). Stąd też w ich przypadku optymalizacja głównie skupia się na redukcji kosztów, podążaniu za wymaganiami klientów oraz ograniczaniu szkodliwych odpadów tak by produkt przynosił możliwie największy dochód, był konkurencyjny na rynku, a także miał pozytywny wpływ na wizerunek firmy [20].

4.2.1 Optymalizacja w chemii - metody *in vivo*

Optymalizacja *in vivo* opiera się przede wszystkim na intuicji i doświadczeniu chemicznym. Studia uniwersyteckie, praca w laboratorium oraz ciągłe rozszerzanie wiedzy są niezbędne do ich zdobycia i wykorzystania w celu poprawy metryk. Podłoże stanowi zrozumienie procesów zachodzących podczas reakcji i umiejętne manipulowanie jej warunkami.

W metodzie OFAT (ang. *One Factor At a Time*) reakcje przeprowadza się iteracyjnie, ustalając najlepszą wartość dla jednego z czynników np. temperatury utrzymując stałą wartość pozostałych [21]. W momencie osiągnięcia optimum, reakcję powtarza się analogicznie do wyznaczenia optimum uwzględniając tym razem kolejny parametr. W wyniku badań otrzymuje wielowymiarową przestrzeń, ograniczoną minimum i maksimum każdego z analizowanych czynników.

Rys.5 Przykładowy schemat przestrzeni wyników dla parametrów reakcji (x, y, z).







Wyniki uzyskane metodą OFAT obarczone są wysokim błędem, ponieważ nie uwzględniają interakcji pomiędzy parametrami. Opierają się na intuicji i wymagają wielokrotnego powtarzania tej samej reakcji, dlatego głównie stosuje się je w laboratoriach akademickich.

Jednym z największych wyzwań optymalizacyjnych *in vivo* jest skalowanie reakcji. Problem ten szczególnie dotyka laboratoria przemysłowe, gdzie stawia się wymóg dostarczania handlowych ilości produktu przy stałym dążeniu do ograniczania kosztów związanych z produkcją w tym kosztów operacyjnych jak również środków przeznaczanych na niwelowanie szkodliwych produktów ubocznych reakcji. Jednocześnie produkt musi spełniać wszystkie wymagania jakościowe, szczególnie istotne w przemyśle farmaceutycznym oraz spożywczym.

W większości przypadków podczas skalowania nie jest możliwe odtworzenie zoptymalizowanych warunków reakcji, wyznaczonych w mikroskali. Jest to spowodowane występowaniem w reaktorach wielkoskalowych szerokiego gradientu temperatury, stężenia reagentów oraz ciśnienia. Po przeniesieniu reakcji na dużą skalę, konieczna jest jej ponowna optymalizacja, co wiąże się z poniesieniem dodatkowych kosztów, często wykraczających poza możliwości finansowe przedsiębiorstwa lub instytutu badawczego.

Największe trudności przy skalowaniu reakcji występują w przypadku reakcji szybkich. Wolniejsze o kilka rzędów wielkości procesy mieszania czy wymiany ciepła, o zmiennych wartościach w zależności od punktu reaktora, prowadzą m.in do ograniczenia dostępności reagenta czy nieosiągnięcia wymaganej temperatury w odpowiednim czasie.

Rys.6 Porównanie czasu mieszania i ogrzewania w skali laboratoryjnej oraz przemysłowej [22, 23].

Czas procesu w skali laboratoryjnej kontra przemysłowej [sekundy]				
	Kolba okrągłodenna	Mikroreaktor	Reaktor przepływowy	Reaktor z mieszaniem
Mieszanie	1 - 10	0.001 – 0.2	0.02 – 0.2	1 - 20
Ogrzewanie		0.01 – 0.5	10 - 150	300 - 600

Optymalizacja reakcji w makroskali nie jest możliwa do osiągnięcia wyłącznie poprzez liniowe zwiększenie parametrów reakcji wyznaczonych w mikroskali. Najefektywniejszym rozwiązaniem problemu jest doświadczalne wyznaczenie modelu kinetycznego, określającego wpływ parametrów reakcji takich jak stężenie substratów, temperatury, ciśnienia, rodzaju i stężenia rozpuszczalnika oraz katalizatora na szybkość reakcji.

Dla ułatwienia procesu wyznaczania modeli kinetycznych wykorzystuje się uproszczone równania analizy wielowymiarowej, umożliwiające pominięcie parametrów niekrytycznych w pierwszej iteracji. Wsparcie stanowią również literaturowe opracowania równań dla danego typu reaktora i ich eksperymentalne przykłady jak np. prace Seana Morana i Klaus-Dieter Henkela [22, 23]. Przedstawiono w nich przykłady przemysłowego zastosowania reaktorów wraz z ich parametryzacją, w zależności od typu przeprowadzanej reakcji oraz stanu skupienia reagentów. W praktyce laboratoryjnej, wspomniane opracowania wykorzystuje się jako punkt wyjściowy optymalizacji, dostosowując je do reakcji, która ma zostać przeskalowana.

Alternatywą dla zwiększania objętości reaktorów jest zastosowanie ich zwielokrotnienia. W instalacji umieszcza się setki lub tysiące analogicznych mikroreaktorów i parametryzuje reakcje zgodnie z pierwotną optymalizacją. Rozwiązanie znacznie skraca czas potrzebny na

skalowanie, jednak wiąże się z wysokim kosztem inwestycyjnym, wynikającym z konieczności zakupu dodatkowych reaktorów, mieszadeł czy układów sterowania. Ze względu na aspekt finansowy, rozwiązanie stosuje się głównie w produkcji leków, dzięki możliwości uzyskania tą metodą produktu o wymaganej, wysokiej jakości.

4.2.2 Optymalizacja w chemii - metody *in silico*

Wraz ze wzrostem dostępności komputerów i oprogramowania chemicznego, gwałtownie wzrósł udział metod *in silico* w strategiach badawczych, również w obszarze chemii. Znajdują one zastosowanie także w optymalizacji, od etapu projektowania reakcji przez odpowiednie projektowanie eksperymentu (DoE, ang. *Design of Experiment*) [21].

W kontekście części materiałowej mojej pracy (Badania Własne, Rozdział 5.4) istotnym problemem optymalizacji *in silico* są metody efektywne w domenie wielkich danych (ang. *Big Data*), których pierwszy etap stanowi przygotowanie bazy danych.

Drugi etap polega na wykorzystaniu zebranych danych do wyznaczenia potencjalnego optimum reakcji, w tym jej produktów oraz warunków. W tym celu stosuje się metody uczenia maszynowego (ang. *Machine Learning*, ML). Choć uczenie maszynowe z powodzeniem stosowane jest w wielu obszarach, w tym w ruchu samochodowym do rozpoznawania znaków i opracowywania autonomicznych pojazdów, w obszarze *public relations* i marketingu do śledzenia i reagowania na zachowanie publiczności oraz klientów czy w przetwarzaniu mowy i translacjach w czasie rzeczywistym, jego zastosowanie w chemii wciąż spotyka się z krytyką i niezrozumieniem [24].

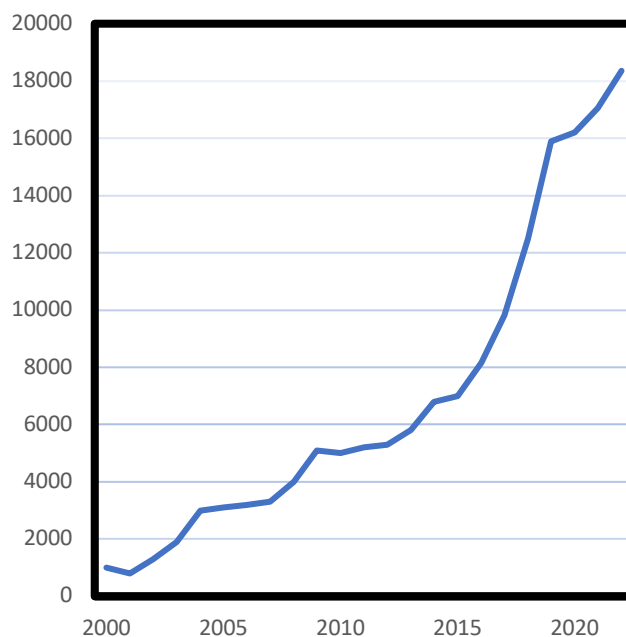
Zespół Johna A. Keitha przedstawia w swoim opracowaniu wynik ankiety dotyczącej obaw związanych z zastosowaniem uczenia maszynowego w chemii, przeprowadzonej wśród społeczności naukowej. Wśród nich znalazły się takie postulaty jak:

1. ML jest coraz częściej stosowany w postaci “czarnej skrzynki”, przy braku zrozumienia mechanizmu jego działania.
2. ML konstruuje się na niepoprawnie przygotowanych zbiorach testowych i treningowych, dyktowanych brakiem doświadczenia zespołu w tym obszarze. Ponadto trudności związane z dostępem do danych (np. płatne bazy) powodują, że jakość danych wejściowych często jest niska.
3. Nie istnieją kryteria jasno wskazujące, w jakim przypadku wykorzystanie ML jest uzasadnione.

Pomimo zgłaszanych zastrzeżeń, ta sama grupa ankietowanych wskazała, iż ML stanowi najbardziej ekscytujący kierunek rozwoju takich obszarów jak kataliza, projektowanie leków oraz peptydów, wyznaczanie przebiegu reakcji jak i odkrywanie nowych związków chemicznych. Przykładem potwierdzającym tezę ankietowanych jest zastosowanie sztucznej inteligencji we wspomaganej komputerowo syntezie organicznej (retrosyntezie) oraz projektowaniu leków, stanowiących przedmiot zainteresowania Centrum Projektowania i Syntezy Leków oraz Materiałów Uniwersytetu Śląskiego [25, 26].

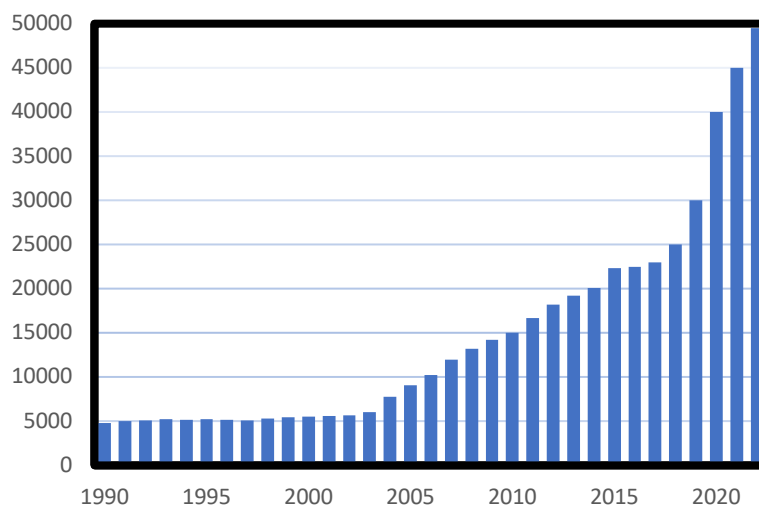
Na potwierdzenie wzrostu znaczenia ML, przedstawiony został również wykres ilustrujący gwałtowny skok zainteresowania tą metodą wyznaczony na podstawie częstości występowania terminu uczenia maszynowego w dowolnym kontekście, w artykułach opublikowanych w czasopiśmie PHYS (*Physical Chemistry, ACS Technical Division*).

Rys. 7 Wykres przedstawiający liczbę wystąpień terminu uczenie maszynowe w artykułach opublikowanych w czasopiśmie PHYS (lata 2000 – 2023), inspiracja [24].



Inne potwierdzenie tezy stanowi analogiczny wykres przedstawiający liczbę artykułów opublikowanych w PubMed w latach 1990 - 2023, w których w tytule, lub abstrakcie wymienione zostały słowa kluczowe takie jak “Machine Learning”, „Artificial Intelligence” lub “AI”.

Rys. 8 Częstość występowania terminów “Machine Learning”, Artificial Intelligence” lub “AI” w artykułach zgromadzonych w repozytorium PubChem w latach 1990-2023, inspiracja [24].



Krytykę zastosowania ML w chemii można porównać do zjawiska technofobii, powstałego w czasie Rewolucji Przemysłowej (XVIII w., Anglia i Szkocja). W obu przypadkach obawy spowodowane są przede wszystkim brakiem wiedzy z obszaru podstaw funkcjonowania technologii. To z kolei prowadzi do błędnego przekonania o możliwości całkowitego wykluczenia czynnika ludzkiego z przemysłu oraz badań, lęku przed utratą zatrudnienia oraz pozycji społecznej.

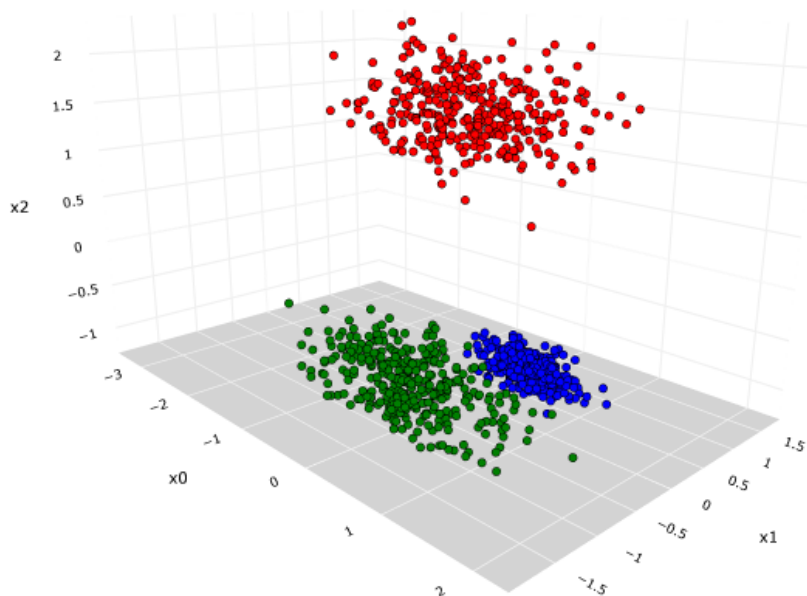
Kluczem do poprawnej implementacji ML jest zrozumienie jego istoty oraz mocnych i słabych stron. Decyzja o zastosowaniu uczenia maszynowego powinna zostać podjęta na podstawie porównania ograniczeń metody względem wyznaczonego celu badawczego. Jednocześnie termin ML w najszerszym znaczeniu obejmuje obecnie rozległy obszar metod numerycznych, w tym na przykład klasteryzację.

Zgodnie z definicją Donalda Michie (1991 r.), uczeniem maszynowym nazywa się “system wykorzystujący zewnętrzne dane empiryczne w celu tworzenia i aktualizacji podstaw dla udoskonalonego działania na podobnych danych w przyszłości oraz wyrażania tych podstaw w zrozumiałej i symbolicznej postaci.” W bardziej ogólnym ujęciu, uczenie maszynowe polega na szacowaniu zależności pomiędzy danymi wejściowymi (uczenie nienadzorowane) lub danymi wejściowymi kontra dane wyjściowe o etykietach nadanych przez człowieka (uczenie nadzorowane) w celu wyznaczenia modelu matematycznego opisującego zbiór [21].

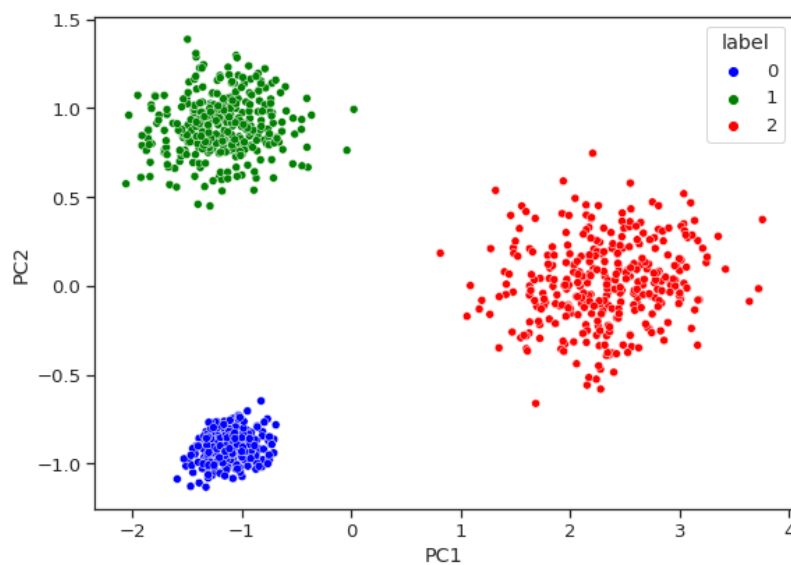
Wynikiem zastosowania metod z obszaru uczenia nienadzorowanego jest uporządkowanie zbioru wejściowego. Jednym z przeprowadzanych w tym celu procesów jest analiza PCA (ang. *Principal Component Analysis*), prowadząca do redukcji wymiarowości danych, poprzez typowanie i odrzucanie cech o najmniejszej wartości informacyjnej przy jednoczesnym odkrywaniu ewentualnych prawidłowości między nimi. W rezultacie otrzymuje się uproszczony, zazwyczaj dwuwymiarowy układ, którego współrzędne stanowią dwie najistotniejsze wariancje. W celu ich wyznaczenia wykorzystuje się z techniki faktoryzacji macierzy, zwanej rozkładem według wartości osobliwych (ang. *Singular Value Decomposition*, SVD), rozkładające macierz danych na iloczyn skalarny. W praktyce jednak, do analizy PCA wykorzystuje się gotowe biblioteki np. w Pythonie czy predefiniowane

funkcje jak np. w MATLABIE, uzyskując w rezultacie łatwiejszą do analizy wizualizację danych.

Rys. 9a – b Przykład uproszczenia danych w wyniku analizy PCA, trójwymiarowe dane wejściowe (a) zostały uproszczone do dwóch wymiarów (b), rysunki zostały odtworzone w Pythonie zgodnie z kodem udostępnionym na <https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d#0226>



9a

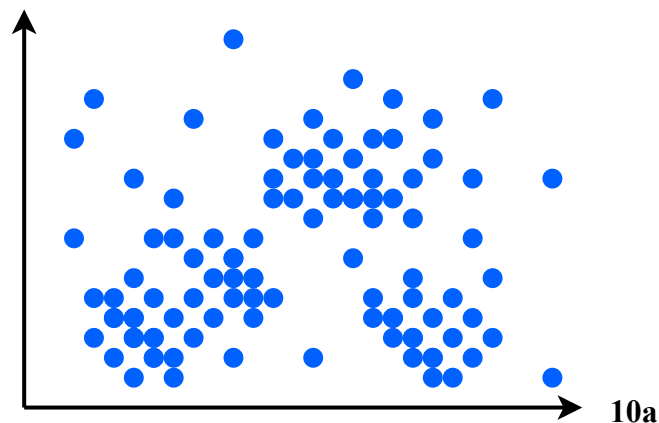


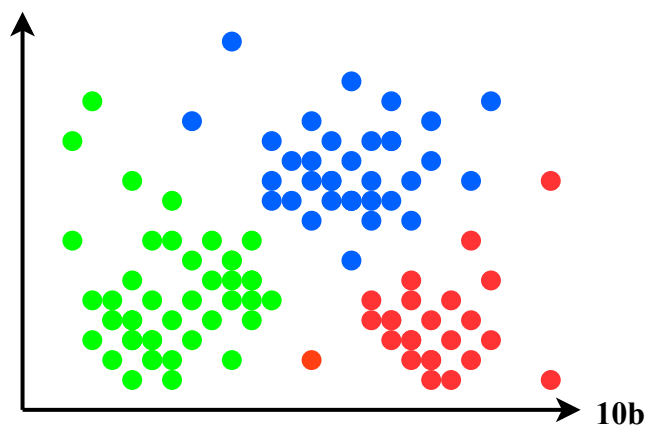
9b

Drugim przykładem uczenia nienadzorowanego są mapy samoorganizujące (ang. *Self-Organizing Maps*, SOM). Analogicznie do PCA umożliwiają zmniejszenie wymiarowości danych, a to z kolei prowadzi do uproszczenia ich analizy. SOM działa w dwóch trybach, treningu (samouczenia) i mapowania zbioru testowego. W pierwszym etapie na podstawie danych wejściowych tworzona jest mapa o dwóch wymiarach, składająca się z komponentów zwanych węzłami (neurony). Neurony reprezentujące podobne klasy znajdują się obok siebie, umożliwiając obserwację relacji między nimi. Drugi etap stanowi przepuszczenie przez SOM pozornie nieuporządkowanego zbioru testowego, w rezultacie uzyskując rozdział między klasy.

Innym przykładem procesu uczenia nienadzorowanego jest klasteryzacja (analiza skupień), w wyniku której pozornie nieskorelowane ze sobą dane łączone są w grupy o wspólnych atrybutach, w celu ekstrapolacji występujących w nich zależności. Umożliwia również wytypowanie cechy odrębnej dla danego rekordu tj. niewystępującej w żadnym z utworzonych zbiorów. Analogicznie jak w przypadku PCA, funkcje do klasteryzacji dostępne są w postaci bibliotek lub predefiniowanych funkcji, wbudowanych w oprogramowanie analityczne.

Rys. 10a – b Schemat klasteryzacji, dane przed analizą, pozornie brak zdefiniowanych klas (a), dane po klasteryzacji, wyraźnie wyodrębnione trzy klasy (b).






Uczenie nadzorowane prowadzi do wyznaczenia modelu predykcyjnego na podstawie zbioru uczącego, który składa się z par danych wejściowych oraz przypisanych im przez nadzorującego (człowieka) wartości wyjściowych.

Jednym z zastosowań uczenia nadzorowanego może być rozwiązywanie problemu regresji, czyli konstrukcja modelu przewidującego wartości wyjściowe. Prostym przykładem jest model utworzony na podstawie bazy danych zawierającej wzrost dzieci oraz ich wiek, zgodnie z poniższym schematem:

Rys. 11 Przykład schematu uczenia nadzorowanego do rozwiązywania problemu regresji

Wiek [lata]	2	4	5	7	9
Wzrost [cm]	98	110	122	134	146

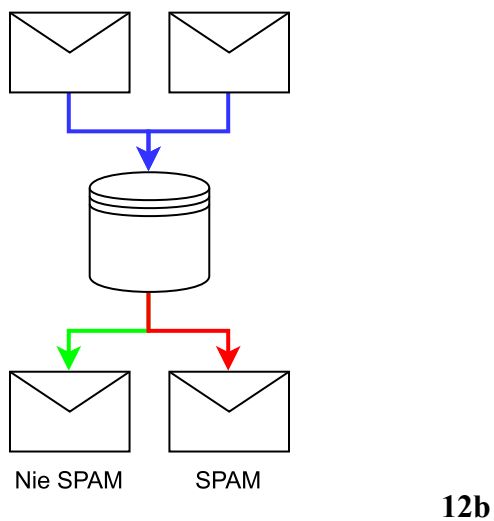
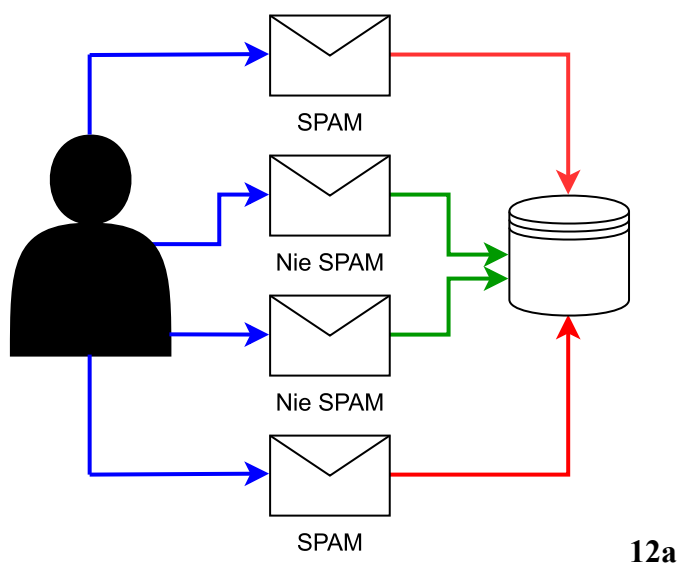

 Ile wzrostu ma 8 letnie dziecko?

Przykładami algorytmów regresji są [13]:

1. regresja liniowa
2. regresja wielomianowa
3. drzewo regresyjne
4. sieci neuronowe

Drugim problemem rozwiązywanym przy użyciu uczenia nadzorowanego jest klasyfikacja. Umożliwia przewidywanie przynależności do predefiniowanych klas, np. kwalifikowanie e-maili jako spam. Pierwszy etap stanowi oznaczanie próbek etykietami, zgodnie z założeniem klasyfikacyjnym, na tzw. zbiorze uczącym. Wraz ze wzrostem rozmiaru zbioru uczącego oraz wysoką precyzją oznaczeń, rośnie moc klasyfikacyjna modelu, tj. poprawne przypisywanie przez niego etykiet na nieoznakowanym zbiorze (tzw. zbiór testowy).

Rys. 12a - b Uproszczony schemat uczenia nadzorowanego do rozwiązywania problemu klasyfikacji, etap uczenia (a), klasyfikacja (b).



Przykładami algorytmów stosowanych do rozwiązywania problemu klasyfikacji są [27]:

1. drzewa decyzyjne
2. metoda k-najbliższych sąsiadów
3. maszyny wektorów nośnych (ang. *Support Vector Machine*, SVM)
4. naiwny klasyfikator Bayesa
5. las losowy
6. regresja logistyczna
7. sieci neuronowe

Podczas optymalizacji procesów chemicznych, typowymi danymi wejściowymi dla ML są parametry takie jak czas reakcji, temperatura, ciśnienie. Dane wyjściowe stanowią cele optymalizacji, mogą być nimi wydajność, czystość produktu lub nadmiar enancjomeryczny. Moc predykcyjna modelu jest następnie sprawdzana na podstawie danych spoza zbioru uczącego.

Obecnie większość zbiorów uczących jest konstruowana na podstawie wyników doświadczalnych, uzyskanych uprzednio np. metodą HTE (ang. *High-Throughput Experimentation*, metoda polegająca na równoległym przeprowadzaniu wielu, analogicznych reakcji w mikroskali, manipulując warunkami reakcji oraz/lub reagentami i umożliwiającą relatywnie szybkie budowanie bazy danych). Otrzymane w ten sposób wyniki tworzą wystandaryzowaną bazę o wysokiej jakości, jednak proces jest czasochłonny i kosztowny. Równie kosztowne jest wykupienie dostępu do bazy danych skonstruowanej przez inne laboratorium.

Alternatywę stanowi *data mining*, czyli eksploracja i wykorzystanie publicznie dostępnych baz danych molekularno-strukturalnych. Jest to jednak proces żmudny i narażony na błędy w przypadku przeprowadzania go w formie manualnej. Forma zautomatyzowana natomiast wymaga odpowiedniej konstrukcji i parametryzacji narzędzia, do czego wymagana jest wiedza i doświadczenie z zakresu robotyzacji i automatyzacji procesów.

4.3 Bazy danych

Wraz ze wzrostem dostępności komputerów oraz rozszerzającym się dostępem do Internetu, gwałtownie wzrasta ilość danych generowanych przez jego użytkowników. Średnio, jeden mieszkaniec Ziemi produkuje 1.7MB danych na sekundę, a sam Facebook zbiera 4PB danych każdego dnia (biliard bajtów, wartość równa 11,000 filmów w jakości 4K [28, 29]). Według IBM zaledwie 15% danych jest oryginalnych, resztę stanowią kopie.

Możliwości jakie płyną z analizy danych sprawiają, że obecnie uważa się je za najmocniejszą walutę [30]. Są uniwersalnym i nieskończonym zasobem, dostarczonym za darmo przez użytkowników *social media*, sklepów internetowych czy platform streamingowych. Celem hurtowni danych jest ich gromadzenie i przekształcanie w sposób, umożliwiający przeprowadzanie statystyk w wybranym obszarze.

Szczególnie pożądane są bazy danych określane mianem *big data*. WHO definiuje je jako szybko gromadzone, złożone i wielowymiarowe zbiory, do przechowywania których, konieczne są dyski o tera, peta, lub zetabajtach pojemności [31]. Aby zakwalifikować obszerną bazę danych jako *big data*, przechowywane w niej dane muszą być zróżnicowane, a ich przetwarzanie możliwie najszybsze (reguła 3V, *Volume - Variety - Velocity* [32])

4.3.1 Chemiczne i farmaceutyczne bazy danych

W ciągu ostatniej dekady nastąpił znaczący wzrost ilości dostępnych danych molekularno-strukturalnych, w tym danych dotyczących aktywności związków. Jednym z powodów jest wprowadzanie nowych technik eksperymentalnych takich jak HTE czy synteza równoległa, umożliwiających dostarczanie większej ilości danych, w krótszym czasie. Za wzrost odpowiedzialny jest również rozwój narzędzi służących do eksploracji tekstu (ang. *text-mining*). Przykładem może być praca Igora V Tetko. et al [33]. Omówiono w niej problem dostępności danych dotyczących punktu topnienia, koniecznych do wyznaczenia modelu predykcyjnego. Dotychczas dostępne bazy zawierały zaledwie 50 tysięcy rekordów, podczas gdy ilość danych generowanych i opisywanych w patentach była znacząco większa. Ostatecznie, po skonstruowaniu i zastosowaniu przez zespół dedykowanego narzędzia do

text-miningu, możliwe było uzyskanie bazy zawierającej niemal 300,000 rekordów i wyznaczenie na jej podstawie modeli o wyższej dokładności predykcyjnej.

W obszarze chemii, analizy przeprowadzane z wykorzystaniem *big data* znalazły zastosowanie przede wszystkim w projektowaniu leków. Jest to skomplikowany, długotrwały i kosztowny proces, gdzie wypuszczenie na rynek nowego leku zajmuje około 10 lat i pochłania średnio 2 miliardy dolarów [34]. Wbrew rozwijającym się technikom eksperymentalnym i ogólnemu postępowi technologicznemu, wiek leku ciągle wzrasta, a nowo powstałe leki nie są w stanie w pełni zastąpić tych, z wygasającymi patentami [35].

Zgodnie z definicją F. Browna, zwiększenie ilości danych skutkuje przyspieszeniem i większą precyzją projektowania leków. Stąd też wynika konieczność stosowania możliwie największych baz danych. Przyspieszenie procesu w tym aspekcie polega przede wszystkim na wstępnym opracowaniu danych biochemicznych, w celu identyfikacji celu i odkrycia związku wiodącego ograniczając jednocześnie ilość badań wstępnych przeprowadzanych eksperymentalnie w laboratorium, w tym syntezy nieudanych związków.

Głównym powodem szerokiego zastosowania *big data* w projektowaniu leków jest konieczność optymalizacji procesu pod kątem redukcji kosztów z nim związanych. Z ekonomicznego i biznesowego punktu widzenia, projektowanie leków jest częścią dwóch najbardziej rozwiniętych sektorów, medycyny oraz służby zdrowia [34]. Podlegają one ciągłej kontroli z wykorzystaniem tzw. *Business Intelligence* (BI), czyli metod umożliwiających podejmowanie decyzji biznesowych na podstawie analizy wyników firmy w celu zapewnienia jej konkurencyjności na rynku. Przykładem narzędzia umożliwiającego analizę BI jest rozwiązanie chmurowe *iServer* projektu OrbusSoftware. Jego główną funkcjonalnością jest utworzenie portfolio aplikacji i oprogramowania posiadanych przez przedsiębiorstwo, wraz z danymi dotyczącymi ich wykorzystania, kosztu czy ilości licencji. Następnie zebrane dane wykorzystuje się do przygotowywania raportów, na podstawie których zarząd oraz menadżerowie podejmują decyzje o przedłużeniu, wygaszaniu lub zakupie licencji. Tym samym, redukuje się koszty związane ze zbędnym oprogramowaniem oraz inwestuje w obszary, w których posiadane zasoby nie jest wystarczające.

Big data oprócz szansy optymalizacji procesów, niesie za sobą również wyzwania związane z jej gromadzeniem oraz zarządzaniem. Uporządkowane, specjalistyczne bazy zawierające wyniki uzyskane eksperymentalne najczęściej zawierają małą liczbę rekordów i nie są dostępne publicznie podczas gdy bazy o wolnym dostępie zawierają dużą ilość danych o niskiej jakości (duplikaty, niepełne dane).

4.3.1.1 PubChem

Baza danych PubChem (<https://pubchem.ncbi.nlm.nih.gov/>) jest jedną z największych publicznie dostępnych chemicznych baz danych, gromadzącą 115 milionów związków, 304 miliony rekordów dotyczących bioaktywności pochodzących z ponad 900 źródeł (pełne statystyki dostępne: <https://pubchem.ncbi.nlm.nih.gov/docs/statistics>). Jest to baza typu *open cooperative*, umożliwiająca przesyłanie danych laboratorium akademickim, agencjom rządowym, producentom chemicznym i farmaceutycznym jak również darmowe ich pobieranie bez konieczności zakupu licencji. W szczytowych okresach użytkowania, PubChem miesięcznie odwiedzało niemal 3 miliony unikatowych użytkowników (2018 r., [36]).

Baza podzielona jest na 3 podstruktury, bazę zawierającą dane dotyczące substancji (*Substance*), dane strukturalne związków (*Compounds*) oraz aktywność biologiczną (*BioAssay*). Bazy są ze sobą powiązane, dzięki czemu możliwe jest płynne nawigowanie pomiędzy poszczególnymi sekcjami. Ponadto, użytkownik może skorzystać z rozbudowanego interfejsu wyszukiwania, który obsługuje złożone zapytania tekstowe oraz strukturalne.

Baza PubChem jest zarządzana przez NCBI (*National Library of Medicine*), który dąży do rozwijania bazy zgodnie z rosnącymi możliwościami technologicznymi i wymaganiami użytkowników. W ostatniej dekadzie zaobserwowano gwałtowny wzrost ilości opublikowanych rekordów:

Tabela 1 Porównanie ilości dostępnych rekordów podstruktur PubChem w latach 2013 oraz 2023

Rok / Ilość rekordów [M]	PubChem Substance	PubChem Compound	PubChem BioAssay
2013 r. [37]	119 000 000	47 000 000	600 000
2023 r. [PubChem]	306 000 000	114 000 000	1 550 000

Oprócz systematycznej aktualizacji i zwiększania ilości dostępnych danych, PubChem dąży również do uzyskania większego stopienia ich zróżnicowania i komplementarności, wprowadzając m.in rozszerzone dane spektralne (13C NMR, 1H NMR, 2D NMR, ATR-IR, FT-IR, MS, GC-MS, Raman, UV-Vis, IR w fazie gazowej dzięki współpracy z SpectraBase, <http://spectrabase.com>), rekordy dotyczące nowych grup związków takich jak pestycydy (dzięki współpracy m.in z Biurem Programów Pestycydowych EPA, <http://www.ipmcenters.org/ECOTOX/index.cfm>) czy dodatków do żywności (współpraca z FDA Center for Food Safety and Applied Nutrition, <https://www.fda.gov/Food/>, EU Food Improvement Agents https://ec.europa.eu/food/safety/food_improvement_agents/, oraz WHO Food Additive Evaluations, <http://apps.who.int/food-additives-contaminants-jecfa-database/>).

4.3.1.2 ChEMBL

ChEMBL (<https://www.ebi.ac.uk/chembl/>) to baza danych typu *open*, gromadząca informacje z zakresu chemii medycznej w całym procesie badań i rozwoju farmaceutycznego [38]. W swoich zasobach posiada niemal 2.5 miliona związków, w tym ponad 14 tysięcy leków. ChEMBL wykorzystuje narzędzia do eksploracji tekstu, do wydobywania danych z publikacji naukowych na temat kandydatów na lek oraz zatwierdzonych leków w tym ich mechanizmy działania i wskazania terapeutyczne (m.in z takich czasopism jak *Journal of Medicinal Chemistry*, *Bioorganic & Medicinal Chemistry Letters* czy *Journal of Natural Products*).

Dane dotyczące aktywności biologicznej są wymieniane z innymi bazami danych, takimi jak PubChem BioAssay i BindingDB, dzięki czemu użytkownicy otrzymują możliwie

najpełniejsze wyniki dla swoich wyszukiwań. ChEMBL ma szeroki zakres praktycznych zastosowań, w tym identyfikację narzędzi chemicznych dla cząsteczki celu, ocenę selektywności związku, szkolenie modeli uczenia maszynowego, pomoc w generowaniu hipotez dotyczących zmiany przeznaczenia leku, ocenę podatności celu i integrację z innymi zasobami odkrywania leków [38].

4.3.1.3 Reaxys

Baza Reaxys powstała w wyniku połączenia bazy Beilstein, Gmelin oraz Patent Chemistry Database (2009 r.). Obecnie jej zarządcą jest Elsevier, a dostęp wymaga wykupienia licencji. Baza w głównej mierze dedykowana jest chemii organicznej, nieorganicznej i metaloorganicznej w tym dane dotyczące produktów naturalnych i ich pochodnych. Obecnie posiada w repozytorium informacje o niemal 180 milionach substancji oraz 57 milionach reakcji.

Niewątpliwą zaletą bazy Reaxys jest jej najbardziej rozbudowany na rynku, ale nadal intuicyjny interfejs. Dzięki niemu możliwe jest przeszukiwanie bazy nie tylko pod kątem interesującej cząsteczki, ale również reakcji lub jej części, połączenie wyszukiwania tekstowego i strukturalnego, czy wartości wybranej właściwości. Ponadto, dane wejściowe są weryfikowane przez specjalistów z obszaru chemii, w celu utrzymania możliwie najwyższej jakości zbioru.

Baza Reaxys umożliwia również planowanie retrosyntezy, budowę modelu SAR (ang. *Structure Activity Relationship*), skorzystanie z map cieplnych ukazujących relacje między substancjami aktywnymi biologicznie i ich celami (ang. *targets*), a także eksport wyników wyszukiwań w wybranym formacie.

4.3.1.4 DrugBank

Bio-chemoinformatyczna baza DrugBank (DB, www.drugbank.ca) zawiera kompleksowe informacje molekularne o lekach (zatwierdzonych przez FDA oraz lekach eksperymentalnych), ich mechanizmach działania, interakcjach oraz celach. W ciągu 10 lat

od założenia, całkowita liczba leków dostępnych w bazie wzrosła o prawie 300%, liczba interakcji lek-lek o niemal 600%, a liczba efektów leków związanych z SNP wzrosła o ponad 3000% [39]. Oprócz gwałtownego wzrostu, baza jest ciągle poddawana ulepszeniom pod kątem zapewnienia wysokiej jakości przechowywanych danych oraz łatwości przeszukiwania repozytorium.

Tabela 2 Zestawienie ilości dostępnych przykładowych grup danych w kolejnych wersjach bazy DrugBank [39]

Kategoria	DB 1.0	DB 2.0	DB 3.0	DB 4.0	DB 5.0
Ilość pól danych na karcie charakterystycznej	88	108	148	208	215
Ilość dostępnych metod wyszukiwania	8	12	16	18	20
Ilość opisów interakcji lek - lek	0	13 242	13 795	14 150	365 984
Ilość opisów interakcji lek - żywność	0	714	1039	1180	1195
Ilość leków z parametryzacją ADMET	0	276	890	6667	6700
Ilość leków z opisem patentu/ceny/producenta	0	0	1208	1450	1820

DrugBank udostępniając publicznie zróżnicowane rodzaje danych, często dostępnych wyłącznie w płatnych bazach jak np. przewidywane lub eksperymentalne metabolity leków czy ich widma MS i NMR, dąży do ułatwienia badań w zakresie farmakogenomiki, farmakoproteomiki, farmakokinetyki, farmakotranskryptomiki, farmakometabolomiki, farmakodynamiki, stanowiąc wsparcie procesu projektowania leków.

4.3.1.5 FDA approvals

Amerykańska Agencja Żywności i Leków (ang. *The Food and Drug Administration*, FDA) jest odpowiedzialna za regulację ochrony zdrowia publicznego w Stanach Zjednoczonych od 1906 r. W zakres jej obowiązków wchodzi monitorowanie bezpieczeństwa oraz skuteczności

leków, produktów biologicznych oraz urzędów medycznych. Oprócz aspektu farmakologicznego, FDA reguluje również produkcję, marketing i dystrybucję wyrobów tytoniowych, dążąc do minimalizacji ich negatywnego wpływu na społeczeństwo.

Podstawą regulacji FDA jest ustawa o czystej żywności i lekach podpisana przez Theodore'a Roosevelta w 1906 r. i stanowiącą punkt rozpoczęcia jej działalności. Do dzisiaj powstało ponad 200 uzupełniających ją ustaw, które uchwalane są zazwyczaj jako odpowiedź na udowodnienie szkodliwego działania np. kosmetyków, palenia papierosów czy stosowania w żywności barwników zawierających metale ciężkie [40].

Określenie FDA *approvals* dotyczy leków zatwierdzonych przez FDA, których korzyści dla pacjenta przewyższają znane i potencjalne ryzyko wiążące się z ich zażywaniem. Opinia o leku jest wystawiana przez CDER (ang. *The Center for Drug Evaluation Research*), jednostkę FDA dedykowaną badaniom nad udzieleniem pozwolenia na wypuszczenie leku na rynek. Proces zatwierdzania składa się z wielu etapów, wchodzących w zakres trzech głównych obszarów [40]:

1. Analiza stanu docelowego oraz dotychczas dostępnych metod leczenia - choroba, dla której jest przeznaczony lek, jest analizowana w celu ustalenia ryzyka związanego z jego zatwierdzeniem. Najszybszą ścieżkę otrzymują leki dedykowane dotychczas nieobjętym farmakologią chorobom zagrażającym życiu, ponieważ ich korzyść znacznie przewyższa ryzyko wynikające z ich zastosowania.
2. Ocena korzyści i ryzyka na podstawie danych klinicznych - producent leku jest zobowiązany do dostarczenia wyników dwóch, niezależnych badań klinicznych. Oprócz rezultatu, ocenie podlega również sposób zaprojektowania badań.
3. Analiza wymagań dotyczących zarządzania ryzykiem - każdy lek obarczony jest ryzykiem wystąpienia niepożądanych skutków ubocznych. Po ich analizie, FDA przekazuje producentowi wymagania dot. etykiety leku, w tym konieczność jasnego przekazania ryzyka związanego z jego zażywaniem.

Ścieżka akceptacji FDA może zostać skrócona w uzasadnionych przypadkach, m.in dla leku targetującego dotychczas nieleczoną chorobę czy posiadającego znacznie wyższe korzyści w porównaniu do dotychczas stosowanej farmakologii.

Baza zawierające wszystkie dotychczasowe FDA approvals jest udostępniona publicznie, na rządowej stronie FDA (<https://www.fda.gov/>).

4.3.1.6 ZINC

Baza ZINC udostępnia komercyjnie dostępne związki, z myślą o procesie *virtual screeningu*. Najnowsza wersja bazy (ZINC-22) posiada około 37 miliardów cząsteczek w 2D oraz 4,5 miliarda cząsteczek w 3D. Obszerne repozytorium podzielone zostało na 4 sekcje w celu ułatwienia przeszukiwania: *Substances* z rozbudowaną wyszukiwarką strukturalną oraz tekstową, *Catalogs* stanowiącą zgrupowane zbiory cząsteczek, pod kątem wybranego parametru np. związki biogeniczne, *Tranches* stanowiącą mapę repozytorium, z której użytkownik może wybierać grupy związków o preferowanej wartości MW, HAC i/lub logP oraz *Biological*, stanowiąca zbiór danych biologicznych, dotyczących genów, ludzkich metabolitów czy aktywności.

Baza ZINC jest nieustannie rozbudowywana, by sprostać wyzwaniom związanym ze stale rozszerzającą się przestrzenią chemiczną. Jednym z nich jest opracowywanie nowych narzędzi do *virtual screeningu*, które będą w stanie przetwarzać rosnące wolumeny danych, z zachowaniem odpowiedniej wydajności. Przykładem takiego narzędzia jest *CartBlanche*, interfejs do najnowszej wersji bazy, umożliwiający wyszukiwanie cząsteczek na podstawie podobieństwa, liczby atomów ciężkich, lipofilowości czy ładunku molekularnego, z wykorzystaniem interfejsu strukturalnego generującego w czasie rzeczywistym kod SMILES oraz wyniki w czasie relatywnie krótkim wobec rozmiaru bazy (dla przykładu, zapytanie o struktury analogiczne do benzenu dało 3278 wyników w 5,5 sekundy).

ZINC jest dostępny bezpłatnie pod adresem <https://zinc.docking.org> natomiast dostęp do CartBlanche - <https://cartblanche22.docking.org>.

4.3.1.7 PharmaCompass

PharmaCompass (<https://www.pharmacompass.com/>) to darmowa platforma internetowa, udostępniająca firmom farmaceutycznym dane dotyczące leków w tym ich cen, producentów, dostawców oraz patentów. Wspiera zespoły sprzedażowe, umożliwiając im rozszerzanie grona potencjalnych klientów, śledzenie trendów cenowych na rynku (w szczególności u producenta konkurencyjnego leku), ułatwiając tym samym podejmowanie współpracy, decyzji biznesowych, a także zwiększając widoczność firm na arenie międzynarodowej.

Dane przechowywane przez PharmaCompass pochodzą z integracji z bazami FDA, CEP (Certyfikaty Zgodności Europejskiej Dyrekcji ds. Jakości Leków i Opieki Zdrowotnej) oraz *DailyMed* (baza leków Narodowej Biblioteki Medycznej Stanów Zjednoczonych), a także z bezpośredniej współpracy z firmami farmaceutycznymi. Co roku PharmaCompass przedstawia zestawienie najlepiej sprzedających się leków, zarówno w wersji interaktywnego panelu będącego integralną częścią platformy, jak również w wersji do pobrania (2022 r. <https://www.pharmacompass.com/data-compilation/top-drugs-by-sales-in-2022-who-sold-the-blockbuster-drugs>).

4.3.2 Materiałowe bazy danych

Materiałowe bazy danych analogicznie do baz leków, zawierają dane eksperymentalne oraz obliczeniowe w tym właściwości czy skład chemiczny oraz wielkości charakterystyczne dla tej grupy takie jak informacje o łączeniu materiałów (np. spawaniu), korozji, krzywe naprężenia i odkształcenia czy plastyczność. Ilość dostępnych baz materiałów jest znacząco niższa w porównaniu do baz danych dotyczących leków, a do większości z nich konieczne jest wykupienie licencji.

Najbardziej kompleksową bazą materiałów jest Total Materia, założona w 1999 r. w Holandii. (<https://www.totalmateria.com/page.aspx?ID=Home&LN=PL>). Zawiera dane dla ponad 450 000 materiałów, w tym metali, polimerów, ceramiki oraz kompozytów. Umożliwia analizę pod kątem fizykochemicznym, identyfikację nieznanego materiału,

porównywanie właściwości, a także dostarcza informację o dostawcach wspomagając nawiązywanie relacji biznesowych.

Jedną z najbardziej popularnych i rozbudowanych baz o częściowo darmowym dostępie jest UL Prospector (<https://www.ulprospector.com/en/eu>). Repozytorium podzielone jest na kategorie odpowiadające obszarom przemysłu w tym między innymi kleje i uszczelniacze, artykuły spożywcze, środki czystości, kosmetyki, plastiki, metale czy farby. Dla każdego materiału udostępniane są dane o jego budowie chemicznej, nazwa INCI, krótki opis działania i zastosowania, ale także historię materiału i pełne arkusze specyfikacji. Ponadto, platforma wspiera dostawców, umieszczając na kartach materiałów krótki opis producenta, możliwość zamówienia próbki czy skontaktowania się z działem sprzedaży.

Na szczególną uwagę zasługuje The Materials Project (<https://materialsproject.org/>), interaktywna witryna poświęcona gromadzeniu właściwości dotyczących materiałów nieorganicznych. Dane udostępniane są bezpłatnie wraz z autorskimi aplikacjami webowymi takimi jak np. *Phase Diagram* służący do generowania diagramów fazowych na podstawie obliczeń DFT, *Synthesis Explorer* umożliwiający wyszukiwanie metod syntezy w literaturze na podstawie zapytań strukturalnych bądź tekstowych czy *Battery Explorer*, dzięki któremu możliwe jest przeanalizowanie typowanego materiału pod kątem zastosowania w bateriach, z predykcją uzyskanego napięcia i utlenienia. W celu wygenerowania wyników wykorzystywane są autorskie modele predykcyjne, wyznaczone na podstawie danych literaturowych.

W celu zwiększenia stopnia wykorzystania narzędzi chemoinformatycznych do projektowania fotoreagentów, czemu między innymi poświęcona jest ta rozprawa doktorska, konieczne jest zwiększenie dostępności i ustrukturyzowania danych literaturowych z obszaru materiałów, co z kolei stanowi jeden z kierunków badań własnych tej pracy.

4.4 Chemoinformatyka jako narzędzie optymalizacyjne

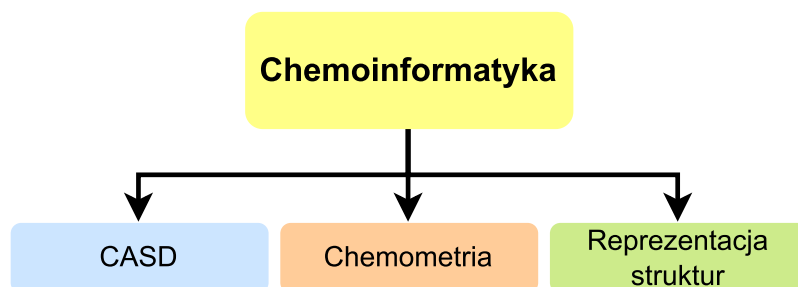
Chemoinformatyka łączy chemię, fizykę, biologię, matematykę z statystyką i informatyką tworząc “nowe rozwiązania dla starych problemów” [41]. Głównym pierwotnym zastosowaniem chemoinformatyki był proces projektowania leków, w celu znaczącej redukcji kosztów z nim związanych. Z czasem jednak zaczęto podejmować próby rozszerzenia go na inne obszary takie jak np. agrochemia czy chemia żywności. Pośrednio, chemoinformatyka wpływa również na takie obszary jak ochrona środowiska, spektroskopia, toksykologia, przepisy dotyczące leków czy kosmetyków.

U podstaw chemoinformatyki leżą bazy danych. Ich rosnące wolumeny wymagają ciągłego opracowywania udoskonalonych metod przechowywania, ekstrakcji oraz przetwarzania tak, by modelowane zależności wykazywały wciąż wysoką precyzję. Procesowanie danych w chemoinformatyce można podzielić na [42]:

1. Pozyskiwanie informacji w drodze generowania i gromadzenia danych pochodzących z eksperymentów *in vivo* lub z symulacji molekularnych (eksperymentów *in silico*).
2. Przechowywanie i zarządzanie posiadanymi informacjami, w zakres którego wchodzi tworzenie i utrzymywanie baz, w tym w szczególności spełniających regułę *big data: volume - variety - velocity*.
3. Wykorzystanie dostępnych informacji do analizy, statystyki, badania korelacji, do rozwiązywania problemów chemicznych i biochemicznych.

Trzema filarami chemoinformatyki jest projektowanie syntez wspomaganých komputerowo, chemometria oraz reprezentacja komputerowa struktur chemicznych [42, 9]. Projektowanie leków i materiałów to inna ważna rola chemoinformatyki.

Rys. 13 Trzy filary chemoinformatyki



Z punktu widzenia tej rozprawy doktorskiej, szczególne znaczenie ma ostatni z nich. Bez możliwości przedstawienia struktur chemicznych w postaci czytelnej dla komputera, niemożliwe byłoby utworzenie chemicznych baz danych oraz ich *skryning*, projektowanie molekularne, obliczenia deskryptorów, a także wyznaczanie zależności struktura - aktywność (QSAR).

Tabela 3 Przykładowe sposoby reprezentacji komputerowej struktury chemicznej kofeiny

Reprezentacja	Nazwa
kofeina	nazwa zwyczajowa
teina, guaranina, mateina	synonimy
$C_8H_{10}N_4O_2$	wzór sumaryczny
1,3,7-Trimethylpurine-2,6-dione	nazwa IUPAC
58-08-2	numer identyfikacyjny CAS
<chem>CN1C=NC2=C1C(=O)N(C(=O)N2C)C</chem>	SMILES
1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	InChI
0000100110100111	<i>fingerprint</i>

5244987098423150

kod *hash*

	1	2	3	4	5	6	7	8	9	10	11
1	0	1	0	0	0	0	0	0	0	2	0
2	1	0	2	0	0	0	0	0	0	0	0
3	0	2	0	1	0	0	0	1	0	0	0
4	0	0	1	0	1	0	0	0	0	0	0
5	0	0	0	1	0	2	1	0	0	0	0
6	0	0	0	0	2	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	2	0	0
9	0	0	0	0	0	0	0	2	0	1	0
10	2	0	0	0	0	0	0	0	1	0	1
11	0	0	0	0	0	0	0	0	0	1	0

tabela połączeń

Wzrastająca ilość oprogramowania chemicznego w połączeniu z szeroko udostępnionymi i aktualizowanymi bibliotekami danych chemicznych sprawia, że prawie każde laboratorium przemysłowe i większość laboratoriów akademickich sięga po metody chemoinformatyczne m.in do [21]:

1. Prognozowanie biologicznych, fizycznych i chemicznych właściwości związków.
2. Przeszukiwania baz danych pod kątem wybranej struktury, podstruktury, podobieństwa lub właściwości pożądaných.
3. Opisu i wyjaśnienia struktury chemicznej, w wyniku analizy danych spektroskopowych.
4. Integracji i nadzorowania układów do HTS, umożliwiających szybkie przesiewanie związków chemicznych w poszukiwaniu pożądaných właściwości, np. aktywności biologicznej.
5. Dokowania molekularnego (projektowania molekularnego), w celu poszukiwania minimum globalnego pola siłowego wyznaczanego dla kompleksu ligand-receptor.

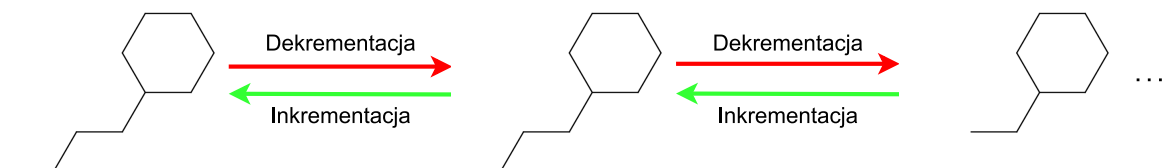
4.4.1 Fragonomika

Fragonomika jest jedną z nowszych metod chemoinformatycznego projektowania leków. W ogólnym ujęciu sprowadza się do analizy przydatności niewielkich molekuł nazywanych fragmentami. Fragmenty takie stanowią mogą elementy podstruktury większych cząsteczek leków. Stanowią więc mogą punkt wyjścia do generowania *in silico* nowych, potencjalnie aktywnych związków, zwanych hitami (ang. *hits*), wykazujących w badaniach przesiewowych pożądaną aktywność, potwierdzoną ponownymi testami [43].

Następnie można wykorzystać je do skringu wirtualnych baz danych. W tym celu wykorzystuje się niewielki fragment cząsteczki o znanej i pożądanej aktywności do wyszukiwania w możliwie najobszerniejszym i najbardziej zróżnicowanym zbiorze cząsteczek, zawierających zadany fragment. Wytypowane cząsteczki poddaje się następnie analizie w celu potwierdzenia lub zaprzeczenia posiadania przez nie pożądanej aktywności biologicznej, przy wykorzystaniu metod eksperymentalnych.

Metody tworzenia fragmentów można podzielić na fragmentację dekrementacyjną i inkrementacyjną. Pierwsza z nich polega na zmniejszaniu wybranej cząsteczki o jeden atom, a przy każdym kroku tworzony jest nowy fragment zawierający $x_n - 1$ atomów. Wyjątek stanowią układy cykliczne, które odłącza się w całości bez rozrywania tworzących je wiązań [37]. Fragmentacja inkrementacyjna jest metodą odwrotną, polega na dodawaniu do fragmentu wyjściowego wybranego atomu.

Rys. 14 Schemat tworzenia fragmentów metodą dekrementacyjną i inkrementacyjną



Ciekawe wyniki uzyskane w wyniku zastosowania fragonomiki podczas projektowania leków spowodowały wzrost zainteresowania użyciem tej metody w celu projektowania innych grup związków np. katalizatorów, fotoreagentów czy związków OLED [14-15, 44].

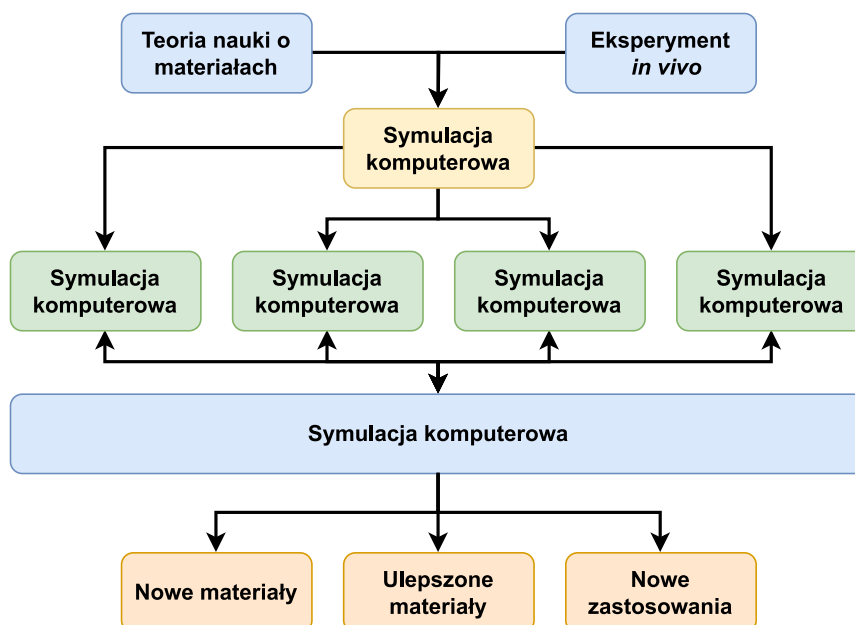
4.4.2 Projektowanie materiałów

Od lat liderem i tym samym źródłem inspiracji postępu technologicznego był przemysł zbrojeniowy. Obecnie podobną pozycję zajmuje tzw. elektronika konsumencka, ze szczególnym udziałem smartfonów oraz tabletów. Rosnące wymagania zarówno konsumenckie i technologiczne powodują konieczność wprowadzenia nowych metod projektowania materiałów, w tym metod inspirowanych innymi gałęziami przemysłu, głównie farmaceutycznego.

Pierwszym etapem projektowania nowej generacji materiałów jest wyznaczenie celu optymalizacji dla dotychczas stosowanych, którym mogą być ich właściwości fizykochemiczne, metody syntezy czy zjawiska (w tym defekty i uszkodzenia, dyfuzja, zjawiska skali i zjawiska zależne od wielkości, rozwój mikrostruktury).

Zastosowanie nowego materiału jest sprawdzane od skali atomowej/nanostrukturalnej, przez mikrostrukturę, aż do skali makroskopowej, przy użyciu technik analitycznych i modelowania komputerowego. Obie metody wzajemnie ze sobą oddziałują, prowadząc do skrócenia czasu i kosztów związanych z projektowaniem materiałów [45].

Rys. 15 Schemat ścieżki projektowania materiałów



5. BADANIA WŁASNE

5.1 *Ligand efficiency* jako szczególny przypadek fragonomiki leków¹

5.1.1 Wstęp teoretyczny

Jednym z zadań projektowania leków jest opracowywanie wysokowydajnościowych narzędzi do wyodrębniania cząsteczek potencjalnych hitów, leadów oraz “kandydatów na lek” (*drug candidates*). W tym celu konstruuje się *big data* zawierające zróżnicowane dane, w tym deskryptory molekularne oraz właściwości fizykochemiczne, by następnie badać zależność pomiędzy nimi (analiza Quantitative Structure–Activity Relationship, QSAR) [34, 46].

W wyniku analizy QSAR leków drobnocząsteczkowych ich podstawowe właściwości takie jak lipofilowość, kształt, właściwości wiązań wodorowych i polarność zostały skorelowane - w różnym stopniu - z rozpuszczalnością [48], przepuszczalnością błon komórkowych [49], toksycznością *in vivo* [50, 51], interakcjami z receptorami, stabilnością metaboliczną [52, 53] i ścieraniem [54, 55] (parametrami określającymi lekopodobność). Lipofilowość oraz liczba donorów wiązań wodorowych zostały wskazane jako kluczowe, ze względu na ich stosunkowo stałą wartość dla cząsteczek leków [56-59]. Ogólnie, projektowanie leków można określić jako optymalizację siły wiązania cząsteczki z receptorem w funkcji jej struktury chemicznej.

Jednym ze sposobów analizy problemu jest eksploracja wpływu MW na moc wiązania [45, 84]. Szanse dopasowania małej, prostej cząsteczki do centrum aktywnego są znacznie większe niż szanse dopasowania do niego cząsteczki o bardziej skomplikowanej budowie. Z drugiej strony jednak, moc wiązania rozbudowanej cząsteczki jest większa, stąd też podwójne znaczenie MW w analizie potencjalnych leków (*drug candidates*) [60].

Koncepcja *ligand efficiency* (LE) powstała w wyniku obserwacji maksymalnej mocy wiązania uzyskiwanej przez ligandy (-1.5 kcal/M/atom niewodorowy, z pominięciem prostych kationów i anionów [61]) połączonej z badaniami nad mocą wiązań poszczególnych

¹ Wyniki omawiane w zakresie Ligand Efficiency zostały przygotowane we współpracy z dr Roksaną Duszkiewicz podczas prowadzenia wspólnych prac badawczych, w toku przygotowywania jej rozprawy doktorskiej.

grup funkcyjnych. Z czasem definicja została uproszczona do badania „stosunku energii wiązania do HAC” [62-64].

LE zostało zaprojektowane jako metoda porównywania cząsteczek, a początkowy koncept rozwinęto o dodatkowe właściwości takie jak lipofilowość [65], masa cząsteczkowa [66], powierzchnia polarna [66], udział grup funkcyjnych [67] i kombinacje właściwości fizykochemicznych [68, 69]. Parametr LE odnosi się wyłącznie do miary siły wiązania wyznaczanych metodą obliczeniową.

Równanie LE opisuje rozkład energii swobodnej Gibbsa na wszystkie atomy cząsteczki z pominięciem wodoru zgodnie z poniższym:

$$LE = \Delta G^0/HAC$$

gdzie:

ΔG^0 – energia wolnego wiązania Gibbsa

HAC – ang. *Heavy Atom Count*, liczba atomów ciężkich, wszystkie atomy cząsteczki z pominięciem atomów wodoru. Należy zauważyć, że w *LE* jest brana pod uwagę tylko ilość atomów, z pominięciem różniących ich właściwości. Parametr nie uwzględnia również możliwości braku udziału danego atomu w wiązanie ligand – centrum aktywne.

Energię swobodną Gibbsa określa równanie:

$$\Delta G^0 = -RT \times \ln(K_d/C^0) = -2.303RT \times \log(K_d/C^0)$$

gdzie:

R = stała gazowa = 1.987×10^{-3} [kcal/K/mol]

T = temperatura w Kelwinach

C^0 = stężenie w warunkach standardowych

K_d = stała dysocjacji

Przyjmując warunki standardowe dla roztworu wodnego w $T = 300\text{K}$, neutralnym pH i przy zachowaniu stężenia w jednostkach molowych otrzymujemy:

$$LE = \Delta G^0/HAC = -(2.303RT/HAC) \times \log(K_d/C^0) = - (1.37/HAC) \times \log(K_d) = \\ 1.37/HAC \times pK_d$$

Zwykle pK_d jest zastępowane przez pAC (logarytm stężenia substancji czynnej, wywołujący określony rezultat, logarytm aktywności) najczęściej pIC_{50} (logarytm stężenia wywołującego połowę maksymalnej inhibicji, miara zdolności inhibicyjnej danej substancji), co prowadzi do równania:

$$LE = (1.37/HAC) \times pIC_{50}$$

lub:

$$LE = (1.37 \times pIC_{50}) / HAC$$

W kolejnej transformacji uwzględnione zostały dwa czynniki. Pierwszym z nich jest wymiar energii wiązania Gibbsa [kcal/mol]. Drugim tożsamość masy cząsteczkowej cząsteczki, MW [Da] (która jest jednostką masy w kg) i masy mola substancji [kg/mol], które wobec tego można zapisać tą obserwacją jako $(MW[Da]/MW[kg/mol]) = 1$.

Po wprowadzeniu do równania otrzymujemy:

$$LE = \Delta G^0/HAC * (MW[Da]/MW[kg/mol]) = (\Delta G^0/MW[kg/mol]) * MW[Da]/HAC$$

Zgodnie z powyższym równaniem, ligand efficiency przyjmuje jednostkę [kcal/kg] * MW[Da]/HAC. Zaskakująco wskazuje to na fizyczne znaczenie LE wyrażone jako energia wiązania kilograma (grama) ligandów przeskalowana do HAC przez Da/HAC. Jak wspomniano wcześniej, LE zostało zaprojektowane jako deskryptor molekularny (obliczony na podstawie reprezentacji molekularnej), w odniesieniu do jednej cząsteczki. Dwoistość „charakteru” LE wyjaśnia wiele paradoksów związanych z zastosowaniem tego parametru w projektowaniu leków, co omówiono w kilku ostatnich publikacjach [70-77].

Niepewność budzi również matematyczna poprawność *ligand efficiency*. W przypadku porównywania dwóch cząsteczek o znacznej różnicy HAC, parametr ten dominuje nad wyznaczonym LE, zupełnie eliminując z równania aktywność liganda. Po drugie,

porównując dwa ligandy o tej samej liczbie HAC, parametr ten można pominąć, gdyż nie uwzględniając różnic pomiędzy analizowanymi atomami ciężkimi, HAC nie wnosi do równania wartości informacyjnej [78].

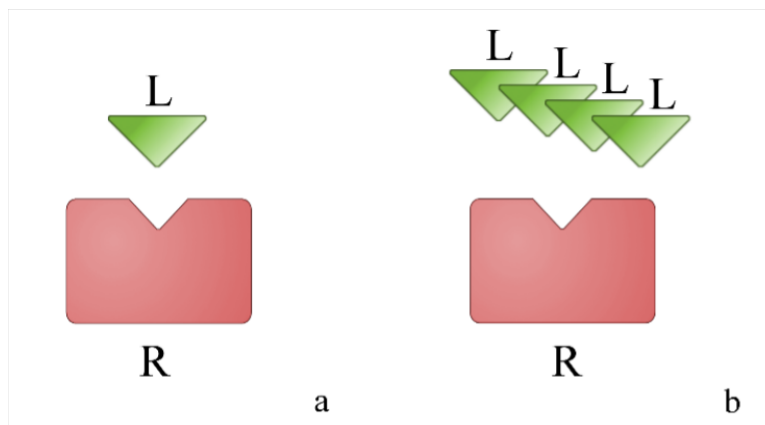
Pomimo kontrowersji, *LE* jest jednym z najczęściej stosowanych parametrów walidacyjnych potencjalnych kandydatów na lek [105 – 108]. Wyjaśnić to można poprzez preferencję *LE* względem małych ligandów, która doskonale odzwierciedla koncepcję tzw. *slim pharmacy*, pozornie potwierdzając poprawność *ligand efficiency*. Drugą przyczyną może być wytyczanie dzięki *LE* interesujących ścieżek optymalizacyjnych *drug candidates*, które *de facto* analizują zmienność *LE* jednego liganda, z uwzględnieniem zmian jego HAC [75, 79-81].

W celu zrozumienia obu, skrajnych opinii o *ligand efficiency*, konieczna jest szczegółowa analiza jej „podwójnej natury”.

Parametr *LE* silnie zależy od wielkości cząsteczki. Ponieważ zależność ta była znacznie wyższa niż spodziewana, stanowiła długo zagadkę [72, 82], którą próbowano rozwiązać modyfikując *LE* formą lipofilową (ang. *Lipophilic Ligand Efficiency*, *LLE*) lub poprzez wprowadzenie do równania zmiennej związanej z rozmiarem cząsteczki (ang. *Size Independent Ligand Efficiency*, *SILE*). Najbardziej precyzyjnie zaskakujący trend *LE* tłumaczy jej interpretacja chemiczna, gdzie *LE* jest miarą wiązania 1 mola HAC (analog 1 mola g) [72, 83]. Ponieważ 1 mol HAC oraz jego analog (1 mol g) istnieją jedynie wirtualnie, stąd przebieg funkcji *LE* vs. HAC wzbudzał w literaturze zdziwienie, bowiem przy HAC dążącego do 1, *LE* „dąży do nieskończoności”.

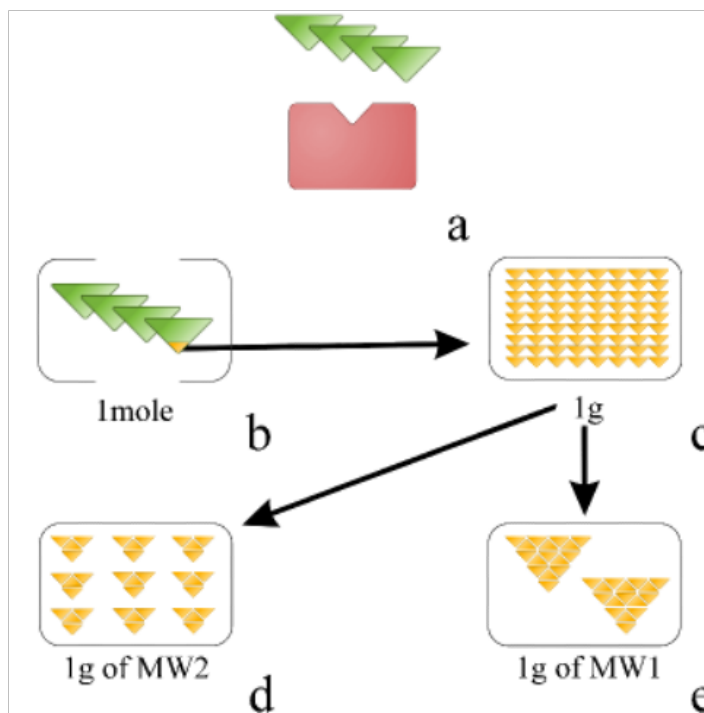
W rzeczywistości znaczenie *LE* jest bardziej złożone niż pierwotnie zakładano. *Ligand efficiency* zaprojektowane zostało jako deskryptor molekularny, tzn. parametr odnoszący się do wirtualnej molekuly, ale jednocześnie ma także znaczenie właściwości odnoszącej się do mola cząsteczek. W takim układzie iloraz IC_{50} odnoszącego się do mola substancji reprezentuje zbiór cząsteczek w mierze wagowej. Na rysunku poniżej przedstawione zostało rzeczywiste fizyczne znaczenie IC_{50} jako deskryptora oraz właściwości. Problem ten omówiliśmy szczegółowo w publikacji [84] oraz [9].

Rys. 16a - b Porównanie teoretycznego pojęcia aktywności (a) z rzeczywistym (b) [84].



Teoretyczne znaczenie aktywności chemicznej w odniesieniu do 1 HAC bądź 1 Da daje wyniki paradoksalne, ponieważ w rzeczywistości aktywność ligand – receptor nie odnosi się do jednej cząsteczki w układzie izolowanym (1 HAC, 1 Da) (17a, b) lecz jest sumą aktywności populacji ligandów z receptorem. Tym samym, 1g substancji (17c) daje różną liczbę cząsteczek (17d, e).

Rys. 17a - d Fizyczne wyjaśnienie nieściśłości w znaczeniu LE [49].



Analiza rzeczywistego znaczenia LE nasuwa na myśl analogię do fragonomiki. W obu przypadkach badany jest wpływ części układu (fragmentu) na końcowy rezultat optymalizacji drug candidates (na siłę oddziaływania ligand – centrum aktywne). Wobec nieintuicyjnego trendu LE, podjęto próbę wyznaczenia alternatywnego narzędzia, w celu zdefiniowania interakcji IC50 i HAC.

5.1.2 Metodologia

W celu analizy zmian LE w obszernym zbiorze leków, kandydatów, struktur wiodących i hitów przeprowadziłam szeroką eksplorację danych takich struktur. W celu utworzenia bazy danych wykorzystałam dwa katalogi chemiczne: PubChem oraz ChEMBL. Pobrałam 2,435,467 rekordów z katalogu PubChem (Sierpień, 2017 r.) oraz 714,791 z katalogu ChEMBL (wersja 22). W analizie ujęłam wszystkie dostępne rodzaje aktywności biologicznych:

- a) IC₅₀ – (ang. *inhibitory concentration*) – mediana stężenia inhibitora hamującego w 50% funkcje biologiczne i biochemiczne organizmów. Parametr ten stosowany jest m.in. do opisu ograniczenia wzrostu bakterii i glonów.
- b) EC₅₀ – (ang. *effective concentration*) – mediana stężenia skutecznego, stężenie obliczone statystycznie, wywołujące określony efekt u 50 % badanych organizmów, w określonych warunkach np. działanie hamujące lub stymulujące procesy fizjologiczne.
- c) CC₅₀ – (ang. *cytotoxic concentration*) – mediana stężenia cytotoksycznego, stężenie obliczone statystycznie, wywołujące efekt cytotoksyczny u 50% badanych komórek, w określonych warunkach.
- d) K_i – stała równowagi dysocjacji ligandu, uzyskana podczas badania procesu tworzenia wiązania z wykorzystaniem oznaczonego ligandu.
- e) K_d – stała równowagi dysocjacji ligandu, uzyskana podczas badania procesu inhibicji.

Następnie, zebrane dane scalałam (z przyjęciem jednego oznaczenia aktywności tj. AC_{50}) tworząc zbiór danych 1. Zbiór 1 został również poddany metodzie „binowania danych” (ang. *data binning*). Umożliwia ona uporządkowanie danych poprzez grupowanie ich według określonego kryterium (w przypadku tej pracy doktorskiej jest nim HAC), a następnie wszystkie wartości AC_{50} przynależące do danego zakresu zostają zastąpione ich medianą.

Ponadto, utworzyłam zbiór 2, zawierający wybrane dane aktywności biologicznej leków określanych mianem fragmentów oraz leków zaakceptowanych przez FDA, pobranych ze strony <https://www.fda.gov>, a brakujące dane zostały uzupełnione z wykorzystaniem innych baz wskazanych we wstępie teoretycznym.

Wykresy przygotowałam przy użyciu darmowej wersji programu MATLAB, dostępnej online na <https://www.mathworks.com/products/matlab.html>.

5.1.3 Wyniki

W praktyce najatrakcyjniejsze cząsteczki pod kątem projektowania leków wykazują niską wartość HAC przy wysokiej aktywności. Korzystne może więc być zdefiniowanie parametru badającego bardziej zrównoważone oddziaływanie AC_{50} oraz HAC. Zaproponowałam by w tym celu wykorzystać iloczyn zmiennych. Takie oddziaływanie zostało odzwierciedlone w *Product Ligand Efficiency* (PLE), zgodnie z poniższym równaniem:

$$PLE = AC_{50} \times HAC \times 1.37$$

Oba czynniki działają kooperacyjnie tzn. ich spadek powoduje zmniejszenie PLE i odwrotnie.

Powszechnie stosowaną formą matematyczną AC_{50} jest jego ujemny logarytm, pAC_{50} . Wyższe wartości pAC_{50} wskazują wykładniczo wyższą aktywność danej substancji. W odniesieniu do PLE również możliwe jest zastosowanie skali logarytmicznej:

$$PLE = AC_{50} \times HAC \times 1.37$$

$$pPLE = -\log(AC_{50} \times HAC \times 1.37)$$

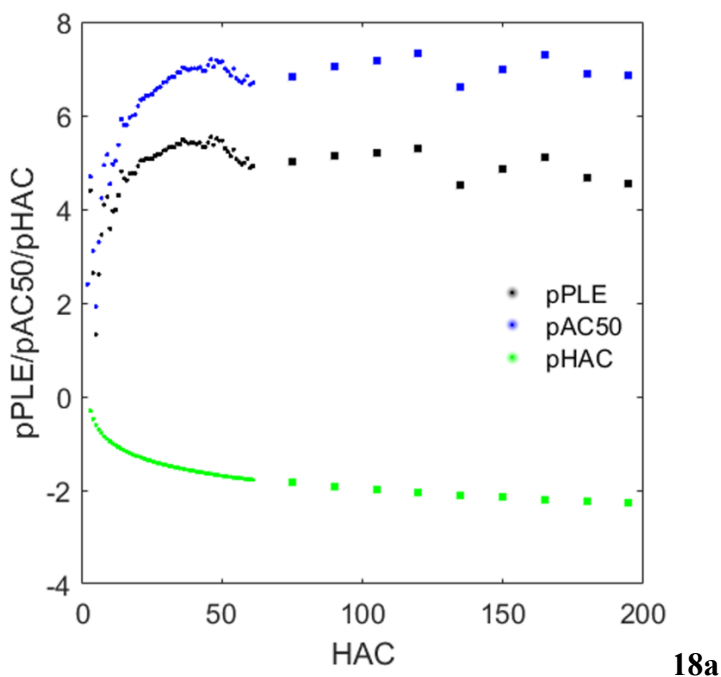
Logarytm iloczynu jest sumą logarytmów, wobec tego:

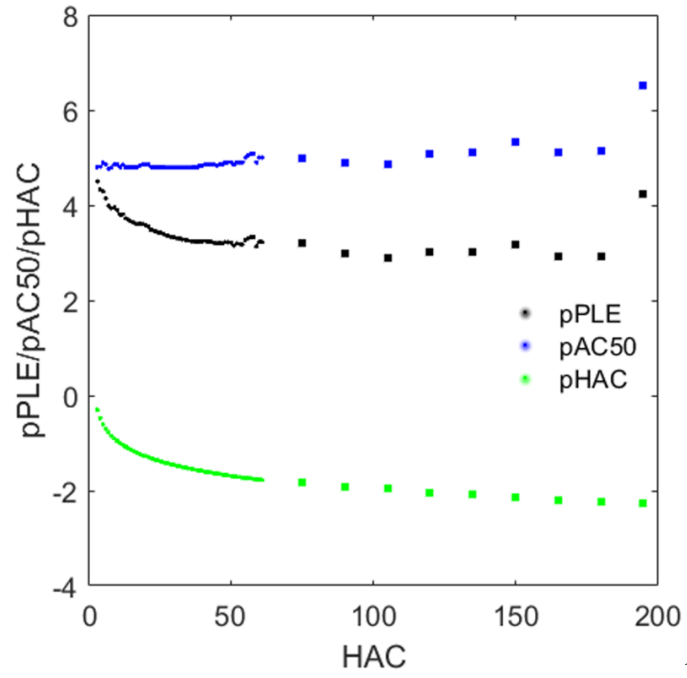
$$pPLE = pAC_{50} + pHAC$$

Poprawność opracowanego narzędzia pPLE została następnie sprawdzona przy użyciu autorskich zbiorów danych aktywności chemicznej.

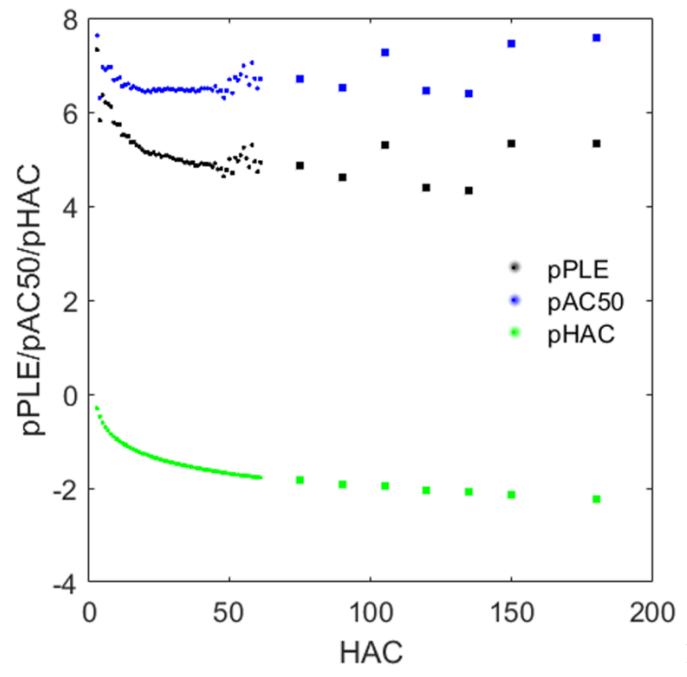
W tym celu dla 3,150,258 rekordów w utworzonej bazie danych 1, wyznaczyłam pPLE, pHAC oraz pAC50. Uzyskane wyniki poddałam binowaniu i przedstawiłam na poniższych wykresach z uwzględnieniem katalogu, z którego pochodzą:

Rys. 18a - d Wykres zależności pPLE od HAC dla katalogu ChEMBL (a), PubChem (b), subpopulacji PubChem o $pAC > 6$ (c). pPLE może zostać rozłożone na pAC₅₀ oraz pHAC (odpowiednio niebieskie i zielone punkty). W celu porównania wyników do LE, utworzyłam wykres (d) przedstawiający zależność LE od HAC z wykorzystaniem tych samych danych jak w przypadku wykresów (a-c) [60].

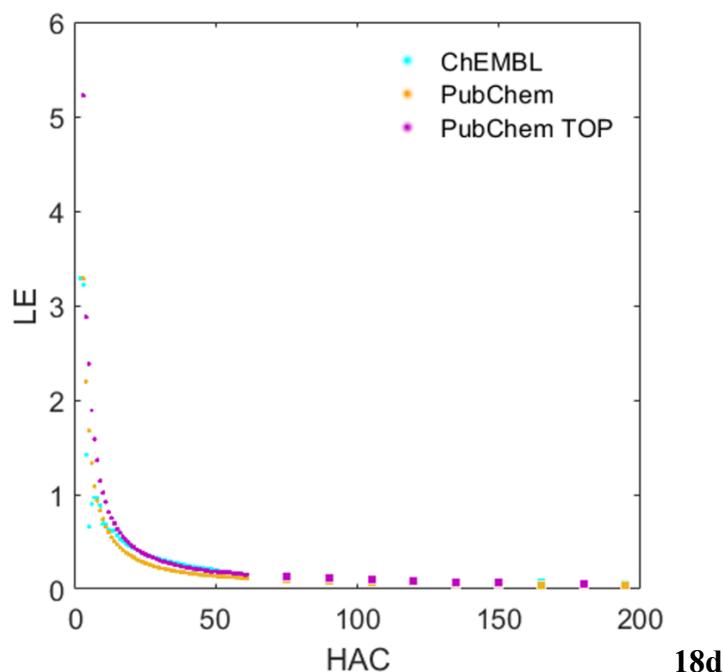




18b



18c



Wykres 18a przedstawia zależność pomiędzy pPLE, pAC_{50} , a HAC dla danych z katalogu ChEMBL. Dane wskazują na wzrost wartości pAC_{50} wraz ze wzrostem HAC do HAC równego 50.

Wykres 18b i 18c w analogiczny sposób przedstawia rozkład danych dla katalogu PubChem. Wykres 18b przedstawia zależność dla całego zbioru danych. Zgodnie z wykresem, pAC_{50} jest stałą funkcją dla HAC co wskazuje, iż aktywność nie jest funkcją wielkości (HAC) cząsteczek, jeśli badana populacja substancji jest wystarczająco duża.

Wykres 18c obejmuje wybrane cząsteczki z katalogu PubChem wykazujące najwyższą aktywność ($pAC_{50} > 6$). Wartość pAC_{50} spada wraz ze wzrostem HAC co świadczy, iż prawdopodobieństwo odpowiedniego dopasowania ligand - centrum aktywne spada wraz ze wzrostem wielkości cząsteczki.

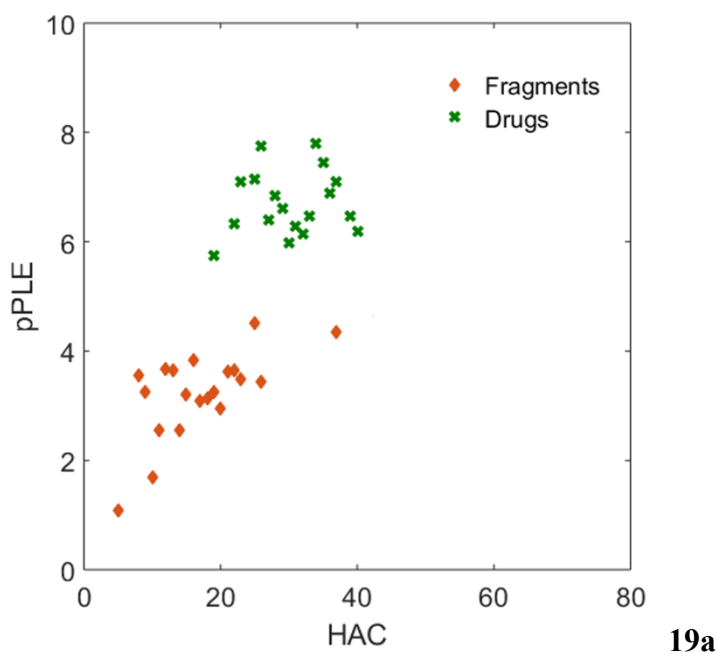
W obu przypadkach, niższe HAC wskazuje optymalny obszar potencjalnych leków, co uwidacznia się także na wykresach pPLE (czarne punkty).

W celu porównania na wykresie 18d została przedstawiona zależność pomiędzy LE a HAC. Analogicznie do poprzednich wykresów wskazuje on niskie wartości HAC jako optymalne

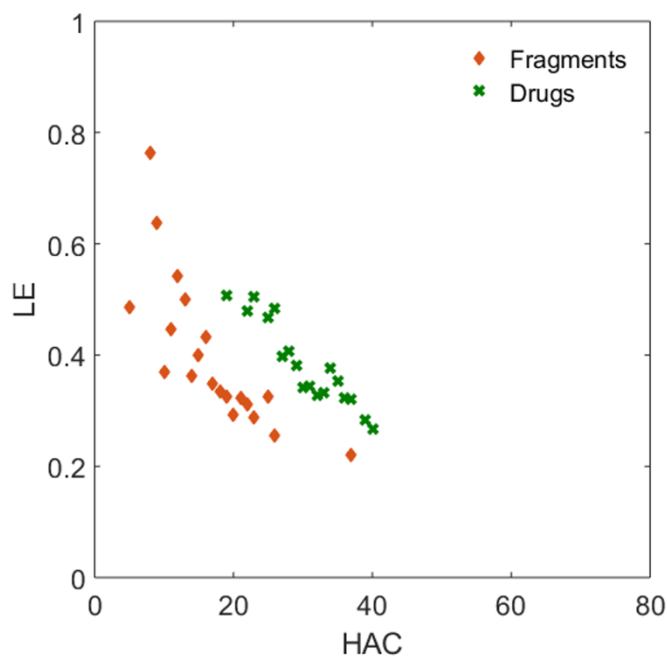
dla *drug candidates*. Pierwszym powodem jest wspomniane wcześniej fizyczne znaczenie LE, które odnosi się nie do jednej cząsteczki, lecz do całej populacji cząsteczek biorących udział w interakcji ligand – centrum aktywne. Drugi powód ma podłoże matematyczne. Wartości HAC najczęściej wahają się pomiędzy 1 a 200, podczas gdy wartości pAC_{50} mieszczą się w zakresie 1 do 7. To powoduje dominację czynnika $1/HAC$ na wykresie przy jednoczesnym maskowaniu wpływu pAC_{50} , zgodnie z dyskusją z rozdziału 5.1.2.

Następnie, pPLE zostało przetestowane na zbiorze zawierającym leki i fragmenty leków. Wynik został porównany do obecnie wykorzystywanego LE oraz przedstawiony na poniższych wykresach.

Rys. 19a - b Zależność pPLE (a) oraz LE (b) od HAC dla leków i fragmentów [60]



19a



19b

Wykres 19a przedstawia rozkład danych dla pPLE. Niezależnie od wartości HAC, pPLE przyjmuje wyższe wartości dla leków niż dla ich fragmentów. Dodatkowo, można zaobserwować wyraźny podział danych na dwie grupy: leki, fragmenty. W celu porównania, na wykresie 19b przedstawiony został rozkład danych dla LE. Dla HAC powyżej 20 LE również przyjmuje wyższe wartości dla leków niż dla fragmentów, natomiast poniżej 20 fragmenty leków wykazują wyższe LE niż leki co świadczy o nieprawidłowym działaniu LE dla cząsteczek w tym zakresie. Tym samym potwierdza tezę, iż *ligand efficiency* nie jest intuicyjnym narzędziem do ewaluacji *drug candidates* i powinien zostać zastąpiony łatwiejszym do interpretacji pPLE, wyznaczonym w wyniku przygotowywania tej rozprawy doktorskiej.

Ponadto, analiza z wykorzystaniem pPLE wykazała optymalny przedział HAC (30-50), dla którego równowaga pomiędzy dopasowaniem molekularnym a siłą wiązania jest najbardziej korzystna z punktu widzenia projektowania leków.

5.2 Badanie możliwości eksploracji wskaźników innowacyjności na podstawie listy najlepiej sprzedających się leków i FDA approvals

5.2.1 Wstęp teoretyczny

Istniejące leki (lub nawet spodziewane ich generyki) nie są w stanie zaspokoić pełnego spektrum potrzeb medycznych, stąd konieczność innowacji w branży farmaceutycznej. Również organy regulacyjne motywują koncerny farmaceutyczne do ciągłych prac nad opracowywaniem nowych leków. Przykładem może być wymuszenie przez prawo patentowe agresywnej aktywności w badaniach nad nowymi kandydatami na lek [85].

Innowacyjna farmacja jest najbardziej lukratywną gałęzią odkrywania leków. Jej celem jest odkrycie potencjalnych leków mogących generować zyski na poziomie miliardów dolarów. Z obietnicą dużych zysków nierozdzielnie związane jest ryzyko, a niepewność powodzenia projektu oraz wskaźnik rezygnacji są w tym obszarze szczególnie wysokie [86].

Innowacje farmaceutyczne są złożoną kwestią, angażującą nauki przyrodnicze, społeczne i ekonomiczne. Rynek farmaceutyczny jest ściśle regulowany, a wysokie standardy konieczne do spełnienia przed wypuszczeniem na rynek, prowadzą do ciągłego wzrostu wieku leków wbrew postępowi technologicznemu [35].

Walka o pozycję lidera na rynku przez lata była powodem niechęci do ujawniania przez koncerny farmaceutyczne danych eksperymentalnych, obliczeniowych oraz ekonomicznych. Sukces pionierskich prób kooperacyjnego projektowania leków (ang. *Collaborative Drug Discovery*, CDD), jak chociażby Wspólna Platforma Badań i Wiedzy nt. Antybiotyków (SPARK) [87] czy Bank Danych Białek RCSB (<https://www.rcsb.org/>) uwidocznili korzyść współpracy i zbiorowej przedsiębiorczości [88]. Zmiany nastawienia koncernów farmaceutycznych do projektów CDD następują jednak niewystarczająco szybko, większość danych ekonomicznych wciąż nie jest ujawniania, a firmy konkurują ze sobą dążąc do objęcia monopolu na terapię wybranych schorzeń, wywołując sprzeciw społeczeństwa [89].

W tym miejscu wartym przytoczenia jest wynik uproszczonego badania stopnia zaangażowania chemoinformatyki w proces projektowania leków zaakceptowanych przez FDA w roku 2017, przeprowadzonego przez autorkę tej rozprawy. Do wszystkich producentów FDA *approvals* 2017 skierowano mailowo prośbę o podanie przybliżonego udziału badań chemoinformatycznych w proces projektowania zaakceptowanego leku. Pomimo obietnicy braku ujawnienia źródła danych oraz braku ujawnienia korelacji udziału chemoinformatyki z nazwą leku, jak również podania celu prowadzenia badania, żaden koncern nie udzielił odpowiedzi. Czy to samo badanie dałoby identyczny wynik w roku 2023? Sądzę, że tak.

5.2.2 Metodologia

W celu utworzenia kompletnej bazy danych bestsellerów leków (TOP100, TOP), połączyłam 3 repozytoria dostępne online. TOP100 dla przedziału lat 2000 - 2013 utworzono na podstawie danych pochodzących ze strony <https://www.drugs.com>, poprzez scalenie wszystkich dostępnych raportów. Do 06/07/2023 r. nie został opublikowany żaden raport zawierający dane ekonomiczne dla okresu po 2013 r.

Dane ekonomiczne dla lat 2014 - 2019 zostały wyodrębnione z list TOP200, dostępnych na stronie <https://www.pharmalive.com>. Na wniosek autorki tej pracy, otrzymano od administratorów PharmaLive zarchiwizowane raporty, które dodatkowo uzupełniono danymi pochodzącymi ze strony <https://www.pharmacompass.com> zapewniając pełne pokrycie wyznaczonego przedziału.

Dane dotyczące FDA *approvals* pobrano z bazy *Compilation of CDER NME and the New Biologic Approvals* dostępnej na stronie <https://www.fda.gov>. Wykorzystano pełen zakres danych, tj. lata 1985 - 2019.

Właściwości chemiczne (MW, logP) oraz kody SMILES pobrałam z bazy DrugBank (<https://go.drugbank.com>). Na podstawie SMILES obliczono deskryptor QED (ang. *Quantitative Estimation of Drug-likeness*), deskryptor molekularny określający stopień

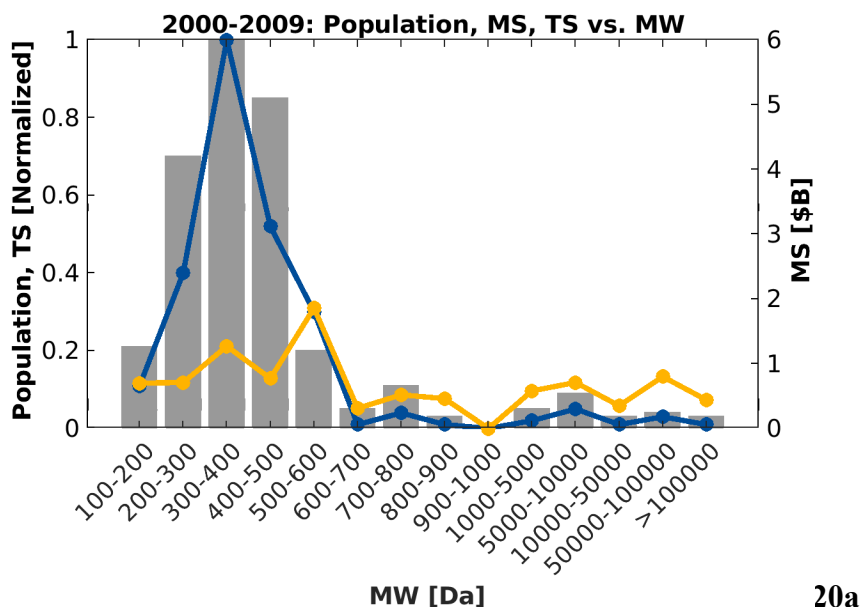
dopasowania cząsteczki do reguł warunkujących powodzenie w charakterze leku tj. odpowiednia masa cząsteczkowa, logP, topologiczna powierzchnia polarna, liczba donorów i akceptorów wiązań wodorowych, liczba pierścieni aromatycznych i wiązań obrotowych oraz obecność niepożądanych funkcji chemicznych [90]. Do wyznaczania QED wykorzystano moduł Chem.QED narzędzia chemoinformatycznego RDKit (oprogramowanie typu *open source*, dostępne online wraz z dokumentacją na stronie <https://www.rdkit.org/docs/index.html>).

Wykresy przygotowałam przy użyciu programu MATLAB.

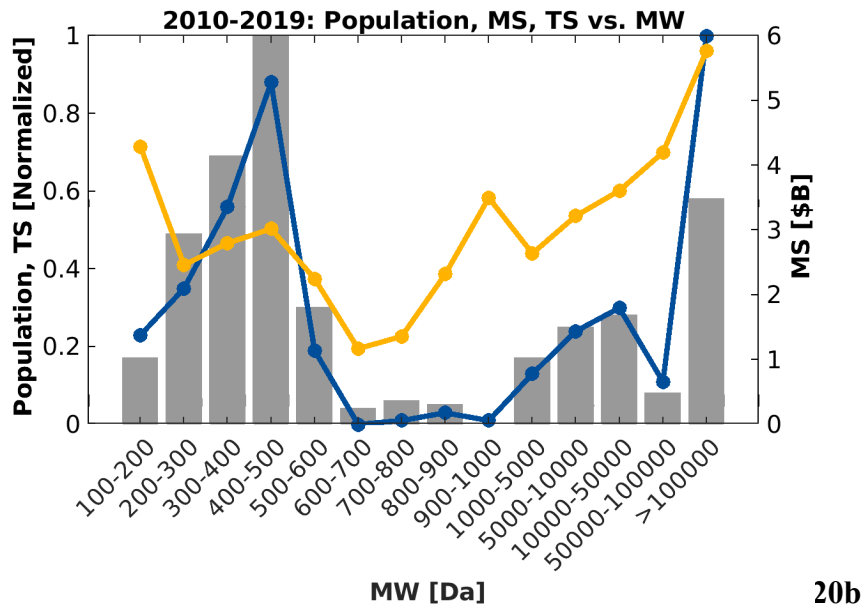
5.2.3 Wyniki

Zebrane dane uporządkowałam, a następnie przeanalizowałam tworząc wykresy 20-24.

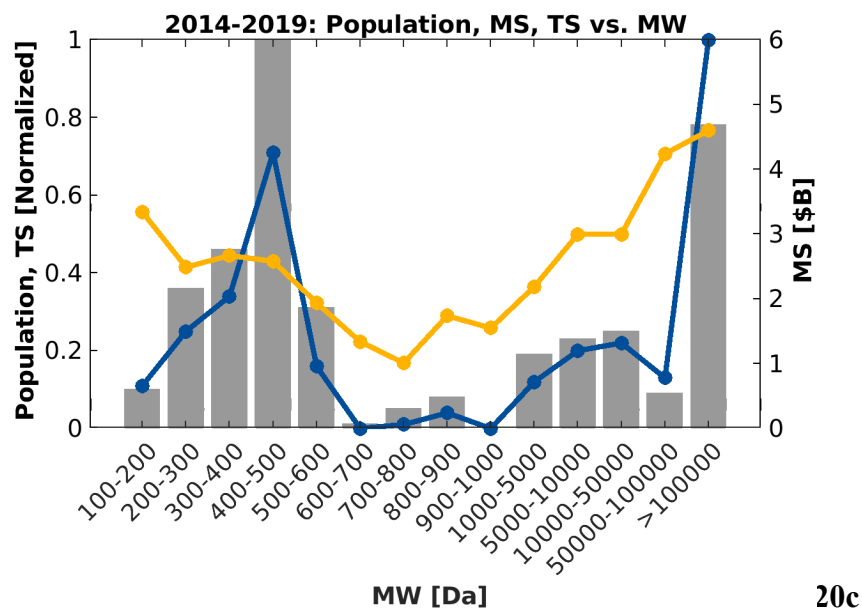
Rys. 20a - c Histogramy częstotliwości (szare słupki), średnia sprzedaż/lek (ang. *Mean Sales*, MS: żółty) oraz całkowita sprzedaż/grupa leków (ang. *Total Sales*, TS: niebieski) w porównaniu do przedziałów MW dla odpowiednich populacji TOP w latach 2000 - 2009 (a), 2010 - 2019 (b) i 2014 - 2019 (c)



20a



20b



20c

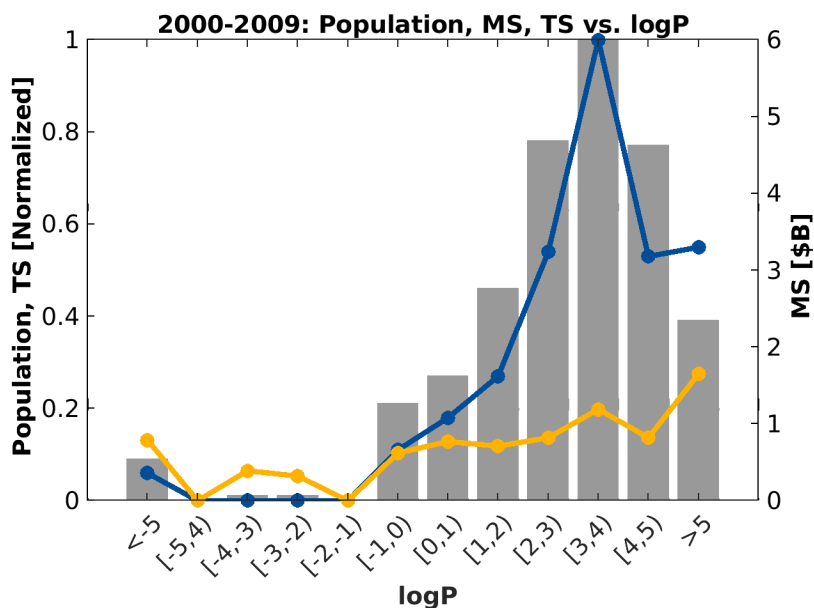
Wykres 20a - c przedstawia rozkład MW w populacji TOP. Zaobserwowano znaczne różnice w częstościach MW pomiędzy okresem 2000 - 2009 a 2010 - 2019 (oraz 2014 - 2019). Po pierwsze, w zakresie MW dyktowanego regułą Lipińskiego (tzw. Reguła Pięciu, preferująca cząsteczki o MW nieprzekraczającej 500 Da [91]) najwyższa częstość przesunęła się z

przedziału 300 - 400 Da (okres 2000 - 2009) w kierunku przedziału 400 - 500 Da (okres 2010 - 2019). Efekt przesunięcia jest jeszcze bardziej wyraźny, gdy porównamy ze sobą okresy 2000 - 2009 i 2014 - 2019.

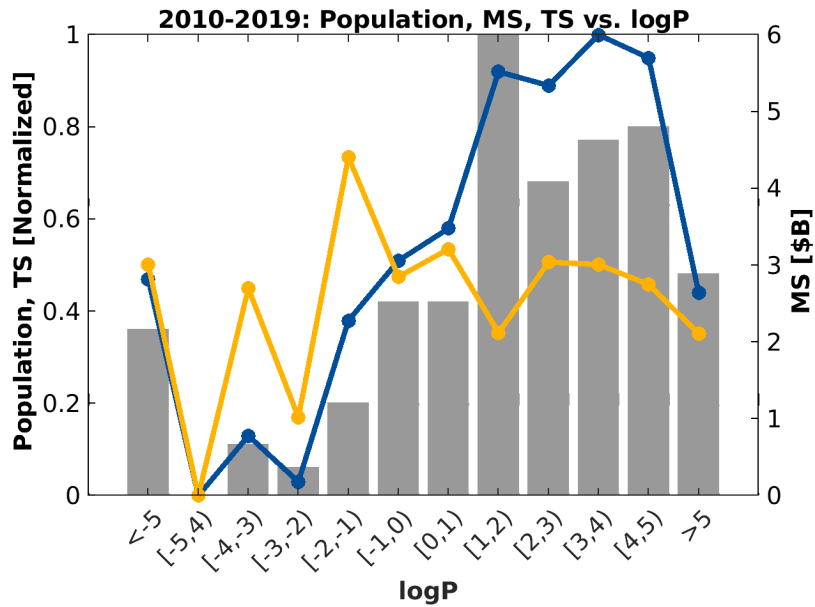
Po drugie, nastąpił znaczny wzrost populacji o wysokim MW (przekraczającym 100 000 Da) w ramach TOP 2010 - 2019 (2014 - 2019) w porównaniu do tego samego regionu (MW > 100 000 Da) w ramach TOP 2000 - 2009. Co ciekawe, dla TOP 2010 - 2019 oraz 2014 - 2019, wysoka populacja MW (powyżej 100 000 Da) była również najbardziej lukratywnym obszarem pod względem sprzedaży pojedynczego leku.

Innym zaobserwowanym efektem podczas analizy danych na wykresach 20a - c jest względny wzrost średniej sprzedaży (MS) dla grupy o najniższym MW (poniżej 300) w latach 2010 - 2019.

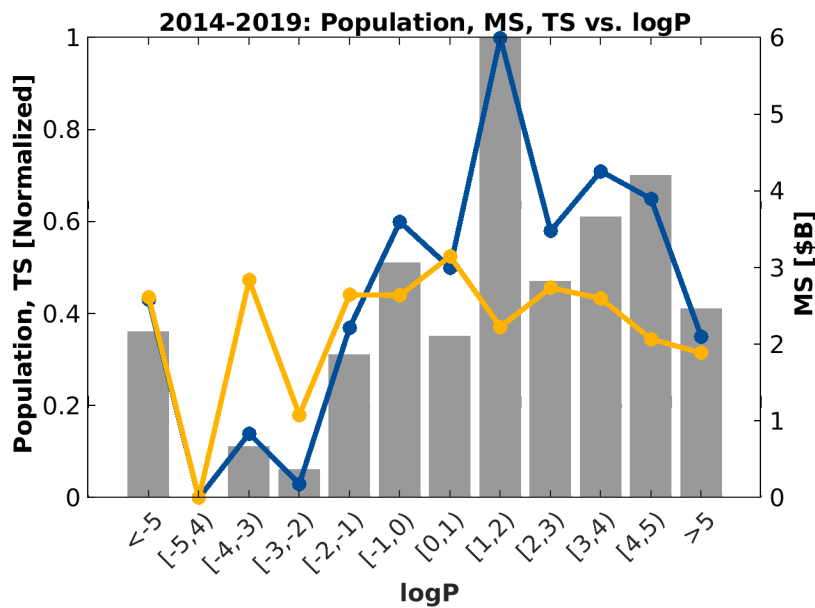
Rys. 21a - c Histogramy częstotliwości (szare słupki), średnia sprzedaż/lek (ang. *Mean Sales*, MS: żółty) oraz całkowita sprzedaż/grupa leków (ang. *Total Sales*, TS: niebieski) w porównaniu do przedziałów logP dla odpowiednich populacji TOP w latach 2000 - 2009 (a), 2010 - 2019 (b) i 2014 - 2019 (c). Częstotliwości i TS zostały znormalizowane, aby śledzić zmieniającą się liczbę leków w różnych okresach.



21a



21b

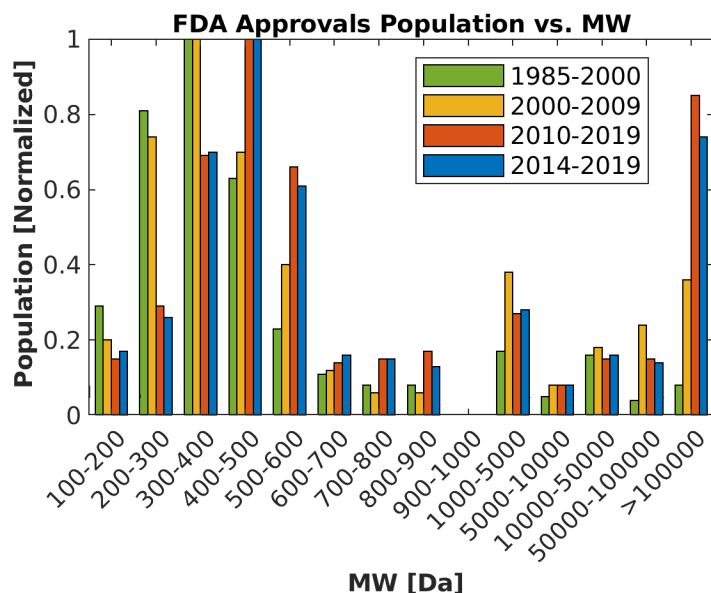


21c

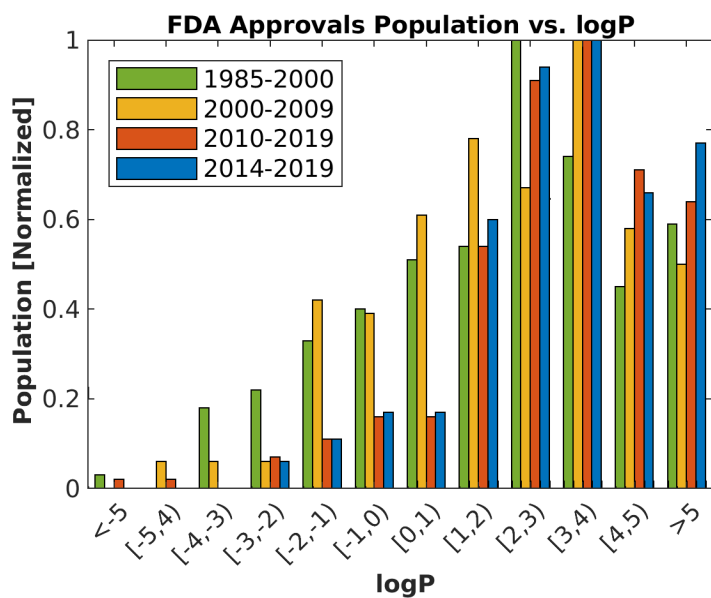
Rozkład częstotliwości populacji TOP vs. logP przedstawiono na rysunku 21a - c. Średnie wartości logP wyniosły 2,41 dla TOP 2000 - 2009 kontra 1,78 dla TOP 2010 - 2019 (1,3 dla TOP 2014 - 2019). Co ciekawe, w ostatnich latach (TOP 2010 - 2019 i szczególnie w TOP

2014 - 2019) najwyższa częstotliwość logP uległa przesunięciu w kierunku niższego logP, z wartości 3-4 (2000 - 2009) do 1-2 (2000 - 2014), Rysunek 21a vs. 21c.

Rys. 22a - b Histogramy częstotliwości populacji FDA w porównaniu do przedziałów MW (a) i logP (b). Kolory słupków oznaczają przedziały czasowe kolejno zielony 1985 - 2000, żółty 2000 - 2009, czerwony 2010 - 2019 oraz niebieski 2014 - 2019.



22a

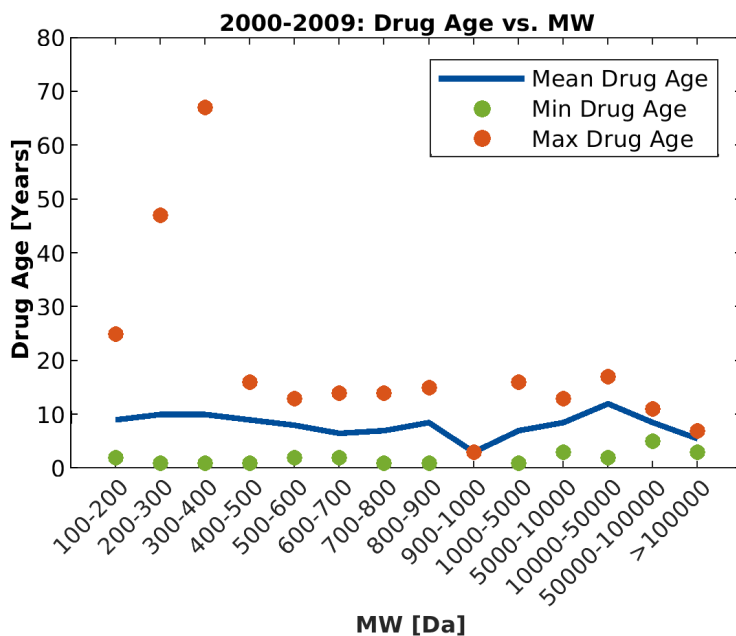


22b

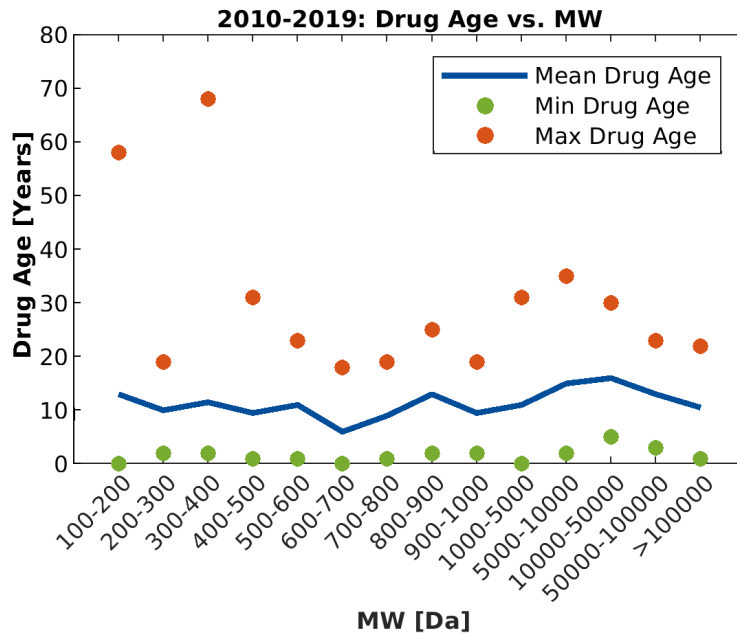
Na wykresie 22a - b porównano rozkłady MW i logP populacji FDA *approvals* przed rokiem 2000 do lat 2000 - 2009, 2010 - 2019 i 2014 - 2019. Choć populacja TOP jest znacznie starsza niż populacja świeżo zaakceptowanych FDA z odpowiadających przedziałów lat, istnieją podobne trendy w histogramach częstości, tj. przesunięcie najwyższych częstości do wyższych MW, zarówno w obszarze zgodnym z regułą Lipińskiego jak i w obszarze najwyższych MW, >100000 Da. Co ciekawe, w przypadku FDA *approvals* nie odnotowano efektu przesunięcia w dół częstości logP analogicznego do zaobserwowanego w przypadku TOP100 (Rysunek 20 vs. Rysunek 22).

Kolejną różnicą pomiędzy bestsellerami a FDA *approvals* jest wzrost populacji o najwyższych MW, występująca dla list TOP z ostatnich lat, w obszarze MW >100,000 Da, podczas gdy dla FDA jest to MW 400 - 500. Ponadto, względne częstości MW 400 - 500 vs. 500 - 600 są różne dla analizowanych grup, a znaczenie MW 500-600 jest niższe dla ostatnich FDA.

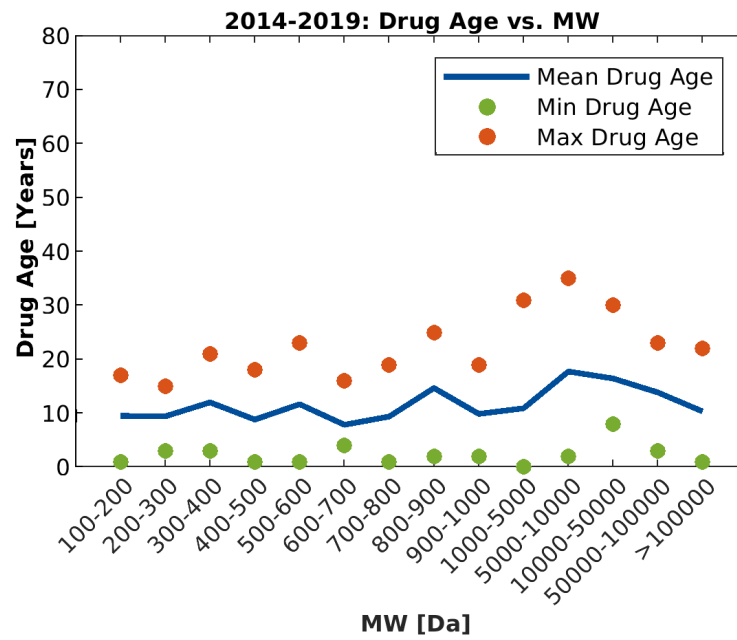
Rys. 23a - d Średni, maksymalny oraz minimalny wiek leku TOP kontra MW i logP



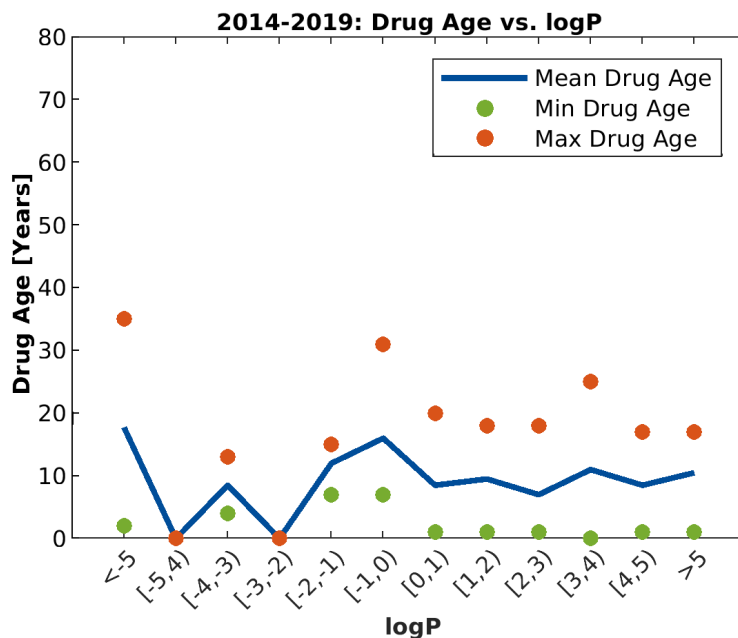
23a



23b



23c

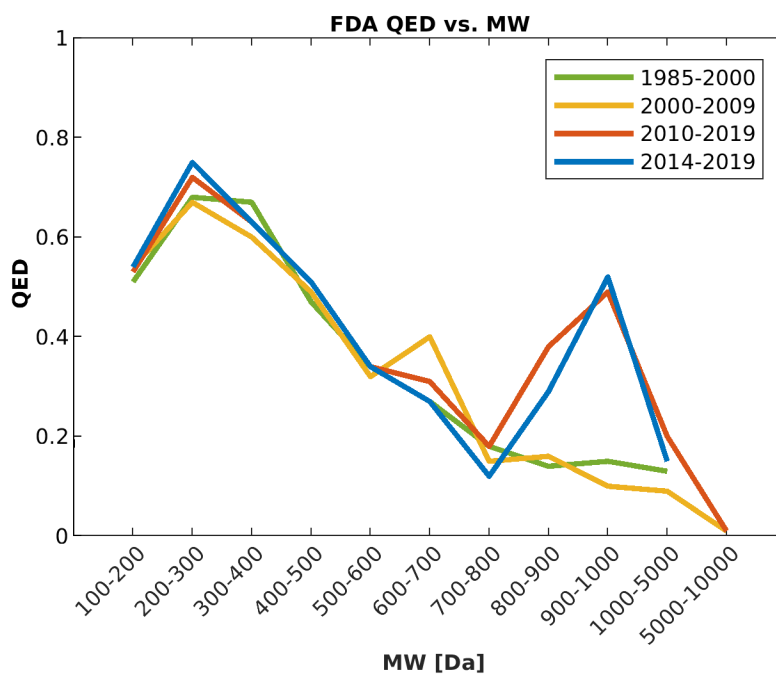


23d

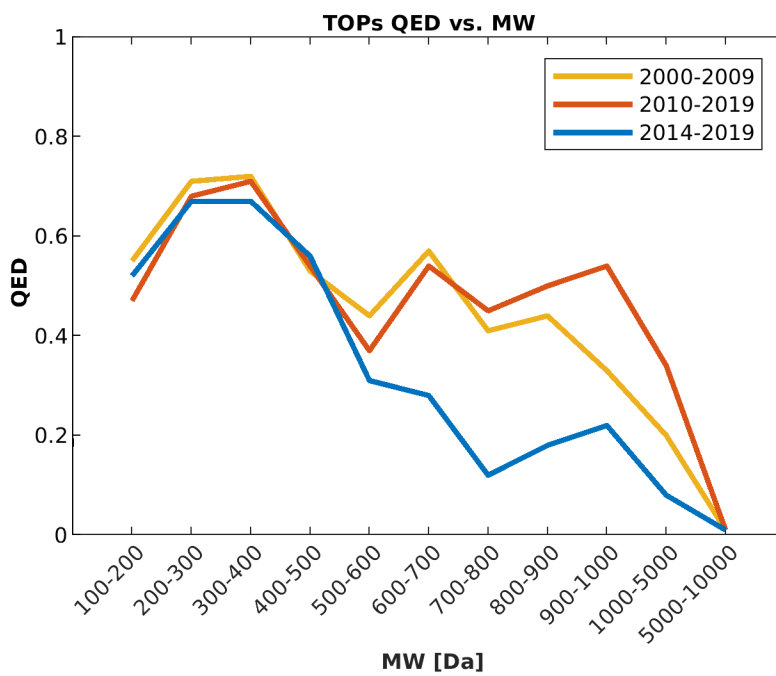
Wykres 23a - d przedstawia średni, maksymalny oraz minimalny wiek leku dla TOP kontra ich MW i logP. Na skróconych wersjach wykresu (23c - d) ukazano analogiczny rozkład dla ostatnich 50 lat. Ogólnie, średni wiek leków jest nieco wyższy w populacjach TOP 2014 - 2019 i TOP 2010 - 2019 (ok. 11 lat) w porównaniu do populacji TOP 2000 - 2009 (około 9 lat). We wszystkich populacjach natomiast, średni wiek leku nie zależy wyraźnie od MW lub logP. Zauważono również wzrost maksymalnego wieku leku dla niskich MW w okresach 2000 - 2009 i 2010 - 2019 (67 kontra 68 lat). Co ciekawe, efekt ten zniknął w ostatniej połowie dekady 2014 - 2019.

Ponadto, nastąpił względny spadek maksymalnego wieku leku dla regionu MW 100 - 200 Da dla TOP 2014 - 2019 (z około 60 lat TOP 2010-2019 do około 20 lat) i dla regionu 300 - 400 Da dla tego samego okresu (z około 70 lat TOP 2000-2009, dla TOP 2010-2019 wyniósł 21). Maksymalny wiek leku dla całej populacji TOP 2014 - 2019 wyniósł 35 lat, kontra około 70 lat dla okresu 2000 - 2009 i 2010 - 2019.

Rys. 24a - b Analiza QED dla TOP oraz FDA approvals



24a



24b

Analiza deskryptorów QED dla TOP wykazała stabilność średniej wartości QED dla wszystkich badanych okresów, wynoszącą kolejno ok. 0,6 dla TOP 2010 - 2019 oraz 0,5 dla TOP 2000 - 2009. Wynik jest korzystny w porównaniu do QED wyznaczonego dla FDA *approvals*, wynoszącego kolejno 0,45 dla okresu 2000 - 2009, 0,44 dla okresu 2010 - 2019 oraz 2014 - 2019 (dla przedziału od 1985 do 2000 roku, średnie QED wyniosło 0,52). Podsumowując, na podstawie analizy bestsellerów leków można zaobserwować dwie główne zależności. Pierwsza z nich jest zakodowana w populacji TOP analizowanej względem deskryptorów molekularnych, MW oraz logP. Drugą stanowi związek między określonymi właściwościami leków, sprzedażą lub wiekiem leku kontra MW lub logP. Podczas gdy histogramy dystrybucji są mapami jakościowymi, ilustrującymi sukces w określonych przedziałach MW lub logP, średnia sprzedaż i wiek leku wspierają te statystyki za pomocą miar ilościowych. Co ciekawe, podczas gdy histogramy częstotliwości zależą od MW i logP, średnia wartość wieku leku nie zależy od MW i logP (rysunek 23).

Zanotowano również znaczące różnice w profilach średniej sprzedaży TOP w latach 2000 - 2009 (wykres 20a) w porównaniu z latami 2010 - 2019 (wykres 20b) i 2014 - 2019 (rysunek 20c). TOP o najwyższych MW (>100,000 Da) miały równocześnie najwyższą średnią sprzedaż/lek zarówno w ostatniej dekadzie, jak i w jej drugiej połowie. Z kolei TOP o najniższych MW (100 - 200 Da) zajęły drugie miejsce w obu okresach. W latach 2014 - 2019, wysokiej średniej wartości sprzedaży/lek dla niskich MW (100 - 200 Da) towarzyszył względny spadek histogramu dla tego obszaru MW, tj. mniejsza liczba leków o najniższym MW (100 - 200 Da) miała wyższą średnią sprzedaż/lek.

Najwyższa średnia sprzedaż na lek umożliwia zidentyfikowanie dwóch obszarów, które były szeroko badane w ostatnich latach. Najniższe MW reprezentują tzw. podejście *slim pharma*, odwrotność *molecular obesity*, według którego niskie MW są najbardziej atrakcyjne dla projektowania leków [92]. „Szczupłe” cząsteczki działają lepiej niż „otyle”, na przykład pod względem wskaźnika ścierania [93]. W związku z tym uproszczenie struktury może odgrywać kluczową rolę w optymalizacji leadów i opracowywaniu nowych leków [12, 94]

Z kolei najwyższe MW to leki biologiczne, często wytwarzane przez żywe organizmy, zawierające ich fragmenty lub konstrukty biomimetyczne, wytworzone w procesach biotechnologicznych. Rośnie świadomość wysokiego potencjału innowacyjnego tych środków terapeutycznych. Strategie przezwyciężenia barier w przyjmowaniu leków biopodobnych i biomimetyków zostały niedawno przeanalizowane i poddane przeglądowi [95, 96].

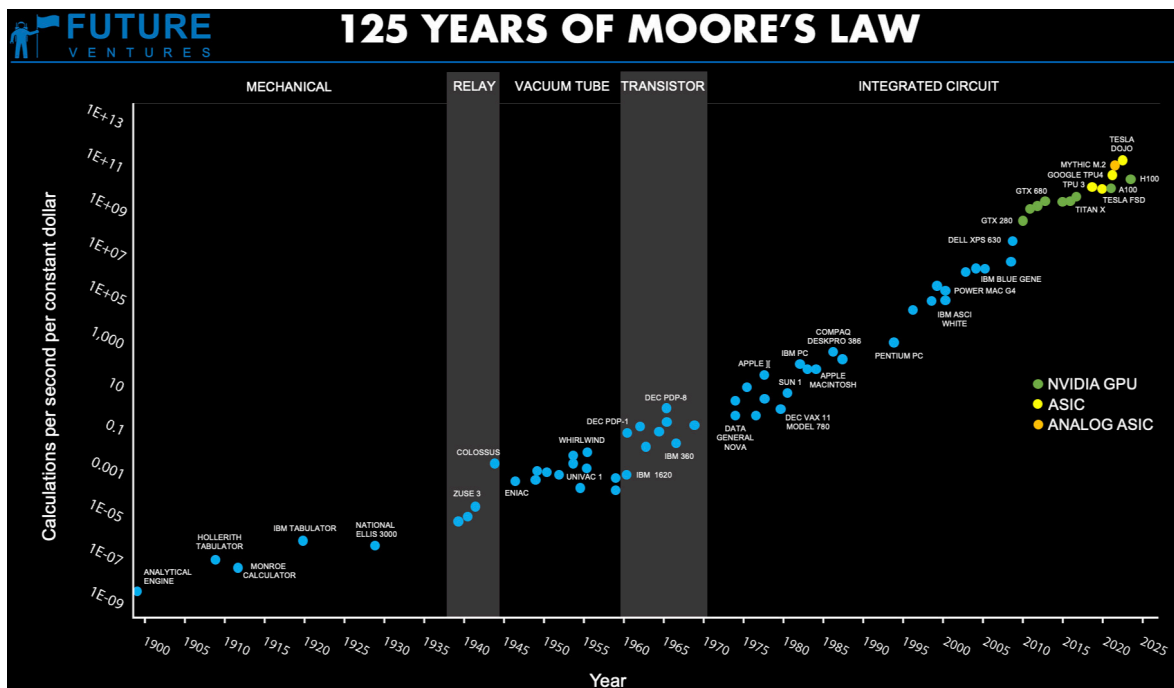
W kontekście farmakoekonomiki, w populacjach TOP średnia sprzedaż odpowiadająca najgęściej zaludnionym częstotliwościom MW utrzymywała się na wysokim poziomie. Z jednej strony oznacza to, że obszar ten może być lukratywny dla producentów. Porównajmy to jednak z podobnymi danymi dotyczącymi rozpoznawania głosu lub systemów biometrycznego, które są powszechnie wymieniane jako jedne z najbardziej innowacyjnych gałęzi IT. Podczas gdy wraz z czasem ich zastosowanie i stopień skomplikowania wzrastają, ceny stale spadają. Ten efekt potwierdza tezę o stałym spadku cen w branży IT wraz z rozpowszechnianiem nowych technologii, ilustrując wyższość innowacyjności w aspekcie IT (prawo Moore'a) przeciwko niższej innowacyjności farmacji (prawo Eroom'a).

Prawo Moore'a powstało na podstawie obserwacji wzrostu optymalnej liczby tranzystorów w mikroprocesorach. Mówi ono o tym, iż w ciągu 2 lat chipy komputerowe zwiększają dwukrotnie swoją złożoność przy niemal stałym koszcie jednostkowym. Wraz ze wzrostem ilości tranzystorów, wrasta moc obliczeniowa oraz gęstość przechowywanych danych. To z kolei powoduje rozwój sztucznej inteligencji, nauczania maszynowego, którego podstawą jest przetwarzanie możliwie największych zbiorów danych, w najkrótszym czasie.

Ogólnie, prawo Moore'a opisuje rygorystyczne warunki rynku IT, wymuszającego na firmach dwukrotne zwiększanie swoich wyników w zaledwie 24 miesiące. W ciągu ostatnich 35 lat, taki efekt udawało się uzyskać firmom co rok [97].

Rys. 25a - b Wykres przedstawiający ścieżkę rozwoju technologicznego, określoną prawem Moore'a, każdy punkt wyznacza granicę wydajności obliczeniowej w danym roku, osiągniętą przez wskazaną firmę (a).

Obecnie na szczycie jest chip D1 superkomputera DOJO (Tesla). Oryginalny wykres obrazujący przewidywane tempo zmian w technologii procesorów, autorstwa Gordona Moore'a (b) [97]



25a

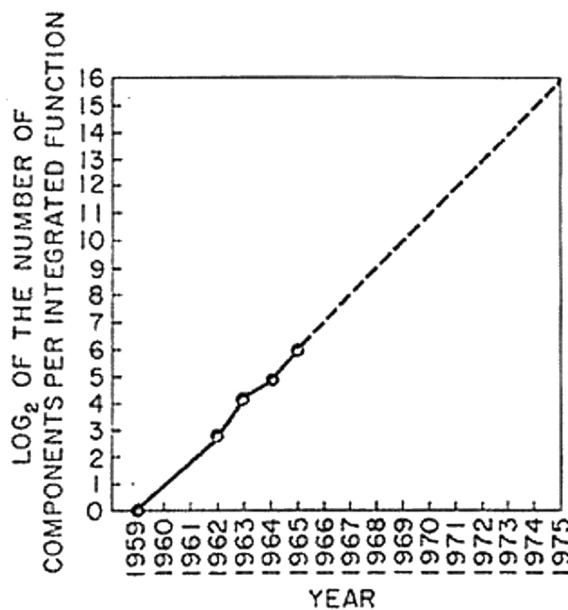


Fig. 2 Number of components per Integrated function for minimum cost per component extrapolated vs time.

25b

Prawo Moore'a napędza elektronikę, komunikację i komputery, które z kolei wkraczają w każdą gałąź gospodarki. Wykładnicze tempo postępu technologicznego, które można zaobserwować na rysunku 25a, jest zobrazowaniem ciągłych zmian na rynku, wywołujących kolejne fale możliwości dla nowych i istniejących firm.

Powodem gwałtownych zmian w technologii jest koncept łączenia nowych i dotychczasowych „pomysłów”, a jego motorem napędowym rozszerzający się dostęp do Internetu. W ciągu najbliższych pięciu lat, swoje pomysły będą mogły dzielić w czasie rzeczywistym mieszkańcy krajów rozwijających się (m.in. Afryka, Azja) dzięki stałemu obniżeniu cen i zwiększaniu dostępności smartfonów. Szacuje się, iż w tym czasie dostęp do Internetu uzyska 3 miliardy osób [97], co z kolei spowoduje bezprecedensowy skok ilości ludzi zaangażowanych pośrednio i bezpośrednio w globalny rozwój gospodarki.

Dysanalogią prawa Moore'a w odniesieniu do rozwoju branży farmaceutycznej jest prawo Eroom'a (Moore czytane wstecz). Według niego, koszt związany z wypuszczeniem na rynek nowego leku podwaja się wykładniczo, wbrew ciągłemu rozwojowi technologicznemu [98].

Spektakularny wynik i zderzenie rozwoju technologicznego z rozwojem branży farmaceutycznej dostarcza analiza spadku ilości roboczogodzin, zapewniających pokrycie kosztu jednej godziny czytania, która spada z 6 godzin w 1800 r. (świeca łożowa), przez >15 minut w 1880 r. (świeca naftowa), >8 sekund w 1950 roku (żarówka świecowa) aż do 1/2 sekundy żarówki CF w 1997 roku [88, 99]. W obszarze leków wciąż nie osiąga się etapu spadku cen innowacji IT obserwowanych od ponad 200 lat.

"Wiedza jest droga w produkcji, ale czasami może się zwrócić. Prawo Amara mówi, że [my] mamy tendencję do przeceniania wpływu nowych technologii w krótkim okresie, ale nie doceniamy go w dłuższej perspektywie". Przykładem może być genomika w farmacji [99]. Pomimo oczywistości celu do którego powinna dążyć branża farmaceutyczna by zrównać się z branżą IT, nadal obserwowana jest odwrotna tendencja ekonomiczna, a leki obejmujące najwyższe lokaty wśród bestsellerów wciąż się starzeją.

Istnieją cztery powody stałego wzrostu kosztu produkcji leków [100]:

1. „Lepiej niż Beatlesi” – firmy farmaceutyczne intensywnie eksploatują możliwości poprawy wyników terapii, próbując zastąpić dotychczasowo stosowane leki (konkurencja). Po latach prac nad udoskonaleniem, nie udaje się uzyskać satysfakcjonującego rozwiązania, a poniesione koszty nie mają zwrotu w sprzedaży nowego leku.
2. „Ostrożny regulator” – wymagania konieczne do spełnienia przez wprowadzany na rynek lek wciąż rosną, powodując konieczność stałego ulepszania i uszczegółowienia procesów badań klinicznych oraz procesów kontroli jakości produkcji. To z kolei generuje dodatkowe koszty m.in. zakupu nowych walidatorów czystości, zwiększania grup badawczych, wprowadzania dodatkowych badań przesiewowych na partiach wyprodukowanych leków.
3. „Rzucanie pieniędzy” – liderzy rynku farmaceutycznego, uzyskujący stałe dochody ze sprzedaży bestsellerów zajmujących listy TOP przez wiele lat, posiadają środki finansowe umożliwiające im gwałtowne badania w rzekomo lukratywnych dziedzinach. Choć dla danej jednostki tej natury straty finansowe są wliczone w ryzyko odkrywania nowych obszarów, mają znaczący wpływ na ogólny wynik ekonomiczny branży farmaceutycznej.
4. Metody „Brutalnej siły” (*brute force*) – niedocenywanie możliwości jakie mogą nieść za sobą badania przesiewowe. Brak dążenia do rzetelnego poznania podstaw chorób z jednoczesnym przeniesieniem zainteresowania na komercjalizację wyników i osiągnięcie potencjalnie najwyższej stopy zwrotu z inwestycji.

Koszt produkcji nowego leku dramatycznie wzrósł w ciągu ostatniej dekady i waha się pomiędzy 314 milionami a 2,8 miliardami dolarów [100]. Krytyczną sytuację uwypuklił okres pandemii COVID-19, w trakcie którego wypuszczono na rynek 53 nowe leki (FDA *approvals 2020*) przy nakładach na badania i rozwój w branży farmaceutycznej rzędu 200

miliardów dolarów (średni koszt jednego leku wyniósł około 3,8 mld dolarów). Żaden z nich nie znalazł się na liście 100 najlepiej sprzedających się leków w dwóch kolejnych latach (badania własne autorki pracy, na podstawie raportu FDA *approvals* na rok 2020 oraz listy najlepiej sprzedających się leków udostępnionej na <https://www.pharmacompass.com>).

Istnieje kilka pomysłów na przełamanie prawa Eroom'a w farmacji. Jednym z nich jest rozpoczęcie aktywnego śledzenia ścieżek niepowodzeń w rozwoju leku leków, na każdym etapie ich projektowania. Dokładnie wytypowane i opisane przyczyny stanowiłyby przestrożę w kolejnych procesach (ang. *red flag*), minimalizując ryzyko porażki nowych leków. Aby rozwiązanie dało rzeczywisty efekt, konieczne byłoby udostępnianie wyników z niepowodzeń firm. To z kolei ciągnie za sobą ryzyko poważnego nadszarpnięcia wizerunku publicznego danego producenta, spadku zaufania społeczeństwa, a co za tym idzie, obniżenie jego wyników finansowych.

Drugim sposobem jest efektywne wprowadzenie metod uczenia maszynowego do wszystkich etapów projektowania leków. Najlepsze wyniki uzyskuje się budując modele obliczeniowe na podstawie precyzyjnych danych dostarczonych eksperymentalnie z żywych tkanek pacjenta. By zmaksymalizować szanse przełamania prawa Eroom'a, należy poprawnie zdefiniować obszary badań klinicznych, które z powodzeniem można wykonać metodami sztucznej inteligencji, a po wprowadzeniu jej w trybie testowym, porównywać uzyskane tą drogą rezultaty wobec istniejącej technologii.

Istnieje kilka firm, które z powodzeniem powierzyły zaprojektowanie cząsteczek leku sztucznej inteligencji. Jedną z nich jest Exscientia (<https://www.exscientia.ai>), firma która opublikowała pierwszą cząsteczkę zaprojektowaną przez sztuczną inteligencję dla immuno-onkologii. Obecnie lek jest na etapie badań klinicznych na ludziach. Exscientia do projektowania leków wykorzystuje autorski system obliczeniowy CentaurAI[®], trenowany na szczegółowych danych eksperymentalnych dotyczących tkanek ludzkich. Ponadto, Exscientia we współpracy z firmą Evotec zbudowała platformę badań Centaur Chemist[®], automatycznie projektującą i profilującą cząsteczki leku pod kątem wymagań w charakterze

drug candidates [101]. Platforma skupia się na opracowaniu antagonisty receptora A2a dla dorosłych pacjentów z zaawansowanymi guzami litymi.

Drugą firmą z powodzeniem wykorzystującą autorskie oprogramowanie do symulacji chemicznych w farmacji jest Schrödinger (<https://www.schrodinger.com>). Niedawno otrzymał zgodę FDA na badanie zaprojektowanej komputerowo terapii chłoniaków nieziarnicznych we wczesnej fazie klinicznej. Platforma firmy opiera się na możliwościach uczenia maszynowego. Podczas prac nad terapią chłoniaków, posortowała 8,2 miliarda potencjalnych związków w ciągu zaledwie 10 miesięcy! Zidentyfikowano i zsyntetyzowano 78 potencjalnych kandydatów. Wyniki przefiltrowano drogą eksperymentalną (badania przedkliniczne) w celu wybrania najbardziej obiecującego kandydata. Obecnie firma planuje rozpocząć badanie kliniczne fazy 1, rekrutując pacjentów z nawrotowym lub opornym na leczenie chłoniakiem nieziarnicznym B-komórkowym.

Ciekawą alternatywą dwóch przytoczonych powyżej firm jest Recursion Pharmaceuticals (<https://www.recursion.com>). Ich działalność można określić recyklingiem farmakologicznym. Wykorzystują sztuczną inteligencję do wyszukiwania nowych zastosowań dla leków należących do innych firm, a na współpracę zdecydowali się już liderzy rynku tacy jak Roche i Genetech. Autorka zauważa tutaj pewnego rodzaju analogię do fragonomiki, która umożliwia odnajdywanie pożądaných właściwości w cząsteczkach, w których pierwotnie nie były one znane. Ostatnie analizy są już bardziej optymistyczne. Być może zbliżamy się do przełamania prawa Erooma [102].

Podsumowując, ranking TOP ilustruje rozwój farmaceutyków pod kątem trendów obowiązujących na rynku farmakoekonomicznym. Efekt przesunięcia w dół najgęściej spopulowanych logP dla TOP nie jest obserwowany w przypadku FDA *approvals*. Oznacza to, że leki o niższych logP wygrywają na rynku, ilustrując *Lipinski's Rule of Five*. W tym samym czasie wzrost maksymalnej częstotliwości MW z 300 - 400 Da do 400-500 Da w regionie Lipińskiego i rosnące znaczenie leków biologicznych w ostatniej dekadzie w porównaniu z populacją TOP z lat 2000 - 2009 to dwie najważniejsze obserwacje badań nad populacjami TOP i FDA *approvals*. Porównanie trendów TOP i FDA ujawniło

nieoczekiwane te same rezultaty, mimo że TOP są wyraźnie starsze niż świeżo zaakceptowane FDA. Oba efekty potwierdzają, że TOP są wyznacznikami trendów dla badań i rozwoju w branży farmaceutycznej.

5.3 Fragmentacja TOP100²

5.3.1 Wstęp teoretyczny

Tematyka TOP100 jak również przyczyny i potencjalne kierunki rozwoju metod chemoinformatycznych w projektowaniu leków zostały omówione w poprzednim rozdziale. Wstęp teoretyczny do fragonomiki został częściowo przedstawiony w rozdziale 4.4.1.

Fragonomika po raz pierwszy pojawiła się w literaturze, podczas poszukiwań antagonistów cholecystokininy [37, 103]. Grupa badawcza Evansa odkryła, iż „w niektórych przypadkach architektura związków oparta na wybranych trzonach molekularnych może stać się punktem wyjścia dla skutecznego projektowania molekuł o pożądanym profilu aktywności farmakologicznej [37, 103]”. Wytypowany wówczas fragment molekularny 1,4-benzodiazepin-2-onu wykazał powinowactwo nie tylko względem obranego celu, ale także względem kilku innych receptorów. Tym samym stanowi on pierwszy w historii wyodrębniony fragment molekularny i fundament popularnej w kontekście projektowania leków fragonomiki.

Analiza fragonometryczna zyskała popularność 10 lat po odkryciu niespodziewanej własności 1,4-benzodiazepin-2-onu [103]. W ogólnym pojęciu, obejmuje generowanie *leadów* (potencjalnych kandydatów na lek) przy użyciu małych cząsteczek. Nie istnieją sztywne reguły co do maksymalnej ilości atomów w fragmencie, jednak większość praktyków ogranicza je do 18 HAC.

Fragmenty również powinny być stosunkowo proste, tzn. posiadać maksymalnie jedno lub dwa miejsca podstawienia. Ograniczenie to znacznie przyspiesza testowanie hipotez

² Opracowano we współpracy z Władysławem Zhdanem. Pełna wersja w publikacji Pedrys A., Zhdan W., et al. *Fragonomics of TOP drug bestsellers: an innovation benchmarks for drug discovery?* (w przygotowaniu)

dotyczących zależności struktura – aktywność, zmniejszając ilość możliwych kombinacji w wyniku podstawiania. Drugim powodem jest ograniczenie miejsc interakcji z celem, co z kolei prowadzi do mniejszego prawdopodobieństwa wystąpienia interakcji niepożądanych [104, 105].

Podstawą badań fragonomicznych jest biblioteka cząsteczek. Nie istnieje jasne określenie jak obszerne powinny być biblioteki ani jakie powinny zawierać parametry. Tworzenie baz regulują dwie ogólne zasady. Pierwszą z nich jest dopasowanie biblioteki do obszaru badań i zapewnienie wysokiej jakości danych, nawet za cenę ograniczenia jej ilości (większość baz zawiera 2000 fragmentów, choć zdarzają się zarówno bazy o kilkuset jak i kilkudziesięciu tysiącach rekordów).

Drugą zasadą, regulującą proces kompletowania fragmentów jest tzw. Reguła Trzech (ang. *The Rule of Three, Ro3*) [106]. Jest analogiczna do Reguły Pięciu Lipińskiego (Ro5) i zgodnie z nią:

1. Masa cząsteczkowa fragmentu nie powinna przekraczać 300 Da
2. Liczba wiązań donorowych fragmentu powinna być mniejsza bądź równa 3
3. Liczba akceptorów wiązań wodorowych fragmentu powinna być mniejsza bądź równa 3
4. Liczba wiązań rotujących fragmentu powinna być mniejsza bądź równa 3
5. Powierzchnia polarna fragmentu powinna być mniejsza lub równa 60 Å

Tak samo jak w przypadku Ro5, dwie ostatnie reguły są opcjonalne. Obiekt krytyki stanowi reguła trzecia, dopuszczająca nieścisłości płynące z dowolnej interpretacji tego czym jest akceptor wiązania wodorowego (dla przykładu, czy jest nim azot w amidach? [106]). Większość praktyków skłania się jednak do klasycznej interpretacji, w myśl której każdy atom azotu czy tlenu liczy się jako akceptor wiązania wodorowego.

Po skonstruowaniu bazy danych, następny krok stanowi identyfikacja struktur uprzywilejowanych tj. wytypowanie najczęściej powtarzającej się sekwencji fragmentu, na

podstawie statystyki ligandów, leków bądź innych grup związków. Przeprowadza się ją na głównej bibliotece, wykorzystując się algorytmy (uczenie maszynowe). W ramach ułatwienia, przyjmuje się formalną koncepcję podziału cząsteczek na stałe fragmenty, takie jak mostki (łączniki) i układy pierścieni wchodzące w skład pionu molekularnego (ang. *scaffold*) oraz łańcuchy boczne [37, 107-108]. Innym przykładem automatyzacji jest zaimplementowanie reguł cięcia wiązań do algorytmu fragmentującego. Trzecia metoda stanowi podstawę omawianej wcześniej działalności firmy *Recursion Pharmaceutical*. Polega ona na tworzeniu fragmentów z zachowaniem narzuconego podobieństwa topologicznego do znanych leków [109].

Trzecim etapem analizy fragonomicznej są badania przesiewowe cząsteczek zaprojektowanych z wykorzystaniem fragmentu uprzywilejowanego. Tu, podobnie jak w przypadku tworzenia bibliotek, fragonomika również nie wskazuje jednego, złotego sposobu osiągnięcia celu. Część decyzji może zostać wsparta AI, nadal jednak większość z nich jest podejmowana w oparciu o doświadczenie jednostki prowadzącej badania. Intuicja chemiczna, odpowiednie przygotowanie „sita” i analiza wyników umożliwiają następnie wytypowanie *candidates*, ich dalsze analizy obliczeniowe lub eksperymentalne. Istnieją również firmy, które zajmują się wyłącznie tworzeniem bibliotek fragmentów i ich ewentualną komercjalizacją np. Enamine (<https://enamine.net>) czy MedChemExpress (<https://www.medchemexpress.com>).

5.3.2 Metodologia³

Do badań fragonomicznych wykorzystałam trzy biblioteki cząsteczek. Dwie z nich stanowiły zbiory TOP oraz FDA approvals utworzone w wyniku badań nad wskaźnikami innowacyjności. W trzecim zbiorze znalazło się 1615 FDA *approvals* do roku 2023, pobranych z bazy ZINC.

Fragmentację cząsteczek przeprowadziłam korzystając z podstawowej biblioteki narzędzia RDKit (Chem), umożliwiającej oprócz tworzenia fragmentów również manipulację

³ Opracowano we współpracy z Władysławem Zhdanem. Pełna wersja w publikacji Pedrys A., Zhdan W., et al. *Fragonomics of TOP drug bestsellers: an innovation benchmarks for drug discovery?* (w przygotowaniu)

strukturami cząsteczek drogą neutralizacji (dodawanie/odjęcie atomu wodoru), generowanie nowych cząsteczek przez addycję predefiniowanego łańcucha bocznego, identyfikację stereochemii, rysowanie struktur i wiele innych. Przyjęto warunek fragmentacji, regulujący długość fragmentu od 6 do 7 atomów.


Wizualizację danych przeprowadziłam korzystając z biblioteki ChemPlot (Python). Na podstawie dwóch bibliotek cząsteczek utworzyłam wykres podobieństwa strukturalnego. W celu usprawnienia analizy danych zastosowałam redukcję ich wymiarowości, korzystając z rozkładu t-SNE.

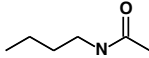

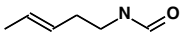
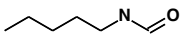
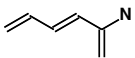
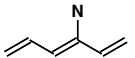
Ponadto, przestrzeń 7-atomowych fragmentów molekularnych zwizualizowałam za pomocą mapy samoorganizującej (SOM). SOM leków zatwierdzonych przez FDA została wykorzystana jako warstwa wejściowa dla SOM TOP.

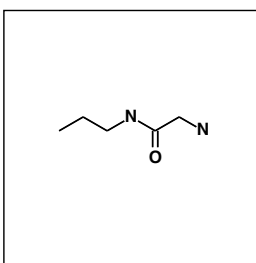
5.3.3 Wyniki

Utworzone fragmenty dla TOP i FDA *approvals* poddałam analizie powtarzalności, w celu wyłonienia struktur uprzywilejowanych. Wyniki zaprezentowałam w poniższej tabeli:

Tabela 4 Zestawienie najczęściej powtarzających się fragmentów dla baz FDA approvals oraz TOP

Fragment	FDA	FDA	TOP	TOP	TOP
	ZINC-2023	1985-2000	2000-2009	2010-2019	2014-2019
	[%]	[%]	[%]	[%]	[%]
	41,91	42,02	45,77	46,49	50,00

	2,96	3,72	2,82	15,79	17,50
	1,97	0,53	1,41	10,53	13,75
	2,11	2,93	4,93	10,53	13,75
	3,49	3,19	4,23	10,53	12,50
	4,87	6,65	9,15	8,77	12,50
	8,55	7,18	6,34	10,53	11,25



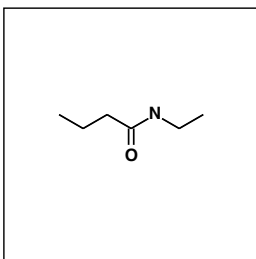
1,64

3,72

2,11

9,65

11,25



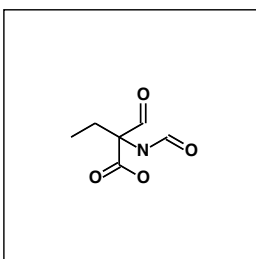
2,76

2,94

2,82

9,65

11,25



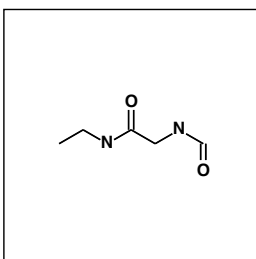
1,18

2,13

2,11

7,98

10,00



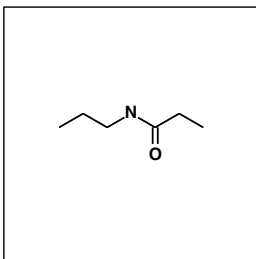
1,91

3,19

2,82

9,65

10,00



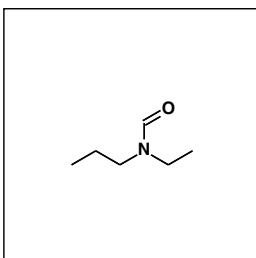
1,58

2,93

2,11

9,65

10,00






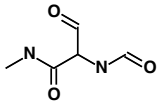


1,71

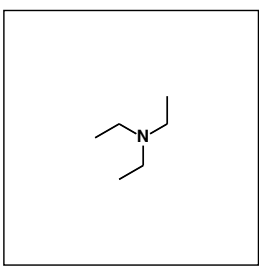
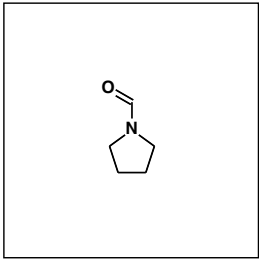
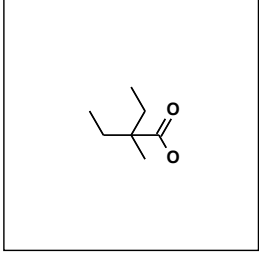
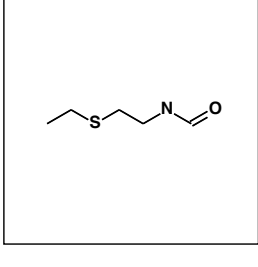
1,06

2,11

7,02

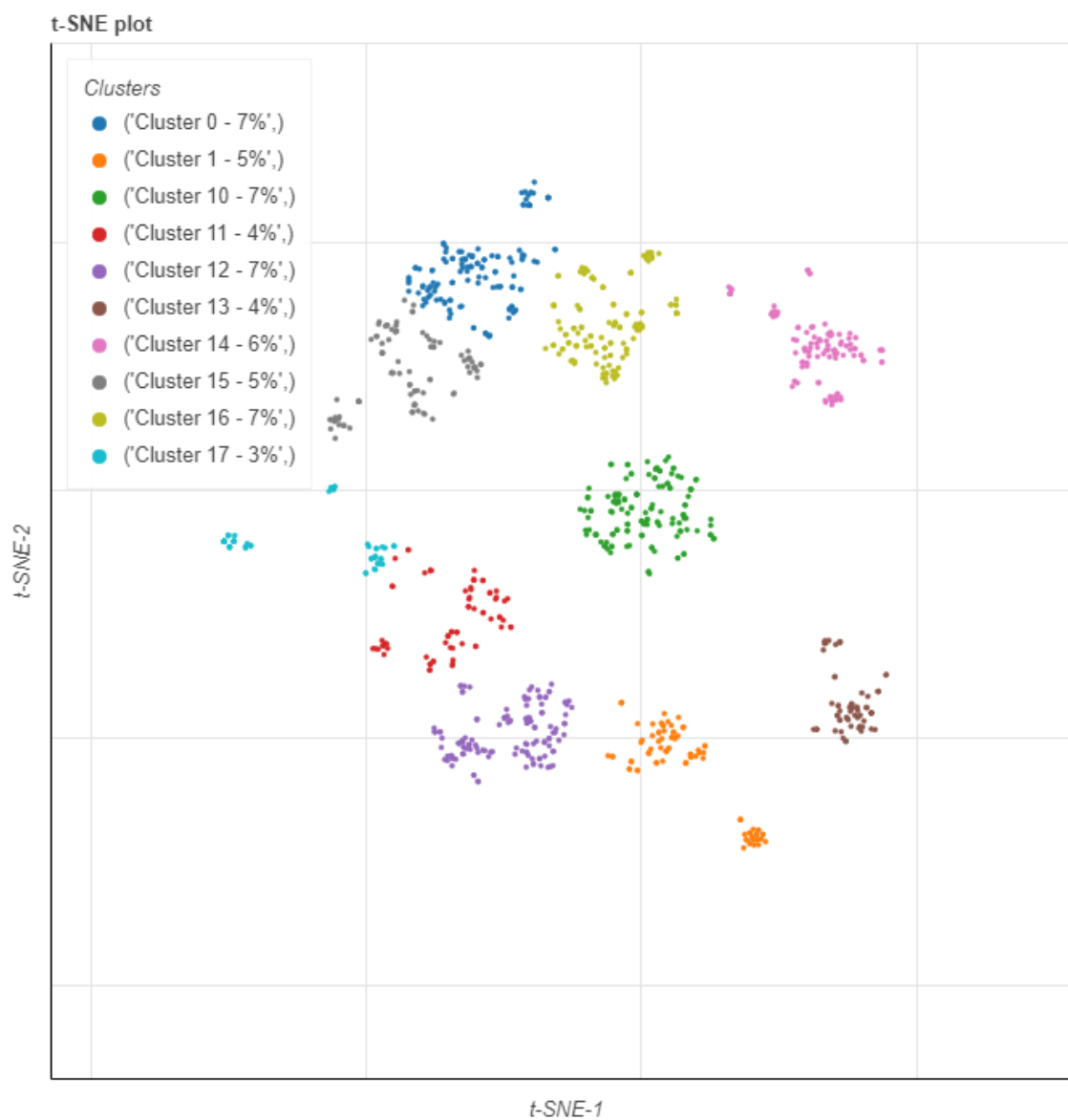
8,75

	7,63	7,98	10,56	7,89	8,75
	4,47	6,65	7,04	7,89	8,75
	5,72	6,38	6,34	7,89	8,75
	0,86	1,86	2,11	8,77	8,75
	9,41	7,18	9,15	8,77	7,50
	7,11	6,12	7,04	8,77	7,50

	8,95	9,04	9,15	5,26	6,25
	1,12	1,06	0	7,89	6,25
	5,26	4,26	7,04	8,77	5,00
	1,71	3,72	1,41	0	0

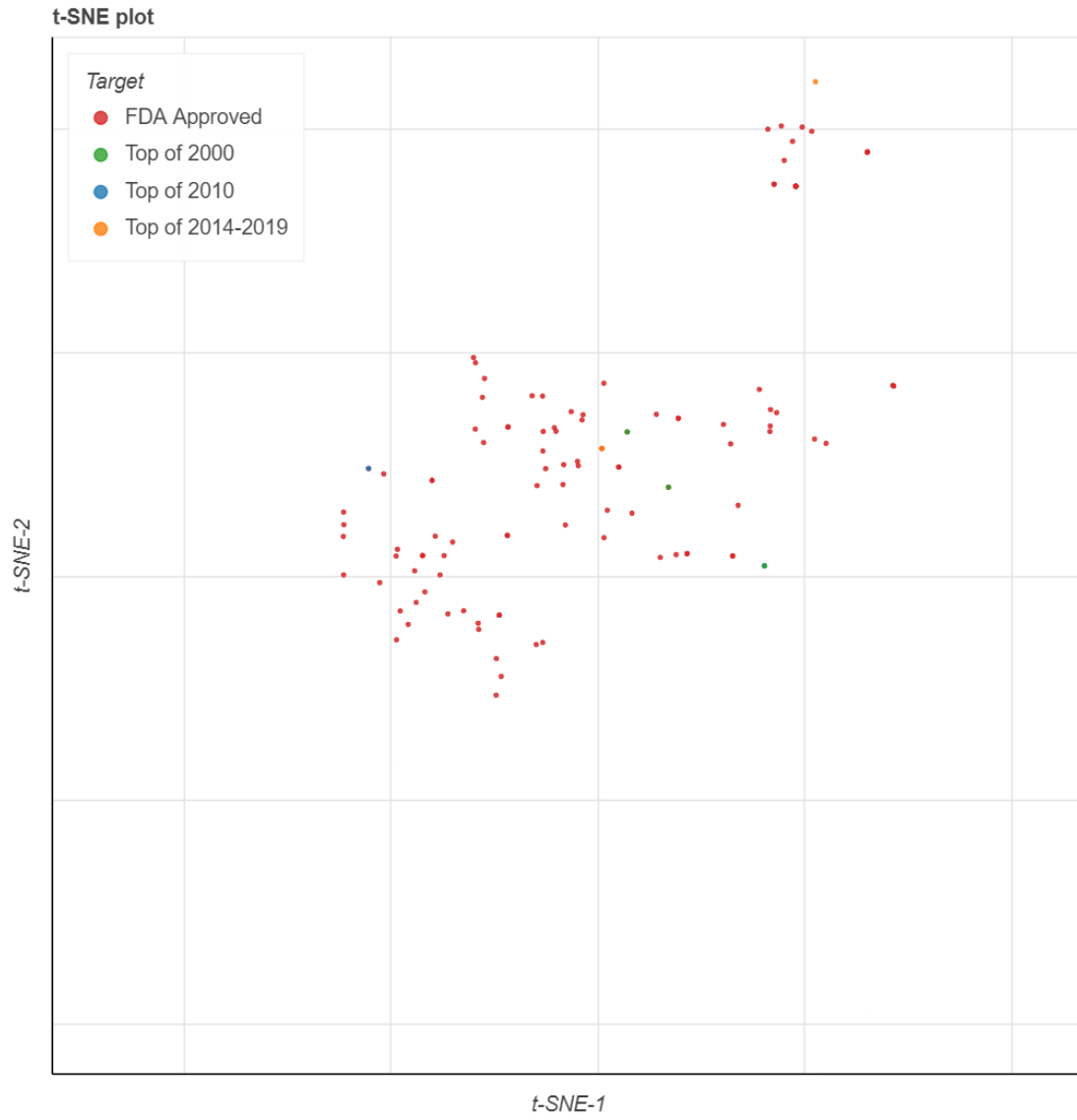
Następnie przeprowadziłam redukcję wymiarowości danych, korzystając z 7-atomowych fragmentów utworzonych w poprzednim kroku oraz rozkładu t-SNE. Dostępna przestrzeń chemiczna FDA/TOP została podzielona na 10 klasterów zgodnie z rysunkami poniżej.

Rys. 26a - k Wynik rozkładu t-SNE dla zbioru FDA approvals (1985-2019) i TOP (a), rozkład danych na poszczególne klasterory wraz ze wskazaniem przykłądów związków poszczególnych klasterów (b – k)



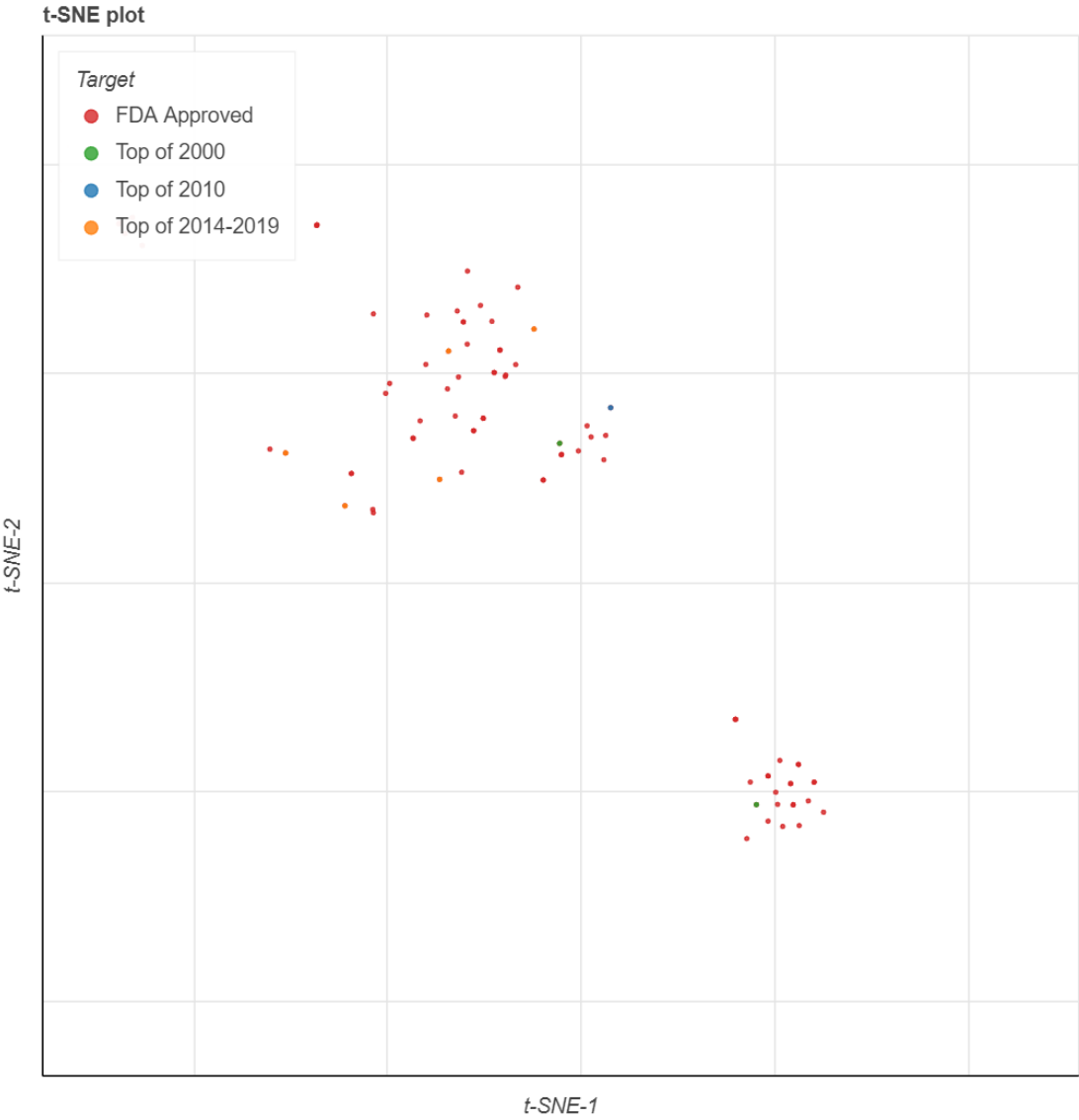
26a

Klaster 1



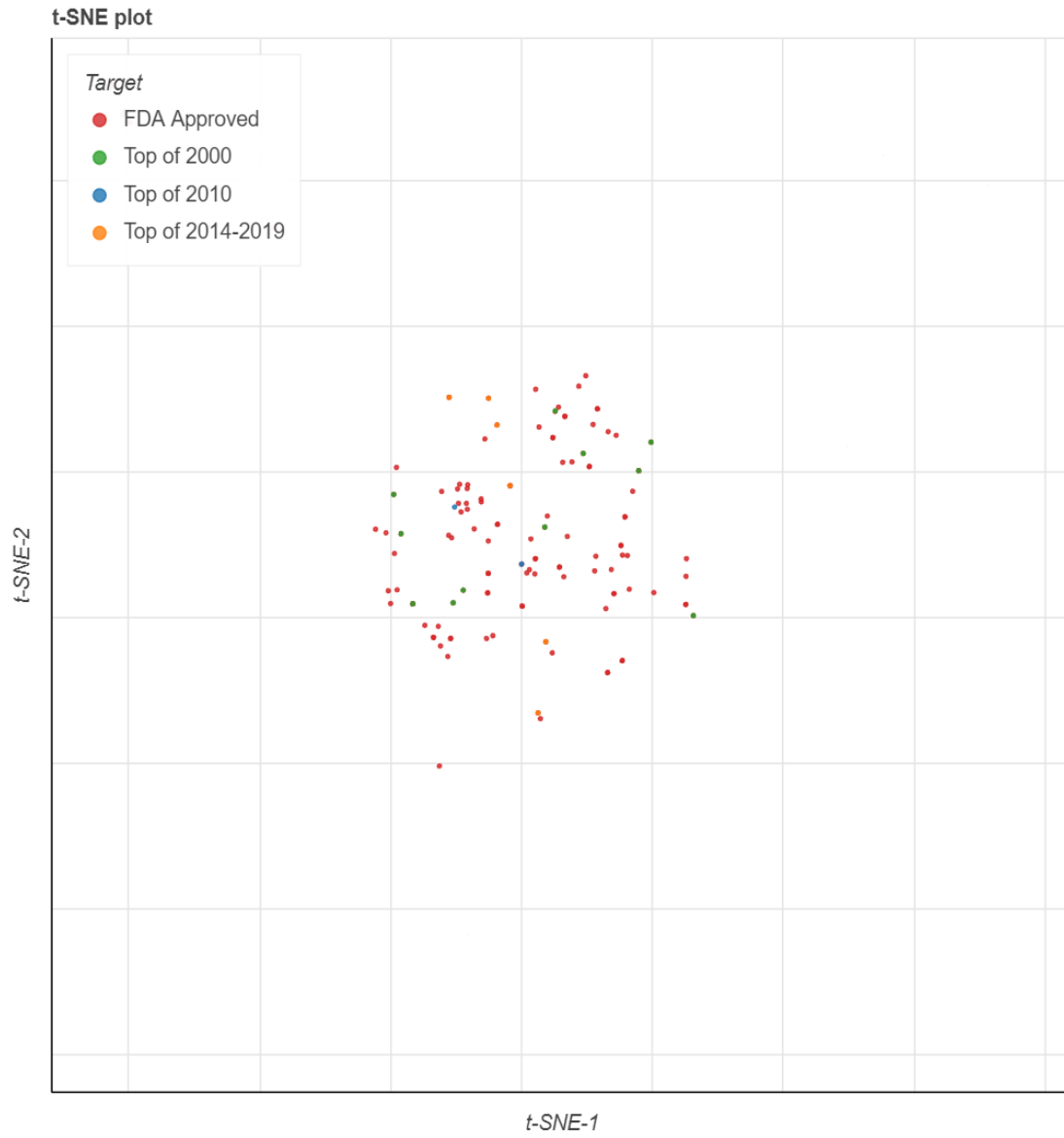
26b

Klaster 2



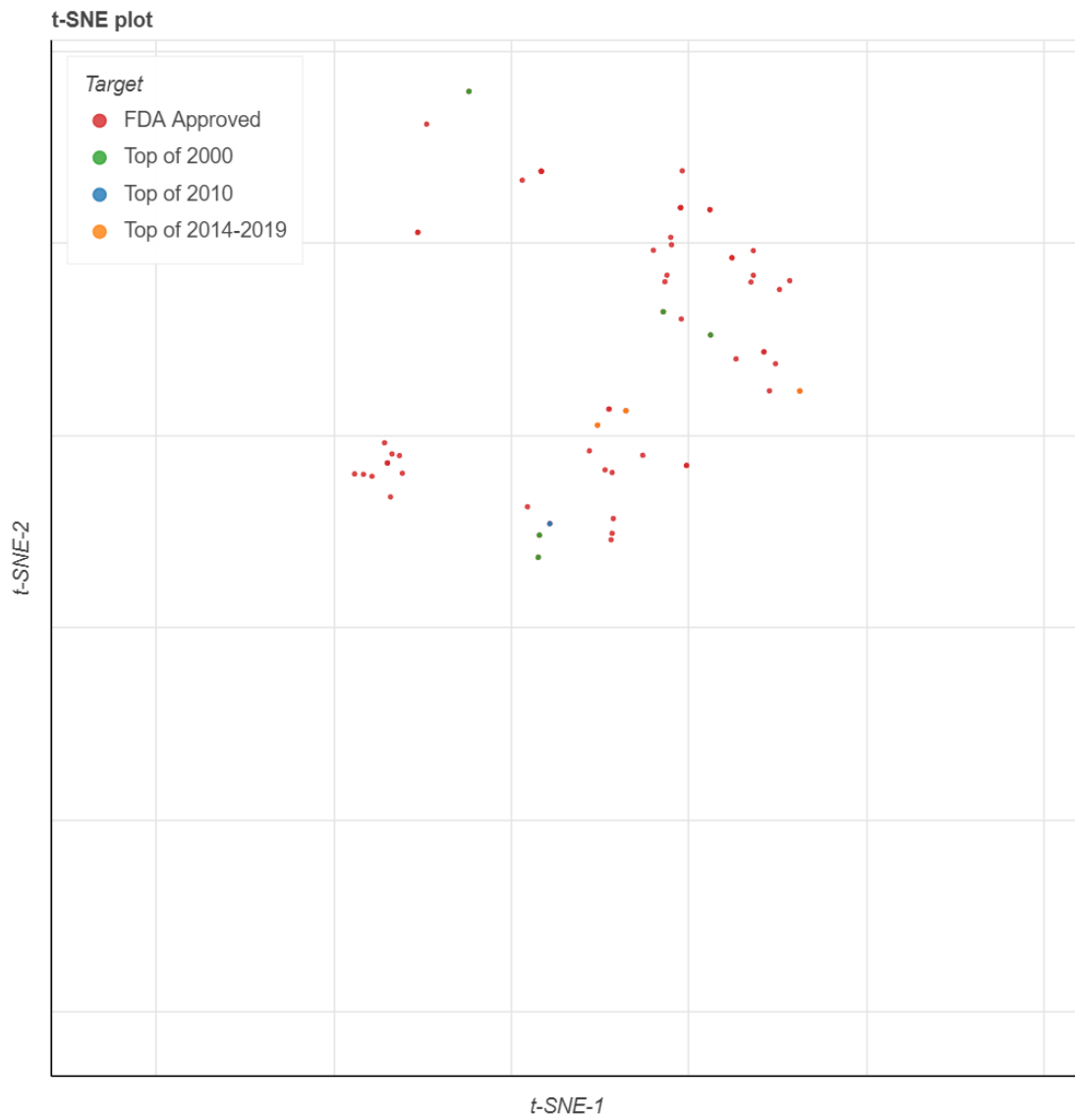
26c

Klaster 3



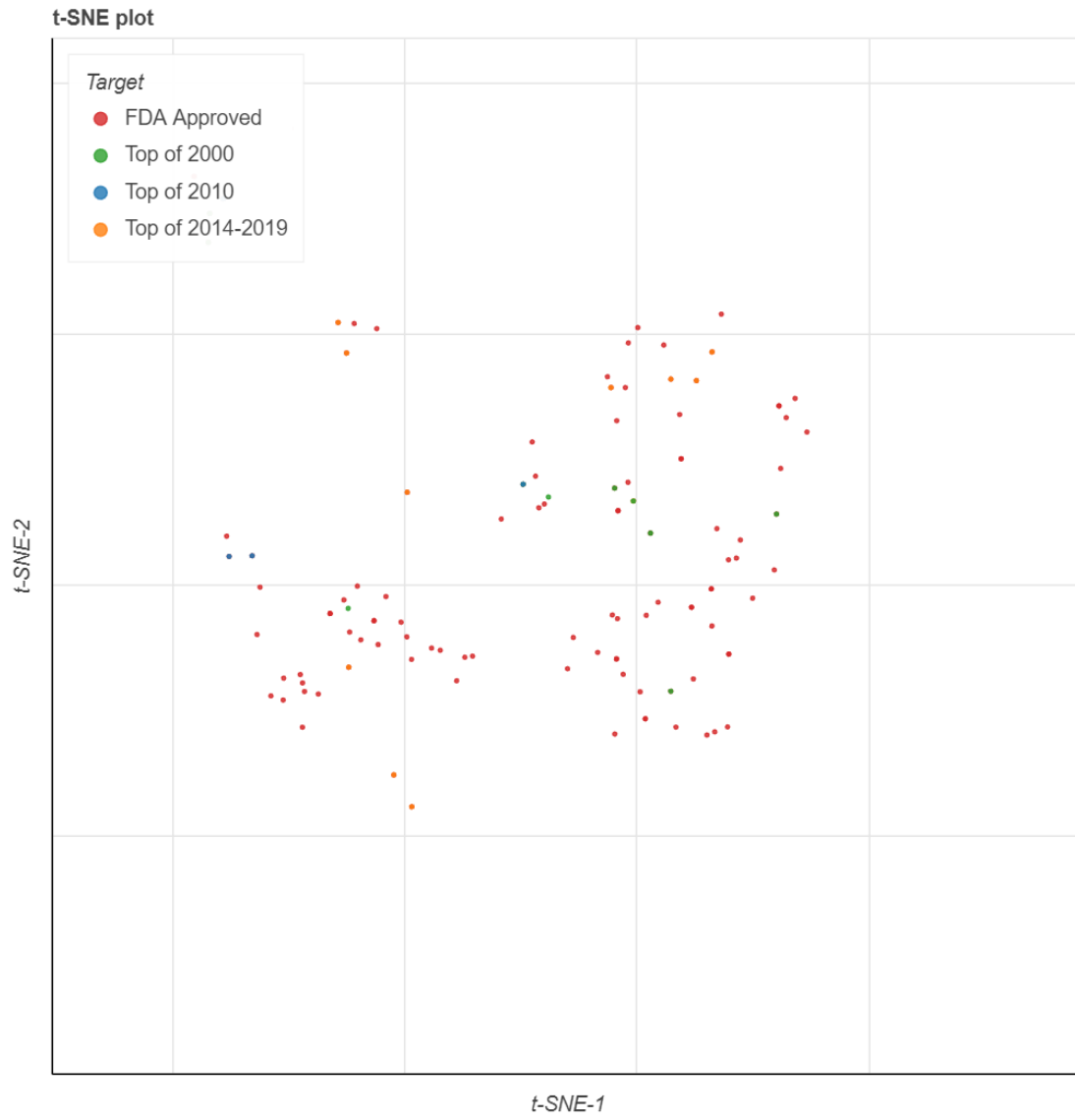
26d

Klaster 4



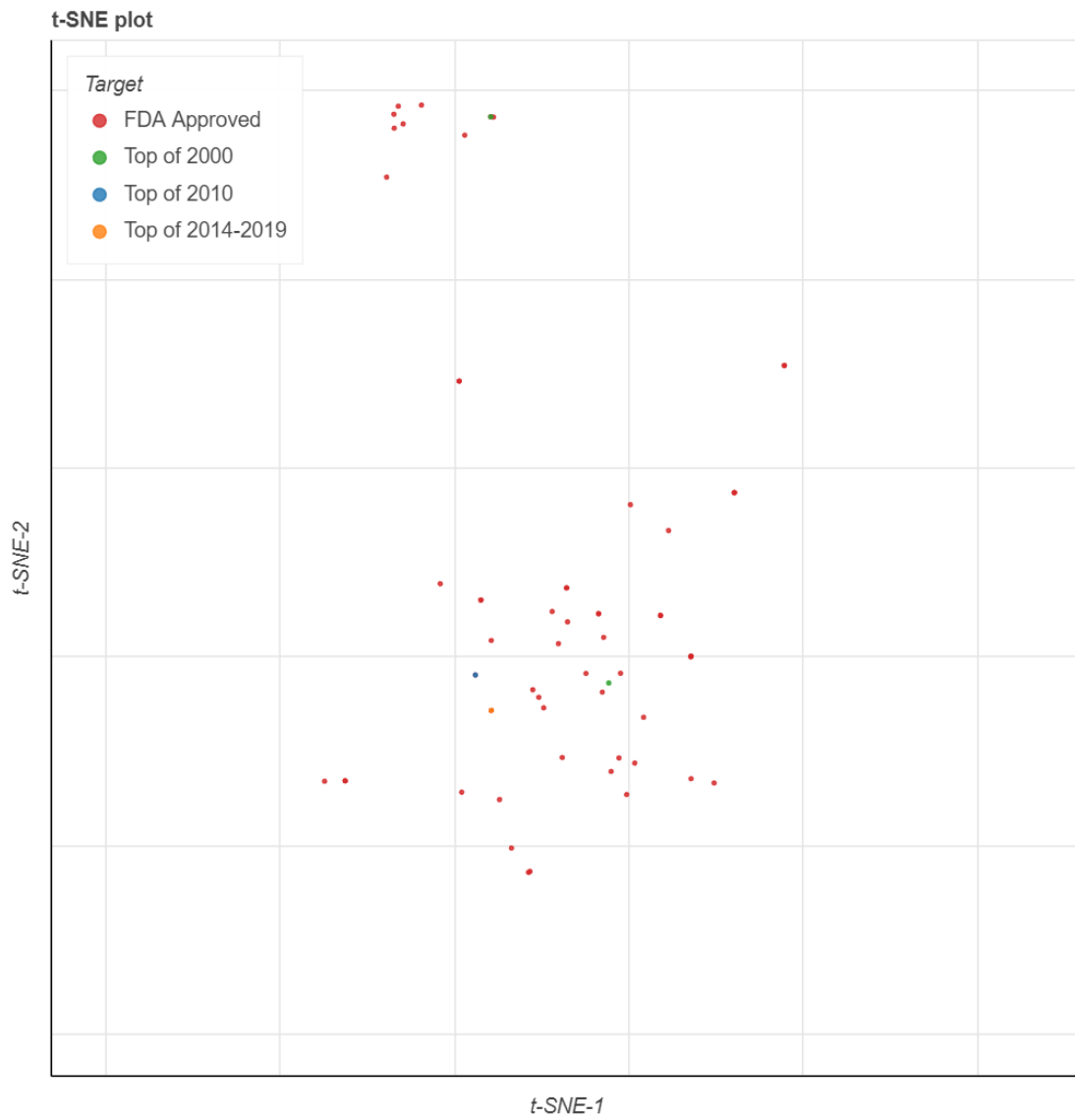
26e

Klaster 5



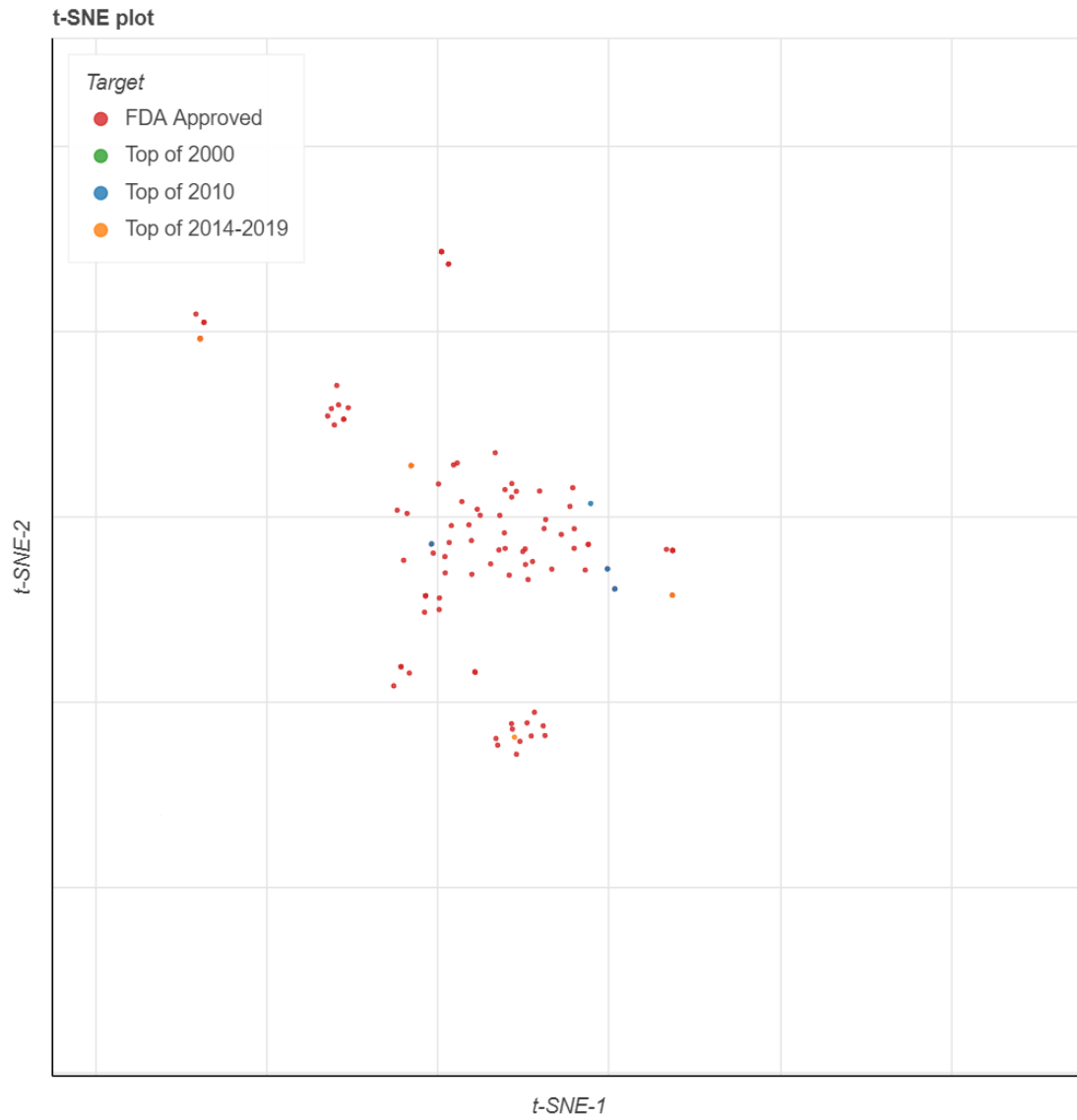
26f

Klaster 6



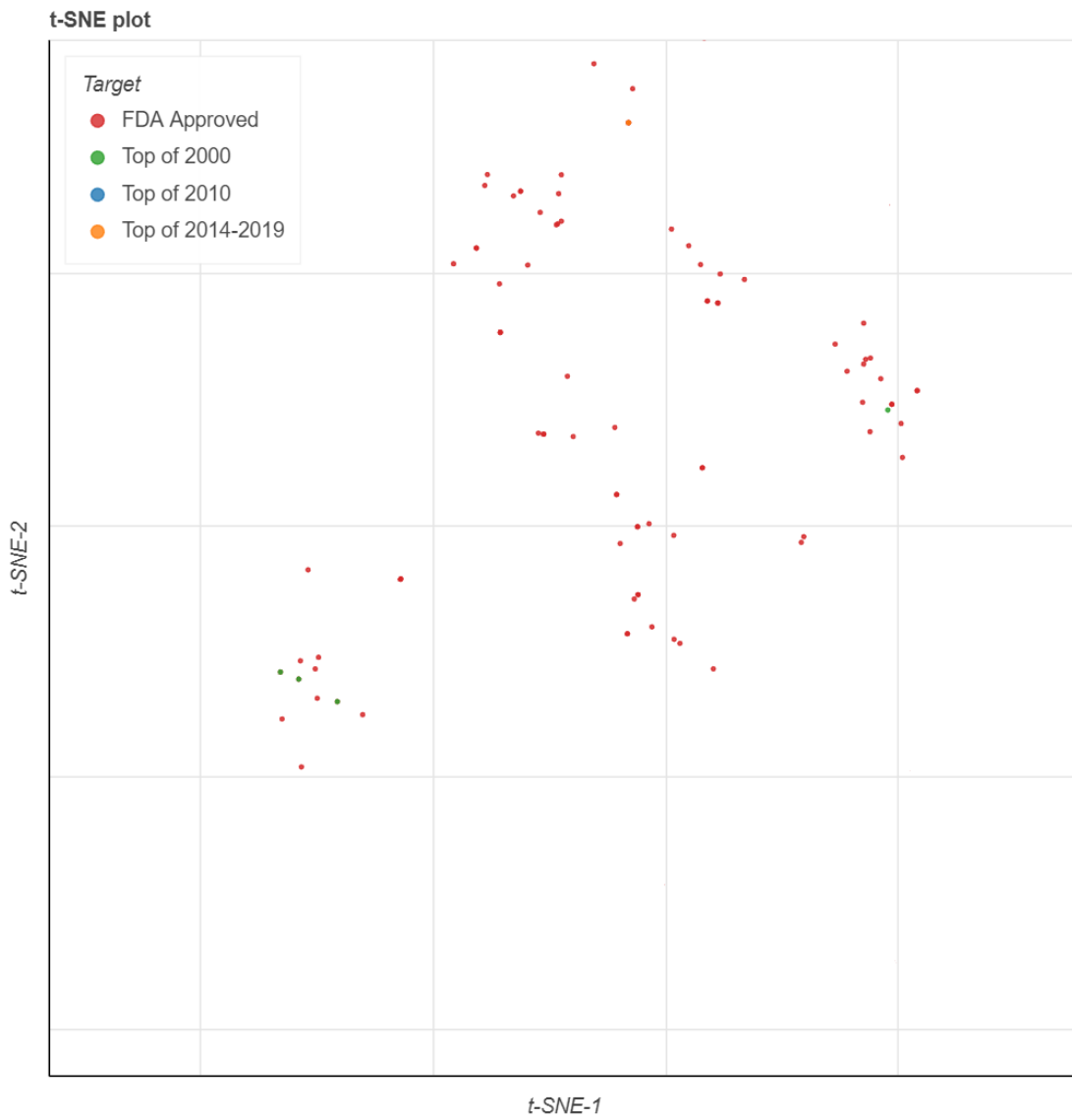
26g

Klaster 7



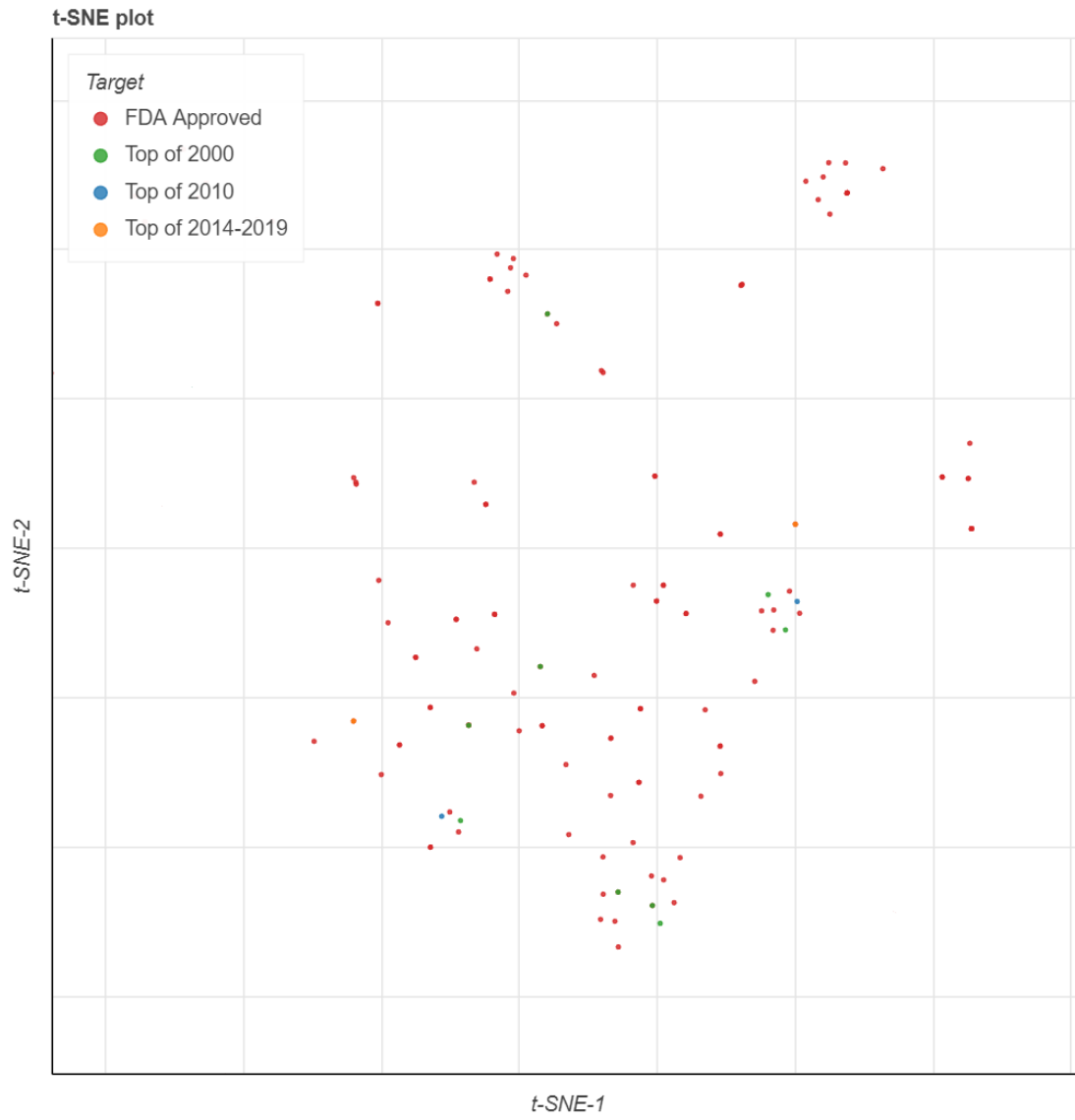
26h

Klaster 8



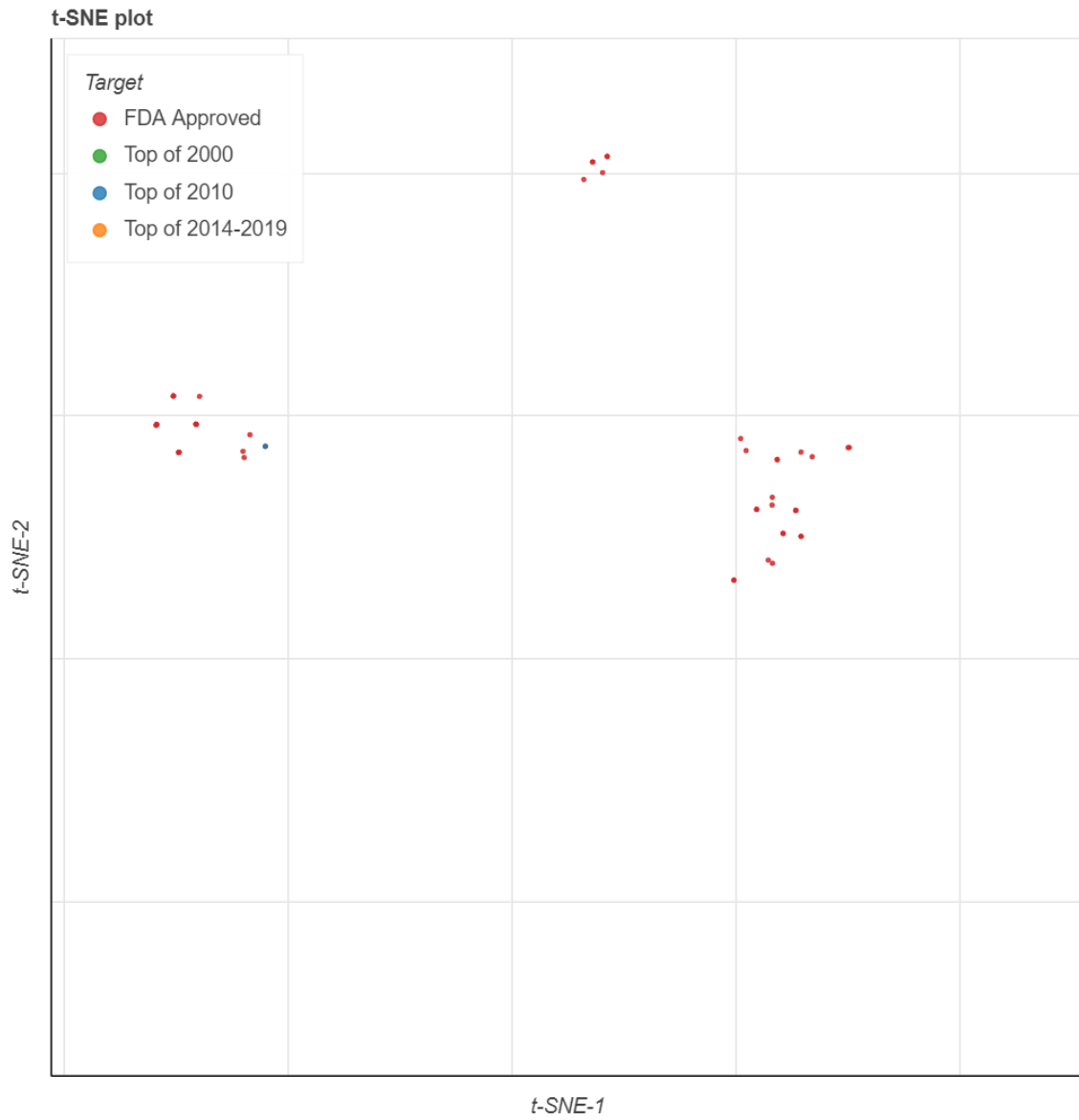
26i

Klaster 9



26j

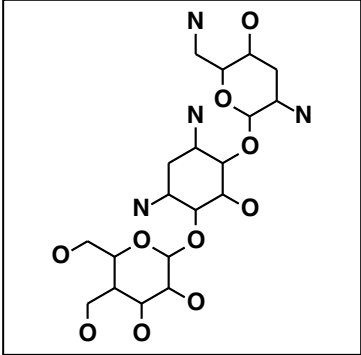
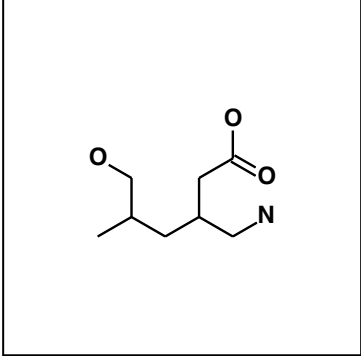
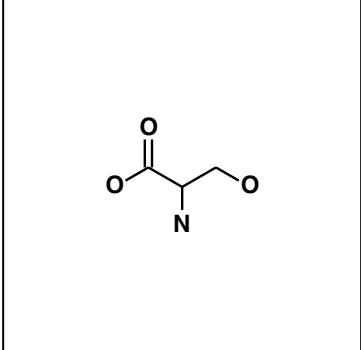
Klaster 10

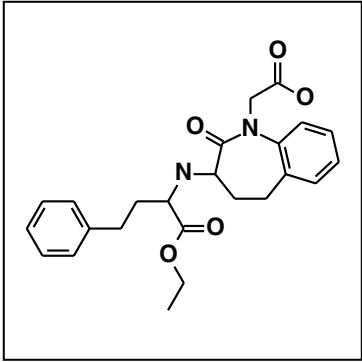


26k

Następnie dla każdego z klastrów wybrano losowo 3 związki (z różnych części klastra), które występują jednocześnie w bazie ZINC FDA oraz przynajmniej w jednej z baz TOP. Wynik przedstawiłam w poniższej tabeli.

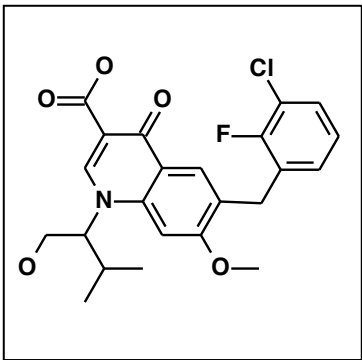
Tabela 5 Przykłady związków dla klastrów 1-10

Związek	Nazwa / numer identyfikacyjny PubChem	Klaster
	Brak danych.	1
	3-(Aminomethyl)-6-hydroxy-5-methylhexanoic acid PubChem CID: 21888190	1
	2-Amino-3-hydroxypropanoic acid PubChem CID: 617	1



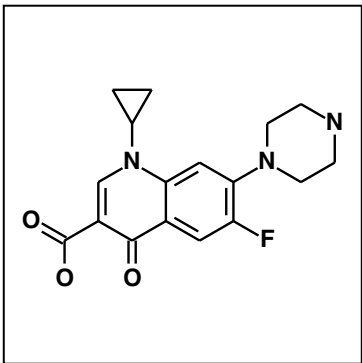
Brak danych.

2



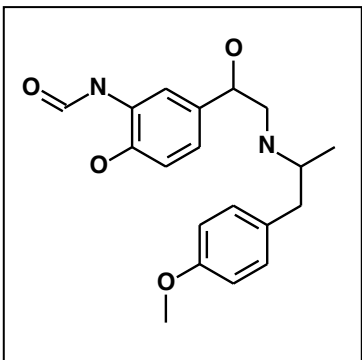
6-[(3-chloro-2-fluorophenyl)methyl]-1-[(2S)-1-hydroxy-3-methylbutan-2-yl]-7-methoxy-4-oxo-1,4-dihydroquinoline-3-carboxylic acid (Elvitegravir-d6, Major)
PubChem CID: 23083982

2



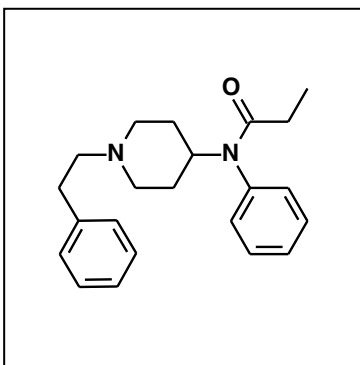
1-Cyclopropyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid (ciprofloxacin)
PubChem CID: 2764

2



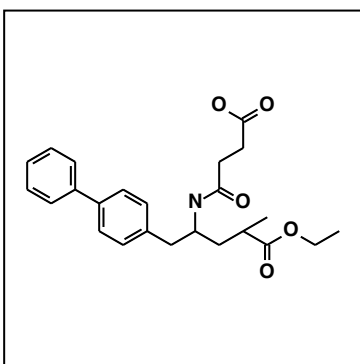
2'-hydroxy-5'-(1-hydroxy-2-(p-methoxy-alpha-methylphenethyl)amino)ethyl formanilide (formoterol)
PubChem CID: 3410

3



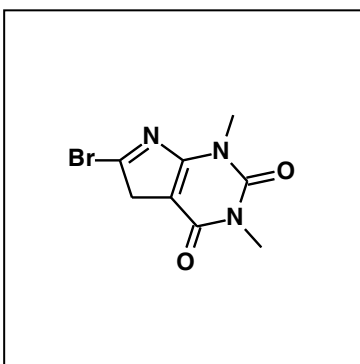
N-(1-Phenethylpiperidin-4-yl)-N-phenylpropionamide (fentanyl)
PubChem CID: 3345

3



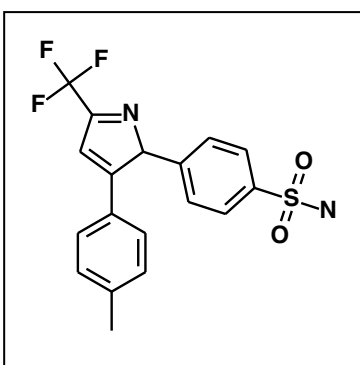
3-[(1-{[1,1'-biphenyl]-4-yl}-5-ethoxy-4-methyl-5-oxopentan-2-yl) carbamoyl] propanoic acid (sacubitril)
PubChem CID: 54214995

3



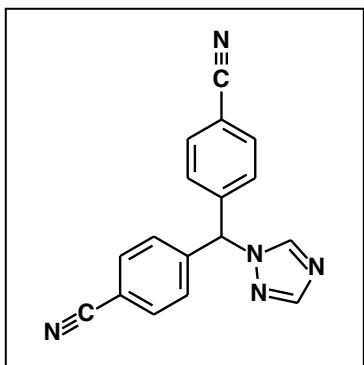
6-bromo-1,3-dimethyl-5H-pyrrolo[2,3-d]pyrimidine-2,4-dione
PubChem CID: 157349162

4



Brak danych.

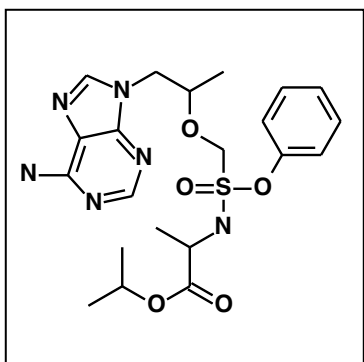
4



4,4'-((1H-1,2,4-triazol-1-yl)methylene)dibenzonitrile
(letrozole)

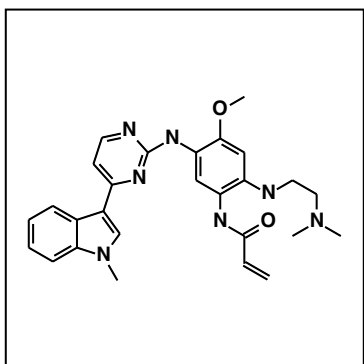
4

PubChem CID: 3902



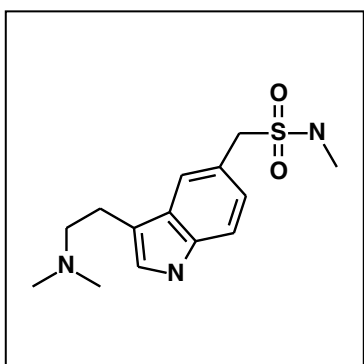
Brak danych.

5



Brak danych.

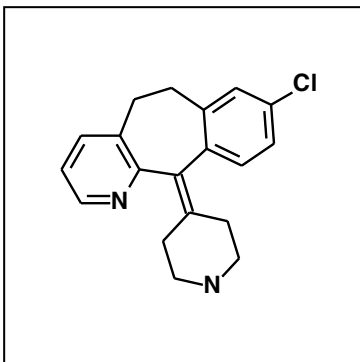
5



3-(2-(Dimethylamino)ethyl)-N-methyl-1H-indole-5-methanesulfonamide (sumatriptan)

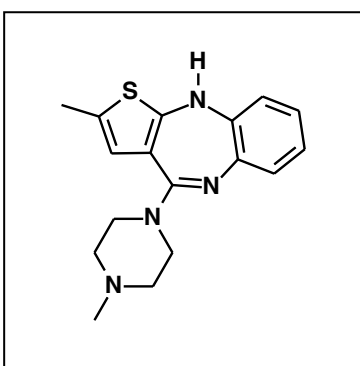
5

PubChem CID: 5358



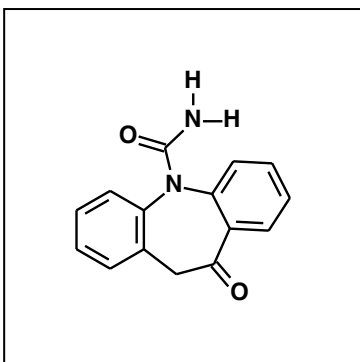
8-Chloro-11-piperidin-4-ylidene-
6,11-dihydro-5H-
benzo[5,6]cyclohepta[1,2-
b]pyridine (desloratadine)
PubChem CID: 124087

6



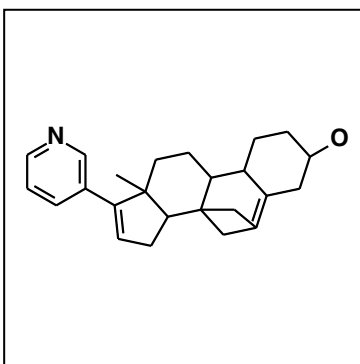
2-Methyl-4-(4-methylpiperazin-1-
yl)-10H-benzo[b]thieno[2,3-
e][1,4]diazepine (olanzapine)
PubChem CID: 135398745

6



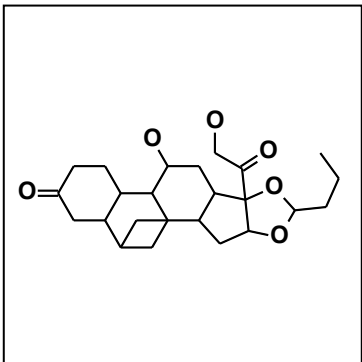
10-Oxo-10,11-dihydro-5H-
dibenzo[b,f]azepine-5-carboxamide
(oxcarbazepine)
PubChem CID: 34312

6



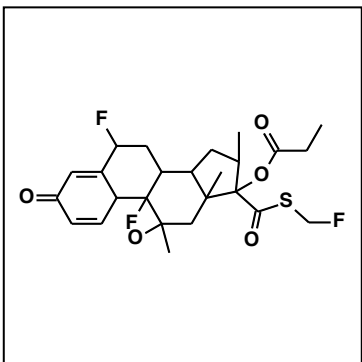
Brak danych.

7



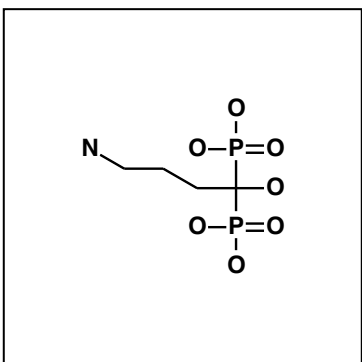
Brak danych.

7



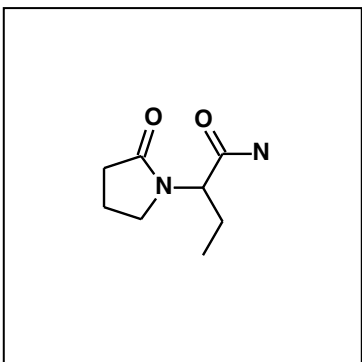
Brak danych.

7



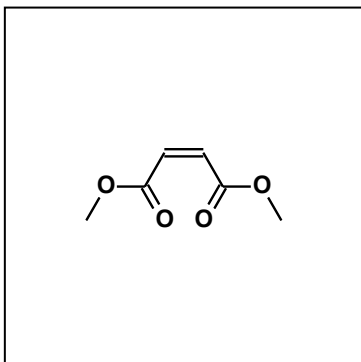
(4-amino-1-hydroxy-1-phosphonobutyl) phosphonic acid
(alendronic acid)
PubChem CID: 2088

8



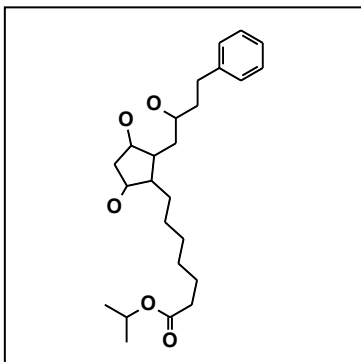
alpha-Ethyl-2-oxo-1-pyrrolidineacetamide
(etiracetam)
PubChem CID: 59708

8



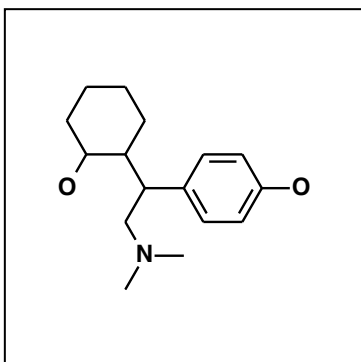
dimethyl but-2-enedioate
PubChem CID: 12215

8



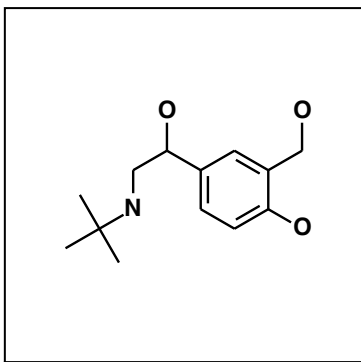
Brak danych.

9



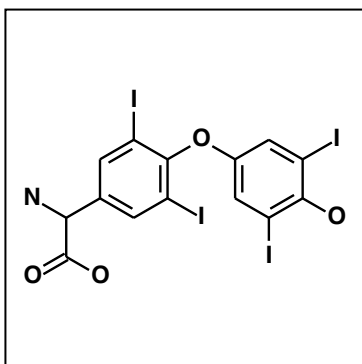
4-[2-(Dimethylamino)-1-(2-
hydroxycyclohexyl)ethyl]phenol
PubChem CID: 59520534

9



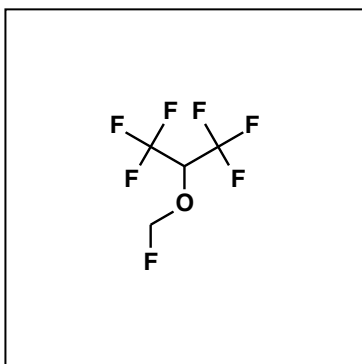
4-[2-(tert-Butylamino)-1-
hydroxyethyl]-2-
(hydroxymethyl)phenol
(salbutamol)
PubChem CID: 2083

9



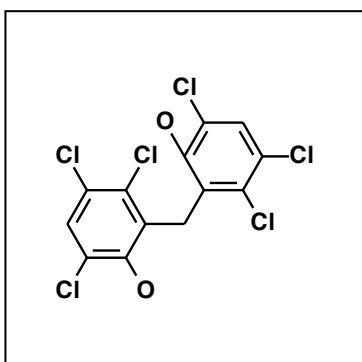
2-Amino-2-[4-(4-hydroxy-3,5-diiodophenoxy)-3,5-diiodophenyl]acetic acid
PubChem CID: 23449

10



1,1,1,3,3,3-Hexafluoro-2-(fluoromethoxy)propane (sevoflurane)
PubChem CID: 5206

10



6,6'-Methylenebis(2,4,5-trichlorophenol) (hexachlorophene)
PubChem CID: 3598

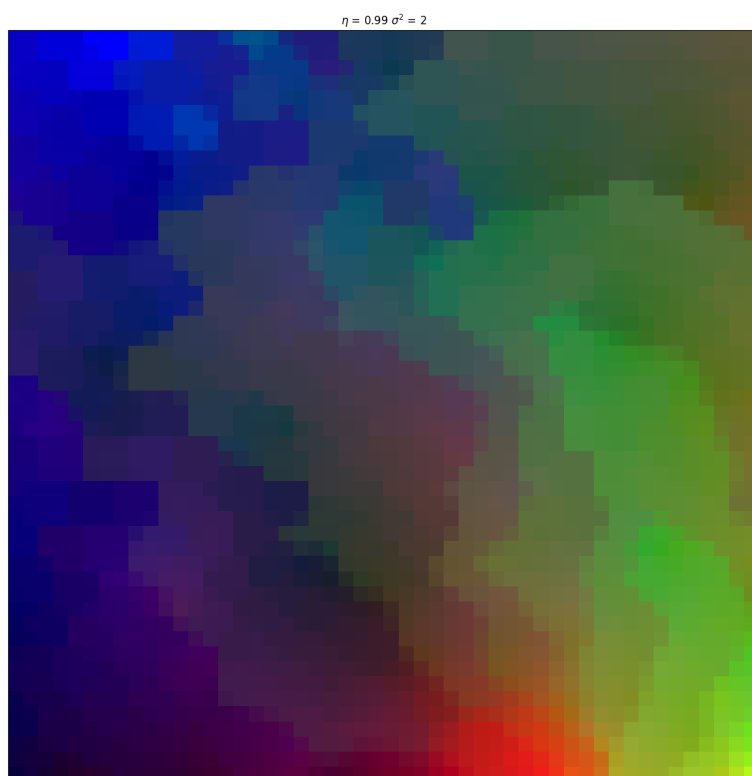
10

Ostatni etap analizy fragmentów FDA i TOP stanowiło przygotowanie map samoorganizujących SOM. Kolor czerwony wskazuje na obecność wielu grup funkcyjnych, zielony na obecność wielu atomów tlenu podczas gdy niebieski oznacza obecność wiązań pojedynczych (rysunek 27a - d).

Analiza uzyskanych profili fragmentacyjnych list TOP stanowić będzie podstawę dalszych badań. Na obecnym etapie trudno rozpoznać możliwość wykorzystania takiej metody do projektowania potencjalnych leków typu TOP. Nie ulega jednak wątpliwości, że profile

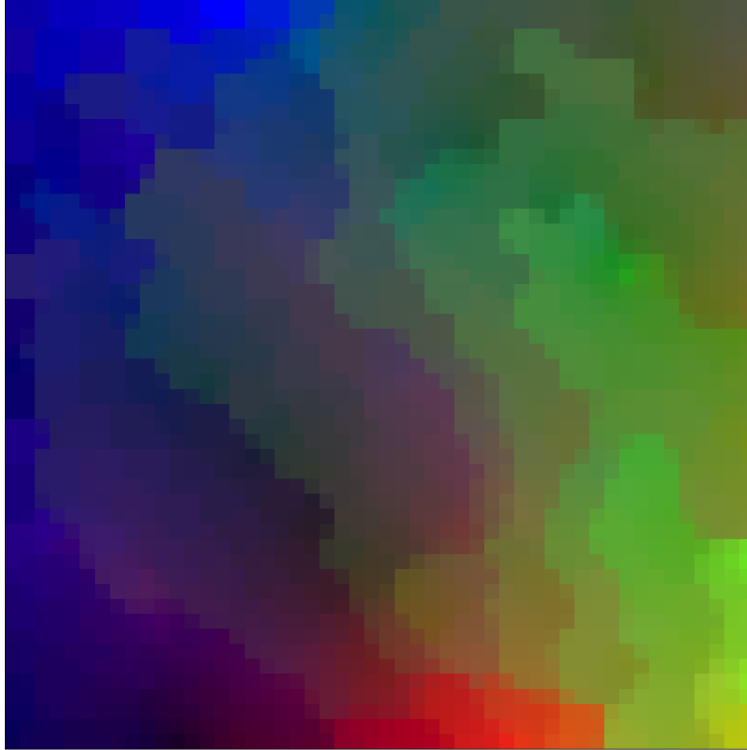
fragmentacyjne TOP oraz innych grup leków różnią się, podobnie jak w przypadku fragmentów uprzywilejowanych w farmacji. Nie jest natomiast jasne, co jest przyczyną występujących różnic.

Rysunek 27a - d Mapy SOM dla zbioru FDA *approvals* (a) oraz TOP 2000-2009 (b) 2010-2019 (c) i 2014-2019 (d). Kolory kodują obecność grup funkcyjnych. Kolor czerwony wskazuje na obecność wielu grup funkcyjnych, zielony na obecność wielu atomów tlenu podczas gdy niebieski oznacza obecność wiązań pojedynczych.



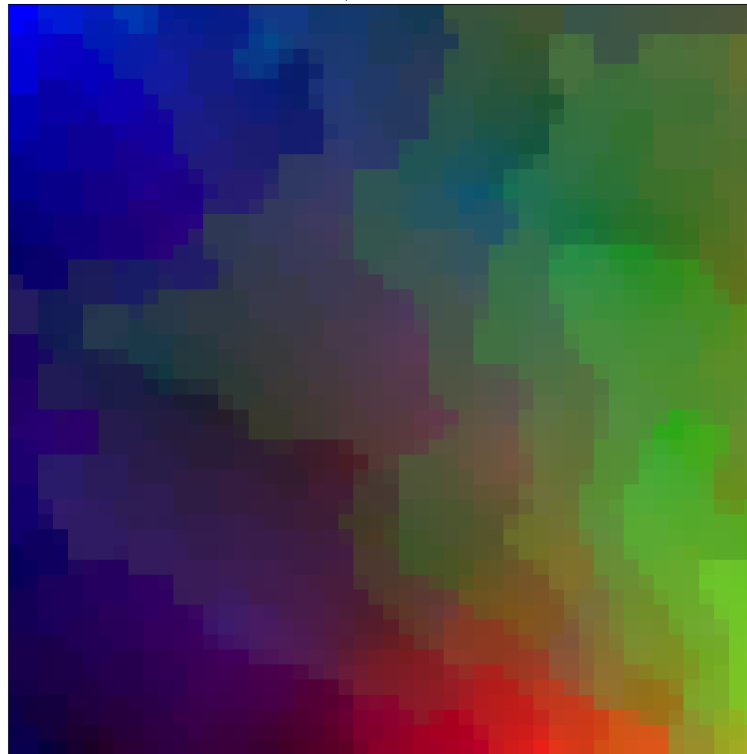
27a

$$\eta = 0.99 \quad \sigma^2 = 2$$

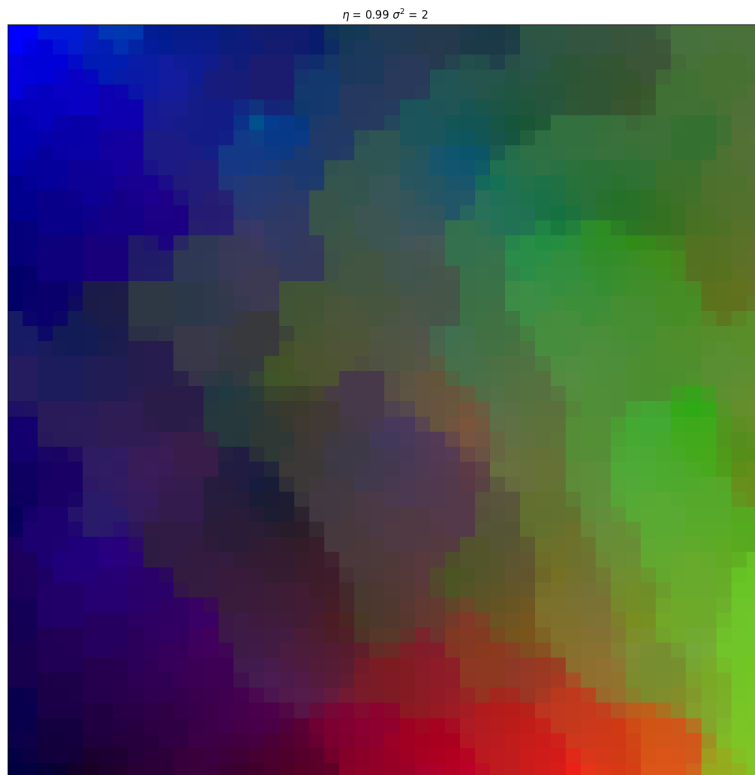


27b

$$\eta = 0.99 \quad \sigma^2 = 2$$



27c



27d

5.4 Fragonomika fotoreagentów

5.4.1 Wstęp teoretyczny

Reakcje kwasowo-zasadowe są podstawą reaktywności chemicznej. Ich „motorem napędowym” jest kwasowość lub zasadowość cząsteczek, wyrażona przez ich pK_a, czyli ujemny logarytm o podstawie 10 stałej dysocjacji kwasu.

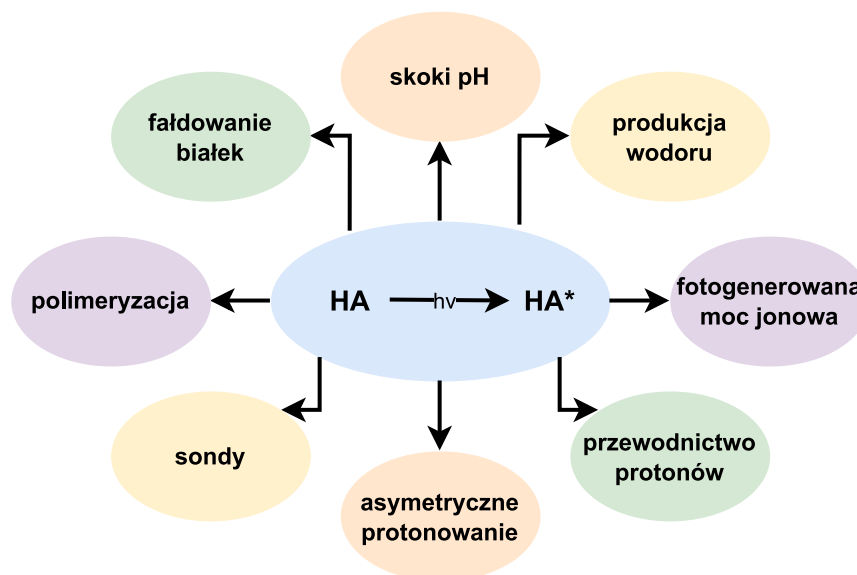
Przeprowadzanie trudnych reakcji, takich jak aktywacja wiązań C-H, C-F lub C-O, wymaga obecnie stosowania silnych kwasów oraz zasad. Te odczynniki są nieprzyjazne dla środowiska, działają selektywnie z ograniczoną gamą rozpuszczalników, wymagają stechiometrycznego dodawania środków neutralizujących i tworzą masowe ilości odpadów trudnych do przechowywania oraz neutralizacji. Alternatywę stanowi zastąpienie klasycznych kwasów i zasad fotoreagentami.

Pojęciem „fotoreagent” (ang. *photoreagent*, PR) określa się każdą cząsteczkę ulegającą fotoreakcji, czyli reakcji absorpcji promieniowania elektromagnetycznego w wyniku której, utworzony zostaje stan wzbudzony (ang. *excited state*) oraz następuje gwałtowna zmiana pH. Wśród fotoreagentów wyróżnia się:

- a) Fotokwasy (ang. *photoacids*, PAHs) – ulegają reakcji fotodysocjacji w sposób odwracalny (przyłączenie protonu, reasocjacja, następuje w wyniku reakcji termicznej). Te z nich, które są wystarczająco silne, aby fotodysocjować w roztworach niewodnych, nazywane są superfotokwasami. W stanie podstawowym wykazują niską bądź średnią kwasowość natomiast w wyniku naświetlania ich kwasowość znacznie wzrasta np. przejście ze stanu podstawowego w wzbudzony 7-hydroksy-4-(trifluorometylo)kumaryny powoduje zmianę jej pK_a z 7.3 do -8.9 [110, 111].

Fotokwasy mają szerokie zastosowanie w procesach technologicznych, m.in. w wywoływaniu fotogenerowanej mocy jonowej (gradienty jonów), produkcji wodoru, czujników chemicznych, laserów przenoszących proton, organicznych diod elektroluminescencyjnych, membran jonowymiennych pobudzanych barwnikami, polimerów przewodzących. Ponadto, wykorzystuje się je w syntezie organicznej (otwieranie pierścieni, polimeryzacja, asymetryczne protonowanie) oraz badaniach przesiewowych (skoki pH, proces fałdowanie białek, wewnątrzkomórkowe wykrywanie pH) [112].

Rys. 28 Zastosowanie fotokwasów [111]



Superfotokwasy są szczególnie interesujące, ponieważ oferują uniwersalną funkcjonalność chemiczną. Związki te niosą duży potencjał rozwoju przemysłu, np. w opracowaniu nowych metod produkcji wodoru w drodze rozszczepiania wody, fotopolimeryzacji i zwiększenia wydajności procesów katalitycznych. Te z kolei pomogą zrewolucjonizować obszar energii odnawialnej, ochrony i rekultywacji środowiska [112].

b) Generatory fotokwasów (ang. *photoacid generators*, PAGs) – ulegają fotoreakcji w sposób nieodwracalny, z utworzeniem kwasu [111]. Można je podzielić na:

- jonowe PAGs, czyli układy kationowo – anionowe. Rolę kationu pełnią najczęściej sole aryldiazoniowe [113], triarylosulfoniowe [114 - 116], triarylofosfonowe [117 - 119], a przeciwjonu – halogenki metali [120]. Ich fotodysocjacja zachodzi drogą rodnikowania i zazwyczaj wymaga obecności protonu z otoczenia. Uwolniony zostaje kwas Brönsteda lub Lewisa. Są rzadko stosowane ze względu na wąski zakres długości promieniowania, w którym możliwe jest ich użycie (obecnie zastosowanie znajdują głównie w litografii).

- niejonowe PAGs – estry benzytowe, iminoestry, spiropirany i teraryleny. Pod wpływem światła uwalniają związki kwasowe, drogą fotolitycznej dysocjacji wiązań CO, SO i NO, z jednoczesnym utworzeniem strukturalnie stabilizowanych rodników. Zazwyczaj w procesie niezbędna jest możliwość ekstrakcji wodoru z rozpuszczalnika protonowego [120]. W porównaniu do generatorów jonowych, wykazują niższą stabilność temperaturową i mogą być stosowane w szerszym zakresie rozpuszczalników oraz matryc polimerowych. Przez ostatnie 3 dekady, niejonowe PAG były najbardziej rozwijaną gałęzią fotoreagentów, stąd też są najczęściej stosowanymi fotoreagentami w przemyśle, w tym do inicjacji polimeryzacji oraz procesów utwardzania powierzchni.

Generatory fotokwasów ulegają reakcjom wzbudzenia w sposób nieodwracalny. Zgodnie z ideą wprowadzania „recyklingu chemicznego” w możliwie największą ilość syntez, należy zastąpić je odwracalnymi fotokwasami, tak by aktywnie zredukować ilość odpadów poreakcyjnych.

- c) Fotozasady i generatory fotozasad (ang. *Photobases*, PBs, *Photobase Generators*, PBGs) – wykorzystują energię fotonów, aby po fotowzbudzeniu gwałtownie zmienić swoją zasadowość ($pK_a^* \gg pK_a$). Stanowią najsłabiej opisaną grupę fotoreagentów w literaturze, o najmniejszym zastosowaniu w przemyśle (dla przykładu, istnieje tylko 1 superfotozasada, FR0-SB [112, 121]). Ich potencjalne zastosowanie w porównaniu do PAG i PAH jest znikome [122], nie stanowią zatem przedmiotu badań tej rozprawy doktorskiej.

Właściwości układów chemicznych - w tym fotoreagentów - można dostosować i znacznie poprawić pod kątem wydajności, opracowując inżynierię ich struktur molekularnych. Przestrzeń chemiczna jest jednak zbyt duża, a ponadto w przypadku fotoreagentów zbyt skąpo opisana w literaturze (liczbę opisanych w literaturze PAHs i PAGs można oszacować na $10^2 - 10^3$, dane według bazy Web of Science oraz PubChem), by można było ją skutecznie

eksplorować wyłącznie za pomocą metod eksperymentalnych. Alternatywę stanowi wykorzystanie nowoczesnej nauki o danych jako obiecującej ścieżki prowadzącej ku złożonym wyzwaniom związanym z projektowaniem w zakresie chemii i materiałów nowej generacji. Pomimo atrakcyjności tej potencjalnej ścieżki rozwoju, analiza *big data* nie stanowi jeszcze powszechnej praktyki w laboratoriach chemicznych i materiałowych.

Problem ten jest przedmiotem ostatniego, czwartego celu badawczego mojej rozprawy doktorskiej. Podjęłam próbę stworzenia dla fotoreagentów eksperymentu fragonomicznego, inspirowanego projektem przeprowadzonym z sukcesem dla materiałów OLED, przez zespół badawczy dr Rafaela Gómez-Bombarelli'ego [44, 123], mający na celu wyznaczenie nowych ścieżek syntezy TADF metodami chemoinformatycznymi (ang. *Thermally Assisted Delayed Fluorescence materials*). Czteroetapowy proces składał się z:

- a) Utworzenia bibliotek cząsteczek oraz bazy fragmentów charakterystycznych dla TADF
- b) Obliczeń chemii kwantowej
- c) *Supervised Machine learning*
- d) Porównania wyników uzyskanych metodą *in silico* z wynikami uzyskanymi eksperymentalnie w celu utworzenia i kalibracji chemoinformatycznego narzędzia.

Rezultatem projektu było wyłonienie spośród 2 milionów związków (z użyciem skryningu wirtualnego) 450 tysięcy potencjalnych kandydatów TADF, na podstawie bazy fragmentów charakterystycznych dla danej grupy. Następnie dla wyłonionych cząsteczek wykonano obliczenia kwantowe (TD-DFT), z wykorzystaniem sieci neuronowej. Uzyskane wyniki zostały zebrane i opublikowane przy użyciu strony internetowej. Zespół liczący od 2 do 10 chemików specjalizujących się w syntezie organicznej oraz inżynierów materiałowych dokonywał ewaluacji opublikowanych cząsteczek *online* przez okres jednego miesiąca (projekt typu CDD). Następnie cząsteczki, które uzyskały najwyższe wyniki zostały zebrane i przedstawione zespołowi w celu ponownej oceny. Przeprowadzono głosowanie i wyłoniono cztery potencjalne cele syntezy (*candidates*), które następnie zsyntetyzowano. Uzyskane związki poddano badaniom, które potwierdziły wcześniejsze wyniki obliczeń. Wszystkie utworzone tą drogą cząsteczki spełniły wymogi przynależności do TADF.

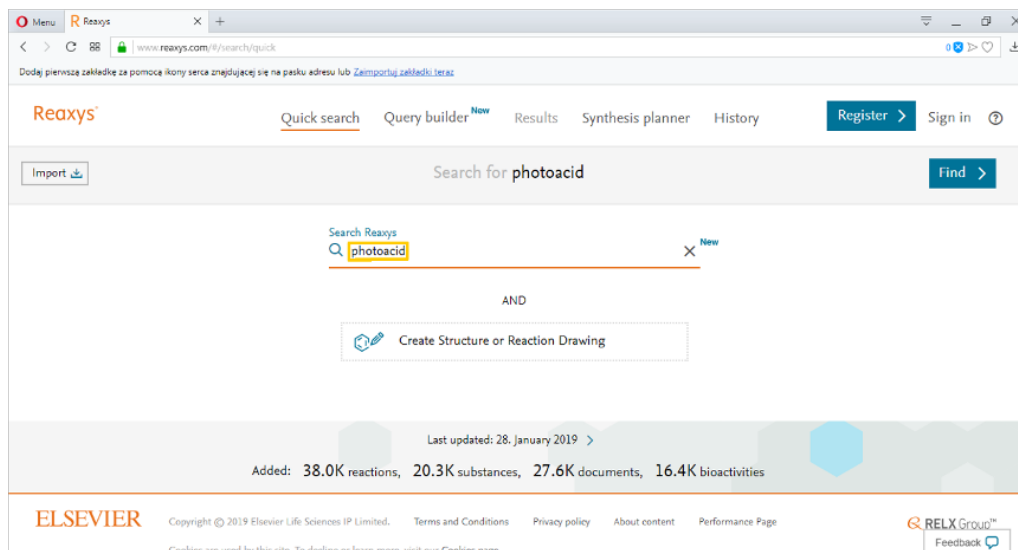
5.4.2 Metodologia

Przy rozpoczęciu gromadzenia danych dla PAHs i PAGs zauważyłam, iż ilość dostępnych baz materiałów jest znacznie niższa w porównaniu do baz danych dotyczących leków. Nie istnieje również żadna baza danych, dedykowana tej grupie związków. Popularne i ogólnodostępne bazy takie jak PubChem, ChEMBL czy Reaxys nie posiadają kart substancji przystosowanych do przechowywania właściwości charakterystycznych fotoreagentów takich jak pK_a stanu wzbudzonego czy ESPT (ang. *Excited State Proton Transfer*, transfer protonu w stanie wzbudzonym). Można je odszukać np. poprzez manualne przeszukiwanie tekstu z wykorzystaniem słowa kluczowego np. *photoacid*, pK_a , ESPT na podstawie publikacji dotyczących fotoreagentów, zawartych w katalogu Reaxys. Choć liczba artykułów o danej tematyce nie jest wysoka (w Reaxys: 2,181 artykułów dotyczących PAH i PAG, na dzień 01.02.2019 r.) proces przeszukiwania jest czasochłonny i wymaga utworzenia przez użytkownika własnej bazy danych np. w arkuszu Excel.

Tak więc celem moich badań opisanych o w obecnej pracy była kwerenda literatury w celu identyfikacji i archiwizacji danych PAHs i PAGs. W celu utworzenia bazy fotoreagentów, zawierającej struktury chemiczne w postaci kodów SMILES oraz właściwości charakterystycznych, wykorzystałam katalog Reaxys (dostępny online na stronie <https://www.reaxys.com/#/search/quick>). Filtrowanie przeprowadziłam w poszukiwaniu obiektów posiadających zadany parametr tj. zawierających w tytule, abstrakcie lub słowach kluczowych hasło „*photoacid*” (tym samym uzyskane wyniki dotyczą zarówno PAH jak PAG, rys. 27).

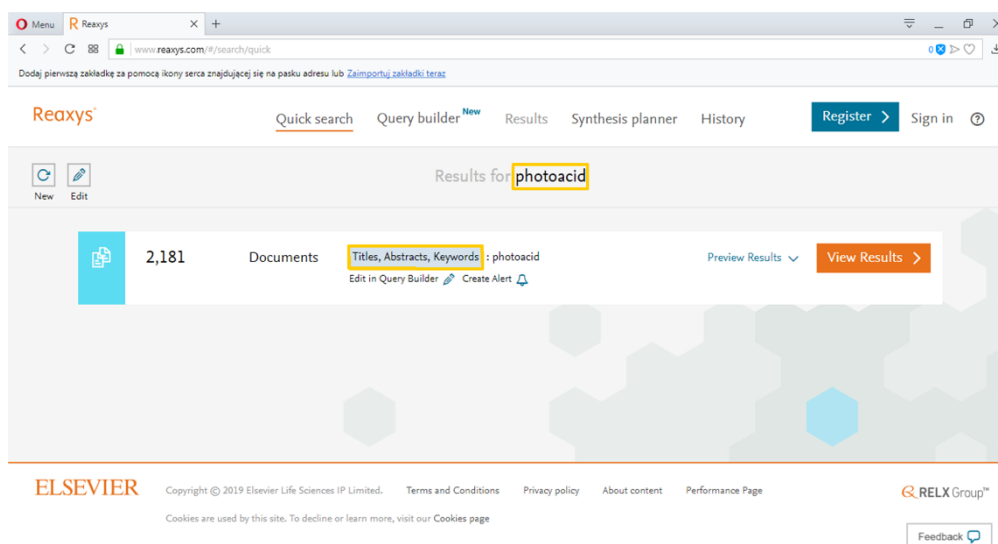
W załączniku 3 przedstawiłam bazę danych związków PAHs i PAGs oraz kodujących ich SMILESów.

Rys. 29 Zrzut ekranu przedstawiający katalog Reaxys wraz z wyszukiwarką i wpisanym parametrem filtrującym

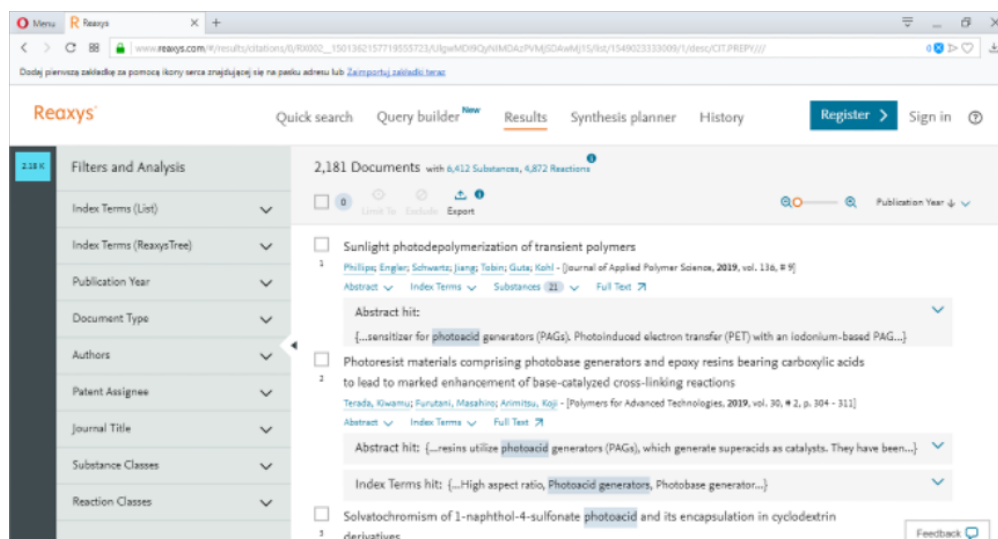


Uzyskałam 2,181 rekordów (Rys. 28, 29), które następnie poddałam manualnemu skryningowi, Rys. 30) w celu uzyskania informacji dotyczących struktury oraz właściwości fizykochemicznych fotoreagentów. Niestety, pomimo iż Reaxys posiada funkcję ekstrakcji z tekstu danych dotyczących wspomnianych w nim substancji i umieszczania ich w zakładce artykułu pt. „*substances*” w większości przypadków metoda okazała się nieskuteczna dla fotoreagentów. Stąd wynikała konieczność manualnego przeszukiwania artykułu, typowania informacji istotnych, a następnie wyszukiwanie brakujących danych (np. w katalogu PubChem) i gromadzenie całości materiału w odrębnej bazie danych (Rys. 30).

Rys. 30 Zrzut ekranu przedstawiający liczbę uzyskanych rekordów



Rys. 31 Zrzut ekranu przedstawiający pełną listę uzyskanych wyników



Rys. 32 Zrzut ekranu przedstawiający przykładowy artykuł przeszukiwany hasłem „photoacid”

Rys. 33 Zrzut ekranu przedstawiający fragment tworzonego, roboczego katalogu fotoreagentów w arkuszu

Excel

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R		
1	SMILES	Substance I	Links to	Data Co	CAS Reg	Chemical	Linear S	Molecu	Molecu	Type of	Type an	InChi K	Compos	Compos	Compos	Field An	Numbe	Nu
2	OC1=CC=C1	1821999	https://w	(2100 of 5	32743-85	1-hydroxy	C10H7O4	C10H7O4	223.229	isocyclic		HGWQOFI				Identificat	11	5
3																		
4	COC1=C(O)	19499640	https://w	(2954 of 5	42436-22	3-(3,4-dim	C17H16N	C17H16N	296.326			QKPPVFIR				Identificat	4	2
5	COC1=CC=C1	180582	https://w	(1516 of 5	5395-56-2	(E)-2-(2-(4	C18H15N	C18H15N	261.323	heterocycl		RKDUZQOI				Identificat	18	34
6	OC1=C(C)	295355	https://w	(1941 of 5	71814-95	(E)-2-(2-h	C22H16N	C22H16N	340.381	heterocycl		UINMFGI				Identificat	9	7
7	BrCOCOC1	26713387	https://w	(3036 of 5		(E)-2-(2-(4	C20H18Br	C20H18Br	368.273			ALWHVFG				Identificat	6	2
8	NC(=N)C1	2833460	https://w	(1774 of 5	58200-88	6-Amidinc	C11H10N	C11H10N	186.213	isocyclic		ULKSSXO				Identificat	65	12
9	O=C(C)C1	2480846	https://w	(2352 of 5	38968-79	2'-<1.5-Dij	C26H20O	C26H20O	380.443	isocyclic		KCHGTHQ				Identificat	1	3
10	O=C1C2=C	10393169	https://w	(2671 of 5		2,6-diphen	C26H20O	C26H20O	380.443	isocyclic	trans	JCRKSTFTC				Identificat	3	2
11	O=C1C2=C	10398223	https://w	(2672 of 5		2-(4-cyanc	C27H19N	C27H19N	405.453	isocyclic	cis	ODPHVGI				Identificat	1	2
12		1563677			876-83-5	2-methyl-l	C10H8O2	C10H8O2	160.172	isocyclic		YBTCNRYWGDPP-UHFFFAOYSA-N						
13		2339963			19294-95	(2R,6S)-2,	C26H20O	C26H20O	380.443	isocyclic		JCRKSTFTQLWFT-ZRZAMGCNSA-N						
14	COC1=CC=C	22235913	https://w	(2995 of 5	1359850-		C27H20O	C27H20O	408.453			ZIAOOPQY				Identificat	2	2
15	OC1=C2C	2196847	https://w	(1263 of 5	27928-00	8-hydroxy	C16H10O	C16H10O	458.447	isocyclic		OBIOZRVS				Identificat	13	91
16																		
17	OC1=CC=C	969616	https://w	(28 of 50	108-95-2	phenol	C6H5(OH)	C6H6O	94.113	isocyclic		ISWSIDIOK				Identificat	40821	34

5.4.3 Wyniki

Z literatury udostępnionej w bazie Reaxys wyekstrahowałam 166 unikatowych fotoreagentów (w tym 140 fotokwasów i 43 generatorów fotokwasów). Korzystając z chemicznego repozytorium online, uzupełniłam brakujące dane w tym kody SMILES, nazwę, strukturę liniową, masę cząsteczkową, identyfikatory InChI i PubChem CID. W utworzonej bazie znalazły się także właściwości charakterystyczne takie jak pKa i pKa* (pKa stanu wzbudzonego). Ponadto, uwzględniłam w niej odnośniki do literatury źródłowej. Tak przygotowane repozytorium zostały przekazane do umieszczenia w ogólnodostępnej bazie *Catalytic Material Database*, w celu jej rozbudowy i zaadresowania problemu braku dostępności repozytorium dedykowanego fotoreagentom.

6. CZĘŚĆ EKSPERYMENTALNA

W części eksperymentalnej zawarto informacje dotyczące oprogramowania wykorzystanego do generowania, przetwarzania, gromadzenia, analizy i wizualizacji danych

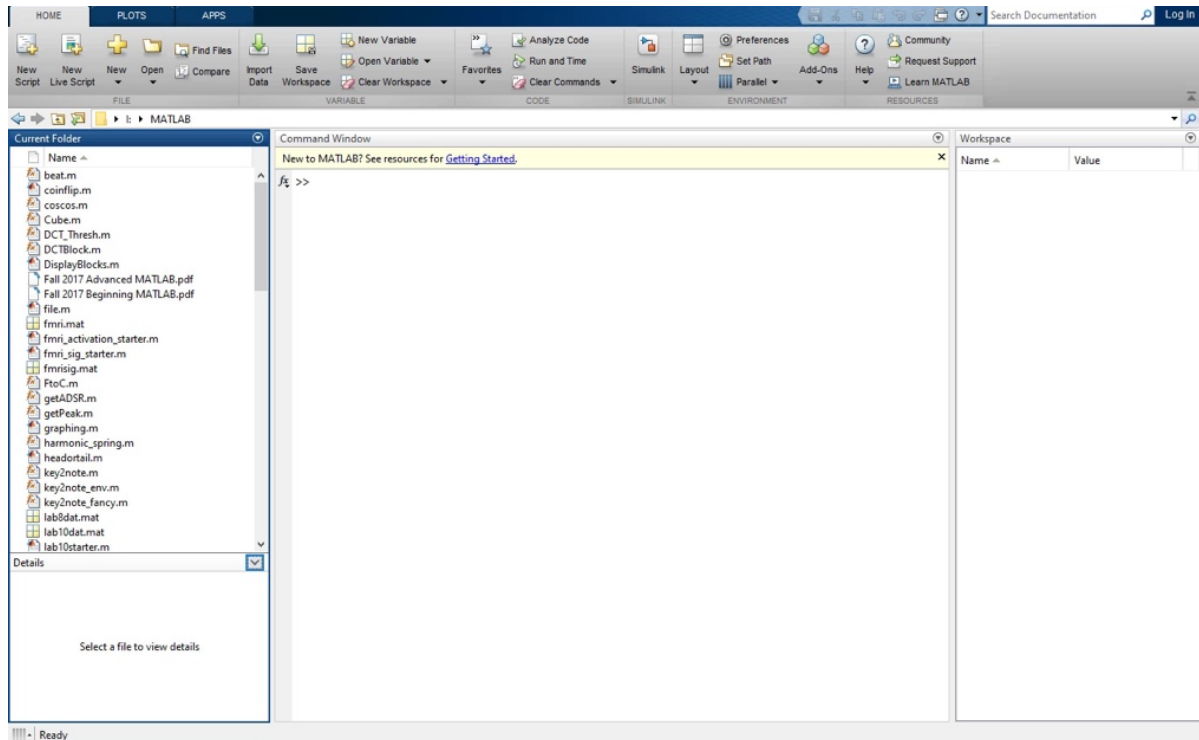
6.1 Charakterystyka oprogramowania

Podczas przygotowania rozprawy doktorskiej korzystałam z komputera osobistego o parametrach Intel Core i5-2410M (2 x 2.3 GHz), 8 GB DDR3 RAM, SSD 250 GB, system operacyjny Microsoft Windows 10 - 64 bit. W analizie danych wykorzystywałam programy takie jak MATLAB R2017a, Microsoft Office, natomiast kod źródłowy tworzyłam w Visual Studio Code, w języku programowania Python z wykorzystaniem pomocniczych pakietów i bibliotek chemoinformatycznych takich jak RDKit, w szczególności moduł Chem.QED czy ChemPlot.

6.1.1 Oprogramowanie MATLAB

Podczas opracowywania wyników korzystałam z oprogramowania MATLAB R2017a, interaktywnego środowiska do obliczeń numerycznych, analizy danych w tym ich wizualizacji, opracowywania algorytmów i tworzenia modeli. MATLAB posiada rozbudowany, intuicyjny interfejs oraz rozbudowaną dokumentację dostępną online (<https://www.mathworks.com/help/matlab/>).

Rys. 34 Interfejs programu MATLAB



Opracowałam skrypty służące do wizualizacji danych w MATLAB, aby zautomatyzować aktualizację wyników wywołaną zmianami w analizowanych zbiorach danych.

Rys. 35 Przykładowy skrypt utworzony w MATLAB, służący do automatyzacji tworzenia wykresów, omawianych w rozdziale 5.2.2

```
%determining the variables
x=[MW_SALES.MW];
y=[MW_SALES.DIG3];
z=[MW_SALES.SALESBYGROUP3];
w=[MW_SALES.SALESBYDRUG3];

%bar projection
b=bar(x,y,0.8);
b.FaceColor = [0.6,0.6,0.6];
b.EdgeColor = [0.6,0.6,0.6];

hold on

%plot by group projection
```

```

p1=plot(x,z);
p1.LineWidth =2;
p1.Color=[0.0,0.3,0.6];
sz=30;
s1=scatter(x,z,sz,'filled','HandleVisibility','off');
s1.MarkerFaceColor = [0.0,0.3,0.6];
yyaxis right;
ylim([0 6])
set(gca,'ycolor','black')
ylabel('\bf RBD [B]');

hold on

%plot by drug projection
p2=plot(x,w);
p2.LineWidth =2;
p2.Color=[1.0,0.7,0.0];
sz=30;
s2=scatter(x,w,sz,'filled','HandleVisibility','off');
s2.MarkerFaceColor = [1.0,0.7,0.0];
yyaxis left;
ylabel('\bf Population, RBG [Normalized]');

%determining the grid&axis rulers
ax=gca;
ax.GridLineStyle = '--';
ax.XGrid = 'off';
ax.YGrid = 'off';
ax.FontSmoothing = 'on';

xticklabels({'100-200','200-300','300-400','400-500','500-600','600-700','700-800','800-900','900-1000','1000-5000','5000-10000','10000-50000','50000-100000','>100000'})
xtickangle(45)

%determining the titles & labels
title('2014-2019: Population, RBG, RBD vs. MW');
xlabel('\bf MW [Da]')

```

6.1.2 Język programowania Python

Do programowania skryptów realizujących opisywane w pracy obliczenia wykorzystałam język programowania Python. Wybierając ten język, kierowałam się przede wszystkim jego uniwersalnym przeznaczeniem, dostępnością bibliotek chemoinformatycznych oraz dokumentacji. Ponadto na decyzję miało wpływ doświadczenie zdobyte przeze mnie w tym obszarze podczas przygotowywania pracy magisterskiej.

6.1.3 Oprogramowanie RDKit (biblioteki podstawowe oraz moduł RDKit.Chem.QED)

Do wygenerowania fragmentów oraz wyznaczenia deskryptorów molekularnych QED dla zbiorów TOP oraz FDA *approvals* korzystałam z pakietu narzędzi RDKit, dedykowanego obliczeniom chemoinformatycznym oraz zagadnieniom z obszaru uczenia maszynowego, opartego o otwarte oprogramowanie.

Ważnym aspektem, dla którego wybrałam RDKit jest fakt posiadania przez niego wrappera do Pythona 3.x, co umożliwiło bezproblemową integrację z wykorzystywanym wcześniej oprogramowaniem. RDKit zainstalowałam używając menadżera pakietów oprogramowania *Conda*, zawartego w dedykowanej celom naukowym dystrybucji Pythona o nazwie *Anaconda*.

Wybór RDKit był również motywowany łatwością wdrożenia tego rozwiązania (instalacja RDKit, uwzględniając odpowiednio przygotowane wcześniej środowisko Python ograniczała się do pojedynczej komendy: `conda create -c conda-forge -n my-rdkit-env rdkit`) oraz rozbudowaną dokumentacją dostępną na stronie <https://www.rdkit.org>.

6.1.4 ChemPlot

ChemPlot analogicznie do RDKit jest łatwym w instalacji (również wymaga wywołania zaledwie jednej komendy przy użyciu menadżera oprogramowania *Conda*), ogólnodostępnym pakietem narzędzi, umożliwiającym wizualizację i redukcję wymiarowości przestrzeni chemicznej posiadanych zbiorów danych (PCA, t-SNE i UMAP).

Ponadto jest to narzędzie łatwe w obsłudze nawet dla osób o małym doświadczeniu, udostępniające dwie metody wyznaczania podobieństwa między cząsteczkami (ang. *structural or tailored similarity*) [124].

Proces analizy danych z użyciem ChemPlot można podzielić na dwa etapy. Pierwszym z nich jest konwersja cząsteczek (notacja SMILES lub InChI) do binarnych macierzy o rzeczywistych wartościach. Każdy z elementów macierzy odpowiada wymiarowi w przestrzeni chemicznej.

W drugim etapie wielowymiarowa przestrzeń chemiczna zostaje zredukowana do przestrzeni dwu lub trójwymiarowej, która następnie jest przekształcana w reprezentację wizualną i wyświetlana dla użytkownika do dalszej analizy również w formie interaktywnej.

Pakiet ChemPlot wraz z dokumentacją jest dostępny na stronie <https://chemplot.readthedocs.io/en/latest/index.html>.

7. WNIOSKI

Głównym problemem jaki został zaadresowany w niniejszej rozprawie doktorskiej było opracowanie sposobu optymalizacji projektowania w zakresie chemii i materiałów, jako efekt przeniesienia na te obszary metod z sukcesem wykorzystywanych w innych, pozornie nieskorelowanych dziedzinach m.in. ekonomii czy projektowaniu leków.

W pracy zidentyfikowano trzy główne wyzwania związane z projektowaniem nowych związków chemicznych. Pierwszym z nich jest stopień skomplikowania zależności struktura - właściwość i struktura - aktywność, które rządzą zachowaniem molekularnym. Próby opisu tych zależności niekiedy utrudniają proces, jak zostało wykazane podczas analizy *ligand efficiency*.

Drugim jest nieskończoność przestrzeni chemicznej, w większości nadal niezbadanej oraz brak udostępniania wyników jej eksploracji. Dobrym przykładem jest ograniczona dostępność baz materiałów, znacznie utrudniająca wprowadzanie metod chemoinformatycznych w tej dziedzinie. Tradycyjny proces odkrywania przestrzeni chemicznej metodą prób i błędów jest czaso-, praco- i zasobochłonny, co ogranicza liczbę i różnorodność *candidates*, które mogą zostać poddane badaniom *in vivo*. Postęp w tym wypadku jest zazwyczaj powolny i stopniowy, szczególnie w przypadku zaawansowanych układów.

Trzecie wyzwanie stanowi wprowadzenie kooperacyjnego podejścia do badań z uwzględnieniem zaangażowania grup badawczych, w tym specjalistów spoza dziedziny chemii. Po dziesięcioleciach ciągłego rozwoju metod, algorytmów i sprzętu komputerowego oraz sprawnego implementowania go w pozostałych obszarach przemysłu, naszedł czas na zaangażowanie ich w procesy projektowania nowych związków chemicznych. Modelowanie i symulacja osiągnęły punkt krytyczny i znajdują się na etapie, w którym są w stanie dokonywać dokładnych, realistycznych i istotnych prognoz dla układów molekularnych. Ukierunkowane predykcje będą prowadziły do zwiększenia efektywności projektów badawczych i umożliwią wytypowanie kandydatów w sposób szybszy i bardziej korzystny pod kątem ekonomicznym. Ponadto mogą one zapewnić unikalny wgląd wykraczający poza

zakres obserwacji empirycznych, a tym samym stanowić solidną podstawę dla nowych odkryć. Zmiana w kierunku paradygmatu odkrywania i racjonalnego projektowania opartego na danych obiecuje złagodzić wiele nieefektywności i niedociągnięć, które wciąż stanowią istotne wyzwanie.

Zakres pracy rozszerzał się w konsekwencji badań nad pierwotnie podjętymi zagadnieniami. Tak oto analiza nieintuicyjnego i trudnego do interpretacji zachowania *ligand efficiency* doprowadziła nie tylko do opracowania alternatywnego narzędzia *Product Ligand Efficiency*, znacznie ułatwiającego analizę diagramów zależności aktywności od liczby atomów ciężkich, ale także zainspirowała autorkę pracy do rozszerzenia badań pod kątem fragmentacji.

Kolejne badania zostały ukierunkowane doświadczeniem zawodowym autorki (praca w charakterze analityka biznesowego w branży FinTech). Podjęto przytaczane kilkakrotnie w tej rozprawie, przyszłościowe podejście, łączące doświadczenie płynące z pozornie nieskorelowanych ze sobą dziedzin takich jak chemia, ekonomia i analiza *big data*. Umiejętności związane z procesowaniem danych w tym ich wyszukiwanie, ocena i porządkowanie umożliwiły utworzenie zbiorów TOP. Te z kolei stanowiły podstawę analizy zależności wyniku finansowego leku względem jego MW czy logP. Zestawienie wyników dla TOP z listą FDA *approvals* z lat 1985 – 2019 umożliwiło obserwacje zbieżnych efektów względem MW, logP i QED.

Podsumowując, w ramach pracy:

1. Zaproponowałam nowy analogiczny do LE parametr PLE i wykorzystałam go do analizy przebiegu pPLE od HAC oraz rozbicia pPLE (pAC₅₀ i pHAC) od HAC dla autorskiej bazy danych, utworzonej w wyniku połączenia rekordów pobranych z katalogu PubChem i ChEMBL oraz katalogów leków i ich fragmentów. Dla porównania, przygotowałam również wizualizację przebieg LE, przygotowaną na podstawie tych samych zbiorów danych.

2. Stworzyłam bazy danych TOP zawierające listę 100 najlepiej sprzedających się leków w latach 2000 – 2019. Następnie przeprowadziłam analizy statystyczno-ekonomiczne utworzonych zbiorów, w tym analizę zależności sprzedaży leków od ich właściwości chemicznych tj. MW oraz logP, będących jednocześnie ważnymi parametrami w projektowaniu leków. Kolejnym etapem było porównanie histogramów częstości MW i logP z wynikami dla FDA *approvals* z lat 1985 – 2019. Wyniki uzupełniono ponadto o analizę wieku leków względem MW oraz logP dla listy TOP oraz wyznaczenie i porównanie deskryptora molekularnego QED dla obu zbiorów (FDA *approvals*, TOP).
3. Dokonałam analizy fragmentarycznej najlepiej sprzedających się leków, typując struktury uprzywilejowane. Wynik analizy porównałam do dwóch zbiorów danych FDA *approvals* typując struktury uprzywilejowane. W kolejnym kroku wygenerowałam mapy SOM, umożliwiające porównanie rozkładu 7-atomowych fragmentów dla zbioru FDA i TOP.
4. Przeanalizowałam dostępność danych literaturowych oraz ogólnodostępnych baz danych dla fotoreagentów. Na podstawie wyniku badania, odkryłam potrzebę utworzenia bazy danych dedykowanej tej grupie związków. W tym celu dokonałam manualnego skryningu literatury dostępnej w katalogu Reaxys, wyodrębniając z niej struktury chemiczne fotoreagentów i zapisując je w postaci SMILES. Stworzona baza danych zawiera również ich właściwości fizykochemiczne. Następnie zebrane dane zostaną zintegrowane z ogólnodostępną bazą *Catalytic Material Database*, tworząc sekcję dedykowaną fotoreagentom.

8. LITERATURA

- [1] Weise T., Zapf M., Chiong R., Nebro A.J. *Why Is Optimization Difficult?* Studies in Computational Intelligence, Chiong R., Eds, Springer, 2009, 193, 1-50.
- [2] Kolmogorov A.N., *On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition* Dokl. Akad. Nauk SSSR, 1957, 114(5), 953-956.
- [3] Balcazar J.L., Gavalda R., Siegelmann H.T., *Computational power of neural networks: a characterization in terms of Kolmogorov complexity* IEEE Trans. Inf., 1997, 43(4), 1175-1183.
- [4] Schmidt-Hieber J. *The Kolmogorov–Arnold representation theorem revisited* NN, 2021, 137, 119-126.
- [5] Simmons B.I., Hoeppeke C., Sutherland W.J. *Beware greedy algorithms.* J Anim Ecol., 2019, 88(5), 804-807.
- [6] Cormen T.H., Leiserson Ch.E, Rivest R.L, C. Stein, *Wprowadzenie do algorytmów*, Wydawnictwo Naukowe PWN, 2012,
- [7] Noto M., H. Sato H., *A method for the shortest path search by extended Dijkstra algorithm* Smc 2000 conference proceedings. 2000 IEEE international conference on systems, man, and cybernetics. Cybernetics evolving to systems, humans, organizations, and their complex interactions, 2000, 3, 2316-2320.
- [8] Haiming L., Qiyang X., Yong W. *Research and Improvement of Kruskal Algorithm.* Int. J. Comput. Commun., 2017, 05(12), 63-69.
- [9] Polanski, J. *Chemoinformatics: From Chemical Art to Chemistry in Silico.* Encyclopedia of Bioinformatics and Computational Biology, Ranganathan S., Gribskov M., Nakai H., Schonbach Ch., Eds, Elsevier, 2017, 2, 601-618.
- [10] Polanski, J., Gasteiger J. *Computer Representation of Chemical Compounds.* Handbook of Computational Chemistry, Leszczynski J., Puzyn T., Reis H., Kaczmarek-Kedziera A., Shukla M.K, Papadopoulos M.G., Eds, Springer, 2017, 1997-2039.
- [11] Williams W.L., Zeng L., Gensch T., Sigman M.S., Doyle A.G., Anslyn E.V. *The Evolution of Data-Driven Modeling in Organic Chemistry.* ACS Central Science, 2021, 7 (10), 1622-1637.
- [12] Polanski J., Kucia U., Duszkiewicz R., et al. *Molecular descriptor data explain market prices of a large commercial chemical compound library.* Sci Rep, 2016, 6, 28521.

- [13] Lach D., Zhdan U., Smolinski A., Polanski J. *Functional and Material Properties in Nanocatalyst Design: A Data Handling and Sharing Problem*. Int J Mol Sci., 2021, 1322(10), 5176.
- [14] Nørskov J.K., Bligaard T., Rossmeisl J., Christensen C.H. *Towards the computational design of solid catalysts*. Nat Chem., 2009, 1(1), 37-46.
- [15] Andersson M.P, Bligaard T., Kustov A., Larsen K.E, Greeley J., Johannessen T., Christensen C.H., Nørskov J.K. *Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts*, J. Catal., 2006, 239 (2), 501-506.
- [16] El-Halwagi M.M. *Overview of optimization*, Process Systems Engineering, Academic Press, 7, 2006, 285-314.
- [17] Praca zbiorowa, *Encyclopedia of optimization*, Floudas C.A., Pardalos P.M., Eds., Springer, 2, 2008.
- [18] Why did the tech savvy Obama administration launch a busted healthcare website? <https://www.theverge.com/2013/10/8/4814098/why-did-the-tech-savvy-obama-administration-launch-a-busted-healthcare-website> [dostęp 15.05.2023 r.]
- [19] HHS failed to heed many warnings that healthcare.gov was in trouble https://www.washingtonpost.com/national/health-science/hhs-failed-to-heed-many-warnings-that-healthcaregov-was-in-trouble/2016/02/22/dd344e7c-d67e-11e5-9823-02b905009f99_story.html [dostęp 10.05.2023 r.]
- [20] Das N., Dhanawat M., Dash B., Nagarwal R.C., Shrivastava S.K., *Codrug: An efficient approach for drug optimization*. Eur. J. Pharm. Sci, 2010, 41(5), 571-588.
- [21] J. Taylor C.J., Pomberger A., Felton K.C., Rachel R., Barecka M., Chamberlain T.W., Bourne R.A., Johnson C.N., Lapkin A.A., *A Brief Introduction to Chemical Reaction Optimization*. Chem. Rev., 2023, 123 (6), 3089-3126.
- [22] Commenge, J.M., Falk, L. *Methodological framework for choice of intensified equipment and development of innovative technologies*, Chem Eng Process., 2014, 84, 109-127.
- [23] Tai R.K., Eberhard W., Buechs J., *Measurement and characterization of mixing time in shake flasks* Chem. Eng. Sci., 2011, 66 (3), 440-447.
- [24] Keith J.A., Vassilev-Galindo V., Cheng B., Chmiela S., Gastegger M., Müller K.R., Tkatchenko A. *Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems* Chem. Rev., 2021, 121 (16), 9816-9872.

- [25] Mikolajczyk A., Zhdan U., Antoniotti S., Smolinski A., Jagiello K., Skurski P., Harb M., Puzyn T., Polanski J. *Retrosynthesis from transforms to predictive sustainable chemistry and nanotechnology: a brief tutorial review* Green Chem., 2023, 25, 2971-2991.
- [26] Polanski J., *Unsupervised Learning in Drug Design from Self-Organization to Deep Chemistry* Int. J. Mol. Sci., 2022, 23(5), 2797.
- [27] Podział modeli uczenia maszynowego wraz z przykładami zastosowania, <https://www.gov.pl/web/popcwsparcie/podzial-modeli-uczenia-maszynowego-wraz-z-przykladami-zastosowania> [dostęp 18.04.2023 r.]
- [28] Petabyte - How Much Information Could it Actually Hold? <https://info.cobaltiron.com/blog/petabyte-how-much-information-could-it-actually-hold> [dostęp 10.05.2023 r.]
- [29] How Much Data Is Created Every Day in 2023, <https://webtribunal.net/blog/how-much-data-is-created-every-day/#gref> [dostęp 10.05.2023 r.]
- [30] Data as Currency: What Value Are You Getting? <https://knowledge.wharton.upenn.edu/podcast/knowledge-at-wharton-podcast/barrett-data-as-currency/> [dostęp 12.05.2023r.]
- [31] Marvin H.J., Janssen E.M., Bouzembrak Y., Peter J. M. Hendriksen P.J., *Staats M. Big data in food safety: An overview* Crit Rev Food Sci Nutr, 2017, 57(11), 2286-2295.
- [32] Sheth A., *Transforming Big Data into Smart Data: Deriving value via harnessing Volume, Variety, and Velocity using semantic techniques and technologies*. IEEE 30th International Conference on Data Engineering (ICDE), 2014.
- [33] Tetko I.V., Lowe D.M, Williams A.J. *The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS* J. Cheminformatics, 2016, 8, 2.
- [34] Polanski J., *Receptor dependent multidimensional QSAR for modeling drug-receptor interactions* Curr Med Chem, 2009, 16(25), 3243-3257.
- [35] Polanski J., Bogocz J., Tkocz A., *Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator* Drug Discov Today, 2015, 20(11), 1300-1304.
- [36] Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q., Shoemaker B.A., Thiessen P.A., Yu B., Zaslavsky L., Zhang J., Bolton E.E. *PubChem 2019 update: improved access to chemical data* Nucleic Acids Res., 2019, 8, 47.
- [37] Kurczyk A., *Polifarmakologiczna analiza leków aktywnych względem wirusa HIV* Rozprawa doktorska, Uniwersytet Śląski w Katowicach, 2013 r.

- [38] Mendez D. et al. *ChEMBL: towards direct deposition of bioassay data* Nucleic Acids Res., 2019, 47(D1), D930–D940.
- [39] Wishart D.S., Feunang Y.D., Guo A.C., Lo E.J., Marcu A., Grant J.R., Sajed T., Johnson D., Li C., Sayeeda Z., Assempour N., Iynkkaran I., Liu Y., Maciejewski A., Gale N., Wilson A., Chin L., Cummings R., Le D., Pon A., Knox C., Wilson M. *DrugBank 5.0: a major update to the DrugBank database for 2018*. Nucleic Acids Res., 2018, 46(D1), D1074–D1082.
- [40] Oficjalna strona FDA <https://www.fda.gov/> [dostęp wielokrotny]
- [41] Engel T. *Basic Overview of Chemoinformatics*, J. Chem. Inf. Model, 2006, 46, 2267–2277.
- [42] Begama B.F., Kumar J.S *A Study on Cheminformatics and its Applications on Modern Drug Discovery* Procedia Eng., 2012, 38, 1264 – 1275.
- [43] Hughes J.P., Rees S., Kalindjian S.B., Philpott K.L. *Principles of early drug discovery* Br J Pharmacol., 2011, 162(6), 1239–1249.
- [44] Gómez-Bombarelli R., Aguilera-Iparraguirre J., Hirzel T et al. *Turbocharged Molecular Discovery of OLED Emitters: From High-Throughput Quantum Simulation to Highly Efficient TADF Devices*. Proceedings of Spie, 2016, 9941, 99410A/1–99410A/8.
- [45] Dobrzanski L.A., Podstawy nauki o materiałach i metaloznawstwo, WNT, 2002.
- [46] Neves B.J., Braga R.C., Melo-Filho C.C., Moreira-Filho J.T., Muratov E.N., Andrade C.H. *QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery* Front Pharmacol., 2018, 13(9), 1275.
- [47] Sweis, R.F. et al. *2-(4-carbonylphenyl)benzoxazole inhibitors of CETP: attenuation of hERG binding and improved HDLc-raising efficacy*. Bioorg. Med. Chem. Lett., 2011, 21, 2597–2600.
- [48] Kallashi, F. et al. *2-arylbenzoxazoles as CETP inhibitors: raising HDL-C in cynoCETP transgenic mice* Bioorg. Med. Chem. Lett., 2011, 21, 558–561.
- [49] Harikrishnan, L.S. et al. *2-arylbenzoxazoles as novel cholesteryl ester transfer protein inhibitors: optimization via array synthesis* Bioorg. Med. Chem. Lett., 2008, 18, 2640–2644.
- [50] Griffith D.A. et al. *Discovery of 1-[9-(4-chloro-phenyl)-8-(2-chlorophenyl)-9H-purin-6-yl]-4 ethylaminopiperidine-4-carboxylic acid amide hydrochloride (CP-945,598), a novel, potent, and selective cannabinoid type 1 receptor antagonist* J. Med. Chem., 2009, 52, 234–237.

- [51] Plowright A.T. et al. *Discovery of agonists of cannabinoid receptor 1 with restricted central nervous system penetration aimed for treatment of gastroesophageal reflux disease*. J. Med. Chem., 2013, 56, 220–240.
- [52] Harikrishnan L.S. et al. *2-arylbenzoxazoles as novel cholesteryl ester transfer protein inhibitors: optimization via array synthesis*. Bioorg. Med. Chem. Lett., 2008, 18, 2640–2644.
- [53] Fernanadez, M.C. et al. *Design, synthesis and structure-activity-relationship of 1,5-tetrahydronaphthyridines as CETP inhibitors*. Bioorg. Med. Chem. Lett., 2012, 22, 3056–3062.
- [54] Darout, E. et al. *Design and synthesis of diazatricyclodecane agonists of the G-protein-coupled receptor 119*. J. Med. Chem., 2013, 56, 301–319.
- [55] Higuero A.P., Schreyer A., Bickerton G.R. J., Blundell T.L., Pitt, W.R. *What can we learn from the evolution of protein–ligand interactions to aid the design of new therapeutics?* PLoS ONE 7, e51742 (2012)
- [56] Valko K., Chiarparin E., Nunhuck S., Montanari D. *In vitro measurement of drug efficiency index to aid early lead optimization* J. Pharm. Sci., 2012, 101, 4155–4169.
- [57] Freeman-Cook K.D., Hoffman R.L., Johnson T.W. *Lipophilic efficiency: the most important efficiency metric in medicinal chemistry* Future Med. Chem., 2013, 5, 113–115.
- [58] Abad-Zapatero C. *Ligand efficiency indices for effective drug discovery* Exp. Opin. Drug Discov., 2007, 2, 469–488.
- [59] Mannhold R., Poda G. I., Ostermann C., Tetko, I.V. *Calculation of molecular lipophilicity: state-of-the-art and comparison of logP methods on more than 96,000 compounds*. J. Pharm. Sci., 2009, 98, 861–893.
- [60] Polanski J., Pedrys A., Duszkiewicz R., Gasteiger J. *Scoring Ligand Efficiency: Potency, Ligand Efficiency and Product Ligand Efficiency within Big Data Landscape* Lett Drug Des Discov, 2019, 16.
- [61] Kuntz I.D., Chen K., Sharp K.A., Kollman P.A. *The maximal affinity of ligands*. Proc Natl Acad Sci USA, 1999, 31, 96(18), 9997-10002.
- [62] Reynolds C.H., Bembenek S.D., Tounge B.A. *The role of molecular size in ligand efficiency* Bioorganic Med. Chem. Lett., 2007, 17, 4258-4261.
- [63] Reynolds C.H., Tounge, B.A., Bembenek S.D., *Ligand binding efficiency: Trends, physical basis, and implications* J. Med. Chem., 2008, 51, 2432-2438.
- [64] Reynolds C.H., Reynolds R.C., *Group Additivity in Ligand Binding Affinity: An Alternative Approach to Ligand Efficiency* J. Chem. Inf. Model., 2017, 57, 3086-3093.

- [65] Leeson, P.D., Springthorpe B. *The influence of drug-like concepts on decision-making in medicinal chemistry* Nature Rev. Drug Discov., 2007, 6, 881–890.
- [66] Abad-Zapatero, C. *Ligand efficiency indices for effective drug discovery* Exp. Opin. Drug Discov., 2007, 2, 469–488.
- [67] Verdonk M.L., Rees D.C., *Group efficiency: a guideline for hits-to-leads chemistry* ChemMedChem, 2008, 3, 1179–1180.
- [68] Keserü G.M., Makara G.M. *The influence of lead discovery strategies on the properties of drug candidates* Nature Rev., 2009, Drug Discov. 8, 203–212.
- [69] Mortenson P.N., Murray C.W. *Assessing the lipophilicity of fragments and early hits*. J. Comput. Aided Mol. Des., 2011, 663–667.
- [70] Shultz M.D., *Improving the Plausibility of Success with Inefficient Metrics* ACS Med. Chem. Lett., 2014, 5, 2-5.
- [71] Shultz M.D. *Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters* Bioorganic Med. Chem. Lett., 2013, 23, 5980-5991.
- [72] Polanski J., Tkocz A., Kucia U. *Beware of ligand efficiency (LE): understanding LE data in modeling structure-activity and structure-economy relationships* J. Cheminform., 2017, 9, 49.
- [73] Polanski J., Tkocz A., *Between Descriptors and Properties: Understanding the Ligand Efficiency Trends for G Protein-Coupled Receptor and Kinase Structure-Activity Data Sets* J. Chem. Inf. Model., 2017, 57, 1321-1329.
- [74] Sheridan P.R., *Debunking the Idea that Ligand Efficiency Indices Are Superior to pIC50 as QSAR Activities* J. Chem. Inf. Model., 2016, 56, 2253-2262.
- [75] Schultes S., de Graaf C., Haaksma E. E., de Esch I.J., Leurs R., Krämer O. *Ligand efficiency as a guide in fragment hit selection and optimization* Drug Discovery Today: Technologies, 2010, 7(3), e157-e162.
- [76] Shultz M.D., *Improving the plausibility of success with inefficient metrics* ACS Med. Chem. Lett., 2014, 5 (1), 2-5.
- [77] Shultz M.D., *Two decades under the influence of the rule of five and the changing properties of approved oral drugs: miniperspective* J. Med. Chem, 2010, 62(4), 1701-1714.
- [78] Scott J.S., Waring M.J. *Practical application of ligand efficiency metrics in lead optimisation* Bioorg. Med. Chem, 2018, 26(11), 3006-3015.

- [79] Mignani S., Rodrigues J., Tomas H., Jalal R., Singh P.P., Majoral J. P., Vishwakarma R. A. *Present drug-likeness filters in medicinal chemistry during the hit and lead optimization process: how far can they be simplified?* Drug Discov. Today, 2018, 23(3), 605-615.
- [80] Meanwell N.A. *Improving drug design: an update on recent applications of efficiency metrics, strategies for replacing problematic elements, and compounds in nontraditional drug space* Chem. Res. Toxicol, 2016, 29(4), 564-616.
- [81] Cavalluzzi M.M., Mangiatordi G.F., Nicolotti O., Lentini G. *Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective.* Expert Opin Drug Discov., 2017, 12(11), 1087-1104.
- [82] Kenny P.W., Leitao A., Montanari C.A. *Ligand efficiency metrics considered harmful* J. Comput. Aided Mol. Des., 2014, 28, 699-710.
- [83] Polanski, J., Tkocz, A. *Between Descriptors and Properties: Understanding the Ligand Efficiency Trends for G Protein-Coupled Receptor and Kinase Structure–Activity Data Sets.* J. Chem. Inf., 2017, 57, 1321-1329.
- [84] Polanski J., Pedrys A., Duszkiewicz R., Kucia U. *Ligand Potency, Efficiency and Drug-likeness: A Story of Intuition, Misinterpretation and Serendipity* Curr Protein Pept Sci. 2019, 20(11), 1069-1076.
- [85] Maryanoff B.E. *Phenotypic Assessment and the Discovery of Topiramate* ACS Med Chem Lett., 2016, 13, 7(7), 662-665.
- [86] Mervin L.H., Johansson S., Semenova E., Giblin K.A., Engkvist O. *Uncertainty quantification in drug design.* Drug Discov Today, 2021, 26(2), 474-489.
- [87] Thomas J., Navre M., Rubio A., Coukell A. *Shared Platform for Antibiotic Research and Knowledge: A Collaborative Tool to SPARK Antibiotic Discovery* ACS Infect. Dis., 2018, 4 (11), 1536-1539.
- [88] Ridley M., *How innovation works: And why it Flourishes in Freedom*, Harper Collins Publishers, 2020.
- [89] Rang H. *Bad Pharma: how drug companies mislead doctors and harm patients* Br J Clin Pharmacol., 2013, 75(5), 1377–1379.
- [90] Bickerton G.R., Paolini G.V., Besnard J., Muresan S., Hopkins A.L., *Quantifying the chemical beauty of drugs* Nat Chem., 2012, 24, 4(2), 90-98.
- [91] Lipinski Ch.A., Lombardo F., Dominy B.W., Feeney P.J., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings* Adv. Drug Deliv. Rev., 1997, 23(1–3), 3-25.

- [92] Hann M.H. *Molecular obesity, potency, and other addictions in drug discovery* Med. Chem. Commun., 2011, 2, 349-355.
- [93] Hann M.H., Keserü G.M. *Finding the sweet spot: the role of nature and nurture in medicinal chemistry*. Nat Rev Drug Discov., 2012, 11, 355-365.
- [94] Wang S., Dong G., Sheng, C., *Structural Simplification of Natural Products* Chem. Rev., 2019, 119, 4180-4220.
- [95] Greene, L., Singh R.M., Carden M.J., Pardo C.O., Lichtenstein G.R. *Strategies for Overcoming Barriers to Adopting Biosimilars and Achieving Goals of the Biologics Price Competition and Innovation Act: A Survey of Managed Care and Specialty Pharmacy Professionals* J. Manag. Care Spec.Pharm., 2019, 25, 904-912.
- [96] Leonard E., Wascovich M., Oskouei S., Gurz P., Carpenter D. *Factors Affecting Health Care Provider Knowledge and Acceptance of Biosimilar Medicines: A Systematic Review* J. Manag. Care Spec. Pharm., 2019, 25, 102-112.
- [97] 125 Years of Moore's Law, <https://www.flickr.com/photos/jurvetson/51391518506> [dostęp 09.07.2023 r.]
- [98] Van Norman G.A. *Overcoming the Declining Trends in Innovation and Investment in Cardiovascular Therapeutics: Beyond EROOM's Law* JACC: Basic to Translational Science, 2017, 2(5), 613-625.
- [99] Nordhaus W. D. *Do Real-Output and Real-Wage Measures Capture Reality? The History of Lighting Suggests Not*. The Economics of New Goods. Bresnahan T.F., Gordon, R.J., Eds., University of Chicago Press: Chicago, 1996, pp. 27-70.
- [100] When Will AI Beat The Eroom's Law In The Pharmaceutical Industry? <https://www.forbes.com/sites/alexzhavoronkov/2022/08/22/when-will-ai-beat-the-erooms-law-in-the-pharmaceutical-industry/?sh=316b87bc5677> [dostęp 10.07.2023 r.]
- [101] Centaur Chemist® AI-based drug discovery platform <https://www.life-sciences-europe.com/product/centaur-chemist-based-drug-exscientia-ltd-scientia-group-2001-30419.html> [dostęp 08.07.2023 r.]
- [102] Ringel M.S., Scannell J.W., Baedeker M., Schulze U. *Breaking Eroom's Law* Nat Rev Drug Discov., 2020, 19(12), 833-834.
- [103] Evans B.E., Rittle K.E., Bock M.G., DiPardo R.M., Freidinger R.M., Whitter W.L., Lundell G.F., Veber D.F., Anderson P.S., Chang R.S., et al. *Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists* J Med Chem., 1988, 31(12), 2235-2246.
- [104] Zartler E.R., *Fragonomics: the -omics with real impact* ACS Med Chem Lett., 2014, 13, 5(9), 952-953

[105] Zartler E.R., Shapiro M.J. *Fragonomics: fragment-based drug discovery* Curr Opin Chem Biol., 2005, 9(4), 366-370.

[106] Practical fragments <http://practicalfragments.blogspot.com/search?q=ro3> [dostęp 03.03.2023r.]

[107] Bemis G.W., Murcko M.A. *The Properties of Known Drugs. 1. Molecular Frameworks* J. Med. Chem., 1996, 39(15), 2887-2893.

[108] Bemis G.W., Murcko M.A. *The Properties of Known Drugs 2. Side Chains* J. Med. Chem., 1991, 42, 5095-5099.

[109] Gianti E., Sartori L. *Identification and selection of "privileged fragments" suitable for primary screening* J. Chem Inf Model, 2008, 48, 2129-2139.

[110] Johns V.K., Patel P.K., Hassett S., Calvo-Marzal P., Qin Y., Chumbimuni-Torres K.Y. *Visible light activated ion sensing using a photoacid polymer for calcium detection*. Anal Chem. 2014, 1, 86(13), 6184-6187.

[111] Raucci U., Savarese M., Adamo C., Ciofini I., Rega N. *Intrinsic and dynamical reaction pathways of an excited state proton transfer* J Phys Chem B., 2015, 12, 119(6), 2650-2657.

[112] Praca zbiorowa, *Accelerated Molecular Discovery (AMD)*, projekt badawczy DARPA, nieopublikowane, 2019.

[113] Smets G., Aerts A., Vanerum, J. *Photochemical Initiation of Cationic Polymerization and Its Kinetics* Polymer Journal, 1980, 12, 539-547.

[114] Saeva F.D., Morgan B.P., Luss, H.R. *Photochemical Conversion of Sulfonium Salts to Sulfides via 1,3-Sigmatropic Rearrangement - Photogeneration of Bronsted Acids* J. Org. Chem. 50, 4360-4362.

[115] Crivello J.V., Lam, J.H.W. *Complex Triarylsulfonium Salt Photoinitiators. The Identification, Characterization and Syntheses of a New Class of Triarylsulfonium Salt Photoinitiators* J. Polym. Sci., 1980, 18, 2677-2695.

[116] Crivello J.V., Lee J.L. *Photosensitized Cationic Polymerizations using Dialkylphenacylsulfonium and Dialkyl(4-hydroxyphenyl)sulfonium Salt Photoinitiators* Macromolecules, 1981, 14, 1141-1147.

[117] Neckers D.C., Abu-Abdoun I.I. *p,p' Bis((triphenylphosphonio) methyl)benzophenone salts as photoinitiators of free radical and cationic polymerization* Macromolecules, 1984, 17, 2468-2473.

- [118] Abu-Abdoun I. I., Aale-ali. *Cationic Photopolymerization of p-Methylstyrene Initiated by Phosphonium and Arsonium Salts* Eur. Polym. J., 1993, 29, 1439-1443.
- [119] Komoto K. et al. *Photopolymerization of Vinyl Ether by Hydroxy- and Methylthio-Alkylphosphonium Salts as Novel Photocationic Initiators* Polymer, 1994, 35, 217-218.
- [120] Martin C.J., Rapenne G., Nakashima T., Kawai, T. *Recent progress in development of photoacid generators* J. Photochem. Photobiol. C: Photochem., 2018, 34, 41-51.
- [121] Sheng W., Nairat M., Pawlaczyk P.D., Mroczka E., Farris B., Pines E., Geiger J.H., Borhan B., Dantus M. *Ultrafast Dynamics of a "Super" Photobase*. Angew. Chem. Int. Ed. 2018, 57 (45), 14742-14746.
- [122] Shirai M., Tsunooka M., *Photoacid and photobase generators: Chemistry and applications to polymeric materials* Prog. Polym. Sci. 1996, 21, 1-45.
- [123] Gómez-Bombarelli R., Aguilera-Iparraguirre J., Hirzel, T., Duvenaud D. K., Maclaurin D., Blood-Forsythe M.A., Sik Chae H., Einzinger M., Ha D.G., Wu T., Markopoulous G., Jeon S., Kang H., Miyazaki H., Numata M., Kim S., Huang W., Ik Hong S., Baldo M., Adams R.P., Aspuru-Guzik, A. *Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach* Nat. Mater, 2016, 15, 1120–1127.
- [124] Sorkun M., Mullaj D., Koelman J., Er S. *ChemPlot, a Python Library for Chemical Space Visualization* Chemistry Methods 2022, 2(7), e202200005.

9. SPIS ILUSTRACJI I TABEL

Rys. 1a – b Grafy przedstawiające problem podejmowania przez algorytm zachłanny decyzji na podstawie lokalnego optimum. Założeniem eksperymentu jest wyznaczenie ścieżki od góry grafu do dołu, tak by osiągnąć maksymalną sumę. Sposób rozwiązania problemu z wykorzystaniem algorytmu zachłannego, suma wynosi 25 (a) Optymalne rozwiązanie, suma wynosi 55 (b), inspirowane [5]6

Rys. 2a – b Przykład wykorzystania danych ekonomicznych do wyznaczenia optymalnej ścieżki projektowania katalizatorów metanizacji. Porównanie stosunku ceny do wydajności katalitycznej jedno- i wieloskładnikowej układu Fe|Ni (a) do wykresu aktywności czystych katalizatorów (b) umożliwia wytypowanie optymalnego katalizatora pod względem ekonomicznym i wydajnościowym [14, 15].....8-9

Rys. 3 Piramida złożoności i precyzji nauk. Im niższy poziom złożoności danej dziedziny, tym wyższa precyzja jej opisu. Ekonomia zajmuje najwyższe miejsce diagramu. Jest najbardziej złożona i najtrudniejsza do modelowania oraz predykcji [9].....9

Rys. 4 Grafika ukazująca awarię rządowego portalu HealthCare, tuż po opublikowaniu [18].....12

Rys. 5 Przykładowy schemat przestrzeni wyników dla parametrów reakcji (x, y, z).....14

Rys. 6 Porównanie czasu mieszania i ogrzewania w skali laboratoryjnej oraz przemysłowej [22, 23]16

Rys. 7 Wykres przedstawiający liczbę wystąpień terminu uczenie maszynowe w artykułach opublikowanych w czasopiśmie PHYS (lata 2000 – 2023), inspiracja [24]19

Rys. 8 Częstość występowania terminów “Machine Learning”, Artificial Intelligence”, lub “AI” w artykułach zgromadzonych w repozytorium PubChem w latach 1990-2023, inspiracja [24]19

Rys. 9a – b Przykład uproszczenia danych w wyniku analizy PCA, trójwymiarowe dane wejściowe (a) zostały uproszczone do dwóch wymiarów (b), rysunki zostały odtworzone w Pythonie zgodnie z kodem udostępnionym na https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d#0226	21
Rys. 10a - b Schemat klasteryzacji, dane przed analizą, pozornie brak zdefiniowanych klas (a), dane po klasteryzacji, wyraźnie wyodrębnione trzy klasy (b).....	22-23
Rys. 11 Przykład schematu uczenia nadzorowanego do rozwiązywania problemu regresji..	23
Rys. 12a - b Uproszczony schemat uczenia nadzorowanego do rozwiązywania problemu klasyfikacji, etap uczenia (a), klasyfikacja (b).....	24
Rys. 13 Trzy filary chemoinformatyki.....	37
Rys. 14 Schemat tworzenia fragmentów metodą dekrementacyjną i inkrementacyjną.....	39
Rys. 15 Schemat ścieżki projektowania materiałów.....	40
Rys. 16a - b Porównanie teoretycznego pojęcia aktywności (a) z rzeczywistym (b) [84]..	46
Rys. 17a - d Fizyczne wyjaśnienie nieścistości w znaczeniu LE [49].....	46
Rys. 18a - d Wykres zależności pPLE od HAC dla katalogu ChEMBL (a), PubChem (b), subpopulacji PubChem o pAC > 6 (c). pPLE może zostać rozłożone na pAC50 oraz pHAC (odpowiednio niebieskie i zielone punkty). W celu porównania wyników do LE, utworzyłam wykres (d) przedstawiający zależność LE od HAC z wykorzystaniem tych samych danych jak w przypadku wykresów (a-c) [60].....	48-50
Rys. 19a - b Zależność pPLE (a) oraz LE (b) od HAC dla leków i fragmentów [60]	51-52
Rys. 20a - c Histogramy częstotliwości (szare słupki), średnia sprzedaż/lek (ang. <i>Mean Sales</i> , MS: żółty) oraz całkowita sprzedaż/grupa leków (ang. <i>Total Sales</i> , TS: niebieski) w	

porównaniu do przedziałów MW dla odpowiednich populacji TOP w latach 2000 - 2009 (a), 2010 - 2019 (b) i 2014 - 2019 (c).....**55-56**

Rys. 21a - c Histogramy częstotliwości (szare słupki), średnia sprzedaż/lek (ang. *Mean Sales*, MS: żółty) oraz całkowita sprzedaż/grupa leków (ang. *Total Sales*, TS: niebieski) w porównaniu do przedziałów logP dla odpowiednich populacji TOP w latach 2000 - 2009 (a), 2010 - 2019 (b) i 2014 - 2019 (c). Częstotliwości i TS zostały znormalizowane, aby śledzić zmieniającą się liczbę leków w różnych okresach.....**57-58**

Rys. 22a - b Histogramy częstotliwości populacji FDA w porównaniu do przedziałów MW (a) i logP (b). Kolory słupków oznaczają przedziały czasowe kolejno zielony 1985 - 2000, żółty 2000 - 2009, czerwony 2010 - 2019 oraz niebieski 2014 - 2019.....**59**

Rys. 23a - d Średni, maksymalny oraz minimalny wiek leku TOP kontra MW i logP...**60-61**

Rys. 24a - b Analiza QED dla TOP oraz FDA *approvals*.....**63**

Rys. 25a - b Wykres przedstawiający ścieżkę rozwoju technologicznego, określoną prawem Moore'a, każdy punkt wyznacza granicę wydajności obliczeniowej w danym roku, osiągniętą przez wskazaną firmę (a). Obecnie na szczycie jest chip D1 superkomputera DOJO (Tesla). Oryginalny wykres obrazujący tempo zmian w technologii procesorów, autorstwa Gordona Moore'a (b) [97].....**66**

Rys. 26a - k Wynik rozkładu t-SNE dla zbioru FDA *approvals* (1985-2019) i TOP (a), rozkład danych na poszczególne klastera wraz ze wskazaniem przykładów związków poszczególnych klastrów (b - k).....**79-89**

Rys. 27a - d Mapy SOM dla zbioru FDA *approvals* (a) oraz TOP 2000-2009 (b) 2010-2019 (c) i 2014-2019 (d). Kolory kodują obecność grup funkcyjnych. Kolor czerwony wskazuje na obecność wielu grup funkcyjnych, zielony na obecność wielu atomów tlenu podczas gdy niebieski oznacza obecność wiązań pojedynczych.....**98-100**

Rys. 28 Zastosowanie fotokwasów [111].....	102
Rys. 29 Zrzut ekranu przedstawiający katalog Reaxys wraz z wyszukiwarką i wpisanym parametrem filtrującym.....	106
Rys. 30 Zrzut ekranu przedstawiający liczbę uzyskanych rekordów.....	107
Rys. 31 Zrzut ekranu przedstawiający pełną listę uzyskanych wyników.....	107
Rys. 32 Zrzut ekranu przedstawiający przykładowy artykuł przeszukiwany hasłem „ <i>photoacid</i> ”.....	108
Rys. 33 Zrzut ekranu przedstawiający fragment tworzonego roboczego katalogu fotoreagentów w arkuszu Excel.....	108
Rys. 34 Interfejs programu MATLAB.....	111
Rys. 35 Przykładowy skrypt utworzony w MATLAB, służący do automatyzacji tworzenia wykresów, omawianych w rozdziale 4.2.2.....	111-112
Tabela 1 Porównanie ilości dostępnych rekordów podstruktur PubChem w latach 2013 oraz 2023.....	29
Tabela 2 Zestawienie ilości dostępnych przykładowych grup danych w kolejnych wersjach bazy DrugBank [39]	31
Tabela 3 Przykładowe sposoby reprezentacji komputerowej struktury chemicznej kofeiny.....	38
Tabela 4 Zestawienie najczęściej powtarzających się fragmentów dla baz FDA approvals oraz TOP.....	74-78
Tabela 5 Przykłady związków dla klastrów 1-10.....	90-97

10. ZAŁĄCZNIKI

Do treści niniejszej rozprawy dołączono trzy załączniki:

1. Życiorys naukowy autorki
2. Trzy publikacje naukowe z dorobku naukowego autorki
3. Baza danych związków PAH i PAG oraz kodujących ich SMILES. Ilość kolumn została zredukowana.

Załącznik 2a

Polanski J., Duszkiewicz R., Pedrys A., Gasteiger J., Scoring ligand efficiency. *Acta Pol. Pharm.*, 2019, 76(4), 761 - 768

Załącznik 2b

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

Załącznik 2c

Polanski J., Pedrys A., Duszkiewicz R., Gasteiger J., Scoring Ligand Efficiency: Potency, Ligand Efficiency and Product Ligand Efficiency within Big Data Landscape, *Lett. Drug Des Discov.* 2019, 16(11), 1258 – 1263.

Curriculum Vitae

Dane osobowe

Imię i nazwisko: Anna Pędrys

Data urodzenia: 9 maja 1991 r.

Miejsce urodzenia: Gliwice

Wykształcenie

2016 – 2022 Studia doktoranckie w zakresie chemii, Uniwersytet Śląski w Katowicach

2013 – 2015 Studia magisterskie w zakresie chemii, Uniwersytet Śląski w Katowicach

2013 – 2015 Studia licencjackie w zakresie chemii, Uniwersytet Śląski w Katowicach

Działalność dydaktyczna

Prowadzenie zajęć dydaktycznych w Zakładzie Chemii Organicznej Instytutu Chemii, Wydziału Matematyki, Fizyki i Chemii Uniwersytetu Śląskiego w Katowicach w zakresie laboratorium chemii organicznej, laboratorium chemii organicznej dla biotechnologii oraz laboratorium z podstaw chemii.

Aktywny udział w organizacji Święta Liczby Pi (prowadzenie laboratorium pokazowego, Uniwersytet Śląski w Katowicach, laboratoria dla odwiedzających, 2019 r.).

Zainteresowania naukowe i badawcze

Nowoczesne metody projektowania leków i ich adaptacja na inne obszary chemii, farmakoekonomia, analiza biznesowa, analiza *big data*.

Udział w konferencjach

13-16.06.2018 r., I International Conference Chemistry for Beauty and Health, Toruń

Poster: PUBCHEM – An innovative way to import Big Data, **Pedrys A.**, Polanski J.

17-20.09.2018 r., 61. Zjazd Naukowy Polskiego Towarzystwa Chemicznego, Kraków

Wystąpienie ustne: Scoring ligand efficiency – paradoks oddziaływania deskryptor – właściwość, **Pedrys A.**

10-14.04.2019 r., ZWSSPTChem, Ustroń

Wystąpienie ustne: Fragonomika materiałów OLED i fotoreagentów, **Pedrys A.**

02-06.2019 r., 62 Zjazd Naukowy Polskiego Towarzystwa Chemicznego, Warszawa

Poster: Fragonomika fotoreagentów, **Pedrys A.**, Polanski J.

19.09.2019 r., Pomiędzy Naukami, Chorzów

Poster: Fragonomic: Fragment-based design of new photoreagents, **Pedrys A.**, Polanski J.

Publikacje

Pedrys A., Zhdan Uladzislau et al. *Fragonomics of TOP drug bestsellers: an innovation benchmarks for drug discovery?* (w przygotowaniu)

Polanski J., Duszkiewicz R., **Pedrys A.**, Gasteiger J., *Scoring ligand efficiency* Acta Pol. Pharm., 2019, 76(4), 761 – 768.

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

Polanski J., **Pedrys A.**, Duszkiewicz R., Gasteiger J., *Scoring Ligand Efficiency: Potency, Ligand Efficiency and Product Ligand Efficiency within Big Data Landscape* Lett. Drug Des Discov., 2019, 16(11), 1258 – 1263.

SCORING LIGAND EFFICIENCY

JAROSLAW POLANSKI^{1*}, ROKSANA DUSZKIEWICZ^{1,2}, ANNA PEDRYS¹
and JOHANN GASTEIGER³

¹Institute of Chemistry, University of Silesia in Katowice, Szkolna 9, 40-006 Katowice, Poland

²Department of Pharmacology, School of Medicine in Katowice,

Medical University of Silesia in Katowice, Medyków 18, 40-752 Katowice, Poland

³Computer-Chemie-Centrum, University of Erlangen-Nuernberg, Erlangen 91052, Germany

Abstract: Ligand efficiency (LE) is a molecular descriptor that probes the ratio of potency vs. heavy atom count (HAC). As an estimator of drug candidates, LE emphasizes a low heavy atom count more than potency. The objective was to design a novel transform where potency and the HAC would be balanced more evenly. A series of novel descriptors SCORE were defined to evaluate the co-influence of potency and the HAC. In particular, the product ligand efficiency (PLE) was designed and tested using the data of the ChEMBL, PubChem as well as the selected series of drugs and drug-fragments.

Keywords: ligand efficiency, product ligand efficiency; heavy atom count, activity, drugs, drug design, PubChem, ChEMBL

In drug design, we are optimizing the binding ability of ligands as a function of their chemical structure. A variety of methods has appeared that focus on this problem. In the simplest approach, the chemical structure can be represented by molecular size. The importance of molecular size in drug development has received much attention (1, 2). Generally, an increase in the molecular size also increases the molecular complexity. It is interesting to analyze how this increase affects the probability of identifying new drugs. In particular, the chances of finding smaller and less complex ligands is higher than it is for larger ones. In turn, an increase in molecular complexity can also increase the potency of ligand-target binding. Eventually, the binding can drop below a measurable level for molecules or fragments that are too small (3, 4). In other words, ligand size plays a dichotomic role in matching and binding a target. Rating these effects is an important tool in the search for a more efficient way to design better drugs. The heavy atom count (HAC) is a simple descriptor measuring molecular size that is related to ligand efficiency (LE), which is commonly used to evaluate the binding ability of ligands. LE is defined as the ratio of binding energy to the HAC (1, 5-8). A variety of LE-based analyses have been proposed (1) despite the fact that the observed trend of

LE seemed to be paradoxical and therefore could not be fully understood. (9-15). Not only has the mathematical validity of LE with its high preference for small ligands been questioned but a number of authors have indicated that this descriptor could have a quite low degree of usefulness (9, 10, 13-15). e.g., being “*uninformative when the changes do not significantly alter the size of the compounds*” (16). In contrast, generally, LE has gained an enthusiastic reception (1). One of the precautions sounds however that LE is size-dependent and therefore we should not compare the compounds of extremely different sizes, which sounds accurately opposite to the warning (16). Accordingly, this is the reason for the uncertainties connected with LE.

Essentially, LE was designed to evaluate the average contribution of one non-hydrogen atom (HAC) to the binding free energy. We can easily understand the importance of the direct relation of LE to binding energy, if we recognize that LE is commonly calculated as a function of the inhibitory concentration pIC_{50} , namely, $LE = pIC_{50} * (1.37/HAC)$ (1). Accordingly, the closely related ligand efficiency index (LEI), which is defined directly by the pIC_{50}/HAC (1), seems to be a simple replacement for LE; however, LEI is rarely discussed in the literature.

* Corresponding author: e-mail: jaroslaw.polanski@us.edu.pl

To explain the paradoxical LE behavior, we should realize the dichotomic nature of the molecular vs. molar representations of chemical compounds (13, 14). Molecular descriptors (MD) and properties (P) are the best illustration here. While molecular descriptors are usually calculated from the representation of a single molecule, the properties are usually measured in experiments for substances, i.e. for ensembles of molecules. Sometimes, these representations are barely distinguishable (17). In this context, on the one hand, LE has been designed to be a molecular descriptor that is connected with a single molecule or, more precisely, with a single HAC that is a part of a single molecule. On the other hand, the binding energy is a property that is related to the ligand-receptor interaction that engages a population (ensemble) of molecules. It is worth mentioning here that recent advances in technology have made the so-called single-molecule biology system more and more popular (18, 19), which extends this problem to also include the property (single molecule vs. molar) representations. Accordingly, the uncertainty of LE comes from the ill-defined chemistry and not mathematics, because fragments (1 HAC or 1 Dalton) do not have a real molar representation. Interestingly, however, a mole of Daltons virtually represents 1 g (13, 14).

A high LE preference for small ligands (13, 14) fits into the recent trends in pharma that favor small molecular ligands (the so-called slim pharma concept) (20), which have advantageous drug-likeness profiles (21). Therefore, LE performs unexpectedly well if used as guideline filters during the hit and lead optimization (22-25) despite the uncertainty in its physical meaning (13, 14). This uncertainty causes effects that have been interpreted as being unexpected and paradoxical. Basically, the LE trend could not be understood. To illustrate the problems, the development of the empirical formula of the so-called size-independent LE (SILE) can be cited here (26). The analogy of LE to car fuel efficiency can even better illustrate a confusion (12).

In this publication, we show that a more complete understanding of the LE enables rational predictors for molecular design to be defined. In practice, a low HAC and high potency indicate attractive drug candidates. This suggests an interaction between the HAC and potency which in the most general meaning will be represented here by pAC_{50} . Statistically, LE probes the reciprocal interaction between pAC_{50} (negative logarithm of the active concentration) and $1/HAC$. In turn, we confronted this with a direct multiplicative HAC and potency association. In particular, we analyzed the physical

meaning of such a product ligand efficiency (PLE) and tested the behavior of the PLE on data from the PubChem and ChEMBL potency databases as well as on the selected series of drug and drug candidates (1, 25). Finally, the most important function of LE is its use as a guideline during the hit and lead optimization (22-25). In such functionality, the association of AC_{50} and HAC is just a scoring function enabling decision making in the drug pipeline. Accordingly, we defined a novel flexible predictor SCORE that is capable of adaptably scoring the development potential of drug candidates that is related to potency and the HAC.

EXPERIMENTAL

Ligand binding: a potency vs. single biology, ligand efficiency (LE) and binding efficiency index (BEI)

Measuring the binding effects of ligands is a complex problem. Accordingly, a ligand molecule interacts with a receptor to produce a signal. Actually, recent technological development has provided us with the possibility to directly measure these systems using the so-called single biology approaches (18, 19). However, historically, this representation has not been available for direct observations. Therefore, the so-called potency is commonly used to describe ligand-receptor systems. A collection of molecules interacts with a collection of receptors. In this model, a signal is produced by the competitive mechanism of ligand and receptor involvement thereby producing a well-known sigmoid-like potency signal, which is usually denoted by the so-called inhibitory concentration IC_{50} . Ligand efficiency, LE, is another metric that has recently been proposed as an intensive representation of ligand binding, which originated in an effort to identify the maximal ligand affinities (5).

LE or BEI are calculated by a simple calculation of the proportion of binding properties to the molecular size that is denoted by the non-hydrogen atoms (the so-called heavy atom counts: HAC) or molecular weight (MW). In practice, LE has been of special interest in drug design. The astonishing successes of LE (1, 12) were confronted with a series of skeptical analyses that argue that the mathematics of LE is incorrect (9-11). The behavior of LE function has been widely investigated. For example, the nature of binding sites and the target class is likely to have an impact on ligand efficiency (for instance, inhibitors of protein-protein interactions versus enzyme inhibitors). Similarly, different mechanisms of activity should be treated separately for ligand efficiency purposes, because covalent inhibitors can

have very high ligand efficiency values. In addition, larger, more optimized molecules may have smaller ligand efficiencies, which is due in part to the fact that such compounds have been optimized for properties other than binding to the target (for instance, pharmacokinetics, solubility, selectivity or cell permeability). It is vital to remember that ligand efficiency has limited value in late-stage lead optimization and that it should primarily be used in hit identification and hit-to-lead settings. Another factor that may have an important role could be assay limitations: many assays have a detection limit of EC_{50} of ~ 1 nM due to signal-to-noise issues. Good ligand efficiency for large molecules would require sub-nanomolar potencies, which may not be recorded (or recorded correctly) because of assay limitations.

Data

We used the largest available potency data in the form of AC_{50} as is defined by the PubChem classification. Accordingly, we called all of the data that was used the active concentration, AC_{50} . For PubChem, AC_{50} stands for the inhibitory concentration, IC_{50} , the effective concentration, EC_{50} , the cytotoxic concentration, CC_{50} , the equilibrium dissociation constant, K_i , for the ligand that is determined directly in a binding assay using a labeled ligand or dissociation constant, K_d , for the ligand determined in inhibition studies. For ChEMBL the AC_{50} values are either the IC_{50} , K_i or K_d values. Binding energy and potency are related as described in (1).

In the context of the used data, we should remember that originally LE has been defined as the contribution of a non-hydrogen atom to the binding free energy. Thermodynamically, K_i and K_d values are therefore adequate measures for calculating LE. IC_{50} and more importantly EC_{50} values are determined in functional measurements and in many cases, they do not change proportionally with the binding affinity. However, in practice usually IC_{50} , EC_{50} and K_i data are used as interchangeable values. Such a practice is especially necessary if we would like to probe the big data type statistics.

Physical meaning of reciprocal (LE) and multiplicative ligand efficiency (PLE)

Formally, LE probes reciprocal interaction of pAC_{50} and HAC. Physical meaning of LE was described in (13, 14). We are exploring here a direct multiplicative interaction of AC_{50} and PLE as the product $AC_{50} \cdot HAC$. Below we clear a meaning of PLE for molar ligand representation.

$$\text{Accordingly:} \\ \text{PLE} = AC_{50} \cdot HAC \quad (1)$$

As $MW \text{ (kg/mole)}/MW \text{ (Da)} = 1$, PLE can be defined by:

$$\text{PLE} = AC_{50} \cdot HAC \cdot (MW \text{ (kg/mole)}/MW \text{ (Da)}) \quad (1a)$$

Because AC_{50} is a concentration-based metric that has the dimension of (mol/L), from eq. 1a, we obtain the PLE dimension (unit), which is (kg/L) \cdot HAC/MW(Da). This means that the physical meaning of the PLE is the minimum inhibitory concentration (MIC) scaled to the HAC. Since the AC_{50} usually relates to the 50% inhibitory concentration multiplicative LE also relates to MIC_{50} , which is given in kg/L.

Size independent LE (SILE)

SILE was calculated from equation:

$$\text{SILE} = AC_{50}/HAC^{0.3} \quad (1b)$$

The SCORE estimator

SCORE is defined as:

$$\text{SCORE} = a \cdot pAC_{50} + b \cdot pHAC \quad (2)$$

if a and $b = 1$, then $\text{SCORE} = pPLE$.

Data acquisition and calculations

All of the records along with their numerical values used in the analyses were downloaded from the ChEMBL and PubChem databases. Records repeating individual compounds were treated as independent entries. From PubChem, 2,435,467 records (download: August 2017, pubchem.ncbi.nlm.nih.gov) and from ChEMBL, 779,714 records (ChEMBL version 24, www.ebi.ac.uk) were downloaded from their Internet sites, respectively. The pAC_{50} data for drugs or drug fragments were taken from the literature as the pIC_{50} values (1, 25). The mean values for the drugs are the HAC 31; pIC_{50} 8.12 and for the fragments, the HAC 15; pIC_{50} 4.41. A series of fragment to lead drug development projects reported in J. Med. Chem. were analyzed after the data presented in the references (27-29). Additionally, the data for the selected series of drugs collected in the Binding Database were analyzed. This includes: BindingBD, 570,927 records, PTaylorLab, 180 records, USPatent, 210,254 records, 5HT, 830 records, or AChE, 726 records (download: December 2018, www.bindingdb.org) and Psychoactive Drug Screening Program PDSP database 22,273 records (download: December 2018, pdsp.unc.edu/databases).

Binning

To plot the $pPLE$ and pAC_{50} vs. the HAC for ChEMBL and PubChem data, we used a binning method. In binning, data that fall into a given inter-

val, a bin, are replaced by a value that is representative of that interval. Individual single HAC numbers define the size of the intervals, while the AC_{50} values are represented by their median value. For a $HAC > 60$, where not enough data were available, the HAC was binned for each 15 HACs.

RESULTS AND DISCUSSION

The physical meaning of LE determines its preference for small ligands. The question is whether we can design a predictor that would balance more evenly the interaction between potency and the HAC. In practice, a low HAC and low AC_{50} (high potency) values indicate attractive drug candidates. Therefore, multiplicative interaction of potency and the HAC (Product LE: PLE) should be an informative estimator of the quality of drug candidates. A simultaneous decrease in the HAC and AC_{50} will decrease the PLE, and vice versa, their increase will increase the PLE. Accordingly, both terms in the PLE act cooperatively. Interestingly, besides their relation to a single molecule, ligand efficiency estimators are evidently associated with substances and their properties (Compare Materials and Methods). Firstly, this proves the design concept for the PLE, which was shown to be related to the minimum inhibitory concentration (MIC) because the MIC is an obvious property measure that is related to the activity of a compound. Secondly, this drew our attention to the fact that in designing a metric that can be used to evaluate drug candidates, we should observe both its role as a synthetic descriptor that optimizes the balance between its activity and the HAC and its potential physical meaning, which originates from the descriptor-property interplay.

An interesting problem will be to analyze LE changes in the drug development pipeline. Johnson et al., observed that for the data published in *J. Med. Chem.* that “LE decreased during optimization for only a minority of examples.” He indicated however, that “this finding should be treated with caution because the data analyzed here are biased toward publishable F2L campaigns.” (27). To conclude LE in the F2L can both increase or decrease. However, can we indicate the key factors limiting this effect? A question is if we should expect any regularity for a relatively small population of ligands described in the reference (27). The similarity paradox claims that even the smallest structural change can result in the substantial activity changes; therefore, the regularity will rather be surprising. In Figure 1 we reanalyzed the F2L data (27-29) plotting the change of AC_{50} related values of ΔpAC_{50} , ΔLE or $\Delta pPLE$ in the course of F2L conversion as a function of HAC of the original fragments. As expected there is no correlation between the ΔpAC_{50} , and $\Delta pPLE$ and HAC for fragments. In turn, ΔLE obviously indicate an increasing trend vs. HAC for fragments. The lower the HAC of the fragment is the lower also is the gain in LE for the resulted lead. For a low HAC fragment any AC_{50} gain by the lead cannot balance the HAC contributing into the LE value of the fragment by means of the hyperbolic $1/HAC$ term. This clearly illustrates the dominating influence of HAC into LE, which can be explained by non-Avogadro LE statistics (13, 14). A small population of the F2L could not necessarily be generally representative; therefore, we focus below on the large potency databases to explore scoring potential of efficiency functions. A common mathematical representation for AC_{50} is its negative log scale, pAC_{50} , in which higher values of pAC_{50} indicate an exponentially greater potency.

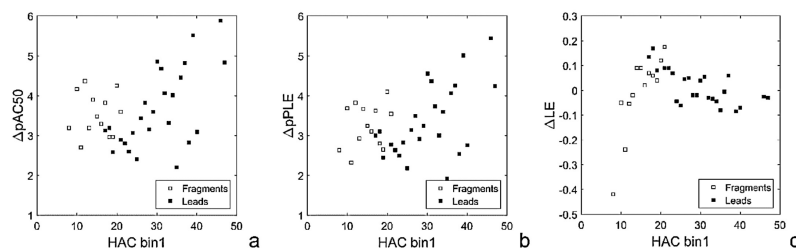


Figure 1. The dependence of the difference between the lead and fragment potency measured as ΔpAC_{50} (a), $\Delta pPLE$ (b) or ΔLE (c) as a function of HAC during F2L development reported in *Journal of Medicinal Chemistry* 2015-2017, indicated both for fragments and leads. Data after (27-29)

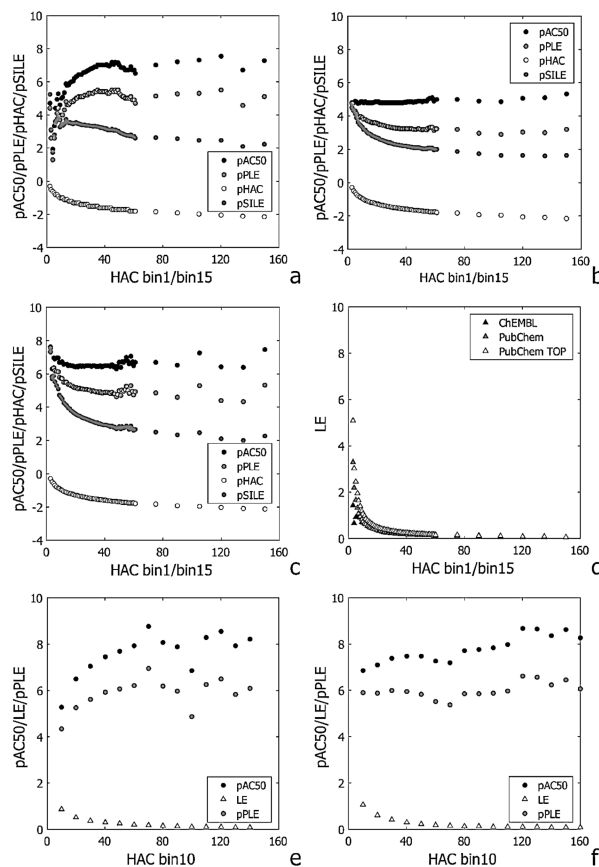


Figure 2. The dependence of the median pPLE on HAC for data from ChEMBL (a), PubChem (b), and a PubChem subpopulation with $pAC_{50} > 6$ (c). pPLE can be defragmented to median pAC_{50} (black dots) and pHAC ($-\log HAC$, white dots) which is additionally illustrated in (a-c). For comparison we show pSILE (size-independent LE) in (a-c), the median LE for all data of ChEMBL, PubChem or a PubChem subpopulation of $pAC_{50} > 6$ (d) and AC_{50} vs. HAC for a series of patented drug candidates (e) or psychoactive PDSP drugs (f)

Accordingly, we used the analogous pPLE scale (a higher pPLE indicates better quality) to analyze the PLE trends. Because the logarithm of a product is the sum of the logarithms, pPLE can be defragmented to its $\log HAC$ (pHAC) and pAC_{50} components as is shown in Figures 2a-c. In Figure 2a, the ChEMBL data show an increase in pAC_{50} until a value of approximately 50 for the HAC. At the same time, the interaction between pAC_{50} and pHAC is clearly

revealed in the pPLE which optimum is shifted slightly towards the lower HAC values in comparison to the maximum of pAC_{50} plot. In particular, for the pPLE plot the maximum at ca. 30-50 HAC is broader while the depression for high HAC (160) is higher. In Figures 2b and 2c, we illustrate the PubChem data, which shows that pAC_{50} is a constant function of the HAC (Fig. 2b), thus indicating that potency is not generally a function of the molec-

ular size, if the probed population of the active compounds is large enough. If so, lower HAC, less complex ligands are better drug candidates, then pPLE properly indicates the lower HAC as statistically better. In turn, for the most active PubChem data (ligands of $pAC_{50} > 6$), pAC_{50} decreased with the HAC (Fig. 2c), thus indicating that the probability of the proper ligand-target fit decreases with an increasing molecular size for this high activity ligands. In both cases (Figs. 2b-c), the lowest HAC indicated the optimum candidates, which was correctly predicted by the pPLE. Accordingly, all possible interaction scenarios of pAC_{50} vs. HAC (Figs. 2a-c) were correctly predicted by the pPLE function. In turn, the LE vs. the HAC relationship for all of the data plotted together was very similar, always indicating the lowest HAC as optimal (Fig. 2d). In other words, as a predictor LE emphasizes a low HAC more than a high degree of potency. A series of pAC_{50} vs. HAC relationships were explored and plotted in Figures 2e-f. Accordingly, we show a series of US patented drug candidates (Fig. 2e). A typical increase trend of AC_{50} and pPLE vs. HAC can be observed for HAC below 50. In turn for psychoactive PDSP drugs the increase of pAC_{50} vs. HAC was not enough strong to result in the increasing trend of pPLE (Fig. 2f). For a comparison in Figures 2 a-c we illustrated a plot of SILE which is somewhere between pPLE and pHAC.

In Figure 3a, we show the application of the pPLE to evaluate a series of drugs (1) and fragment-like drug candidates (25). Regardless of the HAC range, the pPLE value was always higher for the drugs than for the fragments. There was a clear separation of the cluster of drugs from the cluster of fragments. Furthermore, all of the drugs had a higher PLE value than the fragments. Accordingly, the

pPLE could be a well-balanced predictor that can clearly indicate the development from fragments to drugs. In turn, Figure 4b illustrates the LE statistics for the same data. An analysis of the plot of LE vs. HAC shows that for a given HAC above 20, the drugs had a slightly higher value of LE than the fragments. However, for HAC values below 20, LE for the fragments is higher than any drug can achieve.

The most important function of LE is its use as a guideline during the hit and lead optimization (22-25) where it forces slim pharma at the same time preventing molecular obesity. The mechanism for that was shown in Figure 1. In such a functionality the association of AC_{50} and HAC is just a scoring function enabling decision making in the drug pipeline. Below we designed a versatile function can be formed in which the effect of the pHAC and pAC_{50} components can be tuned by additional a and b parameters:

$$\text{SCORE} = a \cdot pAC_{50} + b \cdot \text{pHAC} \quad (3)$$

if a and b = 1, then SCORE = pPLE.

Mathematically, the numerical values of HAC are usually between 1 and 200, while AC_{50} , usually is in the range of 1 to 7; thus, $1/HAC$ clearly dominates in LE. In other words, LE as a predictor emphasizes low HAC more than high potency. In turn, AC_{50} (the range of 10^1 to 10^7) dominates in pPLE. SCORE was designed as a predictor for the more balanced scoring of the interaction between AC_{50} and the HAC. This function can be adjusted flexibly to change the relative SCORE rank of the fragments and drugs themselves as well as fragments and drug clusters. In particular, changing the a and b parameters can model different fragments to drug development strategies. Accordingly, depending on the preferences we can fine tune the SCORE

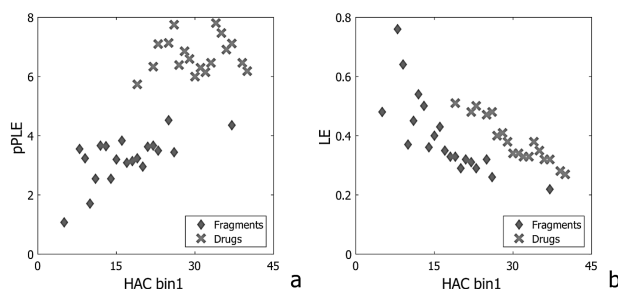


Figure 3. The dependence of pPLE (a), LE (b) on HAC for drug fragments and drugs

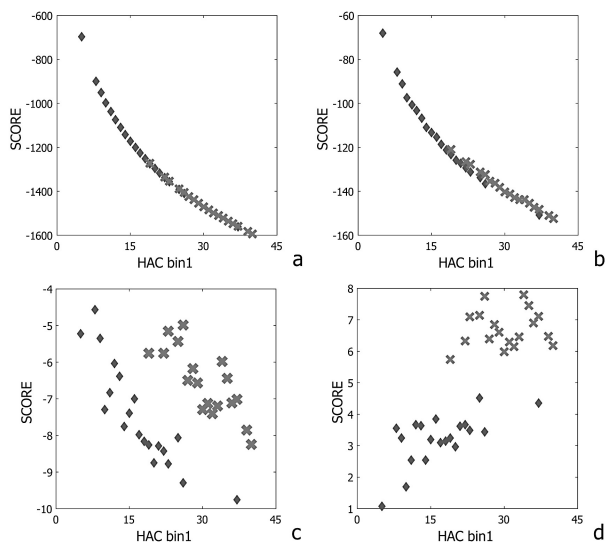


Figure 4. The dependence of SCORE on HAC for drug fragments and drugs for the different a and b values: $a = 1$, $b = 1000$ (a); $a = 1$, $b = 100$ (b); $a = 1$, $b = 10$ (c); $a = 1$, $b = 1$ (d)

function in order to adopt the strategy that is required by a specific drug development project.

CONCLUSION

We explained here the molecular reason for why LE prefers a low HAC to a level that practically masks the influence of AC_{50} . As drug design is a complex problem and drug-likeness prefers small molecules similar to LE, therefore, LE performs unexpectedly well in drug development. An important advantage of LE is that if used as a molecular filter for decision making it prefers small molecules and currently small molecules are among the most promising targets of slim pharma due to the advantageous drug-like profiles. It is however, not only the above mentioned feature of LE that contributed to the efficiency of LE. We have a number of methods for the estimation (prediction) of drug-likeness, e.g., ADMET or Lipinski's rule of five. Although usually these methods are described as drug-like property filters expected to provide a higher success ratio at the advanced development stages based on the values of the molecular feature that are typical

for drugs, precisely speaking, these filters cannot usually be based on the *properties* that are measured because molecules that are being designed at this stage of projects are not available for experimental measurements. Accordingly, they are molecular descriptors that are calculated for the molecular representations in order to predict the desired properties of the respective substances. The most obvious example is logP, in which based on ca. 30000 logP measurements, a regression model is built to provide predictions for any real or virtual molecule. In comparison, LE is a property related function, because we should know a real biological activity to calculate LE. Accordingly, we are not assuming or predicting positive activity, but we do know this fact. This is what makes LE the most reliable marker of drug-likeness.

We showed here that the product ligand efficiency (PLE), which is the product of the pAC_{50} values with the HAC could be an informative drug candidate estimator in which pAC_{50} and the HAC were balanced more evenly. At the same time, the physical meaning of the PLE and other efficiency estimators drew our attention to the relation between vari-

ous chemical representation of chemical compounds, i.e., molar properties, vs. descriptors vs. single molecule properties. On the other hand, finding ligands in drug design can be illustrated by playing between their matching vs. binding ability. In this context, the multiplicative PLE statistics indicate that playing between matching and binding is especially promising at around 30-50 HAC, where PLE takes the maximum value. Moreover, the SCORE predictor was designed for the flexible fine tuning of the ligand HAC vs. the IC₅₀ interaction as may be required by a specific drug development project.

Acknowledgments

Financial support was received from NCBR (Warsaw): ORGANOMET No: PBS2/A5/40/2014, TANGO1/266384/NCBR/2015. The anonymous Reviewer for the indication of the LE function behavior.

Conflicts of interest

The authors declare no competing financial interest.

REFERENCES

- Hopkins A.L., Keseru G.M., Leeson P.D., Rees D.C., Reynolds C.H.: *Nat. Rev. Drug Discov.* 13, 105 (2014).
- Williams G., Ferenczy G.G., Ulander J., Keseru G.M.: *Drug Discov. Today*, 22, 681 (2017).
- Hann M.M., Leach A.R., Harper G.: *J. Chem. Inf. Comput. Sci.* 41, 856 (2001).
- Zartler E.R., Shapiro M.J.: *Curr. Opin. Chem. Biol.* 9, 366 (2005).
- Kuntz I.D., Chen K., Sharp K.A., Kollman P.A.: *Proc. Natl. Acad. Sci. U.S.A.* 96, 9997 (1999).
- Reynolds C.H., Bembek S.D., Tounge B.A.: *Bioorganic Med. Chem. Lett.* 17, 4258 (2007).
- Reynolds C.H., Tounge B. A., Bembek S. D.: *J. Med. Chem.* 51, 2432 (2008).
- Reynolds C.H., Reynolds R.C.: *J. Chem. Inf. Model.* 57, 3086 (2017).
- Shultz M.D.: *ACS Med. Chem. Lett.* 5, 2 (2014).
- Shultz M.D.: *Bioorganic Med. Chem. Lett.* 23, 5980 (2013).
- Zhou H., Gilson M.: *Chem. Rev.* 109, 4092 (2009).
- Murray C.W., Erlanson D.A., Hopkins A.L., Keseru G.M., Leeson P.D. et al.: *ACS Med. Chem. Lett.* 5, 616 (2014).
- Polański J., Tkocz A., Kucia U.: *J. Cheminform.* 9, 49 (2017).
- Polański J., Tkocz A.: *J. Chem. Inf. Model.* 57, 1321 (2017).
- Sheridan P.R.: *J. Chem. Inf. Model.* 56, 2253 (2016).
- Scott J., Waring M.: *Bioorganic Med. Chem.* 26, 3006 (2018).
- Polański J.: in *Silico Encyclopedia of Bioinformatics and Computational Biology*, Ranganathan S., Gribskov M., Nakai H., Schonbach Ch., Eds, Vol. 2, pp. 601-618, Elsevier 2019.
- Knight A.: *Single Molecule Biology*. 1st ed., Academic Press, New York 2009.
- Leake M.: *Philos. Trans. R Soc. Lond. B Biol. Sci.* 368, 1611 (2013).
- Hann M.: *MedChemComm* 2, 349 (2011).
- Shultz M.D.: *J. Med. Chem.* 62, 1701 (2019).
- Mignani S., Rodriguez J., Tomas H., Jalal R., Parvinder P.S. et al.: *Drug Discov. Today* 23, 605 (2018).
- Meanwell N.A.: *Chem. Res. Toxicol.* 29, 564 (2016).
- Cavalluzzi M.M., Mangiatordi G.F., Nicolotti O., Lentini G.: *Expert Opin. Drug Discov.* 12, 1087 (2007).
- Schultes S., de Graaf C., Haaksma E., de Esch I., Leurs R. et al.: *Drug Discov. Today Technol.* 7, e157 (2010).
- Nissink J.: *J. Chem. Inf. Model.* 49, 1617 (2009).
- Mortenson P.N., Erlanson D.A., de Esch I.J.P., Jahnke W., Johnson C.N.: *J. Med. Chem.* 61, 1774 (2018).
- Johnson C.N., Erlanson D.A., Murray C.W., Rees D.C.: *J. Med. Chem.* 60, 89 (2016).
- Johnson C.N., Erlanson D.A., Jahnke W., Mortenson P.N., Rees D.C.: *J. Med. Chem.* 61, 1774 (2018).

Received: 6.01.2019

REVIEW ARTICLE

Ligand Potency, Efficiency and Drug-likeness: A Story of Intuition, Misinterpretation and Serendipity

Jaroslav Polanski^{a,*}, Anna Pedrys^a, Roksana Duszkievicz^b and Urszula Kucia^a

^aInstitute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland; ^bDepartment of Pharmacology, School of Medicine in Katowice, Medical University of Silesia, Medyków 18, 40-752 Katowice, Poland

ARTICLE HISTORY

Received: September 24, 2018
Revised: January 11, 2019
Accepted: April 07, 2019

DOI:
10.2174/1389203719666190527080832

Abstract: The concept of ligand potency is briefly discussed here as well as why this is still a challenge for its complete comprehension. In this context, we explain also the meaning of ligand efficiency (LE), which has been greeted with both enthusiasm and criticism among the drug design audience. A full understanding of LE requires the complex interpretation of the potency concept presenting the uncertainty similar to this of the Zeno paradox. In reality, the efficiency of LE is caused by the high degree of preference for slim pharma drug candidates.

Keywords: Ligand efficiency, potency, biological activity, Big Data, drug design, PubChem, ChEMBL.

1. INTRODUCTION

Ligand potency is an example of an innovative, serendipitous and intuition-based idea that has formed our understanding of biological activity of ligands. Both intuition and serendipity are of significant value in the advancement of science. Breakthrough innovations and theories often not only need outstanding efforts and years of research, but they also often require a change in the dominating beliefs and common sense in science. Routine is a conservative element that controls human activity. This prevents innovation, which in turn, is necessary for developing new ideas and creativity. In turn, new ideas often are imprecise. At the beginning, we usually have only a partial understanding of the processes, mechanisms and models. This forms the perfect environment for serendipitous discoveries. In particular, drug development has been founded on serendipitous discoveries. Basically, what we currently understand by molecular design is the search for molecules that have the potential to interact with biological macromolecular receptors. These interactions should produce signals that are capable of changing the signaling pattern that is destroyed by illness and therefore cure a patient. Organisms are, however, complex systems and we can easily understand that at the time when pioneering drug developers were targeting potential bioactive molecules or drugs, a quantitative description of the drug-receptor interactions was far from trivial, because our understanding and knowledge of the protein receptor structure and functions was quite low.

In reality, the original idea of potency was based on intuition. We will show that although our awareness of the processes that determine the biological activity of ligands has

significantly increased, having a handle on a mathematical and physical formalism of potency is still a challenge to our everyday common sense. One can doubt that this could still be true when we have described many single mechanisms that determine the functions and interactions of proteins. Let us, however, illustrate the limitations of our brains *via* the example of the so-called Zeno's paradox. Accordingly, in this example, Zeno is following a turtle. We know from everyday experience that a human quickly catches up to this slow animal. However, let us change our view slightly using model in which Zeno closes in on the turtle only half of the distance at each time interval, thus never catching up to the animal. Although the second model evidently violates common sense, without a mathematical background, it is not easy to properly understand the problem and answer the question as to why this ancient paradox cannot properly describe the Zeno-turtle system even today.

In this short review, we analyze the development and meaning of potency and its transform ligand efficiency (LE) and show how efficient this kind of representation of biological activity appears to be but also how challenging this has become for a complete and proper understanding and quantitative description of the effects that are related to ligand-receptor interactions. Although we are more and more aware that kinetic data may provide an important improvement in the description of ligand-receptor systems, a simple analysis of the availability of the potency vs. kinetic types of data evidently prove that potency is still the essential estimator that is used by drug designers.

2. POTENCY – THE QUANTITATIVE REPRESENTATION OF THE MATHEMATIC AND PHYSICS OF BIOLOGICAL ACTIVITY.

The idea of potency was pioneered by Hill, who used it to describe the binding of oxygen by hemoglobin [1]. The

*Address correspondence to this author at the Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland; Tel: +48 669 446 570; E-mail: jaroslav.polanski@us.edu.pl

most concise account of this was that Hill had developed a Langmuir-like model to describe the signaling that is activated by the ligand-receptor system. The only problem of this account is that the original Hill model was developed about ten years before the Langmuir equation was developed. We can now ask the question of why we refer to the Hill model as a Langmuir-type model and not vice versa as the Langmuir model as a Hill-like representation. The answer is simple. Langmuir described gas adsorption as a phenomenon that was much closer to the perception of the chemists of that time. This could be represented by a real-life model, and consequently, was understood more easily than the action of a drug. In the early 20th century descriptions of proteins, receptor macromolecules were completely beyond the realm of our observation and comprehension. Interestingly, Hill concluded wrongly, on the basis of temperature dependence, that the time course is limited by receptor interaction rather than diffusion. This misinterpretation continued for decades clearly indicating the serendipitous origin. However, its intuitive basis and even this early misinterpretation did not prevent its usefulness and further function in chemistry and biology [2].

In Fig. (1), we briefly compare the physics and mathematics of the Hill and Langmuir effects. The sigmoid-like inhibitory concentration (IC_{50}) curve that reports a biological effect is a representation that is commonly known in contemporary chemistry and biology. This representation is, however, much more complex than we usually recognize. To explain this, let us compare the potency-type representation of a biological response to the so-called single-molecule model [3, 4] (Fig. 2), in which by the latter, we understand the description of the ligand-receptor system as the microscopic data that is available using the current sophisticated biophysical methods. Both put the formation of a signaling ligand at the center of the model. However, the single-molecule model cannot provide a proper understanding of the Hill representation, which transforms the activity into a concentration scale [mol/L].

Unexpectedly, the difference between the single-molecule vs. potency model resembles a dichotomy between the molecular descriptors that refer to the molecules (usually, but not always, being calculated) and the properties that are usually measured for the molecular assemblies arranging molar representations (moles). The transformation of mole-

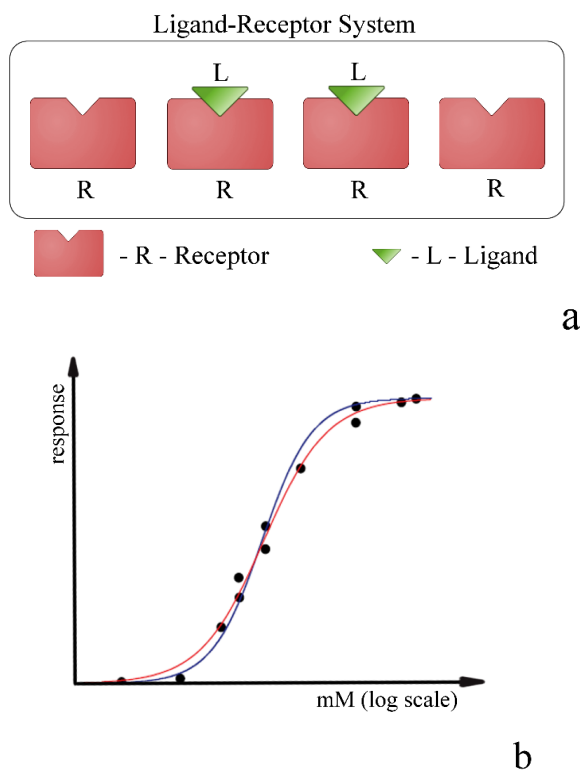


Fig. (1). A physical representation of binding/adsorption with occupied (green) and unoccupied (white) receptor sites (a) and a sigmoid concentration-response curve modeled by the Hill (red) and Langmuir equations (b). Models (b) modified after [2]. (The color version of the figure is available in the electronic copy of the article).

cules to moles, which is defined by the Avogadro statistics, is a foundation of chemistry which broad historical background can be found in the reference [5]. In turn, the property-descriptor dichotomy was discussed in the references [6-10].

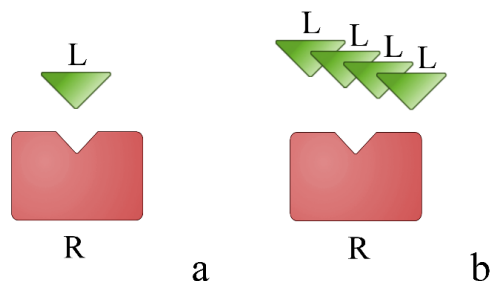


Fig. (2). Single molecule biology (a) vs. potency (b) models of biological activity. (The color version of the figure is available in the electronic copy of the article).

3. BIOACTIVITY AS A BIG DATA PROBLEM

Currently, people are fascinated by the possibility of managing and processing large data sets and the possibility that is offered by computers in this field. In fact, there are good examples of the efficiency of processing large amounts of data in science and everyday life, which can be compared using the prognostic capability of the Google flow simulation [11]. In contemporary chemistry, there is a flood of data that is difficult to process and explain. In reality, however, molecules are not at the top of information complexity.

In Fig. (3), we analyze the complexity of the sciences according to the way in which we explain a certain field of research by reducing its complexity. Therefore, physics offers a less complex representation of chemical objects. Similarly, chemistry reduces the complexity of biology [6, 7, 8, 12]. We can see that chemistry and biology are not at the top of this pyramid. By the way, chemists have only synthesized not more than 200 million chemical molecules, while seven billion human phenotypes populate the world.

In this context, let us consider the biological activity data. Identifying selective ligands and measuring their potency is a complex problem. For a recent review in this field, readers should especially see the reference [13]. The biological activity of chemical compounds, when described on a concentration scale, can be quantified by either their efficacy (maximum effect) or potency, *i.e.*, the amount of the drug that is required to produce a given effect. The most common potency format is the drug dosage or concentration that causes 50% of the maximum effect, which is indicated by an index of 50. However, the chemical compounds can cause different effects. Accordingly, an inhibitory concentration, pIC_{50} , an effective concentration, pEC_{50} , a cytotoxic concentration, pCC_{50} , and so forth are various variants of this data format. In turn, the relationship between the binding constants of the ligands and potency is the next interesting problem that illustrates the collision of theory and practice [2, 14].

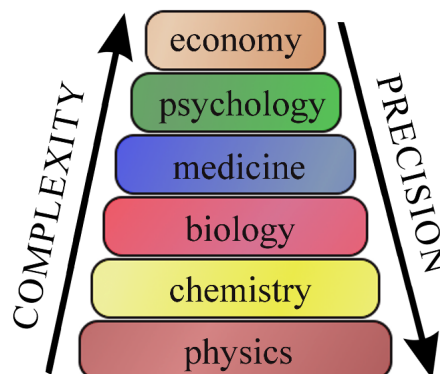


Fig. (3). Complexity vs. precision in sciences. Modified from [7]. (The color version of the figure is available in the electronic copy of the article).

Practically, potency is the only available representation of the properties of the ligand-receptor systems which could be related to big numbers. Potency data are available in several databases. In this context, major chemistry databases have recently been critically reviewed and compared by Southan [15], who indicated that PubChem, UniChem and ChemSpider are the main trio of the “largest public data sources for bioactive chemistry and drug discovery”. These databases have reported significant growth in recent years. For example, PubChem has registered 97 million distinct chemical structures. Practically, the databases include what Southan refers to as identical content, *e.g.*, from the 1.7 million ChEMBL compound records, 0.35 million records were imported from PubChem. This means some of the data are recorded twice.

The ChEMBL and Binding databases are important activity data providers; for example, Binding has registered 69,000 records including 11,000 curated patent entries by Southan [15]. One should compare this with the number of approved drugs, a popular subset in PubChem, which is 1,180 individuals.

The next problem is the specific representation of the type of biological activity in the databases. In the databases, some potency records are represented by descriptions that are difficult to process in large data-type analyses. In Table 1, we compare the databases that provide the large number potency records after [15]. We also presented here an analysis of the downloads of the PubChem and ChEMBL records, which were annotated by numerical potency values.

The larger the amount of data is the less precise and unique is the data format. This relationship can be easily understood, because in order to retain a unique format and precision, we should eliminate some measurements. Accordingly, precision decreases with an increase in data complexity (Fig. 2). The same refers to the potency data. The common label active concentration, AC_{50} , is sometimes used to unify all of the potency data types. ChEMBL, PubChem and ZINC are large databases that register the data on drugs and drug candidates. The ChEMBL data are manually curated and are the most reliable biological activity data for a popu-

Table 1. Potency data vs. big molecular data^a.

Source	Count/Unique ^[a]	Potency data available ^[b]
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	-

[a] As millions of compounds; data modified from [15]; count: the latest available number, Unique [15] (PubChem, ChEMBL).

[b] Available for downloads as numerical values; <https://pubchem.ncbi.nlm.nih.gov/>, downloaded August 2017 [32]; <https://www.ebi.ac.uk/chembl/>, ChEMBL, version 24, May 2018.

lation in the order of 10^5 . About half of these data are given as the standardized values of inhibitory concentration, IC_{50} (approximately 400,000 records). In turn, PubChem registers a much broader population of chemical compounds in the order of 10^6 , and therefore, the data is in different formats. Biological activity data are primarily collected in PubChem under their common active concentration label, AC_{50} , often without identifying which concentration (inhibitory, effective, cytotoxic, *etc.*) this relates to. Thus, a broad spectrum of the biological behavior of chemical compounds is involved at the cost of a lower specificity of the effects that are described. The large numbers of molecular data do not mean the availability of a similar number of BIO records. ZINC is the largest register of commercially available compounds for virtual screening. The population size is of the order of 10^7 , but the compounds are not annotated by activity data.

Several statistical analyses of the PubChem and ChEMBL data are presented in Fig. (4). Interestingly, if we compare the distribution of the pAC_{50} values with the probability of detecting, measuring and ligand-receptor matching [16, 17], we can see an interesting correlation between the individual plots of the distribution of the probability and potency values of the pAC_{50} values.

4. BIOLOGICAL ACTIVITY VS. DRUG-LIKENESS, SLIM PHARMA AND LIGAND EFFICIENCY

How high should the potency for an efficient drug be? After analyzing the properties of the data for the bioactive compounds that have been reported in chemical publications, Walters indicated “an acceleration in the upward trends of several properties, noticeably molecular weight (MW) and $\log P$ ” [18]. By the way, in Fig. (4c), one can compare this trend to the probability of measuring, matching and detecting the binding affinity vs. the molecular complexity [16, 17]. If we ask why medicinal chemistry targeted this area, we can easily realize that determining the highest potency was the

intuitive goal that was sought after. Using common sense, this can be achieved by increasing the complexity of drug candidates. Therefore, the published drug projects involved higher and higher MW and $\log P$. Paradoxically, these values are “moving away from the ranges that are typically found in successful drugs” [18].

An interesting practical question is how high should the optimal drug potency be? The median pXC_{50} values of historical oral drugs indicate that a value of ~ 7.7 (~ 50 nM) of *in vitro* potency is optimal. In turn, a therapeutic dosage correlates with MW ($r^2 = 0.10$) only slightly. Similarly, the correlation between pXC_{50} and MW ($r^2 = 0.16$) is unexpectedly low [19]. What can the highest binding affinity be was the next interesting problem that was addressed by Kuntz *et al.* [20]. This problem was analyzed in a series of further publications by several groups, for review compare references [21, 22], who designed a way to measure the ligand binding performance, the so-called ligand efficiency (LE). Fig. (5) briefly illustrates this concept. Unexpectedly, LE appeared to highly prefer a low molecular weight or the low so-called heavy atom count (HAC) ligands. For a long time, this effect was a puzzle and could not be explained. The mathematical formula for LE has been questioned in an effort to find the reason for this puzzle [23-25].

For example, an empirical formula of the so-called size-independent ligand efficiency (SILE) was proposed to correct the LE vs. HAC relationship for a model [26]. In fact, the LE puzzle appeared so astonishing that the fuel efficiency of a car was used to explain the drug LE trend [22]. It is only recently that the that LE trend could be fully understood and explained based on its physical representation [9, 10].

Fig. (6) explains the actual meaning of the properties of LE and the Binding Efficiency Index (BEI), which is a measure that is related to LE. Potency is a thermodynamically based measure that obeys the Avogadro statistics,

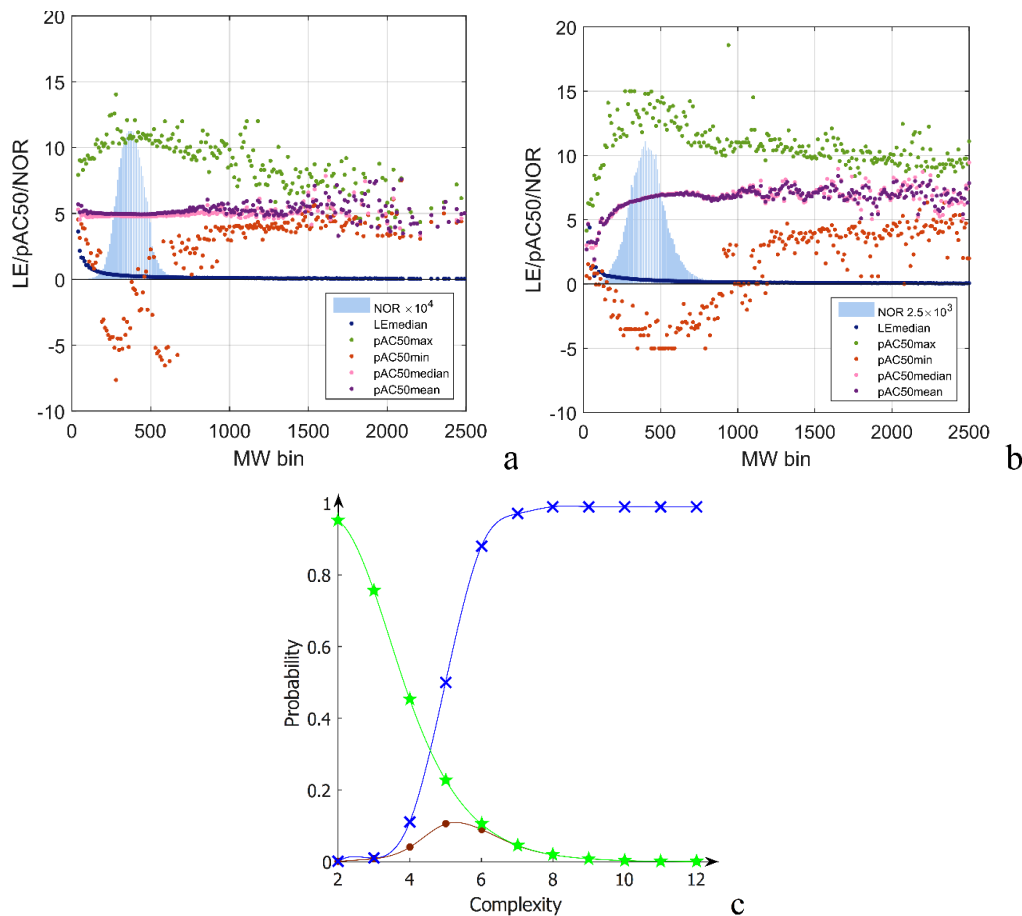


Fig. (4). Statistics of numerical ChEMBL (a); PubChem (b) records compared to the probability of matching and/or measuring ligand/receptor interactions vs. the increasing ligand complexity (c). NOR histogram – number of records; probability of: matching any way (green), measuring binding affinity (blue), detecting a useful event (red). Probability plot (c) modified from [18]. (The color version of the figure is available in the electronic copy of the article).

which preserves the number of molecules as an invariant for molecules of different sizes. Let us focus here on the BEI measure, which is slightly easier for the direct interpretation than LE itself. Unexpectedly, unlike potency, BEI is a measure that is based on the weight invariant of 1 (k)g in which we should observe that the number of interacting molecules depends on the molecular size, *e.g.* it should decrease for higher MW (HAC) molecules [9, 10].

Unexpectedly, unlike potency, BEI is a measure that is based on the weight invariant of 1 (k)g in which we should observe that the number of interacting molecules depends on the molecular size, *e.g.* it should decrease for higher MW (HAC) molecules [9, 10]. Accordingly, a comparison of the molecules of different sizes is skewed by the 1/MW statis-

tics, which define the different numbers of ligands that are interacting with a receptor. We showed that because LE is closely related to BEI, it behaves similarly to LE [9, 10]. Then, the LE vs. HAC (BEI vs. MW) should be illustrated by a hyperbolic-like plot, which, in fact, is what we are observing. Within the context of maximizing LE, a hyperbolic relationship firmly prefers low MW (HAC) ligands. However, this effect is not due the maximization of potency itself, but rather is due to a dramatically changing number of ligands for the small and high MW ligands. The other aspect of this problem could be summed up after Scot and Waring: “We have generally not been examining changes that significantly altered the size of the compounds and hence LE has been uninformative” [27].

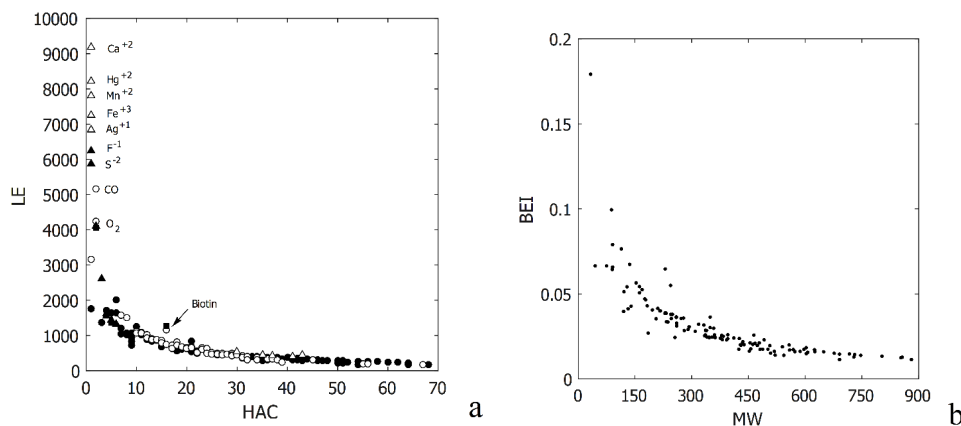


Fig. (5). The dependence of LE vs. HAC and BEI vs. MW for a series of ligands. A plot of free energy of binding per atom in (a) modified after [20] plot in (b) after MW data recovery from [20].

Slim pharma has recently been declared to be the preferred drug-design trend [28]. In this context, we can conclude that LE forms a serendipitous base for the preference of medicinal chemistry that is based on small MW or HAC ligands, thereby following the slim pharma concept. Small ligands usually have better pharmacological, ADMET and lipophilic profiles and therefore form better drug-like candidates. For the recent updates in this topic compare recent review [29]. Small is beautiful [30], which is a synonym for slim pharma. In view of this, LE is a potency transform that optimizes the potency only slightly but is a serendipitous estimator that can be used to cut off high-molecular ligands. For an informative perspective review on the influence of drug-likeness, readers should see the latest reference [29].

As LE appeared useful for drug development, a variety of rules that controls LE have been described. This includes the impact of binding site nature and the target class, *e.g.*, inhibitors of protein–protein interactions versus enzyme inhibitors [21]. In turn, a large survey of the influence of the receptor size on LE a reader can find in the reference [31]. As a mechanism of ligand-receptor interactions influences the potency values, this also influences the LE values, for example, covalent inhibitors can have very high ligand efficiency values. In addition, larger, more optimized molecules may have smaller ligand efficiencies, which is due in part to the fact that such compounds have been optimized for properties other than binding to the target (for instance, pharmacokinetics, solubility, selectivity or cell permeability) and in part to the large HAC values. It is also vital to remember that ligand efficiency has limited value in late-stage lead optimization and that it could primarily be used in hit identification and hit-to-lead conversion. Alternative interpretation of LE size dependency is that that advantageous ligand efficiency range for large molecules would require subnanomolar potency values, which may not be recorded (or recorded correctly) because of assay limitations which is not better than ~1 nm due to signal-to-noise issues [21, 31]. An interesting insight into LE trends for big potency data and various compound series can be found in reference [32]. For novel concepts in

scoring LE see reference [33]. From the theoretical point of view, it is interesting to compare the LE statistics with those that are used in economics in which the weight metric \$/g and not \$/mole is a commonly used measure. However, because large amounts of economic data are rarely available in drug design, as they are among the most secret numbers, in Fig. (7) we re-plotted several analyses for the catalogue of big economy related molecular data [34].

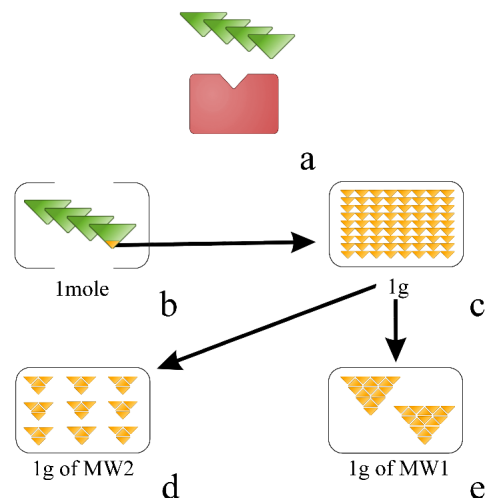


Fig. (6). Physical meaning of BEI (LE). An invariant is a mole of molecules in the potency model (a) according to the standard Avogadro statistics (b) while in the BEI transform a 1D fragment a „virtual” mole of 1 Dalton fragments form an invariant measure of 1g (c), which recovers a different number of ligands for the complete molecules of MW1 and MW2 (d,e) respectively. LE as a simple BEI like representation follows the BEI statistics with a „virtual” mole of 1 HACs. (The color version of the figure is available in the electronic copy of the article).

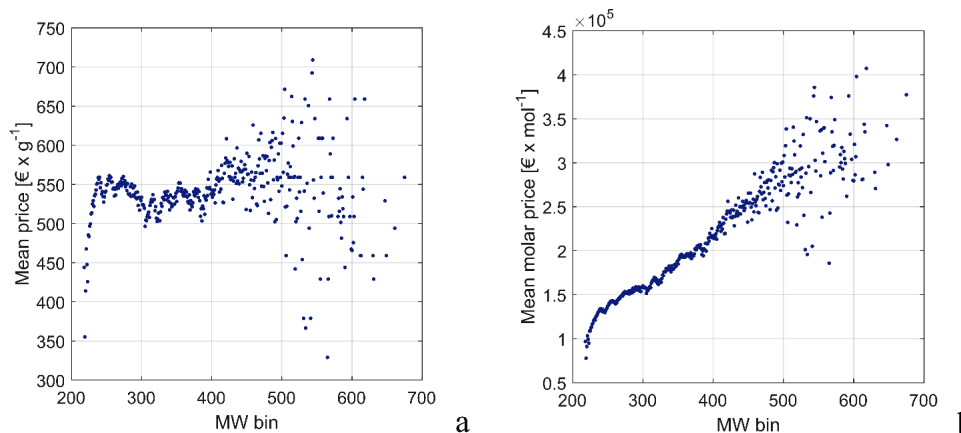


Fig. (7). Economic data in the BEI like (a) and molar scales (b). Modified from [31]. (The color version of the figure is available in the electronic copy of the article).

CONCLUSION

Although we often recognize drug design as being a process whose goal is to maximize the potency value, it is not always a recipe for a successful drug. In practice, a number of the so-called drug-like properties determine the potential of drug candidates, *e.g.*, ADMET and lipophilic behavior. Interestingly, however, although we always measure potency when designing drugs, the ADMET and drug-like parameters are usually predicted *in silico*, and they are very rarely measured. Of course, drug designers will argue that these parameters are measured for the important drug development projects. We agree with this but if one takes an average publication whose aim is to find new bioactive molecules, often potency is then the only property that is measured. Moreover, our experience indicates that the potency level also determines the interest of the audience and thus the potential for the results to be published. Therefore, while we absolutely cannot perform any drug design without measuring the biological activity, the ADMET and drug-like parameters are also very important; however we can do without their measured values. The reason the measurements platform is limited to a single property is clear: measurements are expensive. This effect was described as being a property-deficit environment [6, 12]. It is not only chemistry or drug design but economics that finally determines the current shape of pharmacology [35].

LE is a parameter that is designed to answer the question of what the highest possible value of ligand potency is. If we would like to understand the chemical paradox of LE we should observe that the BEI or LE are based on the 1g or 1 HAC measure, *i.e.*, they are formed by a mole of Daltons or a mole of HACs. As mole of Daltons (HACs) are virtual measures we are observing a paradoxical behavior of these predictors. However, a careful analysis of the physical meaning of the LE parameter has proven that its efficiency in drug design is related to the optimization of the advantageous drug-like profiles that are preferred by the so-called slim

pharma concept. Serendipitously, LE appeared to suit the needs of the single-parameter-based estimator for the optimization of early drug design perfectly.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

The research presented in this publication was supported by NCBR Grant TANGO1/266384/NCBR/2015.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We thank anonymous Reviewer for valuable remarks on LE importance in drug development.

REFERENCES

- [1] Hill, A.V. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. In: *The Journal of Physiology*; Langley, Ed.; Cambridge University Press: London; **1910**; Vol. 40, pp. iv-vi.
- [2] Colquhoun, D. The quantitative analysis of drug-receptor interactions: A short history. *Trends Pharmacol. Sci.*, **2006**, 27, 149-157.
- [3] Knight, A. *Single Molecule Biology*. 1st ed.; Academic Press: New York, **2009**.
- [4] Leake, M. The physics of life: one molecule at a time. *Philos. Trans. R. Soc. Lond., B. Biol. Sci.*, **2013**, 368, 1611.
- [5] Bensaude-Vincent, B.; Simon, J. *Chemistry — The Impure Science*. 2nd ed.; Imperial College Press: London, **2012**.
- [6] Polanski, J. Chemoinformatics: From Chemical Art to Chemistry. In *Silico Encyclopedia of Bioinformatics and Computational Biology*; Ranganathan, Ed.; Elsevier, **2019**; Vol. 2, pp. 601-618.
- [7] Polanski, J.; Gasteiger, J. Computer Representation of Chemical Compounds. In: *Handbook of Computational Chemistry*; Leszczynski, Ed.; Springer: Dordrecht, **2016**, pp. 1-43.
- [8] Rosenblum, B.; Kuttner, F. *Quantum Enigma: Physics Encounters Consciousness*, 1st ed.; Oxford University Press: New York, **2006**.

- [9] Polanski, J.; Tkocz, A. Between descriptors and properties: understanding the ligand efficiency trends for G protein-coupled receptor and kinase structure-activity data sets. *J. Chem. Inf. Model.*, **2017**, *57*(6), 1321-1329.
- [10] Polanski, J.; Tkocz, A.; Kucia, U. Beware of ligand efficiency (LE): Understanding LE data in modeling structure-activity and structure-economy relationships. *J. Cheminformatics*, **2017**, *9*, 49.
- [11] Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature*, **2009**, *457*(7232), 1012-1014.
- [12] Polanski, J. Big Data in Structure-Property Studies—From Definitions to Models. In: *Advances in QSAR Modeling*; Roy Ed.; Springer: Cham, **2017**, pp. 529-555.
- [13] Aldrich, C.; Bertozzi, C.; Georg, G.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz, K.; Schepartz, A.; Wang, S. The ecstasy and agony of assay interference compounds. *ACS Cent. Sci.*, **2017**, *3*(3), 143-147.
- [14] Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem. Int. Ed. Engl.*, **2002**, *41*(15), 2645-2676.
- [15] Southan, C. Caveat USOR: Assessing differences between major chemistry databases. *ChemMedChem*, **2018**, *13*(6), 470-481.
- [16] Hann, M.; Leach, A.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*(3), 856-864.
- [17] Zartler, E.; Shapiro, M. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.*, **2005**, *9*(4), 366-370.
- [18] Walters, W.P.; Green, J.; Weiss, J.; Murcko, M. What do medicinal chemists actually make? A 50-year retrospective. *J. Med. Chem.*, **2011**, *54*(19), 6405-6416.
- [19] Gleeson, M.P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between *in vitro* potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.*, **2011**, *10*(3), 197-208.
- [20] Kuntz, I.D.; Chen, K.; Sharp, K.A.; Kollman, P.A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. USA*, **1999**, *96*(18), 9997-10002.
- [21] Hopkins, A.; Keseru, G.; Leeson, P.; Rees, D.; Reynolds, C. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.*, **2014**, *13*(2), 105-121.
- [22] Murray, C.; Erlanson, D.; Hopkins, A.; Keseru, G.; Leeson, P.; Rees, D.; Reynolds, C.; Richmond, N. Validity of ligand efficiency metrics. *ACS Med. Chem. Lett.*, **2014**, *5*(6), 616-618.
- [23] Kenny, P.; Leitao, A.; Montanari, C. Ligand efficiency metrics considered harmful. *J. Comput. Aided Mol. Des.*, **2014**, *28*(7), 699-710.
- [24] Matta, C.; Massa, L.; Gubskaya, A.; Knoll, E. Can one take the logarithm or the sine of a dimensioned quantity or a unit? Dimensional analysis involving transcendental functions. *J. Chem. Educ.*, **2011**, *88*(1), 67-70.
- [25] Zhou, H.; Gilson, M. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.*, **2009**, *109*(9), 4092-4107.
- [26] Nissink, J. Simple size-independent measure of ligand efficiency. *J. Chem. Inf. Model.*, **2009**, *49*(6), 1617-1622.
- [27] Scott, J.; Waring, M. Practical application of ligand efficiency metrics in lead optimisation. *Bioorg. Med. Chem.*, **2018**, *26*(11), 3006-3015.
- [28] Hann, M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm*, **2011**, *2*(5), 349-355.
- [29] Shultz, M.D. Two decades under the influence of the rule of five and the changing properties of approved oral drugs. *J. Med. Chem.*, **2019**, *62*(4), 1701-1714.
- [30] Williams, G.; Ferenczy, G.; Ulander, J.; Keseru, G. Binding thermodynamics discriminates fragments from druglike compounds: A thermodynamic description of fragment-based drug discovery. *Drug Discov. Today*, **2017**, *22*(4), 681-689.
- [31] Reynolds, C.H.; Reynolds, R.C. Group additivity in ligand binding affinity: an alternative approach to ligand efficiency. *J. Chem. Inf. Model.*, **2017**, *57*, 3086-3093.
- [32] Polanski, J.; Pedrys, A.; Duszkiwicz, R.; Gasteiger, J. Scoring ligand efficiency: Potency, ligand efficiency and product ligand efficiency within big data landscape. *Lett. Drug Des. Discov.*, **2017**, in print.
- [33] Polanski, J.; Duszkiwicz, R.; Pedrys, U.; Gasteiger, J. Scoring Ligand Efficiency. *Acta Pol Pharm*, **2019**, *76*(4) 761-768.
- [34] Polanski, J.; Kucia, U.; Duszkiwicz, R.; Kurczyk, A.; Magdziarz, T.; Gasteiger, J. Molecular descriptor data explain market prices of a large commercial chemical compound library. *Sci. Rep.*, **2016**, *6*.
- [35] Polanski, J.; Bogocz, J.; Tkocz, A. Top 100 bestselling drugs represent an arena struggling for new FDA approvals: Drug age as an efficiency indicator. *Drug Discov. Today*, **2015**, *20*(11), 1300-1304.

RESEARCH ARTICLE

Scoring Ligand Efficiency: Potency, Ligand Efficiency and Product Ligand Efficiency within Big Data Landscape

Jaroslav Polanski^{*a}, Anna Pedrys^a, Roksana Duszkiwicz^a and Johann Gasteiger^b

^aInstitute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland;

^bComputer-Chemie-Centrum, University of Erlangen-Nuernberg, Naegelsbachstrasse 25, 91052 Erlangen, Germany,



Abstract: Background: Potency is the broadest available biological activity data type. In turn, ligand efficiency (LE) is a molecular descriptor that probes the ratio of potency vs heavy atom count (HAC), which emphasizes low HAC more than potency and thus has drawbacks as an estimator of drug candidates.

Objective: The objective was to design a novel transform to probe potency and HAC interaction in which potency and HAC would be balanced more evenly.

Results: A novel descriptor, a product of the pAC₅₀ value with HAC, multiplicative or product ligand efficiency (PLE) was designed and tested using the ChEMBL, PubChem, FDA approvals and drug (fragments) data. In particular PLE was compared with pAC₅₀ and LE vs the HAC statistics for different series of ligands. This indicated that PLE is an informative estimator that can be used to recognize the potential of drugs.

Conclusion: Drug design is a complex problem. Similarly, to drug-likeness, LE prefers small molecules. This makes LE a tool serendipitously improving drug likeness. In this context, LE performs unexpectedly well even despite the uncertainty of its physical meaning. PLE is a more evenly balanced estimator whose physical meaning is the minimum inhibitory concentration (MIC). PLE has a maximum value in the range around 30-50 HAC.

Keywords: ligand efficiency, product ligand efficiency, heavy atom count, activity, drug design, PubChem, ChEMBL

1. INTRODUCTION

Optimizing the binding ability of ligands is the main purpose of drug design. Although a number of molecular descriptors can be involved in this process, the simplest molecular representation in drug design is just molecular size [1, 2]. In this context, the effect of molecular size on the probability of identifying new ligands can be described by the dichotomy between the ability of a ligand to fit the receptor and its potential to bind this receptor [3, 4]. While it is more probable for a smaller molecule to fit the receptor, binding potential can increase with its molecular size. Accordingly, although it is more probable to identify small ligands, the affinity of ligands towards receptors can drop below a measurable level for molecules (fragments) that are too small [3, 4].

Which descriptors can be used to control molecular size of the ligand populations that are broader and broader in contemporary drug design? Heavy atom count (HAC) or molecular weight (MW) are simple molecular representations that are often used in such a function, while the so-called ligand efficiency (LE) is a single number descriptor that explores the interaction between the binding potential of the ligand and HAC or, more precisely, 1/HAC, since LE is defined as the ratio of binding energy to HAC [1, 5-8]. A variety of LE-based projects have appeared to be surprisingly successful [1] despite the fact that the observed trend of LE seemed to be paradoxical and therefore could not be fully understood [9-15].

In LE calculations we use binding energy, which is a property of the molecular ensemble that is usually represented by ligand potency measured in the form of the IC₅₀ (molar concentration) values. In turn, HAC relates to a single molecule and molar representations [16] of 1 HAC is molecular fragment that does not have molar representation.

*Address correspondence to this author at the University of Silesia, Institute of Chemistry, Szkolna 9, 40-006 Katowice, Poland Tel: +48 669 446 570; E-mails: jaroslav.polanski@us.edu.pl

By the way, the analogous molar representation of 1 Da can be interpreted virtually as 1 gram. This brings an unwanted uncertainty into the meaning of LE [13,14]. Therefore, a well-defined and correctly performing predictor that is capable of optimising the ligand binding affinity and monitoring both the binding energy or potency and molecular size would significantly contribute to current drug technology.

In this publication we use the largest available databases of biological activity measurements catalogued by ChEMBL and PubChem to model possible landscapes of potency vs HAC. We show that possible scenarios in the HAC domain, if represented by the LE vs HAC relationship, plot very similarly. As an estimator of drug candidates or fragments LE first of all prefers small HACs [13, 14]. In turn, probing the interaction of potency and HAC using the product ligand efficiency (PLE) appeared to be more balanced when scoring the ligand binding potency vs molecular size with a single number predictor. Because low HAC (low values) and high potency (low values) indicate attractive drug candidates, then, the both quantities (HAC and IC_{50}) are synchronized in PLE. We determined the physical meaning of PLE as the minimum inhibitory concentration (MIC) and also tested its behaviour using the largest amount of molecular potency data available in the PubChem and ChEMBL databases as well as in FDA approvals and a selected series of drugs or drug candidates [1, 17].

2. MATERIALS AND METHODS

2.1 Data

We used potency data in a form of AC_{50} as defined in the PubChem classification. Accordingly, we called all data that were used an active concentration, AC_{50} . For PubChem, AC_{50} stands for inhibitory concentration, IC_{50} , effective concentration, EC_{50} , cytotoxic concentration, CC_{50} , equilibrium dissociation constant, K_i , for the ligand determined directly in a binding assay using a labelled ligand or dissociation constant, K_d , for the ligand, determined in inhibition studies. For ChEMBL the AC_{50} values are either IC_{50} , K_i or K_d values.

2.2 Binding energy, potency and ligand efficiency

Binding energy and potency can be related to each other by a simple equation, as shown in Eqs. 1, 2, and 3 as well as [1]:

$$\Delta G^\circ = -2.303 \cdot RT \log K_d \quad (\text{Eq. 1})$$

or

$$\Delta G^\circ = 1.37 \cdot pIC_{50} \quad (\text{Eq. 2})$$

or more general

$$\Delta G^\circ = 1.37 \cdot pAC_{50} \quad (\text{Eq. 3})$$

where ΔG° is the Gibbs free energy of binding, R is the ideal gas constant, T - temperature in Kelvin, K_d - dissociation constant, pIC_{50} - inhibitory concentration in the negative log scale [1]. Accordingly, potency directly explores the binding energy potential.

LE is commonly calculated from equation [1]:

$$LE = (1.37 \cdot pAC_{50}) / HA \quad (\text{Eq. 4})$$

A meaning of LE as a descriptor is a share of binding energy per HAC; however, since HAC does not have any physical

meaning and binding energy or IC_{50} is a statistical property measured for the molar representation and not for a single molecule representation it has been shown that physical meaning of LE is binding energy for the ensemble of ligands of the weight of 1 gram [13, 14].

2.3 Physical meaning of product ligand efficiency, PLE, and product ligand efficiency index, PLEI

We define product ligand efficiency index, PLEI:

$$PLEI = AC_{50} \cdot MW \quad (\text{Eq. 5})$$

which have the unit of [(mole/l)*(kg/mole)]=[kg/l]. This means that PLEI takes the dimension of minimum inhibitory concentration (MIC).

We define PLE as the product $AC_{50} \cdot HAC$. As a molecular descriptor this probes the interaction of AC_{50} and HAC. Accordingly:

$$PLE = AC_{50} \cdot HAC \quad (\text{Eq. 5a})$$

As $MW[\text{kg/mole}]/MW[\text{Da}]=1$, PLE can be defined by:

$$PLE = PLEI \cdot (HAC/MW) \quad (\text{Eq. 5b})$$

AC_{50} is a concentration-based property having the dimension of [mol/l]. This means that the physical meaning of PLE is the minimum inhibitory concentration (MIC) scaled to HAC.

2.4 Data acquisition and calculations

The FDA approval data (34,415 entries) were collected from the Binding database. All records with numerical values downloaded from the ChEMBL and PubChem databases were used in the analyses. Records repeating individual compounds were treated as independent entries for all data analysed. PubChem, 2,435,467 records (download: August 2017, pubchem.ncbi.nlm.nih.gov) and ChEMBL 714,791 records (ChEMBL version 22, www.ebi.ac.uk) were downloaded from the Internet sites, respectively. pAC_{50} data for drugs or drug fragments were taken from the literature as pIC_{50} values [1, 17]. The mean values for drugs are HAC 31; pIC_{50} 8.12 and for fragments HAC 15; pIC_{50} 4.41.

2.5 Binning

For plotting $pPLE$ and pAC_{50} vs. HAC we used a binning method, in particular high-resolution binning [19]. In binning, data which fall into a given interval, a bin, are replaced by a value representative of that interval. Individual single HAC numbers define here the size of intervals, while AC_{50} values are represented by their mean value. For $HAC > 60$, where not enough data were available, HAC was binned for each 15 HAC.

3. RESULTS AND DISCUSSION

In Fig. (1), the PubChem AC_{50} data were plotted against HAC. In general, for a random population one should expect the binding energy (or in turn also potency) not to be a function of molecular size because the ligand-target fit is a complex entropy-enthalpy interplay [18]. Actually, this appeared to be not far from truth for the large PubChem potency population which, on average, did not depend on HAC when enough records (number of records, NOR) were available. However, when only a small number of NORs was available, e.g., for $HAC > 100$ (Fig. 1) the relationship between AC_{50} and HAC was somehow irregular. For a large range of lower values of HAC the relationship of AC_{50} with

HAC was nearly constant with the mean potency pAC_{50} having a value of approximately five. More precisely, the mean potency slightly decreased from 5.3 to 4.9 when HAC increases from 2 to 50. Usually, in big data statistics, the correlations that are observed are more noteworthy for the small number statistics. Therefore, we may interpret this small decrease as being meaningful and explain this by the fact that the arrangement of the ligand and the target should become more difficult and less probable with an increase in the complexity of the ligand-target interactions that are caused by increasing molecular size. Although for smaller ligands a perfect fit can even happen accidentally, this is more and more unlikely when ligands are larger. This effect is rather minor and the average pAC_{50} only drops ca. 0.4 pAC_{50} units in the range of up to 50 HAC (Fig. 1.a).

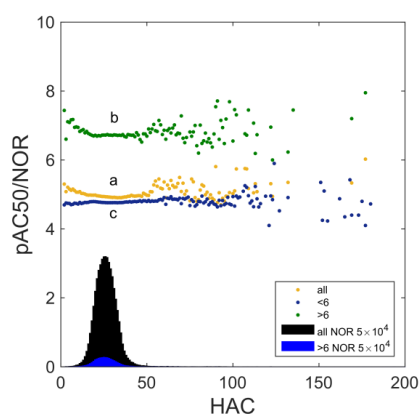


Fig. (1). Mean potency pAC_{50} values binned against HAC for a population of more than 2.3 million potencies recorded in PubChem (a) or for the PubChem subpopulations with pAC_{50} values above 6 (b) and below 6 (c). Histograms below present the number of records (NOR) included.

The effect of increasing complexity on the ligand-target interaction can even be better illustrated if we plot the potency vs HAC separately for the certain potency ranges, e.g., pAC_{50} above 6 (Fig. 1.b) and below 6 (Fig. 1.c). For the highest potency population above 6, we can observe a clear decrease in the average potency up to an HAC of 21 where the mean pAC_{50} amounted to 6.69. The reason for the decreasing trend of AC_{50} within this high AC_{50} subpopulation could be explained similar to the effect that was observed for the statistics that is illustrated in Fig. 1.a. In turn, the statistics for the lower potency ligands (Fig. 1.c) did not record such an effect. As a high AC_{50} value is a rare *elite* quality, a perfect ligand-target fit is especially important for the ligands in this group. In other words, the less active ligands simply did not skew the statistics.

In PubChem, the data sources are highly diversified and therefore, the *systematics* of the measurements and data notation cannot be precise enough. In contrast, the most reliable numerical potency data for drug candidates are recorded in the ChEMBL database, which contains ca. 700 000 processable numerical records if we accept that the AC_{50} format is analogous to PubChem. In Fig. (2), we compare the PubChem data with the ChEMBL population. The ChEMBL data indicates an increase in the mean potency

with an increasing HAC up to a value of 47 for HAC ($pAC_{50}=7.14$). This is in contrast to the constant or decreasing trend of the PubChem population. We can conclude that one of the basic drug design intentions, which is the maximization of potency by increasing the molecular size of drug candidates, is evidently reflected by the ChEMBL potency data. This ChEMBL trend can also be explained. Intentional design, which is supported by intentional and specific data selection of compounds to the ChEMBL database, has provided ligands that fit their targets better than the PubChem average. The number of interacting centres between a ligand and a target can increase with increasing HAC. A precise fitting of these centres can result in an increase in pAC_{50} , which could be higher and higher the larger the molecular size is. After reaching a maximum at a value of 47 for HAC (with a value 7.14) the average ChEMBL potency decreased until a

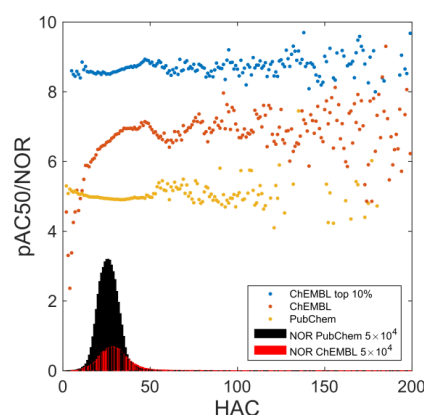


Fig. (2). Mean potency pAC_{50} binned against HAC recorded in PubChem, ChEMBL or ChEMBL subpopulation (with the highest 10% AC_{50} values). A distribution of PubChem or ChEMBL as given by the number of records (NOR) is shown below the plots.

This probably illustrates the fact that *design* becomes more and more difficult to achieve with an increasing molecular size. The statistics of the top 10% of ChEMBL resembles this trend. However, now the plot is flatter with a less sharp maximum. An interesting effect can be observed in the range below a value of 10 for HAC. Here the average potency level was higher for the less reliable but larger PubChem population than for the more reliable but smaller ChEMBL population.

In Fig. (3), we illustrate the PubChem ChEMBL and pAC_{50} vs HAC as a landscape for the analysis of various ligand series. The PubChem average (black dots in Fig. 3.) provides a discriminator for two essential classes in the regions below and above this line. In turn, a fraction of the top 10% of ChEMBL potencies delimits the most potent candidates. Accordingly, we used these statistics as the landscape to plot the potency of FDA approvals between 1939 – 2014 (FDA), a series of drugs [1] or drug fragments [17] (Fig. 3.).

The potencies of the drugs are well above the ChEMBL average, and fragments can be found within almost all of the regions, namely 74% of the potency values are below the PubChem average, 13% are between the PubChem and

ChEMBL average, 13% are above the ChEMBL average and none are above the ChEMBL top 10%.

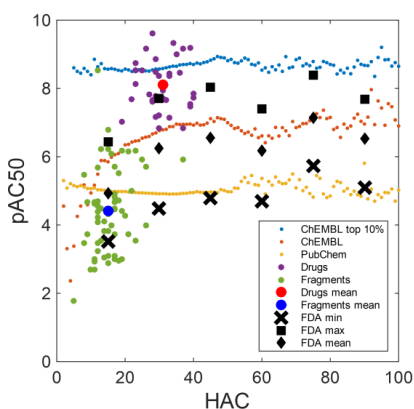


Fig. (3). The potencies of the FDA approvals (black crosses: minimal values, black squares: maximum values, black diamonds: mean values), the series of fragments that were used for hit selection and optimization [17] (green dots) and a series of selected drug potencies [1] (violet dots) (b) shown within the landscape of ChEMBL and PubChem data.

Interestingly, the minimal FDA potencies are surprisingly close to those of the PubChem averages. In particular, FDA approvals with an HAC of less than 45 were below the PubChem average, while those with a value above 45 for HAC (potency=5.29) were above the PubChem average. Consequently, FDA and PubChem statistics intersect with each other. Similar to the ChEMBL average which indicates a maximum potency value at 47 for HAC, the mean FDA potency achieved a maximum potency at 45 for HAC. The average potency of the FDA approvals was close to ChEMBL average and well below the top 10% ChEMBL. Interestingly, the maximum average AC_{50} for ChEMBL and FDA was observed in the relatively low populated HAC area, which suggests that these compounds originated from intentional design rather than from a random search in a huge number of structures.

In Fig. (4). we illustrate the plots of LE vs HAC for all the data of the PubChem, ChEMBL, FDA, drugs and fragments that were investigated. A hyperbolic decrease of LE with HAC was observed for all the data that was analysed. Accordingly, LE is not an informative measure for evaluating binding, rather it shows a high preference for a low HAC.

In practice, low HAC and low AC_{50} (high potency) values indicate attractive drug candidates. Therefore, the product of potency and HAC (Product LE: PLE) should be an estimator that is more balanced against both parameters. A simultaneous decrease of HAC and AC_{50} will decrease PLE, and vice versa their increase will increase PLE. Accordingly, both terms act cooperatively in PLE. A common mathematical representation for AC_{50} is its negative log scale, pAC_{50} , in which higher values of pAC_{50} indicate an exponentially greater potency. Accordingly, we used the analogous pPLE scale (the higher pPLE indicates better quality) to analyse the PLE trends. In Fig. (5), we plotted the dependence of the

pPLE for the PubChem and ChEMBL data. The pPLE nicely diagnoses that when pAC_{50} does not increase with an increase of HAC (PubChem compare Fig. 2.), then, the lower HAC, less complex ligands should be selected as being optimal. Moreover, when pAC_{50} increases with HAC (ChEMBL), then pPLE also follows the same trend, thus clearly indicating the interaction of pAC_{50} and HAC.

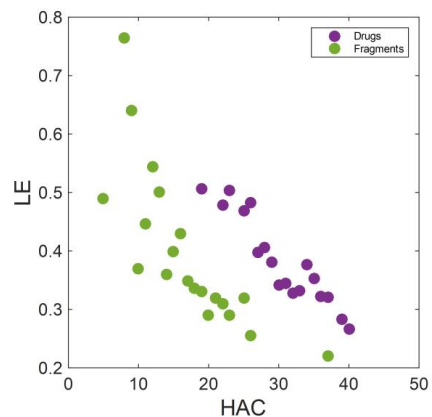
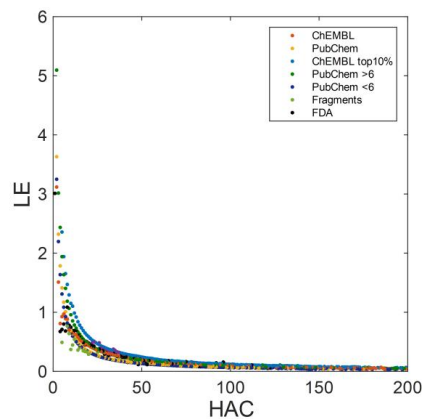


Fig. (4). Mean LE values binned against HAC for all the populations investigated (a). The drugs and fragments plot were shown in higher resolution in (b).

In Fig. (6). we used the big data pPLE landscape that was plotted in Fig. (5). to evaluate the FDA approvals, series of drugs [1] and fragment-like drug candidates [17]. Generally, the pPLE for FDA approvals follow the trend of the ChEMBL landscape, increasing with HAC up to ca. 50 HAC. Regardless of the HAC range, the pPLE value was always higher for the drugs than for fragments. There was a clear separation of the cluster of drugs from the cluster of fragments. In turn, an analysis of the plot of LE vs HAC (Fig. 4b.) showed that for a given HAC above 20, the drugs had a slightly higher value of LE than the fragments. However, for HAC values below 20, LE for the fragments always has higher values than any drug can achieve, which opposes the relation between AC_{50} values in these groups (Fig. 3.).

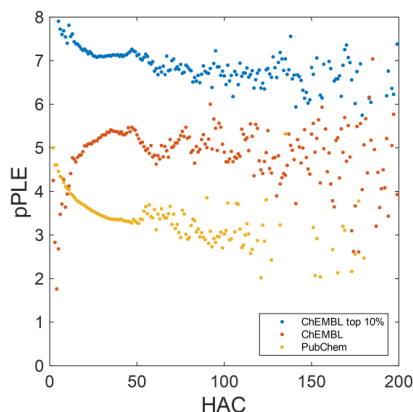


Fig. (5). ChEMBL and PubChem data in the form of the pPLE vs HAC relationship.

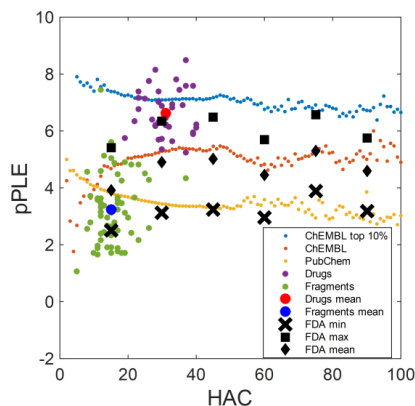


Fig. (6). The dependence of pPLE on HAC for drugs [1], FDA approvals, and selected fragments [17] presented within the ChEMBL and PubChem landscape.

CONCLUSION

In conclusion, we analysed the largest amount of potency pAC_{50} data in the PubChem and ChEMBL databases using a high-resolution binning. In particular, we show how the correlation of the pAC_{50} values vs HAC can be used to landscape drug, drug candidates, or fragments, by evaluating their potential for drug design. At the same time, the landscaped compounds are mapped here only as an example. The large PubChem and ChEMBL data reservoirs illustrate different approaches to data collection. While the first focuses on the widest possible data coverage, the second is focused more on data precision, including manual data curation. An analysis of the mean AC_{50} values indicated that the ChEMBL mean AC_{50} values depend on molecular size (HAC). On the other hand, for the PubChem data the mean AC_{50} values did not depend on HAC for a large HAC range. However, for some populations the relationship between the pAC_{50} and

molecular size was observed. In particular, the trend of the relationship depends, first of all, on the rules for the data collection for the individual populations. When pAC_{50} increases with increasing HAC, this is because with increasing ligand size, the ligand-target interaction area increases if the two moieties fit precisely enough. We observed this trend for ChEMBL data or for FDA approvals, i.e., when the data related to the ligands for which the potency was intentionally maximized. Accordingly, for the carefully designed ligands a value of approximately 47 for HAC indicates a place where the average potency reaches a maximum, which should attract more attention in drug development, especially since, on average this value is slightly higher than the Lipinski MW limit.

While using LE for probing the interaction of LE and HAC we are highlighting low HAC ligands to a level that practically masks the effect of AC_{50} . Because small ligands usually have better drug-like properties, this approach appeared to be surprisingly good. If we would like to evaluate the pAC_{50} and HAC more evenly, then, we could score ligand performance using the multiplicative function of pAC_{50} and HAC. Accordingly, we defined the product ligand efficiency (PLE) here. We tested this using various ligand series for which we compared individual ligands to the landscape of the large amount of potency data of the PubChem and ChEMBL databases. The PLE statistics indicate that the interaction of pAC_{50} and HAC is optimal at around 30-50 HAC, where the PLE has the maximum value. Finally, the current analysis illustrates the potential of molecular big data statistics [19] in drug design.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

Financial support: NCBR (Warsaw): ORGANOMET No: PBS2/A5/40/2014, TANGO1/266384/NCBR/2015.

REFERENCES

- [1] Hopkins, A.L.; Keseru, G.M.; Leeson, P.D.; Rees, D.C.; Reynolds, C.H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.*, **2014**, *13*, 105-121.
- [2] Williams, G.; Ferenczy, G.G.; Ulander, J.; Keseru, G.M. Binding thermodynamics discriminates fragments from druglike compounds: a thermodynamic description of fragment-based drug discovery. *Drug Discov. Today*, **2017**, *22*, 681-689.
- [3] Hann, M.M.; Leach, A.R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 856-864.
- [4] Zartler, E.R.; Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.*, **2005**, *9*, 366-370.
- [5] Kuntz, I.D.; Chen, K.; Sharp, K.A.; Kollman, P.A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.*, **1999**, *96*, 9997-10002.
- [6] Reynolds, C.H.; Bembek, S.D.; Tounge, B.A. The role of molecular size in ligand efficiency. *Bioorganic Med. Chem. Lett.*, **2007**, *17*, 4258-4261.
- [7] Reynolds, C.H.; Tounge, B.A.; Bembek, S.D. Ligand binding efficiency: Trends, physical basis, and implications. *J. Med. Chem.*, **2008**, *51*, 2432-2438.
- [8] Reynolds, C.H.; Reynolds, R.C. Group Additivity in Ligand Binding Affinity: An Alternative Approach to Ligand Efficiency. *J. Chem. Inf. Model.*, **2017**, *57*, 3086-3093.
- [9] Shultz, M.D. Improving the Plausibility of Success with Inefficient Metrics. *ACS Med. Chem. Lett.*, **2014**, *5*, 2-5.
- [10] Shultz, M.D. Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters. *Bioorganic Med. Chem. Lett.*, **2013**, *23*, 5980-5991.
- [11] Zhou, H.; Gilson, M. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.*, **2009**, *109*, 4092-4107.

- [12] Murray, C.W.; Erlanson, D.A.; Hopkins, A.L.; Keseru, G.M.; Leeson, P.D.; Rees, D.C.; Reynolds, C.H.; Richmond, N.J. Validity of Ligand Efficiency Metrics. *ACS Med. Chem. Lett.*, **2014**, *5*, 616-618.
- [13] Polanski, J.; Tkocz, A.; Kucia, U. Beware of ligand efficiency (LE): understanding LE data in modeling structure-activity and structure-economy relationships. *J. Cheminform.*, **2017**, *9*, 49.
- [14] Polanski, J.; Tkocz, A. Between Descriptors and Properties: Understanding the Ligand Efficiency Trends for G Protein-Coupled Receptor and Kinase Structure-Activity Data Sets. *J. Chem. Inf. Model.*, **2017**, *57*, 1321-1329.
- [15] Sheridan, P.R. Debunking the Idea that Ligand Efficiency Indices Are Superior to pIC₅₀ as QSAR Activities. *J. Chem. Inf. Model.*, **2016**, *56*, 2253-2262.
- [16] Polanski, J.; Gasteiger, J. *Computer Representation of Chemical Compounds*. In *Handbook of Computational Chemistry*, Leszczynski, J., Ed.; Springer, Dordrecht, **2016**.
- [17] Schultes, S.; de Graaf, C.; Haaksma, E.; de Esch, I.; Leurs, R.; Krämer, O. Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discov Today: Technol.*, **2010**, *7*, e157-e162.
- [18] Klebe, G. Applying thermodynamic profiling in lead finding and optimization. *Nat. Rev. Drug Discov.*, **2015**, *14*, 95-110.
- [19] Polanski, J.; Kucia, U.; Duszkiewicz, R.; Kurczyk, A.; Magdziarz, T.; Gasteiger, J. Molecular descriptor data explain market prices of a large commercial chemical compound library. *Sci. Rep.*, **2016**, *6*, 28521.

Received: March 20, 2014

Revised: April 16, 2014

Accepted: April 20, 2014

OCC-CC-ClOC-Cl	609970	123.31-9	QGBRXXKICLVML-UHFFFAOYSAN	hydroquin(HO)ZGH4 GHH62	110.11	isopyclic	10		
OCC-CC-ClOC-Cl	609970	123.31-9	QGBRXXKICLVML-UHFFFAOYSAN	hydroquin(HO)ZGH4 GHH62	110.11	isopyclic	9.85		
OCC-CC-ClOC-Cl	289900	353.307-1	WNFDKCCUWVI-UHFFFAOYSAN	2,3,5,6-tet C10H2N4O C10H2N4O2	10.15	isopyclic	0.9	-3.3	-1.1
OCC-CC-ClOC-Cl	289900	353.307-1	WNFDKCCUWVI-UHFFFAOYSAN	2,3,5,6-tet C10H2N4O C10H2N4O2	10.15	isopyclic	1	-1.2	
OCC-CC-ClOC-Cl	289900	353.307-1	WNFDKCCUWVI-UHFFFAOYSAN	2,3,5,6-tet C10H2N4O C10H2N4O2	10.15	isopyclic	3.9		-0.5
OCC-CC-ClOC-Cl	2937461	198.1249-2	ZMMEZFPSBBB-UHFFFAOYSAN	4-O-amino-4' HOCl2N8C C13H9NO	195.22	isopyclic	9.4		
OCC-CC-ClOC-Cl	18558211	104.6782755	QDNISMGOPRTAU-UHFFFAOYSAN	4-hydroxy- C15H13NO C15H13NO2	232.27	isopyclic	10.1		
OCC-CC-ClOC-Cl	2214801	84.877	HGVWDFDUIWCDDA-UHFFFAOYSAN	1-hydroxy- C10H6(OH) C10H8O5	224.24	isopyclic			

