Uniwersytet Śląski
Wydział Matematyki, Fizyki i Chemii

Jacek Bogocz

Metody eksploracji baz danych w poszukiwaniu nowych reguł projektowania leków

Praca doktorska wykonana
w Zakładzie Chemii Organicznej
Instytutu Chemii pod kierunkiem
prof. dr hab. inż. Jarosława Polańskiego

Katowice 2016

Serdecznie dziękuję Panu Profesorowi Jarosławowi Polańskiemu
za inspiracje naukowe, opiekę i wsparcie w trakcie wykonywania prac badawczych,
a także wszelką pomoc i poświęcony czas.

Dziękuję mojej Rodzinie, bez której wsparcia
niniejsza praca nigdy by nie powstała.

Spis treści

Część Eksperymentalna:

1 Cel pracy

Chemia jest jedną z fundamentalnych nauk, powiązaną w różny sposób z niemal każdą dziedziną naukową, w szczególności z farmacją. Niestety na obecnym poziomie wiedzy nie można ustalić ścisłych kryteriów jakie powinny spełniać aktywne związki, a klasyczne sposoby poszukiwania nowych leków wiążą się z dużymi kosztami oraz ryzykiem niepowodzenia. Tak więc, jednym z priorytetowych zadań chemii jest doskonalenie metod projektowania i konstruowania cząsteczek leków. Metody poszukiwania nowych substancji leczniczych w ciągu ostatnich lat przeszły ewolucję, podobną do zmian zachodzących w przemyśle. Postępująca informatyzacja pozwala na wydajniejszą i wygodniejszą pracę, a praktyczne wykorzystanie technologii informatycznych przyczynia się do rozwinięcia lub powstania wielu innych dziedzin nauki. Szczególny wpływ informatyzacji jest odczuwalny w naukach ścisłych np. w chemii w ostatnich latach obserwujemy prężny rozwój nowo powstałej dziedziny jaką jest chemoinformatyka.

Chemoinformatyka (informatyka chemiczna) została zdefiniowana w 1998 roku przez F. Browna jako nauka polegająca na wykorzystaniu zasobów informatycznych koniecznych do przekształcenia danych w informacje a informacje w wiedzę celem szybszego podejmowania lepszych decyzji na polu identyfikacji i optymalizacji leków [1]. Nowa dziedzina nauki wzbudziła duże zainteresowanie dzięki szerokiemu potencjałowi zastosowań. Innymi słowy chemoinformatyka zajmuje się obliczeniami właściwości wirtualnych molekuł oraz wykorzystuje metody *in silico* w projektowaniu leków [2].

Klasyczną chemoinformatykę ogranicza przeprowadzanie analizy stricto do problemów chemicznych, co jest jednak w dużym stopniu redukcjonistycznym obrazem rzeczywistości. Pojęcie leku stanowi bowiem kategorię rynkową. Nie można wobec tego przewidywać losu przykładowych farmaceutycznych w oderwaniu od efektów rynkowych. Z drugiej strony analiza efektów chemicznych w tym zakresie jest bardzo złożona. Przez lata sądzono, że np. modelowanie relacji cen związków chemicznych względem deskryptorów molekularnych na rynku nie jest możliwe [3]. Ostatnio opisano analizę cen rynkowych 2,2 mln związków chemicznych. Badania wykazały korelację pomiędzy ceną a deskryptorami molekularnymi opisującymi budowę cząsteczek chemicznych [4].

Celem niniejszej pracy była próba eksploracji baz danych w poszukiwaniu informacji, które pozwoliłyby wytłumaczyć rolę i wpływ rynku na fragonomikę leków w przemyśle farmaceutycznym.

2 Wstęp

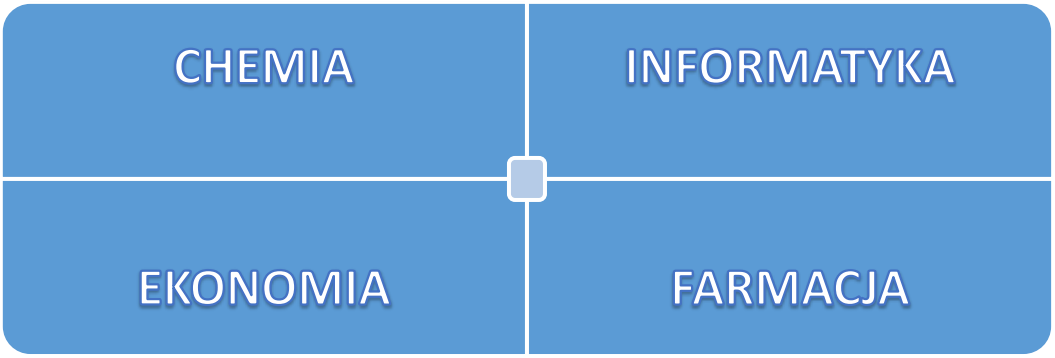
Metody komputerowe *in silico* umożliwiają analizę baz danych, odkrywanie celów białkowych, badanie interakcji pomiędzy lekami a wieloma receptorami (tzw. polifarmakologia) [5]. Nowe metody *in silico* wykorzystują dużą moc obliczeniowa komputerów połączonych najczęściej w sieć.

Rysunek 2.1 Typy oraz zbiory bazy danych w chemoinformatyce.

Odkrywanie wiedzy (ang. knowledge discovery) z wykorzystaniem komputerowych metod *in silico* pozwala dostarczyć wiele cennych informacji na temat badanej przestrzeni chemicznej. Istnieje zatem ciągła potrzeba zbierania i analizy danych dotyczących struktur oraz zależności pomiędzy ich cechami i aktywnością. Eksploracja chemicznych baz danych często jest pierwszym etapem procedury wirtualnego przeszukiwania (ang. Virtual Screening, VS) [6,7].

Idea VS jest związana z filtrowaniem bibliotek związków chemicznych w celu wskazania struktur o potencjalnej aktywności biologicznej, badania lekopodobieństwa [2] lub budowania sieci polifarmakologicznych [2,7]. Na podstawie zebranej wiedzy można symulować i przewidywać zachodzące trendy w przemyśle farmaceutycznym co z kolei może w znacznym stopniu przyspieszyć i obniżyć koszty poszukiwania nowych leków.

Niniejsza praca doktorska wpisuje się w nurt nowatorskich trendów badawczych. Podstawowym celem projektu jest wykorzystanie i zoptymalizowanie metod eksploracji i analizy danych. Przeprowadzone badania obejmują eksplorację chemo- i bioinformatycznych baz danych pod kątem określenia trendów w badaniach R&D przemysłu farmaceutycznego oraz zaproponowaniu strategii pozwalających ułatwić proces poszukiwania i konstrukcji nowych leków. Innowacyjność niniejszej pracy przejawia się poprzez połączenie kilku dziedzin naukowych takich jak: chemia, informatyka, ekonomia oraz farmacja, celem opracowania i rozwinięcia nowych metod badawczych. W badaniach posłużono się analizą scjentometryczną i farmakoekonomiczną.



Rysunek 2.2 Jednym z fundamentalnych założeń rozprawy doktorskiej było wykorzystanie i połączenie metod analitycznych z różnych dziedzin nauki.

Praca doktorska składa się z dziewięciu rozdziałów. Podstawy teoretyczne zostały pogrupowane wg tematyki i przedstawione w następujących rozdziałach 2, 3, 4 i 5. Badania własne wraz z podsumowaniem zostały zaprezentowane w 6, 7, 8, 9 oraz 10-tym rozdziale niniejszej rozprawy doktorskiej. Rozdziały 11, 12 i 13-ty obejmują spis ilustracji, tabel oraz literaturę.

Po krótkim wstępie, w rozdziale drugim omówiono podstawowe pojęcia z zakresu chemoinformatyki, farmacji oraz krótko opisano metody wczesnego projektowania nowych leków.

Rozdział trzeci w całości został poświęcony wybranym metodom chemii kombinatorycznej *in silico*. Składa się z trzech podrozdziałów, w których opisano metody dekrementacji związków chemicznych (RECAP, BRICS i SIMPLEX).

W rozdziale czwartym zapoznamy się z pojęciem scjentometrii oraz parametryzacji efektywności badawczej. Naukę oraz przemysł farmaceutyczny łączy nie tylko wysoka innowacyjność lecz także wysokie uzależnienie od nakładów finansowych.

Rozdział piąty ma na celu przybliżyć czytelnikowi zagadnienia informatyczne. Cześć pierwsza w całości została poświęcona językowi programowania Python, który jest powszechnie wykorzystywanym narzędziem w chemoinformatyce. Sposoby kodowania informacji chemicznych np. kody SMILES zostały opisane w następnym ustępie. Dodatkowo dział obejmuje krótką charakterystykę metod maszynowego uczenia oraz biblioteki "Open Source" języka Python, które zostały wykorzystywane w późniejszych obliczeniach.

Kolejna część pracy obejmuje badania przeprowadzone w Instytucie Chemii Uniwersytetu Śląskiego.

Rozdział szósty poświecono scharakteryzowaniu zakresu i metodyki prowadzonych badań.

Scjentometria jest oparta na analizie danych opisujących osiągniecia naukowe. Dane takie rejestrowane są także w bazach danych. Podobnie jest w przypadku leków. Sądzi się, że osiągnięcia naukowe zależą od czynników ekonomicznych. Dane scjentometryczne mogą także stanowić model zbliżony do opisu leków. W niniejszej pracy wykorzystane zostały w takim właśnie charakterze do testowania metodyki eksploracji baz danych oraz przetwarzaniu dostępnych informacji.

W kolejnej części, przedstawiono więc wyniki analiz scjentometrycznych. Badania skupiały się na porównaniu wydajności pracy naukowej w zestawieniu z ekonomicznymi parametrami takimi jak np. finansowanie szkolnictwa wyższego.

Ósmy rozdział skupia się na analizie "sukcesu" leków na rynku amerykańskim. W podrozdziałach zanalizowano wpływ wielu deskryptorów na sukces ekonomiczny oraz na przyczynę kryzysu w przemyśle farmaceutycznym.

Rozbudowanie zbioru bestselerów o dane statystyczne wszystkich leków dopuszczonych na rynek amerykański przez agencję FDA pomogło zrozumieć oraz zweryfikować wcześniej postawione hipotezy. Badania opisano w rozdziale dziewiątym.

Podsumowanie pracy doktorskiej przedstawiono w przedostatnim, dziesiątym rozdziale pracy.

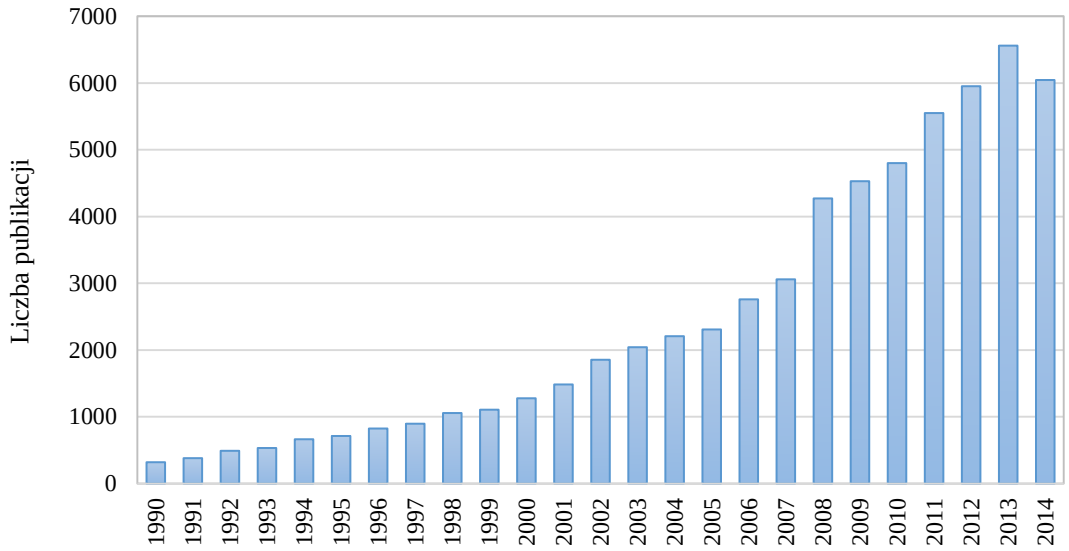
Ostatnia część rozprawy doktorskiej obejmuje spis wykresów, tabel oraz bibliografii.

3 Proces badań nad nowym lekiem

Ostatnie dekady to okres zmian technologicznych. Rewolucja w sposobie komunikacji oraz powszechna dostępność technologii komputerowej spowodowały wyraźne zmiany stylu życia. Wirtualny świat z coraz większym powodzeniem zmienia znaną nam materialną rzeczywistość. Świadczą o tym coraz to popularniejsze cyfrowe odpowiedniki (kryptowaluty, komunikacja bezprzewodowa, internetowe społeczności, wirtualna praca, itd.). Technologia a wraz z nią krocząca cyfryzacja stały się narzędziem, służącym poprawie jakości życia zapewniającym lepszą i wydajniejszą pracę. Początek trwającej do dziś trzeciej rewolucji przemysłowej zwanej naukowo-techniczną datuje się na lata 1943-1945. Jednym z osiągnięć warunkujących tą rewolucję było powstanie pierwszego komputera. Warto wspomnieć, że idea komputerów osobistych zrodziła się dużo później - na początku lat 70-tych.

W 1985 roku miała miejsce premiera systemu Windows 1.0 (obejmującego interfejs graficzny), który zapoczątkował trwającą do dziś erę systemów przyjaznych dla użytkowników prywatnych oraz biznesowych. Nieustanny rozwój i powszechność urządzeń elektronicznych spowodowały ekspotencjalny wzrost zaawansowania nauki, w tym chemii i farmacji. Dodatkowo wykorzystanie technik informatycznych pozwoliło poszerzyć i zmienić zakres dotychczasowych tradycyjnych metod poszukiwań nowych farmaceutyków. Jednak pomimo dostępu do zaawansowanych narzędzi oraz wzmożonego zainteresowania środowisk naukowo-biznesowych, konstruowanie całkowicie nowych leków dalej pozostało trudnym przedsięwzięciem [8]. Rosnąca liczba publikacji i badań poświęconych problematyce

projektowania leków znacząco wpłynęła na różnorodność metod opisujących procesy zachodzące w farmacji oraz ogólnych trendach R&D [9].



Rysunek 3.1 Wzrost liczby publikacji dotyczących projektowania leków w latach 1990 – 2014 (Scopus, data dostępu: 07.05.2015).

Zanim przejdziemy do omówienia problemów i mechanizmów funkcjonowania współczesnej farmacji, musimy zapoznać się z dziejami historycznymi. Poniżej przedstawiono, według kolejności chronologicznej, etapy rozwojowe projektowania leków [10]:

- Projektowanie chemiczne – początkowo poszukiwanie nowych leków w głównej mierze opierało się na przypadkowości odkrycia substancji biologicznie aktywnej. W miarę postępu projektowanie farmaceutyków coraz częściej opierano o wiedzę na temat dostępnych związków wykazujących działanie biologiczne.
- Era genomiki – wykorzystanie technik informatycznych i rozwój biochemii przyczyniły się do powstania nauki pozwalającej zidentyfikować target molekularny. Poznanie miejsca interakcji ligand-receptor drastycznie zmieniło dotychczasowy sposób projektowania nowych bioefektorów.
- Era proteomiczna (postgenomowa) - poznanie całego genomu wraz z genomiką strukturalną dało początek nowym dziedzinom badającym biochemię organizmów żywych na poziomie molekularnym. Należałoby podkreślić, że era proteomiki jest synonimem intensywnego rozwoju i wdrażania biomarkerów wykorzystywanych zarówno w celach nowoczesnej diagnostyki jak również badań nad lekami.

Substancja aktywna będąca kandydatem na lek musi przejść skomplikowany, długotrwały a przede wszystkim bardzo kosztowny proces testów i badań [11]. Projekt taki trwa około 15 lat i kosztuje do kilku miliardów dolarów. Zastosowanie metod chemoinformatycznych *in silico* we wczesnych etapach badań jest obiecującą alternatywą, cieszącą się dużym zainteresowaniem w biznesowym środowisku korporacji farmaceutycznych. Z tego powodu proces badawczo-rozwojowy nowego farmaceutyku przebiega pod kontrolą tzw. regulatora. Jego przykładem jest amerykańska agencja rządowa Food and Drug Administration (FDA) [12]. Jej odpowiednikiem w Unii Europejskiej jest Europejska Agencja Leków (ang. European Medicines Agency, EMA), która funkcjonuje na zasadzie jądra administracyjnego w skład którego wchodzą m.in. komitety, wewnętrzne jednostki audytowe oraz organy rządowe.

Spośród globalnych regulatorów zajmujących się rejestracją leków, największym prestiżem oraz zasięgiem cieszy się amerykańska agencja FDA. Amerykańska jednostka stanowi wyznacznik i wzorzec legislacyjny dla postępowania rejestracyjnego prowadzonego na całym świecie. Zarejestrowany lek FDA można bez problemu wdrożyć na terytorium innych państw zgodnie z procedurą wzajemnego uznania (ang. Mutual Recognition Procedure, MRP).

Agencja FDA została powołana w 1906 roku. Wchodzi w skład Departamentu Zdrowia i Usług Społecznych. Zajmuje się kontrolą żywności, suplementów, leków skierowanych zarówno dla ludzi jak i zwierząt. Ponadto kontroluje urządzenia o przeznaczeniu medycznym, emitującymi promieniowanie, jak również materiały biologiczne i preparaty krwiopochodne. FDA słynie z rygorystycznych przepisów dotyczących dopuszczania leków do obrotu. Stoi na straży procedur związanych z ich stosowaniem [9]. Pozytywna opinia wydana przez FDA dla danego produktu jest uznawana za wyznacznik jakości i gwarancji bezpieczeństwa dla zdrowia.

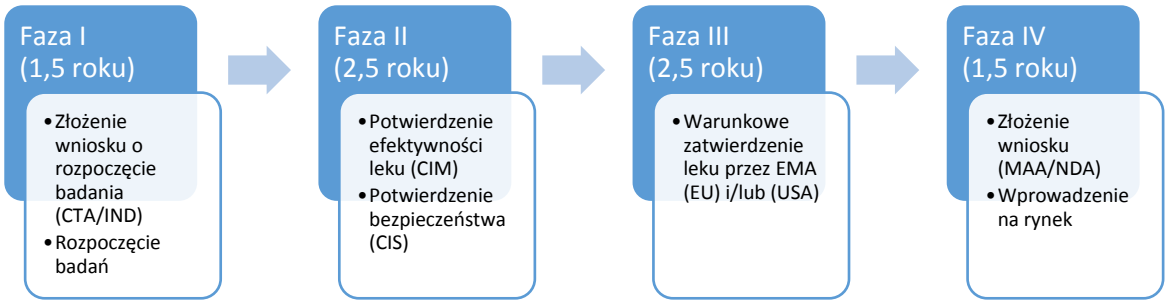
Poszukiwanie struktury wiodącej lub uprzywilejowanej jest ważnym zadaniem metod *in silico* wspomagających projektowanie nowych farmaceutyków. Pozwala na opisanie zbioru wirtualnych substancji o określonych właściwościach biologicznych, fizykochemicznych oraz farmakokinetycznych, które mają wpływ na postać aplikowanego leku [13]. Z kolei badanie zależność między budową fizykochemiczną, a działaniem biologicznym związku (QSAR) [2], pozwala oszacować potencjalny wpływ podstawników na aktywność oraz ewentualne poznanie mechanizmów oddziaływania ligand-białko. Ważnym etapem wielu analiz jest badanie tzw. reguł ADMET (adsorbcja, dystrybucja, metabolizm, eliminacja, toksyczność) [14]. Zaprojektowany lek powinien być selektywny, posiadać właściwą rozpuszczalność w wodzie,

wykazywać brak toksyczności, a przede wszystkim powinien dotrzeć do pożądanego celu działania [15,16].

Dokonanie analizy związków z użyciem technik komputerowych pozwala w dość dobrym zakresie przewidywać właściwości fizykochemiczne. Dzięki temu możliwe jest zredukowanie ryzyka niepowodzenia badań biologicznych. Wytypowane związki mogą z powodzeniem zostać zsyntezowane w laboratoriach chemicznych i następnie skierowane na dalsze badania. Po pomyślnych badaniach przedklinicznych *in vitro* związki biologicznie aktywne kieruje się do testów klinicznych *in vivo*. Na wczesnych etapach badań klinicznych istnieje najwyższe ryzyko niepowodzenia. Badania przeprowadzane są pod kątem oszacowania skuteczności oraz toksycznego działania na organizm, wystąpienia interakcji z innymi receptorami oraz pozostałych czynników, które z kolei mogą przekreślić szanse na wprowadzenie leku na rynek [17]. Według danych FDA tylko 1 na 1000 leków przechodzi do etapu badań klinicznych, a 1 na 5 z tych, które tam dotarły zostaje zarejestrowany i trafia na rynek.

Istnieje szereg ograniczeń prawno-ekonomicznych związanych z wprowadzeniem substancji chemicznej do badań klinicznych. Pierwszym z nich jest uzyskanie pozwoleń oraz uregulowanie wszelkich aspektów prawnych w tym bioetycznych wraz z szczegółowym planem przeprowadzania badań. Instytucje muszą również zainteresować i pozyskać inwestorów, którzy pomimo wysokiego ryzyka pokryją część wymaganego kapitału, który waha się w granicach kilkunastu milionów a niejednokrotnie może sięgać nawet kilka miliardów dolarów.

Badania kliniczne zasadniczo składają się z kilku następujących faz:



Rysunek 3.2 Schemat przebiegu badań klinicznych.

Poniżej pokrótce wyjaśniono proces wdrażania leku na podstawie obowiązujących przepisów oraz regulacji prawnych wydanych przez FDA [9,18].

1. Pierwszym etapem jest złożenie wniosku IND (ang. Investigational New Drug Application), w którym sponsor motywuje swoją prośbę o zgodę na skierowanie substancji do badań klinicznych. W dokumentacji przedstawia szczegółowe wyniki dotychczasowych badań przedklinicznych *in vitro* oraz *in vivo*. Dodatkowo składa pisma przedstawiające plan i sposób przeprowadzania projektu klinicznego. Schemat późniejszych faz często zależy od wyników uzyskanych z poprzednich etapów. Wnioskodawca po pozytywnym rozpatrzeniu wniosku IND może bez przeszkód skierować potencjalny farmaceutyk na szczegółowe badania przeprowadzane na pacjentach (faza I).

2. Badania fazy pierwszej – przeprowadzane na grupie kilkudziesięciu zdrowych wolontariuszy. Celem jest określenie bezpieczeństwa leku, jego skutków ubocznych, ustalenia optymalnej dawki oraz uzyskanie informacji o wchłanianiu i sposobie wydalania z organizmu.

3. Badania fazy drugiej – kontynuowane w przypadku pozytywnej oceny wcześniejszego etapu. Polegają na ocenie skuteczności leczniczej oraz tolerancji organizmu na lek. Istotną badań tej fazy jest oszacowanie efektywności leczniczej w kierunku danego schorzenia. Ponadto precyzowane są zależności pomiędzy bezpieczeństwem a skutecznością podczas krótkotrwałego oraz długotrwałego stosowania terapii. Badane są również skutki uboczne oraz przeprowadzane są badania porównawcze działania nowego preparatu wobec placebo lub innego leku metodą ślepej próby. W drugiej fazie badań bierze udział od kilkudziesięciu do około trzystu uczestników.

4. Po wydaniu pozytywnej oceny FDA na zakończenie drugiego etapu badań klinicznych, wszczynana jest debata dotycząca szczegółowych warunków i obustronnych zobowiązań przeprowadzania ostatniej a zarazem najdroższej fazy badań klinicznych.

5. Badania fazy trzeciej – w badaniach zaangażowanych jest od kilkuset do trzech tysięcy pacjentów, często z uwzględnieniem różnic rasowych. Etap ten wymaga dużych nakładów finansowych. W dalszym ciągu oceniana zostaje zależność pomiędzy dawką

a efektem farmakologicznym oraz tolerancją badanego leku i jego interakcjami z innymi medykamentami.

6. Pozytywne wyniki badań ostatniej fazy klinicznej otwierają drogę do złożenia wniosku o rejestracje nowego leku NDA (ang. New Drug Application) wraz z prośbą o dopuszczenie do obrotu rynkowego na terenie Stanów Zjednoczonych. Wniosek NDA musi zawierać wszystkie szczegółowe dane dotyczące leku, daty badań klinicznych, analizę danych klinicznych, jak również informację o sposobie produkcji. FDA podejmuje decyzję do 60 dni od daty złożenia wniosku NDA.

7. Kolejnym etapem (tzw. IV faza kliniczna) jest sprawdzenie wymogów prawnych przed wdrożeniem gotowego produktu na rynek. FDA dokonuje kontroli w zakresie:
   - Etykietowania leku – FDA analizuje i sprawdza poprawność etykiet leków zgodnie z wymogami i obowiązującym prawem.
   - Kontrola obiektów producenta – obejmuje kontrolę firm, w których lek będzie produkowany.

8. Zatwierdzenie leku – FDA zatwierdza wniosek o rejestracje nowego leku oraz wysyła do producenta list z odpowiedzią. Szacowany koszt wprowadzenia jednego leku na rynek mieści się w granicach od kilkuset milionów do kilku miliardów dolarów.

9. Dopuszczenie do obrotu – pozytywna ocena produktu leczniczego dotyczy wiedzy w chwili wydania decyzji. Stosowanie leku w wielomilionowej lub miliardowej populacji w czasie pozwala na rzeczywistą ocenę skuteczności w danej terapii. Obowiązkiem producenta jest stałe obserwowanie oraz okresowe raportowanie o bezpieczeństwie leku. Niespełnienie wymogów bezpieczeństwa może w następstwie doprowadzić do wycofania leku z rynku.

3.1 Ocena właściwości fizykochemicznych

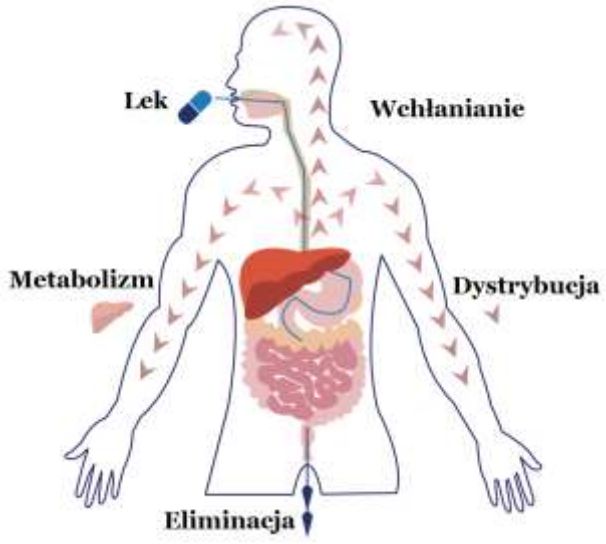
Głównym zadaniem projektowania związków chemicznych jest określenie (prognozowanie) podstawowych właściwości fizykochemicznych. W trakcie opracowywania nowej substancji chemicznej należy poszukiwać struktury o pożądanych właściwościach np. odpowiedniej

rozpuszczalności, lipofilowości, zdolność do przenikania przez błony komórkowe, tworzenia chelatów [19]. Należy pamiętać, że rodzaj atomów budujących cząsteczkę, jej struktura i wielkość decydują o dalszym losie w organizmie po podaniu w postaci gotowego leku.

Substancja aktywna po wprowadzeniu do organizmu człowieka w postaci leku, ulega szeregowi skomplikowanych przemian, które określane są skrótem ADME lub LADME jeśli mamy do czynienia z uwolnieniem aplikowanej postaci leku [19,20].

Skrót LADME pochodzi od pierwszych liter angielskich nazw procesów:

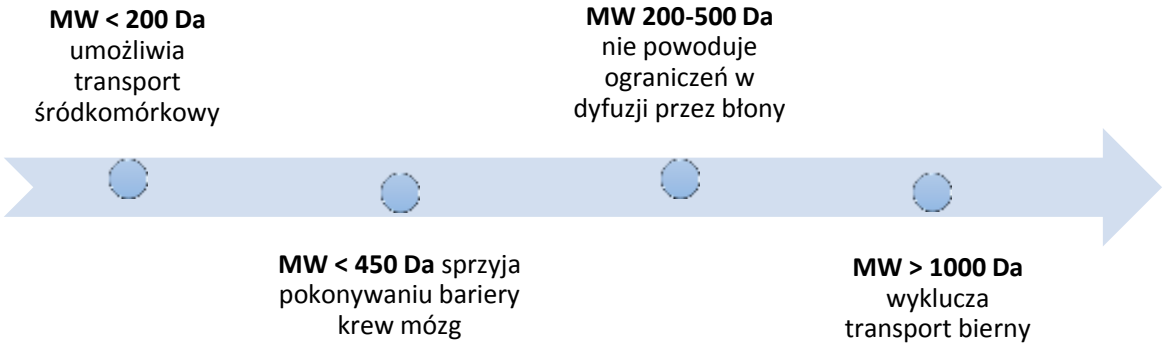
- Uwolnienie (ang. Liberation) – rozpad aplikowanej postaci leki oraz uwolnienie jego cząsteczek.
- Wchłanianie (ang. Absorption) – zachodzi na zasadzie dyfuzji biernej. Obejmuje proces, w którym cząsteczki leku przedostają się do krwi.
- Dystrybucja (ang. Distribution) – proces przenikania leku z krwi do pozostałych tkanek organizmu. Na tym etapie wiele leków zostaje związanych z białkami osocza, co zmniejsza terapeutyczne działanie leku oraz przedłuża proces eliminacji.
- Metabolizm (ang. Metabolism) – struktura cząsteczki leku ulega chemicznej przemianie, co prowadzi do utraty właściwości leczniczych.
- Eliminacja (ang. Excretion) – cząsteczki leku są usuwane z tkanek. Efektem eliminacji jest zmniejszenie stężenia leku w organizmie.



Rysunek 3.3 Procesy ADME na przykładzie układu biologicznego człowieka [21].

Do elementarnych deskryptorów mających bezpośredni wpływ na właściwości fizykochemiczne, które warunkują zachowanie leku w organizmie zaliczamy m.in.:

- **Masa molekularna** (ang. Molecular Weight, MW) to parametr, który w znacznym stopniu decyduje o farmakokinetyce leku. Zgodnie z regułą pięciu Lipińskiego większość substancji leczniczych nie przekracza masy 500 Da. MW w połączeniu z innymi parametrami dostarcza wiele cennych informacji na temat potencjalnego powinowactwa związku chemicznego do jego aktywności biologicznej. Wzrost masy cząsteczkowej determinuje objętość cząsteczki. Na tej podstawie można w dużym stopniu określić prawdopodobieństwo wchłaniania i transportu substancji przez błony biologiczne. W ten sposób umownie wyodrębniono następujące granice:



Rysunek 3.4 Graniczne wartości masy molekularnej w poszczególnych procesach transportu komórkowego.

Na podstawie masy cząsteczkowej można z dużym powodzeniem przewidywać parametry LADMET [22]. Ponadto znajomość masy molowej leku pozwala obliczyć ile cząsteczek leku obecnych jest w organizmie po podaniu określonej dawki substancji aktywnej. Znając liczbę komórek w organizmie człowieka szacowaną obecnie na około 37 trylionów można obliczyć teoretyczną liczbę cząsteczek leku przypadających na jedną komórkę [23]. Jest to wartość pozorna, która jedynie w dużym przybliżeniu informuje o fizykochemicznej stronie cząsteczek leku w organizmie. Biorąc jednak pod uwagę wartość dostępności biologicznej leku oraz jego procent wiązania z białkami osocza można określać trendy, które cechować będą różne substancje w zależności od MW, dostępności biologicznej oraz wiązania z białkami PB (ang. Protein Binding) zastosowanej dawki.

- **Lipofilowość** związku charakteryzuje powinowactwo substancji chemicznej do fazy polarnej (wodnej) i niepolarnej (lipidowej). Parametr ten, znany jako log P, zgodnie z prawem Nernsta jest obliczany eksperymentalnie na podstawie współczynnika podziału miedzy wodą i n-oktanolem.

$$P = \frac{C_o}{C_w}$$

gdzie,

C_o – stężenie molowe substancji w oktanolu [mol/L]

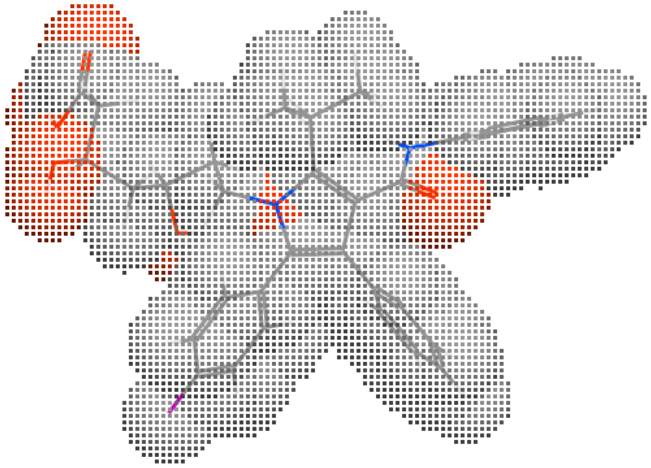
C_w – stężenie molowe substancji w wodzie [mol/L]

Lipofilowość cząsteczki, czyli jej zdolność rozpuszczania się w danym środowisku ma decydujący wpływ na fazę farmaceutyczną (formę i sposób podania leku), farmakodynamiczną (rozprzestrzenianie leku na poziomie układów, narządów i tkane) i farmakokinetyczną leku (transport cząsteczek przez błony biologiczne, barierę krew-mózg i krew-łożysko) [15,16].

- **Rozpuszczalność** ma duże znaczenie na każdym etapie LADME z uwagi na fakt, iż organizm dorosłego człowieka składa się w przybliżeniu z 60% wody. Rozpuszczalność jest podstawową cechą regulującą procesy absorpcji co z kolei określa dostępność biologiczną. Innymi słowy dobra rozpuszczalność leku warunkuje wchłanianie aplikowanej substancji. Prowadzone są badania, których celem jest identyfikacja zależności pomiędzy rozpuszczalnością a innymi właściwościami fizykochemicznymi [24,25]. Parametrami silnie wpływającymi na zmianę rozpuszczalności jest m.in. temperatura, siła jonowa, rodzaj rozpuszczalnika oraz wartość pH roztworu. Analiza rozpuszczalności, zdolności do penetracji błon biologicznych oraz lipofilowość dała podstawy do stworzenia biofarmaceutycznej klasyfikacji leków – BCS (ang. Biophramaceutical Clasification System) [26].

- **Polarna powierzchnia cząsteczki** (PSA) jest jednym z parametrów często wykorzystywanych w metodach *in silico* do określania transportu, dostępności biologicznej oraz toksyczność związku [27]. PSA jest zdefiniowana jako suma pola powierzchni poszczególnych atomów w cząsteczce. Na podstawie analiz korelacji PSA z parametrami farmakokinetycznymi stwierdzono, że [28]:

- PSA < 60 Å – wiąże się z szybkim transferem przez tkanki, co z kolei gwarantuje dostępność biologiczną leku na bardzo wysokim poziomie. Uważa się, że niska wartość PSA decyduje o sukcesie terapeutycznym, stąd tak wiele leków o prostej topologii molekuły [29].
- PSA 140-300 Å – znaczne obniżenie wchłaniania leku, słabe powinowactwo do pokonywania barier biologicznych w trakcie dystrybucji w organizmie.
- PSA > 300 Å – znikoma dostępność (zazwyczaj poniżej 10%) lub brak wchłaniania na drodze transportu biernego.



Rysunek 3.5 PSA przedstawione za pomocą grafu na przykładzie cząsteczki Lipitora [30,31].

W metodach obliczeniowych wykorzystuje się pochodny parametr zwany TPSA (ang. Topological Polar Surface Area). Wartość TPSA obliczana jest za pomocą algorytmu, opartego na podstawie analizy polarnych fragmentów 34810 leków z bazy World Drug Index [28]. Metodę wyróżnia znacznie mniejsze obciążenie sprzętu oraz szybszy czas obliczeń. Parametry PSA, TPSA korelują z masą molową oraz wartościami deskryptorów HBD i HBA. Deskryptor polarnej powierzchni cząsteczki jest parametrem, który nie zawiera żadnej informacji o kształcie i rodzaju atomów w badanej molekule.

Polarne pole powierzchni leku jest parametrem który zmienia się w przypadku nawet niewielkich zmian w strukturze leków. Pokrewnym parametrem jest tzw. fraktalne pole polarne cząsteczki (ang. Fractal Polar Surface Area, FPSA). FPSA wyraża procentowy udział polarnego pola powierzchni cząsteczki do pola powierzchni związku dostępnego

dla cząsteczek rozpuszczalnika. Dodatkowym parametrem opisywanym w przypadku pól cząsteczek leków jest NPSA (ang. Non Polar Surface Area).

Analiza powierzchni związków chemicznych w powiązaniu z innymi parametrami stanowi podstawy nowoczesnego opisu metod *in silico*, w wyniku których możliwe jest wyznaczenie ścieżek i mechanizmów dystrybucji leku w organizmie.

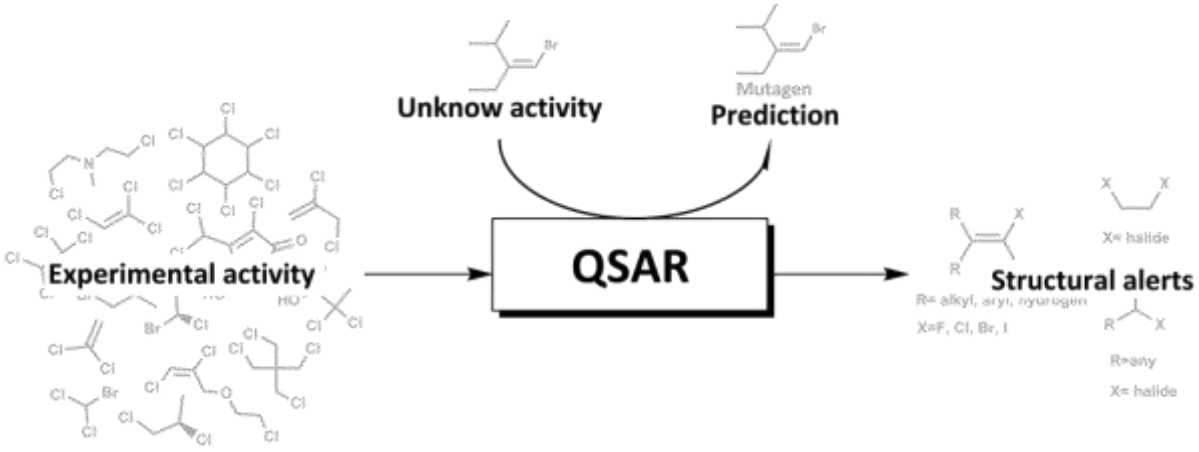
- **Liczba donorów i akceptorów protonu** (HBD, HBA) substancji biologicznie czynnej wpływa na polarność cząsteczki (PSA), a tym samym na lipofilowość, co przekłada się na sposób dystrybucji i transportu wewnątrzkomórkowego leku. Większa liczba HBA decyduje o większej hydrofilowości (możliwości przyłączenia większej liczby cząsteczek wody). Natomiast rozmieszczenia ładunków w obrębie regionów HBA i HBD pozwala określić kwasowość i zasadowość tych regionów oraz całej cząsteczki [32].

- **Orbitale molekularne** HOMO i LUMO wyjaśniają podstawy powstawania i zrywania wiązań chemicznych. Badając orbitale, a dokładniej wielkość przerwy energetycznej HOMO – LUMO można przewidywać reaktywność oraz określać stabilność danego związku [33]. Im większa różnica energetyczna pomiędzy orbitalami tym związek jest bardziej stabilny. Warto dodać, iż energia HOMO jest miarą potencjału jonizacji i nukleofilowości. Im wyższa energia tym cząsteczki łatwiej oddają elektron.

- **RNG, RTB, RGB**, kolejno oznaczają: liczbę pierścieni związku, liczbę wiązań ulegających rotacji oraz liczbę wiązań nie ulegających rotacji. Deskryptory pomagają dokonać orientacyjnego podziału na substancje o właściwościach typowych dla leków od substancji nie posiadających takich właściwości. Wykazano, iż leki posiadające małą liczbę wiązań ulegających rotacji są zbyt sztywne, aby pokonywać bariery błonowe [34].

- **Elektroujemność** jest miarą powinowactwa atomu lub grupy funkcyjnej do elektronów lub pary elektronowej wiązania. Różnica elektroujemności pomiędzy dwoma atomami w cząsteczce determinuje typ wiązania (kowalencyjne niespolaryzowane i polarne, jonowe), tym samym zmieniając właściwości fizykochemiczne związku. Elektroujemność atomów nie tylko wpływa na parametry fizykochemiczne lecz również determinuje powinowactwo do miejsca wiążącego w organizmie [35].

Zależności pomiędzy strukturą a aktywnością substancji określane są mianem SAR (ang. Structure-Activity Relationship). Ich znajomość umożliwia określenie specyficznych cech cząsteczki badanego związku, które decydują i wpływają na aktywność biologiczną. Właściwości fizykochemiczne najczęściej obliczane są za pomocą specjalnych algorytmów, które często korzystają z baz danych eksperymentalnych pomiarów uzyskanych w wyniku badań pokrewnych chemicznie grup związków. W analizach SAR wykorzystuje się często testy *in vitro*.



Rysunek 3.6 Schematyczne przedstawienie metody QSAR [36].

Rozszerzeniem metody SAR jest QSAR (ang. Quantative Structure-Activity Relationship). Dotyczy badań ilościowej zależności między strukturą związku a jego właściwościami biologicznymi. Ma na celu zidentyfikowanie odpowiednich parametrów fizykochemicznych oraz ustalenie przewidywanej aktywności biologicznej danej cząsteczki [37]. Metoda pozwala również na sformułowanie matematycznego modelu tej zależności – równania QSAR. Znajomość równania QSAR pozwala na oszacowanie mechanizmów interakcji ligandów z receptorem, ocenę aktywności niezbadanych związków oraz uporządkowanie i pogrupowanie ligandów pod względem ich aktywności oraz cech fizykochemicznych [2].

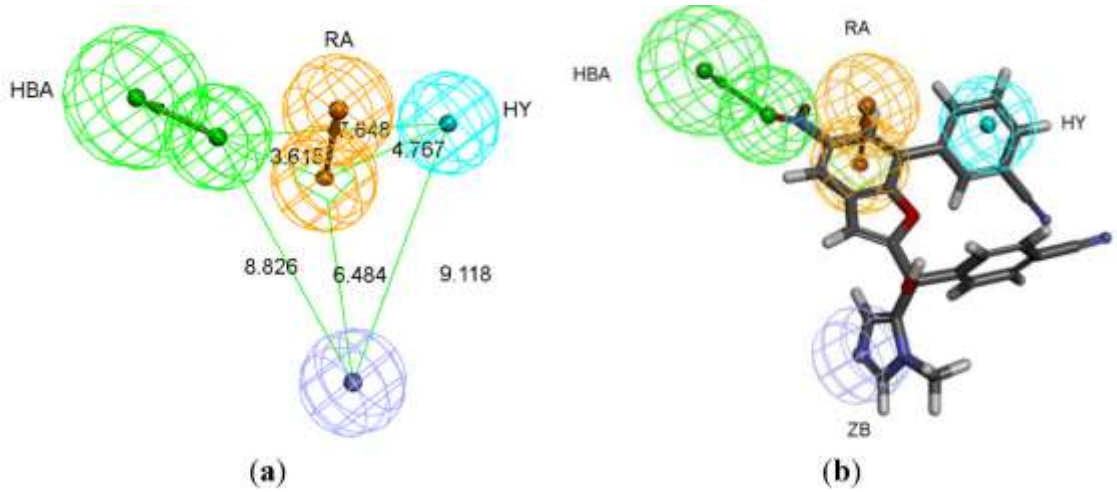
3.2 Projektowanie modelem aktywnego ligandu - farmakofor

Pierwszą metodą projektowania molekuł jest wykorzystanie analizy aktywnych ligandów. Ich porównanie umożliwia syntezę nowych potencjalnie aktywnych związków (drug candidates). Model farmakofora odnosi się w takiej metodzie do przestrzennego opisu właściwości fizykochemicznych danej cząsteczki chemicznej. Ilustruje relacje przestrzenne między wspólnymi elementami strukturalnymi lub elektronowymi dla ligandów oddziałujących z danym receptorem. Modele farmakoforowe wykorzystuje się we wczesnych etapach poszukiwań leków, badając potencjalną toksyczność, interakcje oraz efektywność wiązania z receptorem.

Istnieją dwa zasadnicze rodzaje technik konstrukcji farmakoforów [38]:

- tradycyjne (chemiczne) - układ atomów połączony konkretnymi wiązaniami chemicznymi.
- elektronowe (teoretyczne) - odpowiedni rozkład pola elektrycznego niezależnie od wytwarzającego go układu atomów i ich wiązań.

W klasycznym podejściu z jednym miejscem docelowym może oddziaływać wiele farmakoforów (struktur chemicznych). W przypadku metody elektronowej dla danego celu molekularnego istnieje tylko jeden farmakofor, ale może on być realizowany na wiele sposobów poprzez struktury chemiczne. Podejście elektronowe jest bardziej obiecujące dla poszukiwania i projektowania leków, gdyż poznanie struktury pola elektrycznego i wymogów strukturalnych farmakoforu stanowić może podstawę do poszukiwania (doświadczalnie lub na drodze obliczeń) nowych struktur chemicznych spełniających niezbędne wymogi [2,39]. Tworząc model farmakofora definiuje się maksymalną i minimalną ilość miejsc farmakoforycznych. Jest to bardzo ważne, ponieważ zbyt duża ilość miejsc farmakoforycznych daje zbyt dokładny farmakofor i w efekcie może skutkować negatywnym wynikiem przeszukiwania bazy molekuł. Z drugiej strony liczba zdefiniowanych miejsc farmakoforycznych nie może być zbyt mała. Koncepcja farmakofora jest powszechnie akceptowana i stosowana w chemii w procesach modelowania molekularnego.

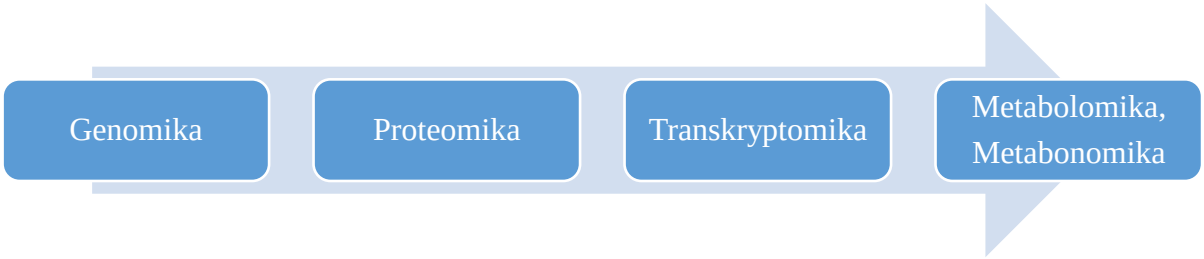


Rysunek 3.7 Relacje przestrzenne między elementami wspólnymi dla ligandów oddziałujących z tym samym receptorem [40].

Mapy farmakoforowe są trójwymiarowymi wykresami. Osie stanowią odległości pomiędzy trzema centrami i każdy z punktów reprezentuje odrębny farmakofor.
Model farmakofora wyjaśnia w jaki sposób strukturalnie różne ligandy mogą wiązać się w tym samym miejscu receptora. W przypadku, gdy struktura miejsca wiążącego jest nieznana, jako postawa do konstrukcji mapy oddziaływań może posłużyć farmakofor wygenerowany w oparciu o zbiór znanych ligandów.

3.3 Target-based design

Dynamiczny rozwój komputerowych metod wspomagania projektowania leków CADD (ang. Computer-Aided Drug Design) stopniowo przyczynia się do zmian schematu projektowania leków. Szybkość, optymalizacja i precyzja przeprowadzania procesów przewidywania właściwości molekuł chemicznych oraz ich interakcji pozwoliły na lepsze zrozumienie mechanizmów biochemicznych. Następstwem powstania metod CADD było powstanie wielu dziedzin opisujących mechanizmy biochemiczne np. systeomika, w skład której wchodzą [41]:



Rysunek 3.8 Rozwój systeomiki w czasie.

- **Genomika** to nauka zajmująca się badaniem materiału genetycznego (genomu). Pozwala zrozumieć prawa rządzące genomami. Opisuje wszelkie zależności i interakcje wewnątrz organizmu oraz analizuje ich zawartość. Poznanie genomu ludzkiemu otworzyło drogę do poszerzenia wiedzy o funkcjonowaniu naszego organizmu.
- **Proteomika** powstała jako następstwo genomiki. Jest bardziej rozbudowaną i szerszą dziedziną. Dostarcza informacje o procesach fizjologicznych. Zajmuje się badaniem organizacji, składu oraz budowy białek w komórkach i tkankach.
- **Transkryptomika** jest dziedziną zajmująca się określaniem miejsca i badaniem aktywności poszczególnych genów, w szczególności transkryptomu (cząsteczek mRNA).
- **Metabolomika** pozwala na identyfikację oraz ilościową analizę metabolitów (niskocząsteczkowych produktów naturalnych) występujących w żywej komórce, tkance czy organizmie.
- **Metabonomika** jest błędnie mylona z metabolomiką. W odróżnieniu od niej zajmuje się identyfikacją profili metabolicznych na poziomie całego organizmu.

Integracja genomiki, transkryptomiki i proteomiki wraz z informacjami uzyskanymi na poziomie metabolomu pozwala na zrozumienie mechanizmów i określenie przebiegu procesów życiowych w komórkach i organizmach.

Wnikanie w szczegóły procesów biochemicznych organizmu człowieka znacznie utrudniło proces poszukiwania nowych leków i wpłynęło na wzrost wymogów jakie musi spełnić lek aby został dopuszczony na rynek. Systeomika przyczyniła się do zidentyfikowania niezliczonej ilości nowych celów molekularnych tym samym otwierając drogę dla komputerowego projektowania leków w oparciu o cel molekularny (ang. target/structure-based design) [41].

Jak wcześniej wspomniano metody opierające się na zdolności leku do osiągnięcia docelowego miejsca działania wymagały reorganizacji klasycznych metod poszukiwania nowych leków [42]. W zależności od informacji oraz wiedzy, którą dysponuje się przystępując do początkowego etapu zmierzającego do konstrukcji modeli lub zbioru preferencyjnych molekuł wykazujących pożądane właściwości zarówno fizykochemiczne jak i przestrzenne możemy rozróżnić dwie główne metody [43,44]:

- W oparciu o znaną strukturę receptora (ang. structure-based design) – obejmuje techniki analityczne miejsca wiążącego. Wśród nich wyróżniamy konstrukcję ligandów *ab-initio*, dopasowanie ligandów do miejsca aktywnego receptora (dokowanie) oraz badanie dynamiki kompleksu ligand-receptor.
- W oparciu o znaną strukturę liganda (ang. ligand-based design) – wykorzystanie modeli QSAR, budowanie farmakoforu oraz skriningu chemoinformatycznych baz danych.

Pomimo dużych oczekiwań sukcesy obu łatwo są kwestionowane w praktyce R&D farmaceutycznego.

Tabela 3.1 Podział komputerowych metod wspomagania projektowania leków (CADD).

	Znany ligand	***Nieznany ligand***
Znana struktura receptora	Metoda structure-based design (dokowanie ligand-receptor)	Metody *de novo*
Nie znana struktura receptora	Metoda ligand-based design (szukanie podobieństwa strukturalnego, farmakofor, QSAR)	analiza parametrów ADMET

W szczególności techniki wykorzystujące projektowanie ukierunkowane na cel molekularny, uzupełnione o dodatkowe techniki na przykład uwzględniające polifarmakologię (badanie zależności i interakcji lek-lek lub z innymi receptorami) paradoksalnie przyczyniły się do znacznego utrudnienia procesu konstrukcji nowych farmaceutyków [45]. Dotychczas prowadzone prace ukierunkowane na "target-design" pozwoliły nam również zrozumieć, że na obecnym poziomie wiedzy ta metoda wciąż nie jest uniwersalnym kluczem do sukcesu [46]. Co ciekawe ostatnie doniesienia pokazują, że technika ta bardziej narażona

jest na niepowodzenie niż projektowanie w oparciu o właściwości fizykochemiczne czy określanie lekopodobieństwa potencjalnych leków [47].

Zmiana fenotypowej strategii na projektowanie oparte o strukturę całkowicie zdominowała badania w R&D firm farmaceutycznych [48]. Efektem były wzmożone inwestycje na reorganizację struktury oraz wprowadzenie wysokowydajnych badań przesiewowych. Ponadto zmiana koncepcji napotkała na ogromne komplikacje na poziomie zrozumienia mechanizmów biochemicznych całego organizmu.

3.4 Eksploracja baz danych

Zaawansowane technologie w oparciu o dużą moc obliczeniową niewątpliwie odgrywają kluczową rolę w badaniach naukowych. Z tego powodu sięganie po różnego typu rozwiązania informatyczne staje się coraz bardziej powszechne w wielu dziedzinach naukowych w tym również chemii i farmacji. Chemia jako nauka eksperymentalna dostarcza wielu danych, których przetwarzanie, usystematyzowanie i udostępnianie umożliwia ich efektywne wykorzystanie w oparciu o systemy informatyczne. Tworzenie nowych związków chemicznych, badania teoretyczne i eksperymentalne wymagają ciągłego przeszukiwania danych oraz umiejętności ich przetwarzania. Według danych szacunkowych całkowita przestrzeń chemiczna może zawierać ponad 10^{60} struktur molekularnych spełniających regułę Lipinskiego [16,49] oraz 10^{20}-10^{24} molekuł do 30 atomów [50,51]. Spośród możliwych wariantów istotnym jest wybranie tylko tych, które w pełni spełniają nasze oczekiwania. Zatem korzystanie z chemoinformatycznych baz danych stało się nie tylko nieodłączną częścią pracy chemika, ale także niezbędnym narzędziem wymagającym posiadania pewnej wiedzy.

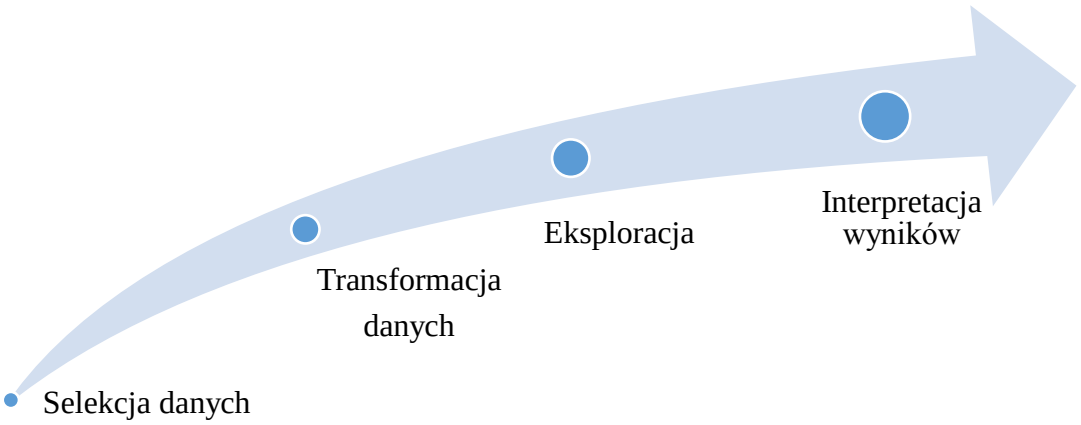
Dynamiczny rozwój systemów bazodanowych oraz konieczność ich analizy i interpretacji zapoczątkowały powstanie nowych dziedzin wspomagających procesy eksploracji danych (ang. data mining), odkrywania wiedzy w bazach danych (ang. knowledge discovery) i eksploracji baz danych (ang. database mining) [6,7,52]. Techniki te pozwalają na efektywne wyszukiwanie nieznanych zależności i schematów, które mogą być wykorzystane do podejmowania decyzji, opisu danych lub określania zachodzących trendów.

Upowszechnienie zaawansowanych technologii pozwoliło na generowanie i zbieranie informacji w formie cyfrowych danych, które z kolei stały się fundamentalnym elementem funkcjonowania każdej jednostki. Nieustanne gromadzenie informacji wielu placówek badawczych, przedsiębiorstw i urzędów doprowadziły do przekształcenia baz danych w ogromne zbiory danych gromadzone przez systemy informatyczne zwane "big data" lub zwyczajowo serwerem. Dla przykładu, zebrane zapisy dotyczące dotychczasowej działalności przedsiębiorstwa, poziomie i strukturze działalności oraz cechach klientów mogą być wykorzystane do wspomagania podejmowania decyzji o dalszym kształtowaniu kierunków strategii i rozwoju przedsiębiorstw. Coraz częściej za pomocą zebranych informacji o klientach dokonuje się precyzyjne profilowanie oferty dla poszczególnych grup klientów bądź jednostek. W niektórych sieciach sklepów podanie kodu pocztowego w kasie pozwala określić preferowane produkty w zależności od lokalizacji przydatne do zabiegów marketingowych, pośrednio pomaga w ocenie konkurencyjności innych sieci. Innym typem wykorzystania zbiorów "big data" jest system klasyfikacji często wykorzystywany przez bankierów i ubezpieczycieli, którzy za pomocą tzw. scoringu kredytowego oceniają wiarygodność klienta. Podobne zabiegi wykonują inwestorzy, którzy z wykorzystaniem narzędzi "due diligence" poddają szczegółowej, wielopłaszczyznowej analizie przedsiębiorstwa w celu określenia istniejącego i potencjalnego ryzyka związanego z inwestycją.

Innym zastosowaniem zasobów "big data" jest również poszukiwanie nowych zależności funkcyjnych oraz nieznanych dotąd trendów np. wyjaśniających przyczyny kryzysu w przemyśle farmaceutycznym [53]. Wreszcie odkrywanie wiedzy ułatwia znajdowanie anomalii w danych, czyli informacji które zaburzają bądź odbiegają od statystycznie dominującej charakterystyki całego zbioru danych [54,55].
Odkrywanie wiedzy jest złożonym procesem selekcji i transformacji danych w wiedzę, korzysta z wielu doświadczeń i metod dziedzin sztucznej inteligencji oraz uczenia maszynowego. Interpretacja wyników za pomocą metod statystycznych powinna polegać na przedstawianiu czytelnych schematów i wzorców o potencjalnym zastosowaniu.

Niestety pomimo możliwości analitycznych tak dużych wolumenów danych, metody analityczne wciąż pozostają daleko w tyle w porównaniu do możliwości ich zbierania i przechowywania. Główne problemy odkrywania wiedzy wiążą się z koniecznością przetwarzania bardzo dużych informacji oraz potrzebą szerokiej interakcyjnej współpracy

wielu wykwalifikowanych specjalistów. Metoda knowledge discovery jest procesem złożonym, którego realizacja polega na przygotowaniu danych, ich eksploracji oraz interpretacji otrzymanych wyników. Proces ten można przedstawić w kilku następujących krokach [56]:



Rysunek 3.9 Etapy odkrywania wiedzy z baz danych.

1. **Wybór bazy** – pierwszym krokiem jest wybór rzetelnego i kompletnego źródła danych. W dużym stopniu decyduje o sukcesie projektu.
2. **Formułowanie zapytań** – właściwe definiowanie zapytań pozwala na precyzyjne wyszukiwanie informacji. Skuteczne formułowanie operatorów pozwala na wyeliminowanie znacznej części szumu zbieranych danych.
3. **Selekcja i transformacja danych** – niejednokrotnie istnieje potrzeba selektywnego wyboru interesujących informacji pomimo precyzyjnych zapytań. Następnym krokiem jest ekstrakcja w postaci odpowiedniego pliku, który wykorzystamy do analizy.
4. **Analiza danych** – wiąże się z wyborem odpowiedniej metody analitycznej adekwatnej do oczekiwanych celów.
5. **Odkrywanie wiedzy** – wyciąganie wniosków i interpretacja otrzymanych wyników na podstawie wyznaczonych zależności i schematów.

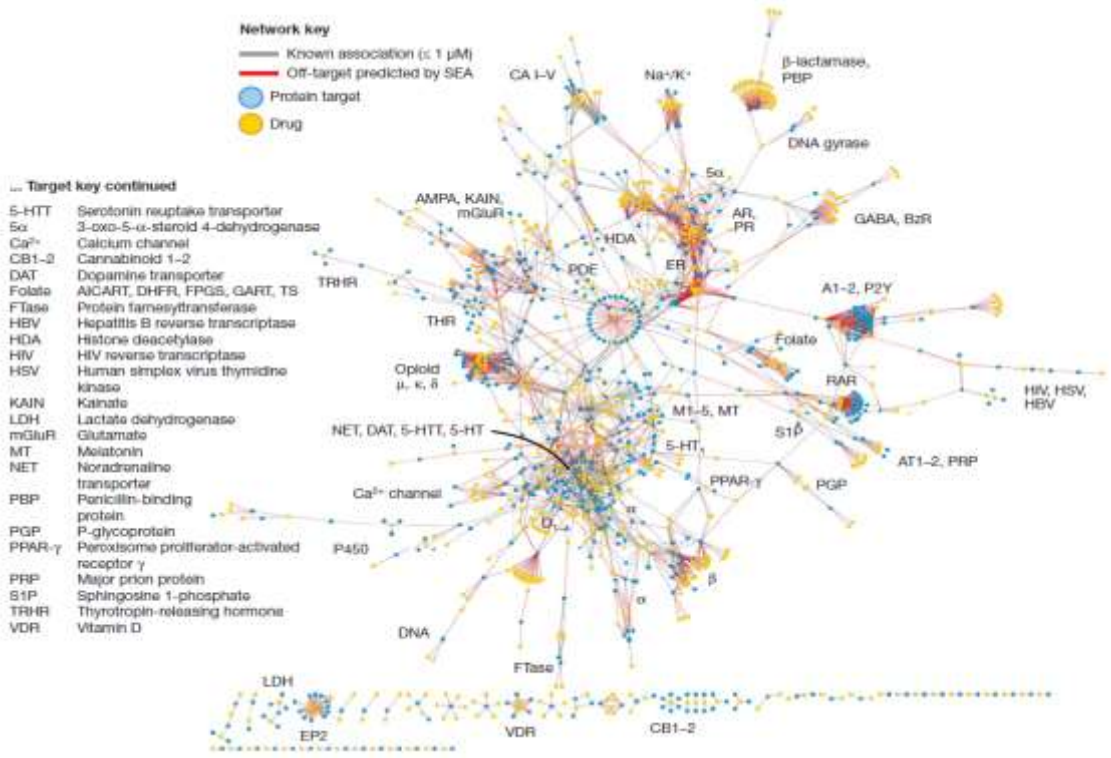
We wczesnych etapach badań *in silico* wykorzystuje się metodę wirtualnego skriningu cyfrowych bibliotek związków chemicznych. Proces VS polega na przeszukiwaniu zdeponowanych związków na drodze kombinatorycznej, w wyniku podmiany atomów, grup funkcyjnych czy rdzenia w znanych związkach aktywnych [6,7]. Często wykorzystuje

się narzędzia analizy parametrów farmakokinetycznych typu (Q)SAR oraz wcześniej omówione narzędzia analizy grup farmakoforowych.

Podobnie jak w CADD, metody wirtualnego skriningu klasyfikujemy wg stosowanego kryterium jak:

1. **Znana struktura ligandów** – dwuwymiarowe binarne fingerprinty molekularne, modele farmakoforowe 3D, fizykochemiczne deskryptory molekularne
2. **Trójwymiarowa struktura miejsca wiążącego (receptora)** – oddziaływania kompleksu ligand-receptor, dynamika molekularna.

Znaczna część prowadzonych badań *in silico* koncentruje się na właściwościach wyizolowanego związku pomijając potencjalne interakcje z wtórnymi celami molekularnymi. Podczas testów biomedycznych często okazuje się, że wiele związków (mających wykazywać pożądane działanie farmakologiczne) jest nieskutecznych lub toksycznych [57]. Trudnym zadaniem jest przewidywanie selektywność działania, które powinno cechować nowoczesne farmaceutyki. Na rynku jest wiele leków, które pomimo oddziaływania z wieloma receptorami wykazują dobre właściwości terapeutyczne jak również wysoce selektywne z niską skutecznością kliniczną.
Informacje o przeprowadzonych badaniach eksperymentalnych *in vitro* lub *in vivo* często są umieszczane w tematycznych chemo- i bioinformatycznych bazach danych. Dane te mogą służyć również do tworzenia i analizowania sieci interakcji lek-receptor tj. sieci polifarmakologicznych. Z punktu widzenia teorii polifarmakologicznej sieć taka pozwala na znacznie zwiększenie kompleksowości chemoinformatycznych danych. Natomiast duża kompleksowość jest charakterystycznym wyznacznikiem zbiorów "big data".



Rysunek 3.10 Polifarmakologiczna sieć interakcji ligand – receptor [58].

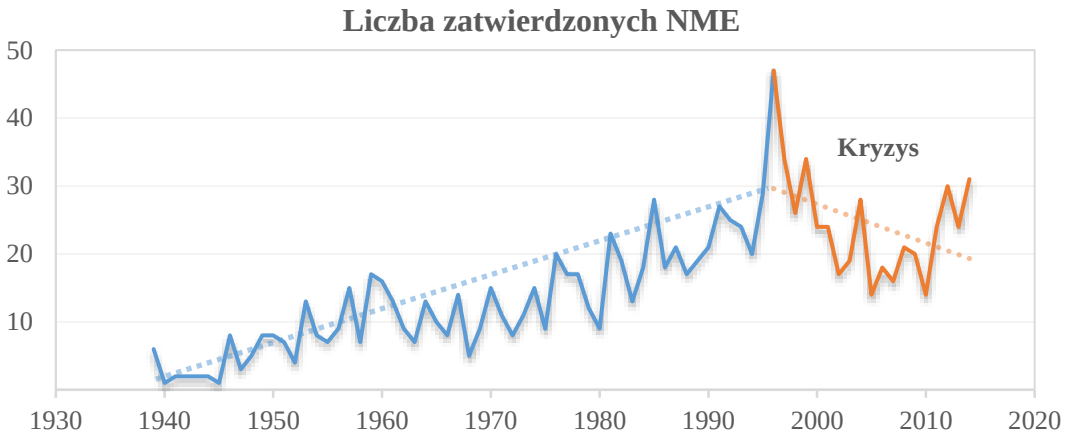
Konstruowanie i analizowanie sieci połączeń i interakcji lek-receptor może mieć istotny wkład w zrozumienie podstaw skuteczności leków, działań niepożądanych, odporności na leki a także odkrywać nowe zastosowania terapeutyczne znanych już leków [58,59]. Ponadto, strategia polifarmakologii ułatwia znajdowanie związków o potencjalnie większej aktywności terapeutycznej [45].

3.5 Rozwój przemysłu farmaceutycznego – koncepcja fast-followers i leki me-too

Przemysł farmaceutyczny należy do jednego z najbardziej dochodowych na świecie. Obecnie boryka się z licznymi problemami i pomimo wykorzystania najnowszych technologii nie zaobserwowano wyraźnego wzrostu, który zaowocowałby wdrożeniem nowych, lepszych i bardziej innowacyjnych farmaceutyków [53]. Kompleksowy charakter problemu nie pozwala jednoznacznie określić przyczyn takiego stanu rzeczy [42]. Uważa się, że rozwój genomiki, proteomiki czy metabolomiki wyraźnie wpłynął na wzrost trudności projektowania nowych bioefektorów [60,61]. Ponadto zmiana strategii firm farmaceutycznych, pragnących

zrewolucjonizować rynek odkryciem przełomowych farmaceutyków w myśl zasady: większe ryzyko większy zysk również przyczynia się do spadku produktywności.

Przemysł farmaceutyczny stanowi integralną część światowej gospodarki ekonomiczno-politycznej. Początkiem XX wieku powstała amerykańska rządowa agencja FDA [62]. Celem instytucji jest kontrola żywności, suplementów, leków, materiałów biologiczny oraz urządzeń medycznych na rynku. Od 1827 roku aż do roku 2013 zatwierdzono 1453 nowych leczniczych substancji chemicznych NME (ang. New Molecular Entity) [63]. Na wykresie strzałki czasu (rysunek 3.11) przedstawiono liczbę zatwierdzonych molekuł NME dostępnych w bazie FDA "Orange Book" (dane od 1939 roku) [64,65]. Dzięki temu możliwe było wyodrębnienie okresów o wzmożonej jak i obniżonej produktywności.



Rysunek 3.11 Lista wszystkich zatwierdzonych molekuł NME na podstawie "Orange Book" (FDA, data dostępu: 4.05.2016).

Według M.S. Kincha istnieje zasadniczo kilka czynników (przytoczonych poniżej), które wpłynęły na obserwowane historycznie trendy w biofarmacji [66].

Początki rozwoju nowych farmaceutyków można datować na lata 30-te XX wieku. W owym czasie ludzkość zmagała się z plagą gruźlicy, grypy oraz innych chorób zakaźnych. Przełom nastąpił w 1935 roku za sprawą Gerharda Domagka, który odkrył iż leki z grupy sulfonamidów wykazują działanie antybiotyczne [67,68]. Podczas leczenia myszy zakażonych bakteriami zauważył on, że czerwony barwnik używany podczas eksperymentu wykazywał biologiczną aktywność wobec drobnoustrojów. Domagk stał się ojcem pierwszych leków antybiotycznych, które z powodzeniem uporały się z morderczymi chorobami.

Następstwem odkrycia antybiotykowego działania amidów kwasu sulfanilowego były wzmożone prace nad syntezami różnych konfiguracji związków z tej grupy. Napływ dużej ilości nowych leków doprowadził w 1937 roku do katastrofy w wyniku której zginęło ponad 100 osób. Przyczyną tego zdarzenia było spożycie toksycznego roztworu sulfanilamidu (ang. Elixir Sulfanilamide), w którym jako rozpuszczalnik zastosowano roztwór glikolodietylenowy (DEG). W celu zapobiegnięcia podobnym zdarzeniom rząd zatwierdził regulującą ustawę o żywności i lekach kosmetycznych (ang. Federa Food, Drug and Cosmetic Act, 1938) [69].

W roku 1941 przeprowadzono badania, które rzuciły nowe światło na proces konstrukcji nowych leków. Za sprawą modyfikacji jednego z atomów sulfapirydyny (leku objętego ochroną patentową) otrzymano nowy, nie objęty patentem związek, który dodatkowo wykazywał skuteczniejsze właściwości terapeutyczne. Odkrycie spowodowało zmianę strategii w metodach poszukiwania nowych farmaceutyków w myśl zasady "fast-follower" lub "me-too drug". Prace nad znanymi już lekami wykazują tą zaletę, iż najczęściej prowadzą do ulepszenia formuły, skuteczności oraz zniwelowania skutków ubocznych. Leki "me-too" szybciej adaptują się na rynku z powodu krótszej drogi poszukiwania struktury, przebadania jej aktywności oraz przetarcia dróg w fazach klinicznych. Dodatkowa redukcja kosztów i adaptacja na rynku w tym poznanie pojemności rynkowej otwiera drogę do skutecznej konkurencyjności. Negatywny wpływ koncepcji "fast-followers" i leków "me-too" obserwuje się zarówno w sektorze innowacyjności jak również niekorzystnego wpływu na kondycję finansową pionierów innowacyjności. Często zdarza się, że firmy kierujące się taktyką "fast-followers" przejmują większą cześć rynku, a w skrajnych przypadkach wypierają całkowicie pierwotny farmaceutyk. Stosunkowo niskie koszty wdrożenia leku "me-too" oraz wysokie profity z pewnością przyczyniły się do spadku produktywności nowych bioefektorów [70-74].

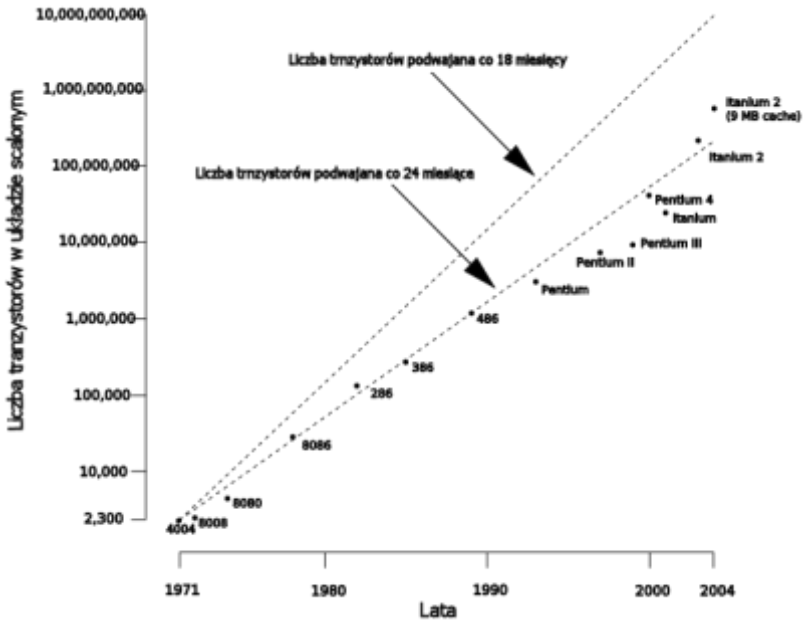
Obecną sytuacje pogarsza polityka gospodarcza niektórych państw np. Chin, które w dużej mierze zbudowały swą potęgą na kopiowaniu najlepszych rozwiązań często wbrew ochronie patentowej. Dzięki liberalnej polityce państwo środka w chwili obecnej jest jednym z głównych dostawców surowców farmaceutycznych i leków generycznych na świecie.

Na sytuację rynkową farmaceutyków bez wątpienia wpływ mają patenty. Po upływie okresu ochrony patentowej na dany lek, który trwa 20 lat obserwuje się wzrost liczby preparatów

leczniczych o takim samym składzie lecz innym pochodzeniu. Leki takie nazywamy generycznymi lub odtwórczymi. Jest to określenie preparatu farmaceutycznego będącego zamiennikiem oryginalnego preparatu z taką samą ilością substancji aktywnej. Mogą zawierać jednak inne substancje pomocnicze, pod warunkiem, że nie zmieniają one właściwości leku i nie wpływają na skuteczność. Wprowadzenie zamienników na rynek jest zwolnione z obowiązku przeprowadzenia kosztownych badań klinicznych. Leki generyczne są tańsze, ponieważ ich producenci nie są monopolistami [70,74]. Co najważniejsze nie są w żaden sposób gorsze od leków oryginalnych. Identyczność substancji czynnej w obydwu preparatach nie jest gwarantem takiego samego składu chemicznego pozostałych substancji pomocniczych oraz indywidualnego wpływu na pacjenta. Środki wypełniające, stabilizujące i zanieczyszczenia, które wchodzą w skład leku generycznego są pozyskiwane odmiennymi metodami [75,76]. Według badań różnica w przyswajalności leków generycznych i oryginalnych wynosi ok. 3,5%, co jest porównywalne z różnicami występującymi dla różnych partii tego samego leku oryginalnego [75]. Producenci leków generycznych, opierają się na wcześniej przeprowadzonych badaniach klinicznych potwierdzających skuteczność i bezpieczeństwo stosowania odpowiedniej dawki substancji leczniczej.

3.6 Prawo Erooma

W latach 1980-tych zauważono, że proces poszukiwania nowych leków staje się coraz bardziej kosztowny i powolniejszy (pomimo ciągłego postępu technologicznego). Paradoksalnie w 1965 roku Gordon Moore sformułował prawo, wynikające z obserwacji, że ekonomicznie optymalna liczba tranzystorów w układzie scalonym rośnie liniowo w kolejnych latach [77]. Obecnie przyjmuje się, że liczba tranzystorów w mikroprocesorach od wielu lat podwaja się co ok. 24 miesiące. Prawo Moorea jest stosowane do określenia praktycznie dowolnego postępu technologicznego.



Rysunek 3.12 Wzrost liczby tranzystorów w czasie. Liniami przerywanymi projekcja okresu podwajania dla 18 i 24 miesięcy [78].

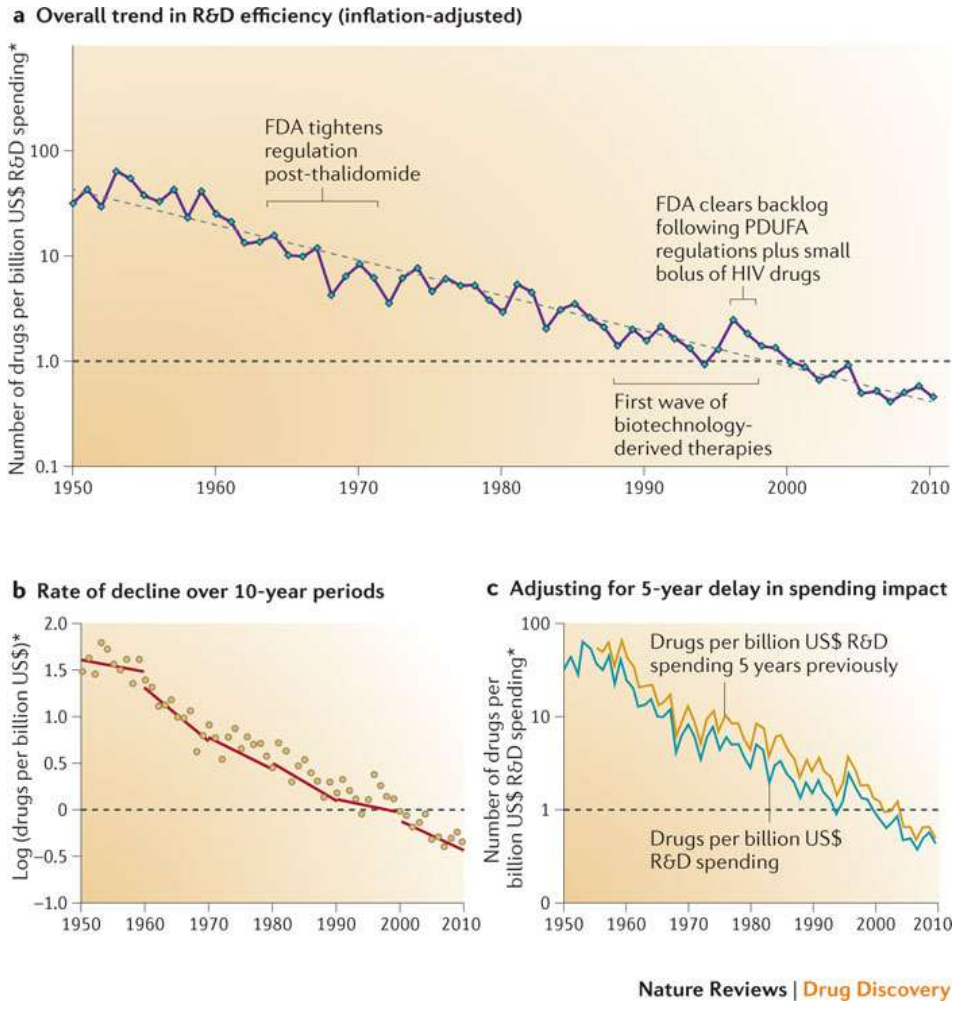
Jak wcześniej zauważono postęp technologiczny spowodował, że proces poszukiwania nowych leków staje się coraz wolniejszy i bardziej kosztowny. Koszty poszukiwania, projektowania, analizy i rozwoju nowych farmaceutyków w przybliżeniu podwajają się co dziewięć lat. Pragnąc wyjaśnić przyczynę pojawienia się kryzysu w przemyśle farmaceutycznym, sformułowano prawo Erooma (Moora czytane wspak).

Prawo Erooma wynika z kilku prostych reguł ekonomicznych [79]:

1. **"Lepszy niż the Beatles"** – debiut oraz światowy sukces zespołu the Beatles, spowodował, że konkurencyjność w branży muzycznej stawała na coraz to wyższym poziomie. Nawet samym autorom ciężko przychodziło komponowanie nowych utworów, które osiągałyby jeszcze większą popularność. Analogiczną sytuację zaobserwowano na rynku farmaceutyków, na przykład bestseller leków Lipitor. Lek odniósł tak spektakularny sukces, iż prawdopodobieństwo odniesienia większego sukcesu na rynku farmaceutyków jest mało prawdopodobne.
2. **Zaostrzenie regulacji prawnych i bezpieczeństwa** – stopniowe podwyższanie norm bezpieczeństwa oraz wprowadzanie nowych regulacji prawnych sprawiają, że z rynku wycofywane zostają starsze leki np. Vioxx. Z drugiej strony opracowanie

bezpieczniejszego leku wiąże się z wyższymi kosztami i dłuższym czasem badania substancji aktywnych pod kątem potencjalnej toksyczności.

3. **Tendencja "wydajmy pieniądze"** – rozrost sektora R&D niejednokrotnie prowadzi do przekroczenia zakładanego budżetu. Wyższe koszty operacyjne przenoszone są na produkt końcowy za który płaci klient.
4. **Przecenienie możliwości** – zmiana sposobu projektowania leków z tradycyjnego podejścia (fenotypowe projektowanie leków) na zapoczątkowane w latach 1990-tych docelowe projektowanie (target-based) spowodowało konieczność reorganizacji prowadzonych badań. Pomimo wykorzystania nowoczesnych metod *in silico* oraz przeprowadzania badań HTS metody te okazują się mniej skuteczne niż oczekiwano. Ograniczone zrozumienie i poznanie biochemicznych mechanizmów z pewnością negatywnie wpływa na skuteczność metod w projektowaniu leków.



Rysunek 3.13 Prawo Erooma w farmacji [79].

Przemysł farmaceutyczny ciągle zmaga się ze zrozumieniem procesów zachodzących na poziomie molekularnym. Skuteczne projektowanie spersonalizowanych farmaceutyków (leki projektowane w oparciu o genotyp) wciąż pozostaje sprawą przyszłości. W ostatnich latach coraz częściej bada się grupę nieselektywnych leków zwanych "brudnymi lekami", które pomimo że oddziaływają z wieloma celami molekularnymi okazują się przydatne w leczeniu niektórych chorób [80,81].

3.7 Idea badań translacyjnych

Pogłębiający się kryzys w przemyśle farmaceutycznym przyczynił się do ukształtowania nowej dyscypliny naukowej określanej mianem medycyny translacyjnej. Badania translacyjne zakładają nową jakość postrzegania roli naukowca. Zmieniają dotychczasowe relacje pomiędzy laboratorium a kliniką w myśl dwukierunkowej zasady "from bench to bedside", co skutkuje lepszą integracją wyników badań podstawowych z dziedzinami medycyny tym samym redukując ryzyko nieskutecznego poszukiwania substancji leczniczych. Głównym celem badań translacyjnych jest szybsze i skutecznie wdrażanie leków spersonizowanych poprzez budowanie struktur, które umożliwią współpracę pomiędzy jednostkami badawczymi i biomedycznymi. Stworzenie centrów medycyny translacyjnej w oparciu o model interdyscyplinarny, w których istnieje koegzystencja i współpraca pomiędzy laboratoriami badawczymi a klinikami "from bench to bedside" oraz "from bedside to bench" staje się kluczem do terapii spersonalizowanej i jednocześnie daje wyraz praktycznego wykorzystania osiągnięć naukowych w zastosowaniu medycznym.

Ostatnie 20 lat związane z rozwojem biologii molekularnej wniosło istotny wkład w poznanie rozmaitych zjawisk i procesów wyjaśniających mechanizmy wielu chorób. Niestety tylko niewiele z tych odkryć dało się przełożyć na odkrycie skutecznych leków. Wykorzystanie narzędzi biologii molekularnej, oceny biomarkerów genetycznych i molekularnych pozwala identyfikować pacjentów, u których ryzyko wystąpienia danej choroby jest największe i tym samym kształtować jej prognozę. Wcześniejsze podjęcie leczenia wiąże się ze zwiększeniem prawdopodobieństwa przeżycia i zahamowaniem wznowy procesu nowotworowego. Obecnie obserwuje się wzmożone inwestycje w rozwój ośrodków czy centrów medycyny translacyjnej [82,83].

4 Chemia kombinatoryczna *in silico*

Kombinatoryczne sposoby generowania nowych związków chemicznych za pomocą metod chemoinformatycznych okazują się być stosunkowo skuteczną metodą projektowania leków. Konstruowanie bibliotek fragmentów molekularnych ze zbioru substancji leczniczych jest jednym ze sposobów odkrywania informacji o uprzywilejowanych motywach mających największy wkład w aktywność biologiczną. Z kolei zastosowanie metod chemii kombinatorycznej [84] z wykorzystaniem wirtualnych bibliotek jako zestawu bloków budulcowych pozwala wyznaczać zasady ich łączenia ze sobą. Umożliwia również określenie, czy połączenia pomiędzy kolejnymi blokami są możliwe. Dzięki temu proces tworzenia nowych leków można przyspieszyć oraz znacznie obniżyć koszty operacyjne.

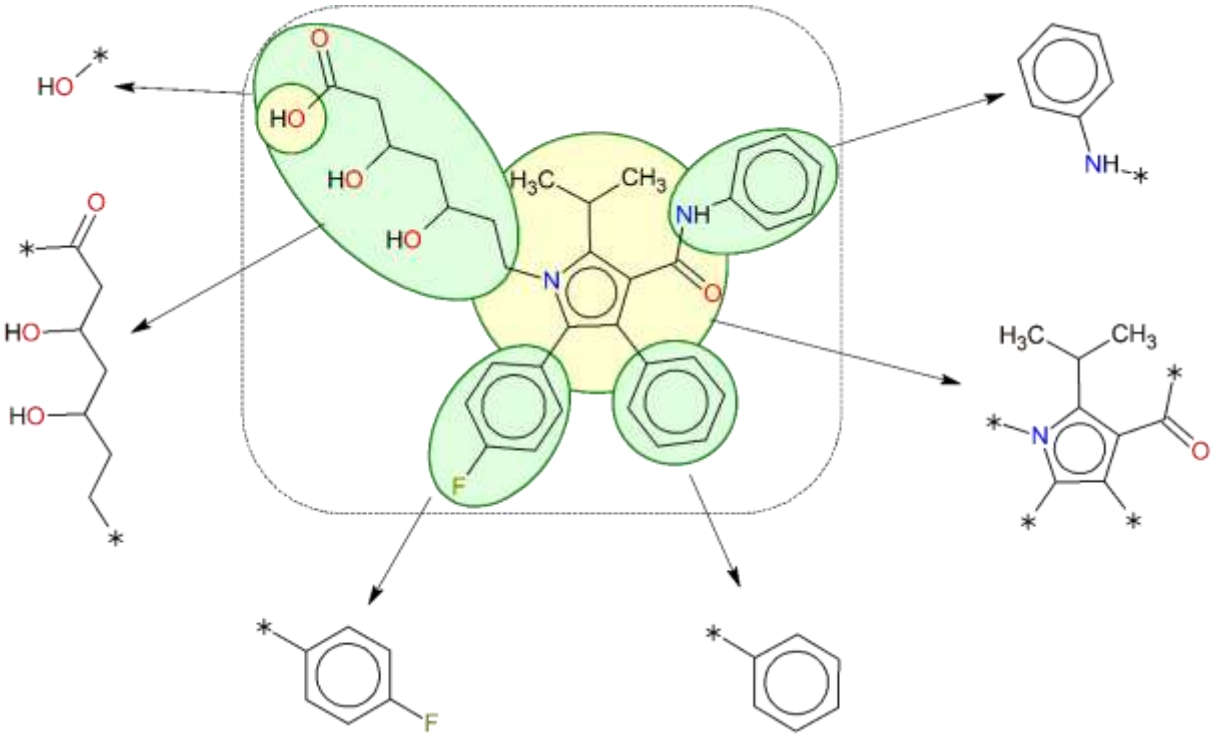
Najliczniejszą klasą metod kombinatorycznych są metody opierające się na dzieleniu cząsteczki na fragmenty. W zależność od zastosowanej metody uzyskuje się różne fragmenty. Retrosyntetyczna przestrzeń kombinatoryczna opiera się więc na fragmentacji cząsteczki poprzez cięcie jej przez ściśle określone rodzaje wiązań [84]. W wyniku tego procesu otrzymuje się określone grupy atomów, które można otrzymać za pomocą rzeczywistych i znanych reakcji. Są to tak zwane markery, które wykorzystywane są następnie do tworzenia nowych substancji aktywnych poprzez łączenie ich w nowy sposób.

W tworzeniu przestrzeni fragmentów *in silico* ważne jest, aby wybrać odpowiedni zestaw zasad dekompozycji w zależności od przyjętych oczekiwań. Otrzymanie i wybór odpowiednich fragmentów pozwala ułatwić proces projektowania i syntezy leków. W pracy doktorskiej wykorzystano trzy metody opierające się na różnych algorytmach dekompozycji struktury molekularnej. Są to RECAP [85], BRICS [86] i SIMPLEX [87,88].

4.1 RECAP

Fragmentacja związków oparta na zdefiniowanych regułach rozłamu wiązań (ang. Retrosynthetic Combinatorial Analysis Procedure, RECAP) została zaproponowana przez Lewella w 1998 roku [85]. Analiza RECAP polega na pomiarze częstości występowania danego fragmentu cząsteczki. Otrzymane grupy atomów są następnie klasyfikowane w klastery, co ułatwia identyfikację wzorców fragmentów. W metodzie RECAP cząsteczki

są cięte na fragmenty, gdy zawierają przynajmniej jeden z 11 typów wiązań chemicznych. Jeżeli fragmenty terminalne zawierają proste grupy funkcyjne takie jak wodór, metyl, etyl, propyl oraz butyl, pozostają one nie naruszone. Dodatkowo struktury pierścieniowe pozostają niezmienione, ponieważ są charakterystyczną cechą topologiczną danego szkieletu cząsteczki.

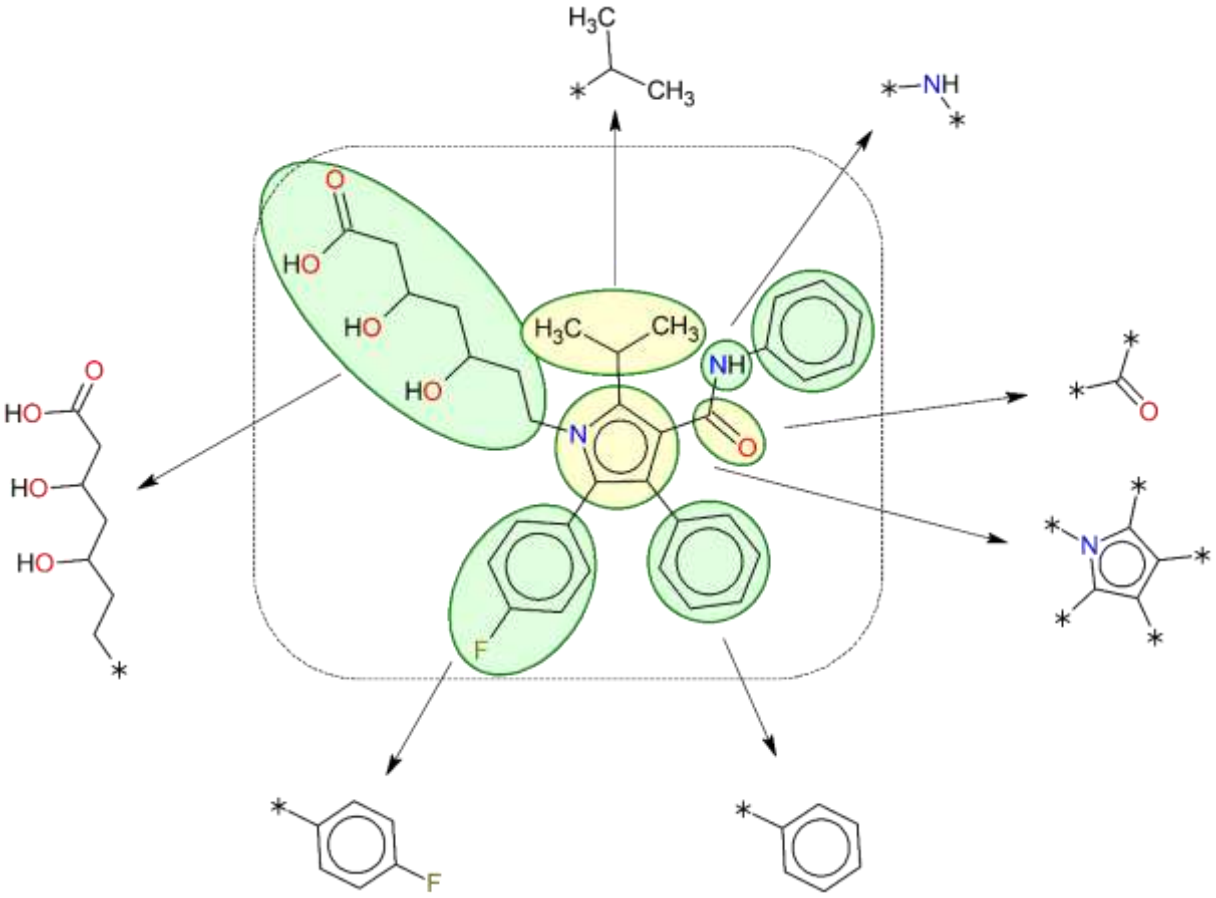


Rysunek 4.1 Fragmentacja Lipitora (najlepiej sprzedającego się leku 2003-2013) metodą RECAP.

Metoda RECAP znajduje zastosowanie w projektowaniu i syntezie substancji o selektywnym działaniu (na konkretny cel biologiczny). Umożliwia identyfikację biologicznie uprzywilejowanych motywów i fragmentów cząsteczek, które można wykorzystać jako składniki w bibliotekach kombinatorycznych. Fragmentacja pojedynczej cząsteczki obywa się w jednym etapie. Oznacza to, że wszystkie wiązania są cięte równocześnie. Dzięki temu otrzymane grupy atomów są możliwie najmniejsze oraz nie pojawiają się struktury przejściowe.

4.2 BRICS

Metoda BRICS (ang. Breaking of Retrosynthetically Intereting Chemical Substructure) została zainspirowana metodą RECAP, zatem jej algorytm jest analogiczny [86]. Cząsteczki zostają dzielone na fragmenty wzdłuż 16 typów wiązań chemicznych. BRICS również stosuje wszystkie możliwe retrosyntetyczne cięcia jednocześnie, co zapobiega generowaniu niepotrzebnych fragmentów takich jak pojedynczy atom wodoru i chlorowców, grupę hydroksylową, nitrową oraz karboksylową, metylową, etylową i grupę izopropylową.



Rysunek 4.2 Fragmentacja Lipitora (najlepiej sprzedającego się leku 2003-2013) metodą BRICS.

Ogólnie BRICS jest w stanie wygenerować około 10% więcej cząsteczek niż RECAP. W wyniku fragmentacji cząsteczek zostaje zachowana informacja o strukturze 3D. Dzięki temu możliwe jest uzyskanie większej liczby kombinacji łączonych fragmentów [84].

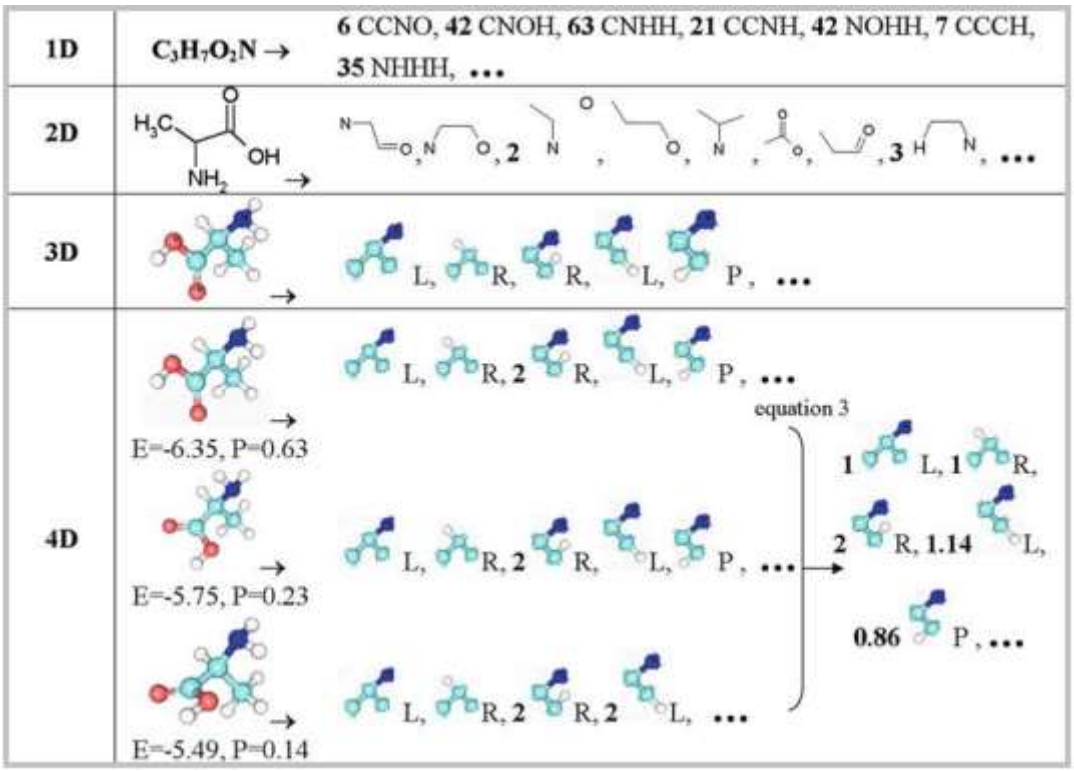
4.3 SIMPLEKS

Metoda generowania fragmentów molekularnych o liczbie atomowej równej cztery została opracowana i opisana przez zespół prof. Kuz'mina na łamach czasopisma Journal of Molecular Modeling [88]. Ze względu na topologię autorzy opisują cztery modele generowania simpleksów (1D-4D).

- Simpleks 1D - obliczany na podstawie wzoru sumarycznego związku. Nie odzwierciedla jednoznacznie konstytucji i stereochemii związku.
- Simpleks 2D - opisuje topologię fragmentu, dostarcza informacji o połączeniach pomiędzy atomami simpleksu oraz rodzaj atomu. Simpleksowi 2D można przypisać następujące deskryptory molekularne:
    - rodzaj atomu
    - ładunek cząstkowy atomu
    - lipofilowość atomu
    - zdolność do rozpraszania światła przez atom
    - donorowość lub akceptorowość atomu wodoru

- Simpleks 3D - uwzględnia zarówno topologię, stereochemię, symetrię jak i konfigurację odpowiadających mu cząsteczek.
- Simpleks 4D - przedstawia prawdopodobieństwo utworzenia simpleksów z modelu 3D.

W procesie generowania simpleksów uzyskujemy liczbę fragmentów uzależnioną od przyjętego schematu definiowania simpleksów (rysunek 4.3).



Rysunek 4.3 Schematy generowania różnych simpleksów [88].

W procesie dekompozycji rozerwaniu ulegają wszystkie rodzaje wiązań z uwzględnieniem wszystkich możliwych kombinacji (bez powtórzeń). W ten sposób otrzymuje się unikatowe czteroatomowe fragmenty zwane simpleksami. Zaletą tej metody jest generowanie niewielkich czteroatomowych fragmentów, które w małym stopniu opisują wyjściową molekułę [87,88]. Ze względu na niewielką budowę fragmenty lepiej dopasowują się do receptora, tworząc pewnego rodzaju farmakofor. Wytypowanie fragmentów czteroatomowych przez zespół prof. Kuz'mina zostało oparte m.in. o możliwość dostarczenia informacji o konfiguracji przestrzennej (np. w przypadku chiralnego atomu węgla).

W pracy badawczej zmodyfikowano metodę prof. Kuz'mina w następujący sposób:

- Generowano fragmenty o długości 2 - 8 atomów – pozwalają uzyskać większą różnorodność. Większe simpleksy zachowują informacje charakterystyczne dla danych grup związków.
- Zachowano topologię wiązań aromatycznych – wraz ze wzrostem długości fragmentu wykładniczo wzrasta liczba simpleksów. Struktury pierścieniowe są charakterystyczną

cechą topologiczną danego szkieletu cząsteczki. Ponadto zachowanie wiązań aromatycznych pozwala na zapisywanie struktury z wykorzystaniem kodu SMILES.

5 Specjalistyczne narzędzia w chemoinformatyce

W tym rozdziale opisano środowisko programowania Python. Język z względu na szereg zalet znajduje szczególne miejsce w chemoinformatyce, ponadto w niniejszej pracy był podstawowym elementem prowadzonych badań.

5.1 Język programowania Python

Początki Pythona sięgają lat 80-tych, gdy Guido van Rossum rozpoczął prace modernizacyjne języka ABC. Po ukończeniu projektu Python spotkał się z dużym zainteresowaniem społeczności, to z kolei przyczyniło się do rozwoju i wydaniu kolejnych wersji. Python jest projektem Open Source, rozprowadzany na otwartej licencji GPL [89]. Jest to dynamiczny obiektowy język programistyczny, który jest stosunkowo łatwy w nauce (w porównaniu do innych języków, np. z rodziny C), lecz mimo to jest bardzo potężny [90,91].

Python jest klasyfikowany jako język skryptowy, interpretowany, oznacza to, że najpierw piszemy skrypt a następnie wykonujemy go za pomocą interpretera. Język posiada rozbudowaną bibliotekę standardową, umożliwiającą jego stosowanie do wielu zadań. Większa część biblioteki standardowej dostępna jest na wszystkich platformach, dzięki czemu nawet duże aplikacje mogą często być uruchamiane bez konieczności modyfikacji na Uniksach, Windows i innych platformach. Oprócz Pythona (napisanego w C) dostępne są także implementacje Pythona w Javie (Jython) i .NET Microsoftu (IronPython) działające wszędzie tam, gdzie dostępne są te platformy [91].

Struktura języka pozwala na szybkie implementowanie nowych funkcjonalności, dodatkowo czytelny kod, który łatwo rozszerzać i aktualizować przyczyniły się do wykorzystania języka m.in. w dużych projektach i aplikacjach serwerowych. Międzynarodowe korporacje takie jak: Google, Yahoo, Nokia, IBM korzystają z możliwości Pythona implementując go w swoich aplikacjach i projektach. NASA od wielu lat stosuje język do zarządzania i kontroli pracy

wahadłowców kosmicznych. Również developerzy systemów operacyjnych (Microsoft, Apple, Unix) w swoich platformach programistycznych oferują pełne wsparcie dla tego języka. Warto wspomnieć, iż wiele popularnych stron internetowych (np. YouTube) w większości jest napisanych w Pythonie [92].

Według indeksu popularności języków programowania TIOBE (z dnia 18.04.2016) [93], Python odnotowuje wysokie noty co przekłada się na szóstą pozycję (na 100 notowanych) w rankingu 2016. Dla porównania w roku poprzednim (2015) zajął czwartą pozycję w rankingu popularności. W chwili obecnej Python jest bardziej popularny niż skryptowy język do generowania stron internetowych PHP czy JavaScript, ustępuje jednak językom z rodziny C (kolejno C, C#, C++) oraz Javie, która utrzymuje pierwszą pozycję [93].

Tabela 5.1. Lista rankingowa TIOBE 10 najpopularniejszych języków programowania (TIOBE, data dostępu: 18.04.2016).

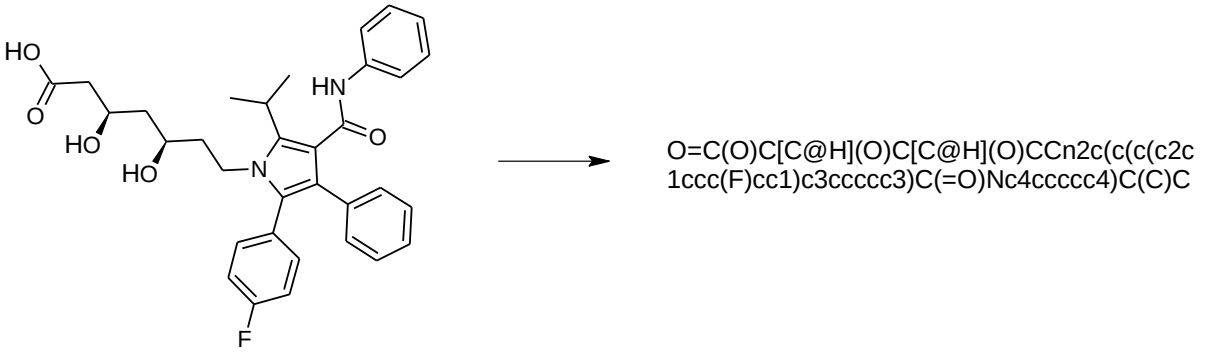
Maj 2016	Maj 2015	Język programowania	Popularność	Zmiana
1	1	Java	20,96%	4,09%
2	2	C	13,22%	-3,62%
3	3	C++	6,70%	-1,18%
4	5	C#	4,48%	-0,78%
5	6	Python	3,79%	0,06%
6	9	PHP	2,99%	0,27%
7	7	JavaScript	2,34%	-0,79%
8	15	Ruby	2,34%	1,07%
9	11	Perl	2,33%	0,51%
10	8	Visual Basic .NET	2,33%	-0,64%

5.2 Kodowanie informacji

Poznanie właściwości określonego związku wciąż stanowi trudne wyzwanie dla chemika. Struktura topologiczna związku bezpośrednio determinuje jego właściwości. W celu transformacji informacji chemicznej zawartej w strukturze cząsteczki do formy numerycznej wykorzystuje się operacje matematyczne i logiczne otrzymując tzw. deskryptory molekularne. Deskryptor może być reprezentowany w postaci wektora lub macierzy. Ze względu na różnorodność danych, podział deskryptorów może być bardzo zróżnicowany. Jednym ze sposobów klasyfikacji deskryptorów jest ich podział w zależności od zastosowania.

Wyróżniamy dwa rodzaje deskryptorów, kodujące i niekodujące. Pierwszy z nich rozwiązuje problem przedstawienia cząsteczki. W sposób przejrzysty koduje stereochemię i konstytucję w postaci np. kodu SMILES. Natomiast drugi rodzaj określa cechy molekularne. Przykładem może być masa molowa, liczba atomów, ładunki cząstkowe, log P, TPSA [2,94,95].

Sposób kodowania cząsteczki w systemach bazodanowych musi być jednoznaczny i unikatowy. W tym celu stworzono system umożliwiający przetwarzanie i wymianę informacji strukturalnej w postaci tzw. kodu liniowego SMILES. Za pomocą notacji liniowej można w kompletny sposób zapisać topologię całej cząsteczki (SMILES), jej fragmentu (SMARTS) oraz reakcji chemicznych (SMIRKS). Do kodowania struktury związku używany jest system kodowania ASCII, wykorzystujący znaki alfanumeryczne (odpowiednie sekwencji liter i cyfr) [96].



Rysunek 5.1. Przykład przekształcenia cząsteczki Lipitora w liniową notację SMILES.

Najprostsza reprezentacja struktury nazywana jest kodem ogólnym lub generycznym. Atomy przedstawiane są poprzez ich symbole atomowe zapisywane w nawiasach kwadratowych. Nie dotyczy to atomów wchodzących w skład związków organicznych (B,C,N,O,S) tworzących typową liczbę wiązań kowalencyjnych np. C(4). Atomy pochodzące od pierścieni aromatycznych zapisywane są małą literą. Atomy wodoru mogą być zaznaczone jako "n" w nawiasach kwadratowych jako [nH]. Natomiast ładunek formalny atomu reprezentowany jest jako "+" dla kationu lub "–" dla anionu z podaną wartością liczbową. Wiązanie pojedyncze przedstawiane jest jako "-", wiązanie podwójne "=", a potrójne "#". Symbole te można pominąć, gdy sąsiednie atomy są połączone wiązaniem pojedynczym lub aromatycznym. Podstawniki lub łańcuchy boczne dołączone do łańcucha głównego zaznaczone są jako ciągi znaków w nawiasach okrągłych. Układ cykliczny sprowadzany jest do postaci liniowej przez rozerwanie jednego z wiązań pojedynczych. Atomy tego wiązania

zaznaczone są arbitralnie dobieranymi liczbami naturalnymi (od 1 do n). Jeśli atomy nie są połączone wiązaniem walencyjnych to w kodzie SMILES oznacza się je jako kropka. Odmiany izotopowe atomów pierwiastków koduje się przy pomocy symbolu pierwiastka poprzedzonego wartością liczby atomowej np. [12C]. Stereoizomeria geometryczna przedstawiana jest w postaci znaków specjalnych, "/" i "\" np. Cl/C=C/Cl lub Cl\C=C\CL dla trans-dichloroetanu [96].

5.3 Biblioteki chemoinformatyczne

Narzędzia chemoinformatyczne (ang. toolkits) są zestawami kodu, funkcji, zaimplementowanych interfejsów bądź aplikacjami komputerowymi. Chemoinformatyczne toolkity znajdują zastosowanie w obliczeniach biochemicznych, przewidywaniu właściwości fizykochemicznych, przy przeszukiwaniu wirtualnych baz bio- lub chemoinformatycznych (wirtualny skrining) oraz innych metodach wykorzystywanych we wczesnych etapach projektowania leków. Pomimo zaawansowanych rozwiązań programistycznych wciąż wiele specjalistycznych tookitów nie jest wpierana interfejsem GUI. Z biegiem czasu coraz więcej udostępnianych programów oferuje graficzny interfejs np. MATLAB, KNIME, ChemAxon, itp. Z drugiej jednak strony rozwój chemoinformatyki jest niezbędny bez ciągłych modyfikacji i ulepszania istniejących już funkcji i algorytmów. Dlatego podstawowa znajomość przynajmniej jednego języka programowania (np. Pythona) jest niezbędna w pracy naukowo-badawczej chemoinformatyka.

Pisanie programów lub funkcji od podstaw jest czasochłonne i wymaga sporej wiedzy użytkownika, ponadto nad bardziej skomplikowanymi projektami pracują całe zespoły specjalistów. W celu zaimplementowania istniejących już rozwiązań z pomocą przychodzą biblioteki programistyczne. Idea bibliotek została specjalnie stworzona z myślą o przechowaniu i ponownym wykorzystaniu danych, funkcji oraz podprogramów z poziomu kodu źródłowego [91,92]. Dobór odpowiedniej biblioteki programistycznej może ułatwić zaimplementowanie niemalże dowolnego zadania, tym samych skracając czas i wysiłek na pisaniu algorytmów od początku.

6 Scjentometria – kategoria sukcesu i prestiżu

Nauka podobnie jak projektowanie leków (R&D w przemyśle farmaceutycznym) jest wyznacznikiem wysokiego prestiżu i poziomu innowacyjności kraju. Postępująca globalizacja oraz możliwości swobodnego transferu danych pomiędzy ośrodkami badawczymi zmieniły charakter pracy uczonych. Generowanie i magazynowanie ogromnych ilości informacji przyczyniły się do stworzenia repozytoriów zwanych "big data". Aktywne przeszukiwanie bibliotek informacji z biegiem lat stawało się coraz bardziej czasochłonne, co w efekcie skutkowało eskalacją kosztów badań. Dotychczasowe systemy oceny pracy naukowej stały się niewystarczające pod zwiększonym napływem publikacji naukowych (rozdział 3 – wykres 3.1). Wielokrotnie podejmowano próby nad wprowadzeniem rozwiązania, w oparciu o które możliwa byłaby odpowiednia dystrybucja środków finansowych na rzecz nauki, a także efektywniejsze gospodarowanie dostępnymi zasobami ludzkimi. Wprowadzenie skutecznej metody oceny pracy naukowej pozwoliłoby na kompleksową analizę nauki, która jest istotnym wyznacznikiem ekonomicznym stanowiącym w dużej mierze o potędze gospodarczej kraju [97-99].

Derek Price jako pierwszy opracował model matematyczny, który posłużył do oszacowania ilościowych indykatorów w ocenie polityki naukowej. Swoją pracę pt. "Little Science, Big Science" opublikował w 1963 roku [100]. Kilka lat później Nalimov zidentyfikował metodykę oceny wartości i efektywności pracy naukowej w oparciu o system metryczny, który nazwał pojęciem scjentometria (ang. scientometrics) i termin ten funkcjonuje po dziś dzień [101,102]. Niestety zimna wojna, żelazna kurtyna oraz słaby dostęp do Internetu w bloku wschodnim spowodowały, iż praca Nalimova nie dotarła do szerszego grona naukowców. W środowisku naukowym pojęcie scjentometria zaczęło funkcjonować dopiero po 1978 roku, kiedy ukazał się pierwszy numer czasopisma Tibora Brauna pt. "Scientometrics". Tytuł publikacji będący zapożyczoną translacją terminu zaproponowanego przez Namilova spowodował, iż określenie to pokonało barierę polityczno-geograficzną i stało się mianem dziedziny naukowej rozpoznawalnej w ujęciu globalnym [103]. Eugene Garfield pomysłodawca systemu zwanego indeksem cytowań (ang. Citation index) wraz z Derekiem Pricem stali się ojcami scjentometrii.

Współcześnie scjentometria jest dziedziną zajmującą się badaniem i oceną wydajności pracy naukowej (badania ilościowe), jak i zaistniałych relacji pomiędzy nimi. Scjentometria leży na pograniczu nauk matematycznych (w szczególności metod statystycznych), bibliometrii

oraz informatyki [104]. Do roku 2014 obok naukowego indeksu cytowań pojawiło się wiele innych narzędzi parametrycznej oceny wydajności naukowej m.in. impact factor, indeks H, algorytm Pagerank, czy różnego rodzaju rankingi uczelni, funkcjonujące w oparciu o cyfrowe bazy danych.

Obecnie istnieje wiele narzędzi i parametrów pozwalających wykonać analizę scjentometryczną. Do najważniejszych wskaźników należą: impact factor, indeks Hirscha, naukowy indeks cytowań lub ranking ARWU.

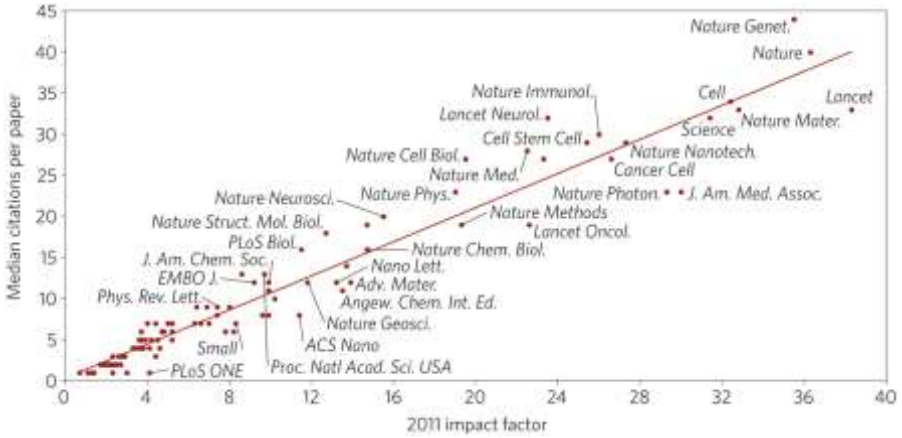
W 1964 roku E. Garfield na podstawie serwisu ISI (ang. Institute for Scientific Information) znanym również jako Instytutem Filadelfijskim, opracował naukowy indeks cytowań (ang. Science Citation Index, SCI), który obecnie należy do korporacji wydawniczej Thomson Reuters. SCI jest cyfrową bazą przechowującą informacje zarówno na temat artykułów, jak i powiązanych z nimi publikacji cytowanych. W oparciu o dane ISI oblicza się wskaźnik cytowań, który ocenia siłę i jakość pracy naukowej to z kolei pomaga porównywać i zestawiać w rankingi poszczególnych naukowców oraz instytucje badawcze [105,106].

Dostęp do baz danych tworzonych przez ISI można uzyskać za pośrednictwem internetowego serwisu Web of Science, będącego częścią baz Web of Knowledge [107]. Dostęp do bazy jest udzielany wyłącznie na podstawie licencji. W sieci istnieje wiele podobnych serwisów. Jednym z najbardziej popularnych jest Scopus należący do wydawnictwa naukowego Elsevier [108]. Do ogólnodostępnych darmowych baz można zaliczyć Google Scholar oraz ResearchGate, który spełnia rolę serwisu społecznościowego zrzeszającego naukowców z całego świata.

Impact factor (ang. impact factor, IF) jest jednym z najpopularniejszych wskaźników cytowań pozwalających oszacować siłę oddziaływania i prestiż czasopism naukowych. Wyraża się on stosunkiem łącznej liczby cytowań wszystkich artykułów danego czasopisma w roku kalendarzowym do liczby cytowanych w nim publikacji, które ukazały się w ciągu ostatnich dwóch lat, zgodnie ze wzorem**:**

$$IF = \frac{A}{B}$$

gdzie; A – liczba cytowań wszystkich artykułów w danym roku, B – liczba cytowanych artykułów, które ukazały się w danym czasopiśmie w ciągu ostatnich dwóch lat.



Rysunek 6.1 Zależność przeciętnej liczby cytowań od impact factora czasopism naukowych [109].

Czasopisma skupiające większą atencję czytelników posiadają proporcjonalnie większy wskaźnik IF, który można utożsamić z większą siłą oddziaływania naukowego. Innymi słowy publikacje o wysokim IF wyznaczają kierunki atrakcyjnych badań tym samym mają wpływ na zachodzące trendy w nauce.

Pokrewnym miernikiem do IF jest indeks h (indeks Hirscha), który został zaproponowany przez J. Hirscha w 2005 roku [110]. Wskaźnik h pomaga w sposób numeryczny zmierzyć siłę oddziaływania i jakość prac naukowych danego autora. W odróżnieniu od IF, indeks charakteryzuje nie pojedynczą publikację lecz całkowity dorobek naukowy danego autora. Używając parametru Hirscha, warto pamiętać o zasadzie porównywania dorobków naukowców w obrębie jednej dziedziny, ponieważ liczba cytowań publikacji jest z nią silnie skorelowana.

Na podstawie parametrów scjentometrycznych stworzono wiele rankingów porównujących jakość pracy naukowej. W ten sposób można porównywać zespoły lub nawet same jednostki badawcze. Przykładem takiego zestawienia jest np. międzynarodowa Szanghajska lista szkół wyższych (ARWU), ranking tajwański (NTU Ranking) czy leideński (CWTS). Do oceny jakości uniwersytetów zestawienia wykorzystują wiele parametrów scjentometrycznych takich jak m.in. liczbę absolwentów czy pracowników, którzy otrzymali prestiżowe nagrody (np. Nobla lub medal Fieldsa), liczbę cytowań i publikacji, stosunek studentów do wykładowców, odsetek studentów międzynarodowych, liczbę artykułów publikowanych w prestiżowych czasopismach (Nature, Science, Lancet), itp.. Wyższa pozycja na liście oznacza większy prestiż, zwiększa szanse otrzymania dotacji, nagród, przyciąga lepiej

wykwalifikowanych pracowników i studentów, co bezpośrednio przekłada się na jakość i ilość publikowanych prac.

Tabela 6.1 Dziesięć państw uszeregowanych malejąco wg rankingu ARWU (data dostępu: 24.07.2016).

Państwo	*Top20*	*Top100*	*Top200*	*Top300*	*Top400*	*Top500*	*Suma*
1. USA	16	51	78	102	125	146	518
2. UK	3	9	21	28	33	37	131
3. Szwajcaria	1	4	6	7	7	7	32
4. Niemcy	—	4	13	21	28	39	105
5. Francja	—	4	8	15	18	22	67
6. Australia	—	4	8	11	19	20	62
7. Holandia	—	4	8	10	12	12	46
8. Japonia	—	4	7	9	12	18	50
9. Kanada	—	4	6	16	18	20	64
10. Szwecja	—	3	5	7	10	11	36

7 Omówienie wyników badań

Praca doktorska została zrealizowana w Zakładzie Chemii Organicznej Uniwersytetu Śląskiego, gdzie od kilkunastu lat prowadzi się badania skoncentrowane na eksploracji baz danych oraz architektury leków.

W pracy badawczej wykorzystano specjalistyczne programy i chemiczne bazy danych (np. FDA [64], PubChem [111,112], ChEMBL [52,113]) idealne do badania, projektowania nowych i potencjalnie aktywnych związków chemicznych. Wykorzystano deskryptory chemiczne jak również ekonomiczne, które poddano szerokiej analizie statystycznej. Badania wykonywano na komputerze klasy mikro (komputer osobisty) o parametrach: Intel Core 2 Duo CPU 2x2,13GHz, RAM 4GB, SSD 500GB używając systemy Linux oraz Windows. W części pierwszej analizowano scjentometryczny model efektywności R&D z wykorzystaniem wyselekcjonowanych leków FDA. Przeprowadzono analizę porównawczą trendów i zmian występujących zarówno w nauce jak i przemyśle farmaceutycznym. Dodatkowe wykorzystanie metod dekompozycji związków pozwoliło zrozumieć topologię oraz jej wpływu na aktywność badanej populacji leków.

W pierwszym etapie badań dokonano analizy scjentometrycznej w oparciu o parametry ekonomiczne. Zbadano wydajność pracy naukowej w skali makro (na poziomie państw) jak i mikro (na poziomie jednostek badawczych) [98]. W następnej części dokonano przeszukiwania baz w oparciu o wcześniej ustalone kryteria (ekonomiczno-prestiżowe). Zebranie możliwie największej liczby leków wraz z odpowiednimi deskryptorami było trudnym i czasochłonnym zadaniem. Eksploracja obejmowała liczne chemo- i bioinformatyczne bazy, literaturę, dokumentację FDA, patenty oraz źródła informacji o prowadzonych badaniach klinicznych. Na tym etapie badań wykorzystano metody programowania celem maksymalizacji efektywności przeszukiwań. Cenna umiejętność programowania pozwoliła na szybkie przetwarzanie i analizowanie dużej liczby danych. Zaprojektowano kilkadziesiąt skryptów chemoinformatycznych. Dodatkowa znajomość MySQL oraz innych języków programowania: Python, PHP, Java Script, oraz systemu Linuks znacząco ułatwiła procesy analityczne. Wyszukane związki opisano różnymi parametrami m.in. strukturą w kodzie SMILES, datą rejestracji FDA, masą molową, lipofilowością, TPSA, itp. Na ich podstawie możliwe było określenie pewnych podobieństw oraz wyodrębnienie grupy o wspólnych cechach. Określenie miejsca działania związku

pozwoliło na sklasyfikowanie ich w odpowiednie grupy. Tak przygotowane i pogrupowane dane w dalszej pracy zostały poddane szeroko pojętej analizie.

W trakcie stopniowego "odkrywania wiedzy" tkwiącej w zgromadzonych danych o lekach zaobserwowano wiele ciekawych zależności i trendów m.in. zmian jakie zaszły w projektowaniu farmaceutyków na przestrzeni ostatnich lat. Zbadano i scharakteryzowano zebrane związki pod względem fizykochemicznym, zaproponowano nowy parametr wiek leków (charakteryzuje efektywność projektowania) [114] oraz dokonano analizy topologicznej z wykorzystaniem wcześniej wspomnianych metod fragmentarycznych.

8 Badania scjentometryczne

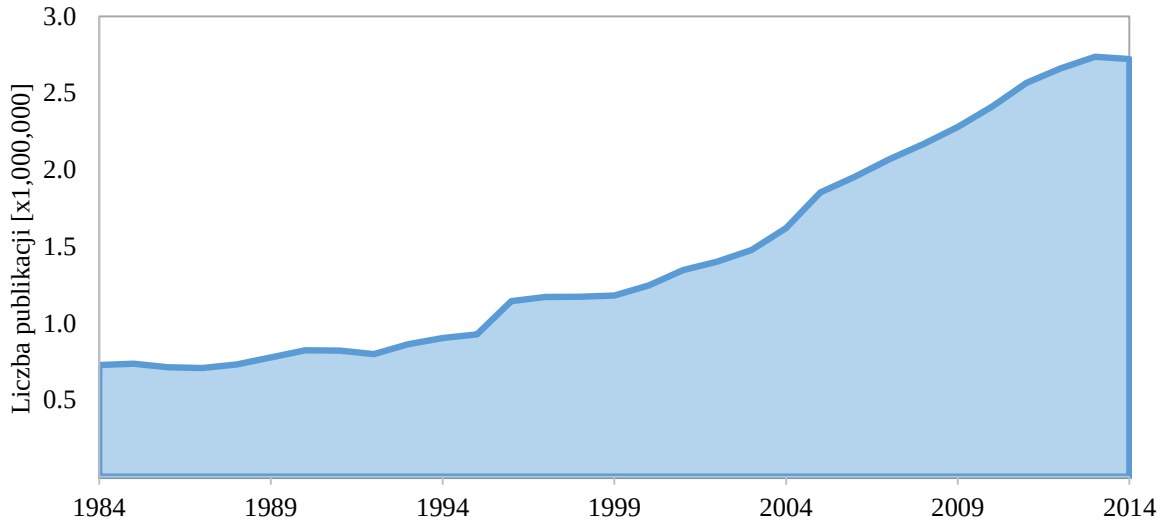
Wkład nauki, czyli badań prowadzonych w jednostkach uniwersyteckich w rozwój przemysłu technologicznego jest nieoceniony. Bez innowacyjności, ciągłego badania, poszukiwania i odkrywania wiedzy nie można liczyć na postęp. Na przestrzeni ostatnich lat coraz więcej firm podejmuje współpracę z zespołami naukowców w celu ulepszania swoich produktów bądź wprowadzania innowacyjnych rozwiązań. Instytuty naukowe mocno zabiegają o dodatkowe fundusze w postaci grantów, projektów lub pozyskania patronatów biznesowych. W celu uzyskania wsparcia finansowego jednostka badawcza musi zainteresować potencjalnych inwestorów. Duży wpływ przy ocenie instytutu mają dotychczasowe osiągniecia. Na podstawie wcześniej przytoczonych informacji można wywnioskować, iż wybór odpowiedniego zespołu można obliczyć za pomocą parametrów scjentometrycznych. Im lepsze wskaźniki tym wyższy poziom bezpieczeństwa. Na podstawie rankingów szkół wyższych wyraźnie widać dysproporcje pomiędzy silnymi gospodarczo państwami takimi jak np. USA, UK, Niemcy. Z drugiej strony kraje cechujące się słabo rozwiniętą gospodarką np. Bangladesz, Zimbabwe nie mają w swoim portfolio innowacyjnych osiągnięć.

Hipoteza badawcza, która została postawiona przed przystąpieniem do analizy brzmiała następująco: "Czy uniwersalny i skalowalny parametr ekonomiczny jakim jest poziom finansowania, ma wpływ na jakość pracy naukowej?"

Następne podrozdziały dotyczą analizy oraz ocenie wpływu finansów na wydajność pracy naukowej w skali makro (krajowej) oraz mikro (poszczególnych uniwersytetów).

8.1 Analiza wydajności pracy naukowej: sukces i prestiż naukowy

Znaczny wzrost ilości materiałów naukowych, spowodowany postępem technologicznym przyczynił się do zmiany charakteru pracy uczonych. Postępująca globalizacja oraz możliwości swobodnego transferu danych pomiędzy ośrodkami badawczymi spowodowały przepływ ogromnych ilości informacji. Jak wcześniej wspomniano, dotychczasowe systemy oceny pracy naukowej stały się niewystarczające pod napływem mnogości publikacji naukowych [97].



Rysunek 8.1 Przyrost ilości publikacji naukowych w latach 1984-2014 (Scopus, data dostępu: 17.04.2015).

Obliczanie efektywności pracy naukowej jest poważnym i kontrowersyjnym problemem. "Publikuj lub giń" stało się strategią (złotą regułą), jednostek akademickich, które czynią wysiłki w publikowaniu prac w najlepiej cytowanych czasopismach. Indeks Hirscha i światowy ranking uniwersytetów ARWU są przykładami narzędzi klasyfikacji, które starają się imitować powszechnie znane rankingi np. ranking klubów piłkarskich.

Analizując popularne rankingi (klubów sportowych, firm, państw, biznesmenów) parametrem korelujący z pozycją na liście jest potencjał finansowy. Jak ważne jest oszacowanie funduszy i połączenie ich z rozwojem nauki? Uniwersytety amerykańskie są zwycięzcami nagród Nobla, częściej publikują w najlepszych czasopismach typu Nature i Science, są motorem innowacyjności, itd.

Stany Zjednoczone rocznie przeznaczają 150 mld dolarów na sektor nauki. Źródła finansowania szkolnictwa wyższego w Ameryce Północnej są szerokie. Na przykład Uniwersytet Stanforda dysponował budżetem większym niż 4 mld dolarów, dodatkowo otrzymał dotacje w wysokości 17 mld (dane dotyczące roku 2012) [98]. Uniwersytet Stanford uzyskał 17 nagród Nobla. Oprócz Stanfordu jest jeszcze kilkunastu równie znaczących graczy na samym rynku amerykańskim. W ostatnich latach zaobserwowano zwiększające się znaczenie krajów Azji m.in. Chiny, Singapur, Japonia. Wraz z zwiększającą się liczbą publikacji naukowych na sile zyskały prestiżowe czasopisma (Nature, Science), coraz częściej nazywane "biblią nauki". Co ciekawe, współczynnik korelacji pomiędzy całkowitą liczbą publikacji a wielkością budżetu przeznaczonego na naukę waha się w granicach od 0,7 do 0,9 [115]. Dotychczas pojawiło się kilka publikacji opisujących silną zależność pomiędzy zapleczem finansowym a jakością prac w różnych sektorach nauki [116-119]. Szczególnie silną zależność zaobserwowano w naukach medycznych, gdzie współczynnik korelacji Pearsona pomiędzy budżetem poszczególnych instytutów a ich publikacjami wyniósł 0,9 [120].

W niniejszym rozdziale przeanalizowano zależność pomiędzy liczbą publikacji w Nature a wydatkami na naukę przez indywidualne państwa i uniwersytety. Celem była weryfikacja hipotezy, że sukces naukowy w dużym stopniu zależy od poziomu finansowania badań naukowych. Nature jest najczęściej cytowanym czasopismem na świecie, jego impact faktor wynosi 41.458 (5-letni IF, 2016). Na podstawie badań K. Kaneiwy [121] przyjęto, że liczba publikacji w Nature jest miarą efektywności pracy naukowej. Nature swoim zasięgiem obejmuje cały glob. Publikacje w Nature są więc źródłem prestiżu. Daje szansę i przewagę na zdobycie dalszych funduszy na badania, co więcej dla wielu naukowców jest celem na miarę Nobla [122]. Science wydaje się być bardziej regionalnym czasopismem, w którym częściej publikują organizacje Ameryki Północnej.

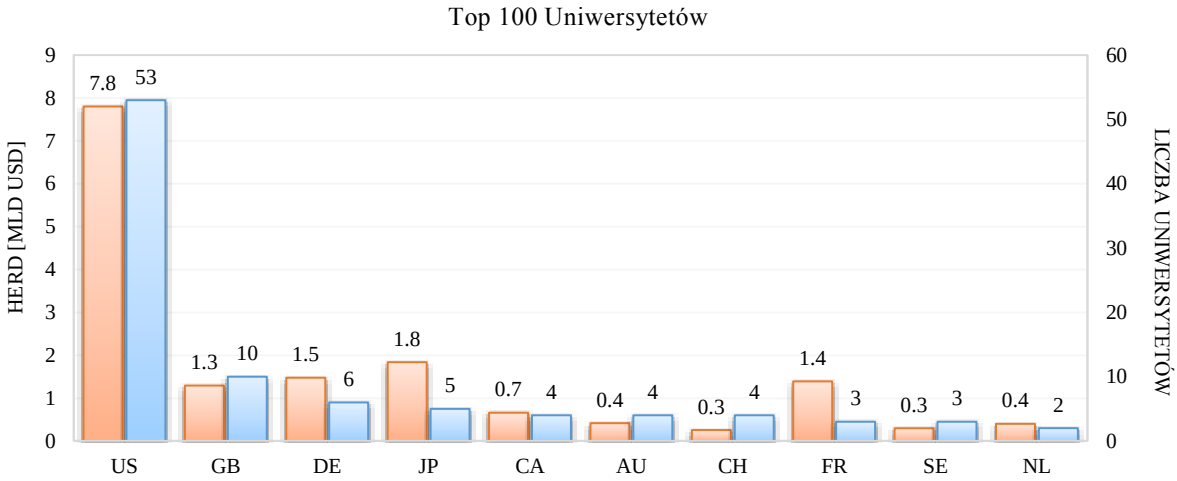
8.2 Ekonomiczny wyznacznik sukcesu

W celu określenia realnych wydatków na szkolnictwo wyższe w rozprawie wykorzystano kilka parametrów ekonomicznych, takich jak:

- **Produkt krajowy brutto** (ang. gross domestic product, GDP) – jest podstawowym ekonomicznym miernikiem wielkości gospodarki państwa. W uproszczeniu GDP określa łączą wartość dóbr i usług wytworzonych na terenie danego kraju w określonej jednostce czasu (najczęściej w ciągu roku). Wzrost lub spadek realnego PKB oraz dynamika tych ruchów stanowi miarę wzrostu gospodarczego.

- **Wydatki krajowe brutto na badania i rozwój** (ang. gross domestic expenditure on research and development, GERD) – całkowite wydatki na prace badawczo-rozwojowe realizowane na terytorium danego kraju (w określonym czasie). Wskaźnik ten obejmuje finansowanie z zagranicy działalności badawczej i rozwojowej prowadzonej w danym kraju, jednak nie uwzględnia funduszy przekazywanych na ten cel za granicę.

- **Wydatki krajowe brutto na badania i rozwój w szkolnictwie wyższym** (ang. high education expenditures on research and development, HERD) – wyrażone jako suma wszystkich wydatków, również zagranicznych przeznaczonych na prace badawczo-rozwojowe w sektorze szkolnictwa wyższego na terenie danego kraju.



Rysunek 8.2 Średnie wydatki HERD w latach 2001-2011 oraz liczbę uniwersytetów w liście ARWU top 100 na rok 2011.

- **Rządowe finansowanie badań** (ang. total research expenditures) – wydatki, które uniwersytet przekazuje do Krajowej Rady Nauki (ang. National Science Board, NSB).

- **Roczne dotacje pozarządowe** (ang. annual fund/giving) – wskaźnik składek pozyskanych przez prywatnych inwestorów, korporacje i inne pozarządowe organizacje. Fundusze najczęściej przeznaczane są na innowacyjne projekty, marketing lub pomoc materialną dla studentów.

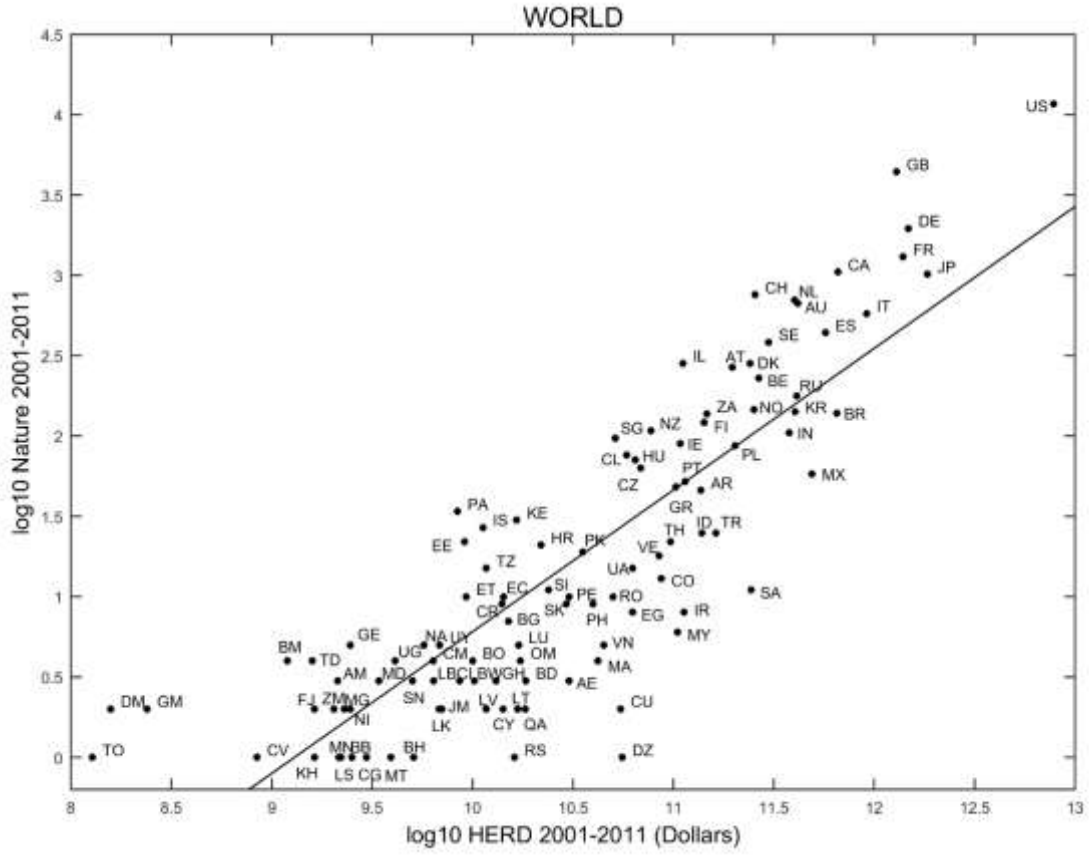
- **Darowizny** (ang. endowments) – parametr ten odzwierciedla długoterminowe zasoby finansowe pozyskane od jednostek pozarządowych. Zazwyczaj darowizny są wysokie zatem głównym przeznaczeniem tego źródła są inwestycje. Przekłada się to na kondycję finansową a to z kolei na pozycję danego uniwersytetu. Ponadto, część budżetu darowizn w granicach 4,5-5% pokrywa bieżące wydatki. W ten sposób jednostka zabezpiecza się przed ewentualnym regresem ekonomicznym np. w przypadku cięć budżetowych. Trzeba zaznaczyć, iż część darowizn pokrywa również koszty innowacyjnych badań. Im większy budżet tym większy nacisk na innowacyjne badania a to z kolei otwiera większą szansę na przełomowe odkrycia.

Źródłem danych były dostępne w sieci internetowej specjalistyczne bazy danych:

- **World Bank Database** – baza zawierająca ponad 8000 różnych ekonomiczno-gospodarczych wskaźników dla ponad 200 krajów świata [123].
- **Eurostat** – największa europejska baza statystyczna [124].
- **Annual Report on Research Universities in the US (ARRU)** – roczny raport zawierający szczegółowe informacje finansowe ponad 200 uniwersytetów na terenie Stanów Zjednoczonych [125].
- **Web of Knowledge (Web of Science)** – należąca do Thomson Reuters jedna z największych dostępnych baz publikacji naukowych [107].
- **Nature Publishing Index (NPI)** – roczny ranking top 200 instytucji ze względu na całkowitą liczbę publikacji w Nature (artykuły, listy, komunikaty) [126].

W prowadzonych badaniach wszystkie użyte parametry ekonomiczne zostały ujednolicone tzn. wyskalowane w stosunku do dolara amerykańskiego. Z uwagi na ograniczony dostęp lub opóźnienia w podawaniu do publicznej wiadomości informacji finansowych, niektóre wskaźniki dotyczą krótszego okresu np. ARRU 2014 zawiera dane dla darowizn do 2010 roku.

W pierwszym etapie dokonano skrining a następnie ekstrakcję danych ekonomicznych i liczby publikacji dla 105 państw świata. Dla Chin, Jordanu, Gabonu czy Libii nie znaleziono informacji dotyczących wydatków na szkolnictwo wyższe. Dlatego nie uwzględniono tych państw w dalszej analizie. Dziesięcioletni okres obejmujący lata od 2001 do 2011 (badania prowadzone w okresie maj-czerwiec 2013 roku) był spowodowany opóźnieniem w zbieraniu informacji i uaktualnianiu internetowej bazy danych.



Rysunek 8.3 Korelacja pomiędzy liczbą publikacji w Nature afiliowanych przez poszczególne państwa a wydatkami na szkolnictwo wyższe (HERD) w latach 2001-2011 (skala logarytmiczna). Współczynnik korelacji Pearsona wynosi 0,96 [98].

Wyniki badań wskazują silną korelację (R = 0,96) dwóch zmiennych tj. skumulowanej liczby publikacji w czasopiśmie Nature oraz nakładów na szkolnictwo wyższe (HERD) w latach 2001-2011. We rankingach (ARWU, top 100 uniwersytetów, liczbie publikacji Nature) dominują Stany Zjednoczone (US), które w tym okresie przeznaczyły 3,80 mld dolarów na R&D

jednocześnie publikując 11648 artykułów w Nature. Dla porównania Japonia (JP, druga pozycja) w tym samym okresie wydała o 55% mniej (1.69 mld dolarów) publikując niecałe 12% w porównaniu do US (1019 artykułów w czasopiśmie Nature).

Podobną analizę wykonano na podstawie danych z Eurostatu [124]. Do analizy włączono 32 państwa europejskie, zebrano dane dotyczące HERD, GERD, jak również liczbie publikacji Nature w tym samym okresie (2001-2011). W wyniku analizy stwierdzono silną korelację pomiędzy parametrami (R > 0,9), dane zaprezentowano w tabeli 8.1.

Tabela 8.1 Wartości parametru korelacji Pearsona dla danych z World Bank oraz Eurostatu.

Źródło danych ekonomicznych	Parametr ekonomiczny	Źródło danych bibliometrycznych	Liczba państw	Współczynnik Pearsona
World Bank Database	HERD	Web of Knowledge	105	0,96
	GERD			0,91
Eurostat	HERD	Web of Knowledge	32	0,92
	GERD			0,91

Wielka Brytania (UK), Niemcy (DE) i Francja (FR) dominują w Europie pod względem wydatków i osiągnięć w sektorze szkolnictwa wyższego. Interesującym wydaje się być fakt, iż Unia Europejska liczona jako jeden organ dominuje nad wszystkimi pozostałymi krajami, publikując 12306 artykułów (dla porównania 11648 w US). Jednakże przegrywa w zestawieniu liczby ludności, która wynosi 508 mln w UE oraz 312 mln w US. Stany Zjednoczone mają przewagę w przeliczeniu publikacji per capita (~35% więcej). Z drugiej jednak strony, kraje Europy wschodniej zaniżają średnią państw zachodnich. Dla porównania afiliacja zachodnich członków UE w Nature wynosi 11829, co daje 96,1% wszystkich publikacji UE (państwa Europy wschodniej publikują 3,9% w tym samym okresie). Warto zauważyć, iż kraje które przeznaczają najwięcej na sektor szkolnictwa wyższego (US, JP, UK, DE, FR, itd.) wiodą również globalny prym w innowacyjności oraz charakteryzują się najwyższym poziomem zaawansowania technologicznego. Według danych NSB (National Science Board) [127], globalne wydatki na badania i rozwój w ostatniej dekadzie rosły szybciej niż średni światowy PKB. Obecnie 42,8% aktywności naukowej jest skoncentrowanej w Europie, 40,6% w US oraz niecałe 3,6% w Japonii. Podobną dystrybucję można zauważyć analizując listę rankingową ARWU (top 500 uniwersytetów). 151 (30,2%) z listy stanowią uniwersytety US, 40,8% jest zlokalizowanych w Europie, a 11,6% z Japonii. Celem dokładniejszego zbadania efektu

finansowego do dalszej analizy poddano indywidualne jednostki badawcze (top 50 uniwersytetów amerykańskich).

8.3 Efekt św. Mateusza – prestiż w nauce

W rozdziale 7.2 przedstawiono analizę porównawczą liczby publikacji w Nature z wydatkami na szkolnictwo wyższe (HERD) indywidualnych państw. Badanie wykazało silnie dodatnią korelację (parametr Pearsona R = 0,96). Kolejnym etapem pozostała analiza uniwersytetów z listy top 50 USA w celu zaobserwowania podobnych zależności. Po wnikliwej analizie dostrzegliśmy, że w przypadku instytucji mamy do czynienia z dodatkowymi czynnikami, które w znaczący sposób wpływają na wyniki naukowe.

Tabela 8.2 Wartości parametru korelacji Pearsona dla danych z ARRU z podziałem na źródła bibliometryczne. Dane w nawiasie dla total research obrazują korelację bez pierwszych trzech uniwersytetów z listy top 50.

Źródło danych ekonomicznych	Parametr ekonomiczny	Źródło danych bibliometrycznych	Liczba uniwersytetów	Współczynnik Pearsona
ARRU	Endowment	Web of Knowledge	50	0,74
	Annual Giving			0,66
	Total Research			0,30 (0,47)
ARRU	Endowment	NPI	50	0,76
	Annual Giving			0,68
	Total Research			0,34 (0,64)

Uniwersytet Harwarda jest liderem listy top 100 ARWU (zarówno obecnie – 2016, jak również w trakcie przeprowadzania badań – 2013). Nr 1 listy otrzymuje najwięcej dotacji (endowment) sięgających 25 mld USD (2001-2011), w tym samym okresie opublikował również najwięcej prac w Nature (1243). Z drugiej strony zajmuje 24-te miejsce (433 mln USD) w dotacjach na rozwój (total research funds), podczas gdy zwycięzcą w tej kategorii okazuje się Uniwersytet Pennsylvania otrzymując 619 mln USD z łączną liczbą 248 prac w tym samym okresie.

Zaskakującym jest fakt, iż dotacje w odróżnieniu od całkowitych wydatków na badania lepiej opisują wcześniej wspomniane zależności. Analizując strukturę oraz wielkość każdego z trzech parametrów stwierdzono, że dotacje w największym stopniu opisują budżet uniwersytecki.

Dla porównania dotacje Uniwersytetu Harwarda stanowią 96%, dla Uniwersytetu Stanford (nr 2) – 91% oraz dla Uniwersytetu Pennsylvania (nr 14) 83% wszystkich źródeł finansowych.

W przypadku analizy porównawczej wydatków na prace badawcze a efektywnością pracy naukowej zauważono, że po eliminacji pierwszej trójki tj. Uniwersytetu Harward, Stanford oraz MIT korelacja zaskakująco rośnie osiągając wartości R = 0,47 oraz 0,64 dla WoK (Web of Knowledge) i NPI. Najlepsze jednostki przy stosunkowo niższych rządowych funduszach publikują znacznie więcej. Przyczyną tej anomalii jest prestiż jednostki. Renoma przyciąga inwestorów, czyli alternatywne źródła finansowe – dlatego obserwujemy większy udział dotacji w całościowym budżecie.

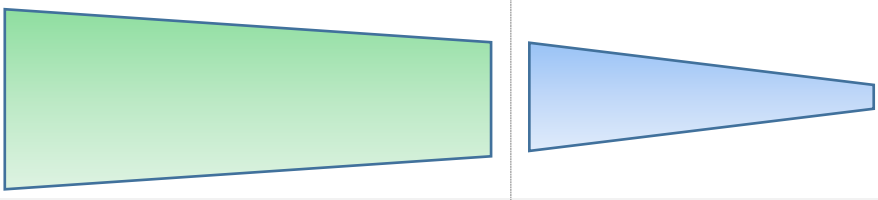
Względy przyrost liczby publikacji wyjaśnić można także przez tzw. efekt św. Mateusza. Według Biblii Wujka z Ewangelii wg św. Mateusza w rozdziale 25, wersecie 29 "*Albowiem wszelkiemu mającemu będzie dano, i obfitować będzie, a temu, który nie ma, i to, co się zda mieć, będzie wzięto od niego.*". Na podstawie powyższej definicji socjolog R. Merton spopularyzował zasadę św. Mateusza o zubożeniu osób biednych i bogaceniu się bogatych. Przeprowadzona analiza potwierdza regułę, że jednostki o wysokim prestiżu publikują najwięcej w Nature [122]. Tworzy się specyficzna pętla sprzężenia zwrotnego, która ułatwia publikowanie i jednocześnie buduje prestiż jednostki.

Warto zwrócić uwagę na kilka innych problematycznych aspektów w przeprowadzonej analizie. Znacznie uproszczono parametr określający efektywność pracy naukowej poprzez analizę publikacji ograniczonej do Nature. Czasopismo z reguły skupia się na kilku dyscyplinach naukowych, tym samym ograniczając dostępność dla szerszego grona odbiorców. Kolejną kontrowersyjną kwestią są afiliacje. W pracy zespołowej zdarza się, że afiliacja jednego artykułu czasami dotyczy innych państw lub uniwersytetów, tym samym jeden artykuł liczony jest we wszystkich instytucjach równocześnie. Innym ważnym czynnikiem jest ograniczony dostęp do danych finansowych w szczególności dotyczy to uniwersytetów. Z tego powodu badania ograniczono do top 50 uniwersytetów amerykańskich dla których scharakteryzowano trzy źródła dochodów: darowizny, rządowe finansowanie badań oraz dofinansowania z innych źródeł. Szczegółowo problemy te omówione zostały w publikacji [98], która została załączona w niniejszej pracy (załącznik 2).

Na podstawie rankingów dokonuje się wyboru najlepszych cech, wskaźników. Dlatego w kolejnych częściach skupiono uwagę na badaniach określających analogię między analizą scjentometryczną (rankingiem prac naukowych) a rankingiem leków.

9 Wprowadzenie leku na rynek – rejestracja przez regulatora FDA.

Związek biologicznie aktywny zyskuje miano leku w chwili uzyskania akceptacji amerykańskiej agencji ds. żywności i leków. FDA uchodzi za bardzo rygorystyczną pod względem bezpieczeństwa i jakości farmaceutyków, dlatego jest ikoną gwarancji i jakości.

	Target-to-hit	Hit-to-lead	Optymalizacja struktury	Faza przedkliniczna	Faza I	Faza II	Faza III	Złożenie wniosku NDA	NME
									
Szansa przejścia do kolejnego etapu	80%	75%	85%	69%	54%	34%	70%	91%	4%
Liczba prac (WIP) na NME	24,3	19,4	14,6	12,4	8,6	4,6	1,6	1,1	1
Czas trwania (lata)	1,0	1,5	2,0	1,0	1,5	2,5	2,5	1,5	13,5
Koszt fazy badawczej (mln $)	$24	$49	$146	$62	$128	$185	$235	$44	$873
Koszty na NME [%]	3%	6%	17%	7%	15%	21%	27%	5%	
Koszt kapitału	11%								
Kapitalizacja kosztów (mln $)	$94	$166	$414	$150	$273	$319	$314	$48	$1778

Rysunek 9.1 Etapy interakcji R&D z rynkiem farmaceutyków [128].

Interakcje leku z rynkiem rozpoczyna jego rejestracja przez regulatora leku np. FDA. Jest to też wyznacznik sukcesu jaki osiągnął projekt R&D firmy farmaceutycznej.

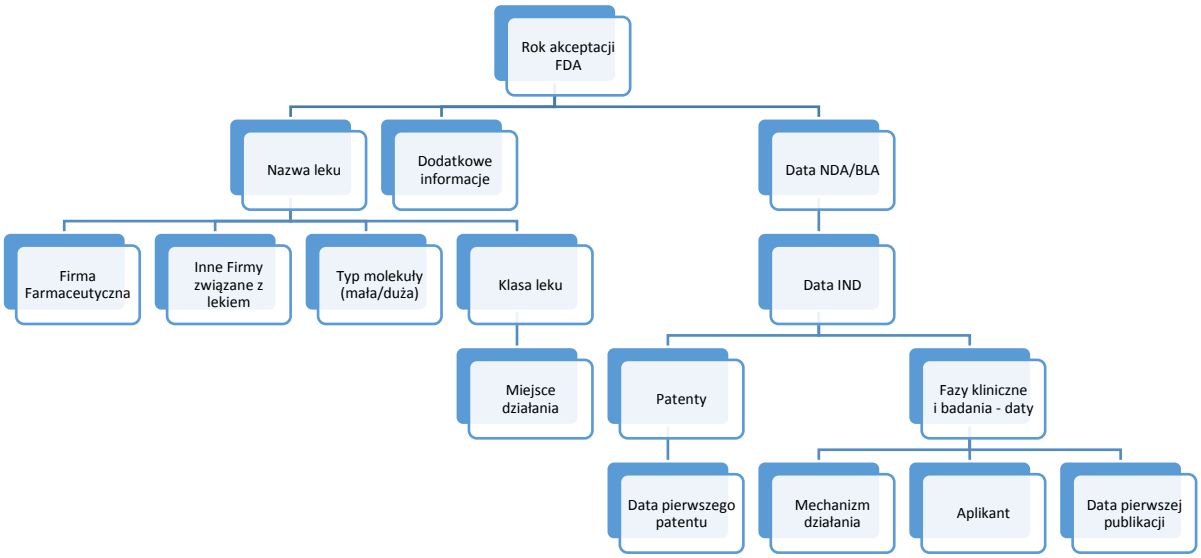
Rysunek 9.2 Ewolucja od kandydata na lek do prestiżowej listy top 100.

Na rysunku 9.2 przedstawiono ewolucję od kandydata na lek przez akceptację FDA do listy top 100, która jest wyznacznikiem "prestiżu" leku. Jakie czynniki decydują o obecności na liście bestselerów?

W niniejszej pracy powiązano tezę występowania zależności między parametrami opisującymi potencjał ekonomiczny oraz badawczy FDA approvals (NME). Tę część pracy przeprowadzono w postaci ekspertyzy oceny wszystkich NME wykonanej we współpracy z:

1. Prof. Jarosławem Polańskim – Uniwersytet Śląski
2. Dr Simone Fishburn – Edytor BioCentury Publications, Centrum Medyczne Uniwersytetu Stanforda
3. Dr Agatą Kurczyk – Politechnika Śląska

Wykonane ekspertyzy opierały się na analizie zbioru szczegółowych informacji (graf 9.3) dla każdego indywidualnego leku. Wybrane dane zestawiono w tabeli 17.1 (załącznik 1).



Rysunek 9.3 Graficzne przedstawienie zbioru informacji, na podstawie których dokonywano oceny parametrów translacyjności.

Pierwszym problemem jest identyfikacja leków zorientowanych translacyjnie od tych które nie spełniają tych reguł. Paradygmat oceny nie może być opisany dyskretnym rozkładem prawdopodobieństwa zwanym zero-jedynkowym. Zatem każdy z leków oceniono używając trzystopniowej skali "- -" (-2), "+ -" (0) lub "+ +" (+2). Tym sposobem szacowano wkład

począwszy od całkowitej niezgodności poprzez bierność aż do całkowitej zgodności z założeniami medycyny translacyjnej [41,82,83,129]. Skalę oparto o kilka elementów, które decydują o translacyjności. W metodyce tej uwzględniono:

1. **Selektywność** związku biologicznie aktywnego związana jest z wywołaniem efektu farmakologicznego z wybranym receptorem. Natomiast leki o niskiej selektywności wywołują ten efekt z wieloma receptorami. Od początków lat 1980-tych obserwujemy wzrost liczby bardziej selektywnych leków. Niska selektywność obecnych leków jest stosunkowo rzadką cechą.

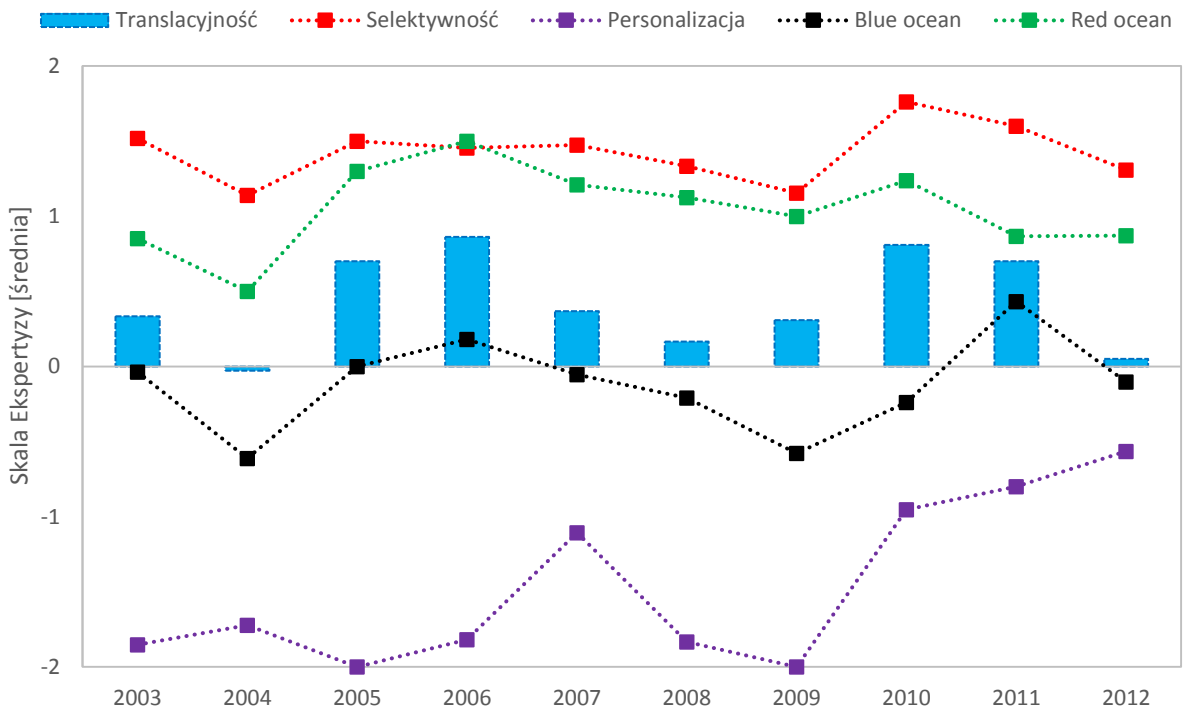
2. **Personalizacja** – relatywnie nowa koncepcja w metodach projektowania leków (zdobyła popularność w latach 2000-nych). Bardzo atrakcyjna dla nowoczesnych metod biologii molekularnej. Oferuje zastosowanie odpowiedniego leku do indywidualnego pacjenta. Diagnostyka chorób w oparciu o jej molekularny obraz otwiera drogę dla przełomowych leków i terapii. Personalizację określa się na poziomie:
   - Organizmu
   - Komórki
   - Celu molekularnego (receptora)

   Obecnie termin ten odnosi się głównie do ostatniego, najbardziej złożonego poziomu (receptora), który jest ściśle opisany przez genom danego organizmu.

3. **Poziom innowacyjności "blue ocean"** (strategia błękitnego oceanu) – polega na tworzeniu nowych rynków gdzie nie ma konkurencji. Podstawą jest kreowanie przez leki wolnej i niezagospodarowanej przestrzeni rynkowej skoncentrowanej na długofalowej, ściśle określonej wizji działania celem zbudowania świadomości konsumenckiej [130].

4. **Poziom innowacyjności "red ocean"** (strategia czerwonego oceanu) – obejmuje istniejące rynki, które działają według ustalonych zasad. Miarą innowacyjności czerwonego oceanu jest konkurencyjność nowego leku względem istniejących farmaceutyków, jego siła przebicia oraz pozycja na rynku.

Uśrednione wyniki ekspertyzy leków NME z lat 2003-2012 przedstawiono na rysunku 9.4 oraz w tabeli 9.1.



Rysunek 9.4 Parametry określające translacyjność (oś rzędnych, linie przerywane ze znacznikiem) w jednostce czasu (oś odciętych). Kolumny przedstawiają średnią wartość translacyjności obliczoną dla danego roku.

Tabela 9.1 Średnie wartości parametrów translacyjności dla leków NME.

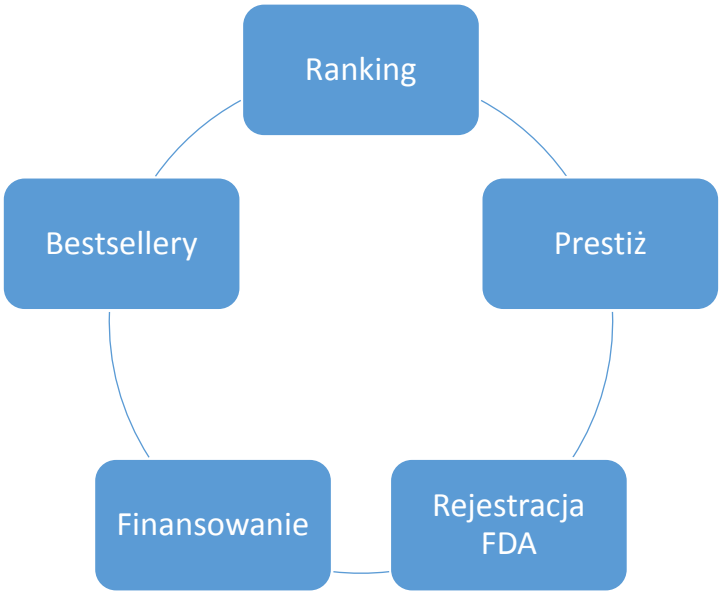
Rok	*Selektywność*	*Personalizacja*	*Blue ocean*	*Red ocean*	*Translacyjność*
2003	1,52	-1,85	-0,04	0,85	**0,33**
2004	1,14	-1,72	-0,61	0,50	**-0,03**
2005	1,50	-2,00	0,00	1,30	**0,70**
2006	1,45	-1,82	0,18	1,50	**0,86**
2007	1,47	-1,11	-0,05	1,21	**0,37**
2008	1,33	-1,83	-0,21	1,13	**0,17**
2009	1,15	-2,00	-0,58	1,00	**0,31**
2010	1,76	-0,95	-0,24	1,24	**0,81**
2011	1,60	-0,80	0,43	0,87	**0,70**
2012	1,31	-0,56	-0,10	0,87	**0,05**

W populacji NME 2003-2012 obserwujemy, że leki wykazują wysoką selektywność (w przedziale od +1 do +2). Zainteresowanie lekami spersonalizowanymi w badanym okresie wyraźnie wzrosło. Jednakże analiza dostępnych na rynku leków wykazała, że obecność farmaceutyków wysoce spersonalizowanych jest znikoma. Nie mniej jednak coraz więcej nowych leków charakteryzuje się dobrym profilem personalizacji.

Pomimo dużego wysiłku badawczego nie udaje się jednak odkryć znaczących zależności w oparciu o wykonane analizy eksperckie.

10 Wieloczynnikowa analiza rynkowego sukcesu leków

W rozdziale 8 niniejszej rozprawy przedstawiono współzależność ekonomicznych parametrów z rangą/prestiżem z jednej strony a efektywnością pracy naukowej z drugiej. Posługując się analogią postawiono następującą hipotezę badawczą: "Efektywność farmaceutyku pozostaje w relacji do jego sukcesu ekonomicznego?"



Rysunek 10.1 Kryteria wykorzystywane we wstępnym etapie skriningu baz leków.

Połączenie kryteriów przedstawionych na wykresie 10.1 wydaje się być dobrym wyznacznikiem poszukiwania najbardziej pożądanych i dochodowych struktur molekularnych.

Za źródło danych posłużyła internetowa baza Drugsite Trust (drugs.com) [131], która jest listą bestselerów wg danych IMS Health (tabela 10.1). Amerykańskie przedsiębiorstwo IMS jest światowym liderem usług technologicznych i rynkowych, związanych z sektorem ochrony zdrowia. Kadra firmy liczy ponad 15000 pracowników w tym 7000 ekspertów, z czego 1200 z branży informatycznej ochrony zdrowia (ang. healthcare informatics experts). Ponadto korporacja posiada nowoczesne serwerownie, systemy bazodanowe oraz biura w ponad 100 krajach.

Tabela 10.1 Struktura danych bazy Drugs.com (top 100 i top 200)

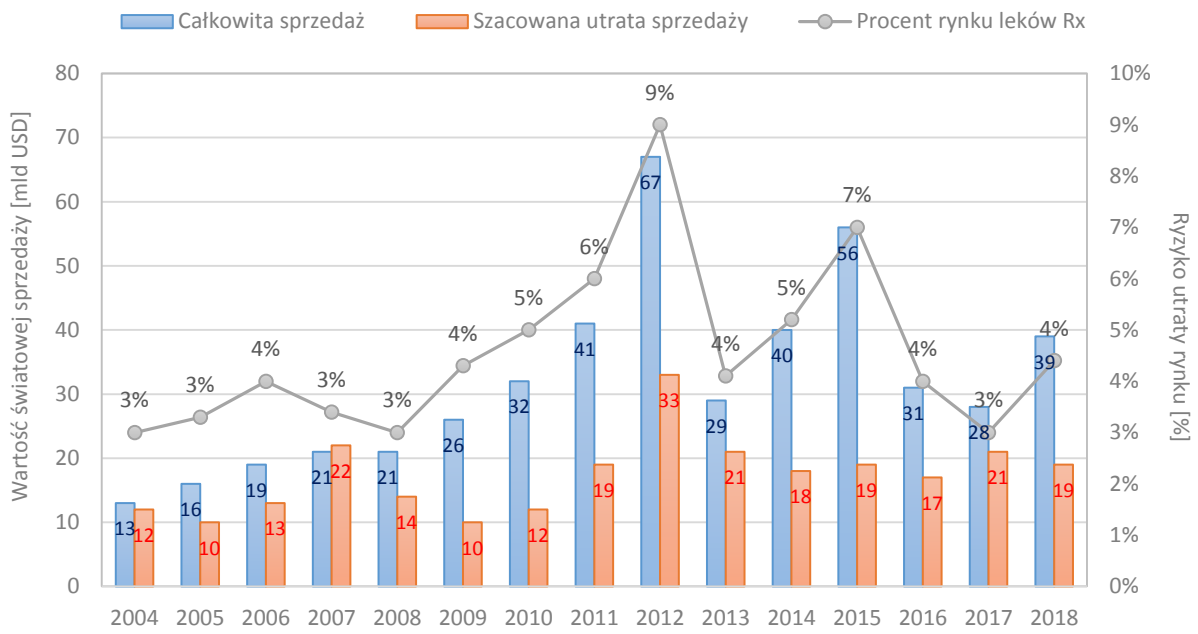
Źródło danych	Baza danych	Parametry	Badany okres	Liczba leków na liście
IMS Health Inc.	Drugs.com	Wartość sprzedaży Nazwa leku	2003-2010	Top 100
			2011-2013	Top 200

Ogólnodostępna baza danych oraz jej struktura (tabelaryczne zestawienie danych) umożliwiła w prosty sposób wyekstrahować dane, a następnie zapisać je w postaci pliku na lokalnej stacji roboczej.

W następnych krokach dane zostały poddane grupowaniu, hierarchizacji, transformacji oraz szerokiej analizie statystycznej. Wyniki badań opisano w kolejnych podrozdziałach.

10.1 Innowacja jest tam, gdzie młodszy wygrywa

Innowacja jest pojęciem immamentnie związanym z przemysłem farmaceutycznym. Rotacja ochrony patentowej wymaga od firm farmaceutycznych ciągłego poszukiwania nowych oryginalnych leków. Pozostaje problem związany z rzetelną miarą wieku leku. W analizie danych regulatora FDA badano datę pierwszej publikacji lub patentu opisującego daną molekułę i/lub substancję chemiczną. Jednak precyzyjność takich danych jest na bardzo niskim poziomie. Przemysł farmaceutyczny dąży bowiem do ukrycia wszelkich danych dotyczących nowych projektów (informacje poufne).



Rysunek 10.2 Szacowana wartość rynku oraz jego utrata w wyniku upływu czasu ochrony patentowej [132].

Nowa prosta idea pomiaru wieku leku, którą opracowano dla potrzeb analiz przeprowadzonych w ramach niniejszej pracy, polegała na porównaniu daty przypisanej przez cenzurę rejestracji i przez regulatora FDA.

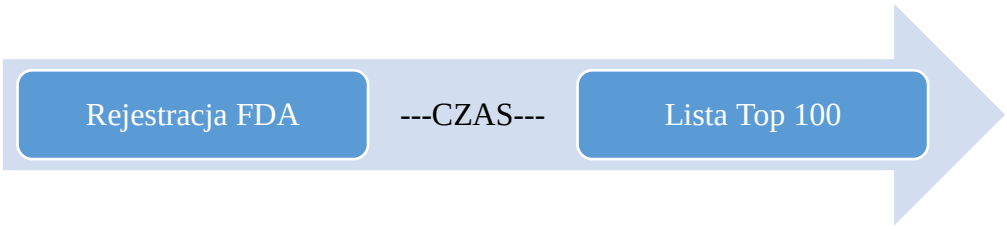
Tak więc cyfrową bibliotekę danych (zawierających listę najlepiej sprzedających się leków w okresie 2003-2013 wraz z wolumenami sprzedaży) uzupełniono o kilka dodatkowych parametrów:

- **Data rejestracji FDA** – termin nadania pozwolenia wprowadzenia leku na rynek. Od tego dnia lek może zacząć przynosić korzyści finansowe patronatowi. Informacje pozyskano z oficjalnej bazy FDA [64].
- **Reprezentacja liniowa SMILES** – strukturę leków przedstawiono w postaci kodu SMILES. Posłużył on do określenia oraz analizy właściwości fizyko-chemicznych. W tym celu wykorzystano metody programistyczne.
  Kody SMILES pobrano z baz:
  1. Europejski Instytut Bioinformatyczny (ChEMBL) [113]
  2. Bio- cheminformatyczna baza DrugBank [133]

Tabela 10.2 Zestawienie dziesięciu najlepiej sprzedających się leków w okresie 2003-2013.

Nr	Nazwa Handlowa	Substancja Aktywna	Firma	Akceptacja FDA	Śr. roczna wartość sprzedaży w okresie 2003-2013 (mld USD)
1	Lipitor	atorvastatin	Pfizer Inc	1996	6,06
2	Nexium	esomeprazole	AstraZeneca	2001	4,56
3	Remicade	infliximab	Janssen	1998	3,69
4	Advair Diskus	fluticasone salmeterol	GlaxoSmithKline	2000	3,51
5	Plavix	clopidogrel	Bristol-Myers	1997	3,44
6	Neulasta	pegfilgrastim	Amgen	2002	3,34
7	Prevacid	lansoprazole	Takeda	1995	3,22
8	Rituxan	Rituximab	Genentech	1997	3,08
9	Zocor	simvastatin	Merck	1991	2,93
10	Abilify	aripiprazole	Bristol-Myers	2002	2,85

Analizując listę rankingową bestselerów przedstawioną w tabeli 10.2 zauważyć można, że przeważającą część stanowią leki z lat 90-tych.



Rysunek 10.3 Definicja wieku leku jako czasu, który upłynął od dnia rejestracji FDA do chwili debiutu na liście top 100 bestselerów.

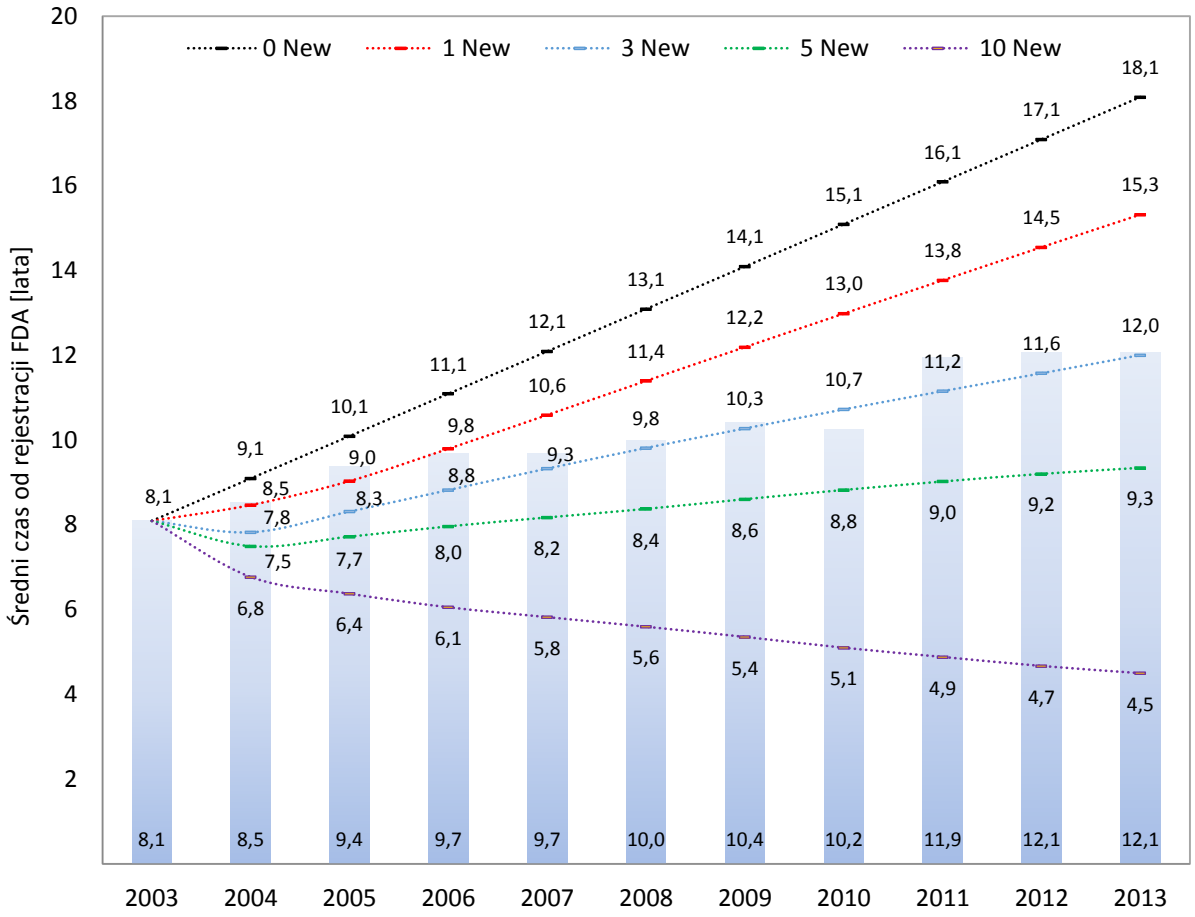
Kinch oraz współpracownicy przeprowadzili analizę wszystkich zatwierdzonych leków FDA [63]. Z danych wynika, iż agencja FDA od 1827 roku (pozyskane z danych archiwalnych) aż do 2013 zatwierdziła 1453 nowych jednostek molekularnych (ang. new molecular entities, NME). W latach 50-tych średnio 15 nowych leków rocznie wchodziło na rynek. Trend ten pozostał niezmienny do lat 70-tych. W następnej dekadzie średnia liczba NME wzrosła, oscylując w granicach 25-30 leków rocznie i na tym poziomie utrzymuje się po dziś dzień. Maksymalny roczny wzrost został odnotowany w okolicach połowy lat 90-tych (55 NME zostało dopuszczonych do obrotu). Po tym czasie nastąpił gwałtowny spadek do poziomu z poprzednich dekad (70' - 80'). Bazując na zbiorze NME, Kinch i współautorzy opisali pewne

zaobserwowane trendy. W szczególności zwrócili uwagę na dużych graczy rynku farmaceutycznego, którzy kontrolują większość leków dopuszczonych przez FDA. W ostatnich latach dostrzegli również wzrost liczby NME kontrolowanych przez organizacje farmaceutyczno-marketingowe (np. Pfizer), które w odróżnieniu od jednostek innowacyjnych dokonujących szeregu przejęć i akwizycji mających na celu zdobycie praw do leku. Dla przykładu jednym z największych transakcji Pfizera było przejęcie firmy Wyeth za kwotę około 86 mld USD. Badaną populację FDA podzielili na podgrupy ze względu na klasy leków: HIV/AIDS i inne infekcyjne [134], onkologiczne [135] oraz antybiotyki [67].

W niniejszej pracy przebadano inną subpopulację leków FDA, tzn. listę bestselerów na rynku amerykańskim (top 100). Zakładając, że sukces rynkowy jest miernikiem doskonałości farmaceutycznej, to tutaj szukać należy idealnego leku. Tak więc podejście takie jest swego rodzaju ewolucyjnym modelem idealnego leku.

W ten sposób możliwym stało się oszacowanie produktywności oraz konkurencyjności na rynku farmaceutyków. Jeżeli nowe leki szybciej osiągają sukces rynkowy, wówczas obserwujemy pozytywny efekt produktywności i tym samym wyraźną przewagę skuteczności nowszej generacji leków. Analogicznie spadek produktywności świadczy o większej trudności w wypieraniu starych leków przez nowe odpowiedniki.

W pracy posłużono się terminem "wiek leku", który jest łatwo mierzalnym parametrem numerycznym. Potrzeba opracowania tego parametru nasunęła nam wiele trudności związanych z rzetelnością danych (m.in. pierwszych informacji dot. odkrycia danej cząsteczki czy wzmianki o pierwszych patentach). Zadecydowało to o wyborze najbardziej wiarygodnej metody oszacowywania wieku leku (jako czasu, który upłynął od daty rejestracji FDA – rysunek 10.3). Za datę debiutu przyjmuje się rok występowania leku na liście bestselerów 2003-2013 w zależności od badanego okresu. Natomiast data rejestracji FDA jest stała i niezmienna dla każdego z badanego leku, tzw. data urodzin. W ten sposób badając każdy z leków oddzielnie, obliczono dokładnie ile czasu lek potrzebował od akceptacji FDA, żeby znaleźć się na badanej liście bestselerów. Następnie dla każdej grupy najbardziej dochodowych molekuł obliczono średni wiek. Porównując wszystkie grupy zauważono trend, korelację pomiędzy dwoma zmiennymi (wiekiem leku a skalą czasu, R = 0,96), świadczącego o powolnym starzeniu się farmaceutyków (rysunek 10.4).



Rysunek 10.4 Histogram przedstawia średni czas jaki upłynął od momentu rejestracji FDA (wiek leku) najlepiej sprzedających się farmaceutyków. Linie przerywane przedstawiają hipotetyczny scenariusz zmieniającego się wieku w wyniku wymiany kolejno: 1, 2, 3, 5 i 10 nowych leków z listy top 100 [114].

Z przeprowadzonej analizy wynika że 49,6% bestselerów zostało zatwierdzonych przed 2000 rokiem. Pozostałe 50,4% pojawiło się po roku 2000. Statystycznie średni wiek leku wyniósł 10,2 lat. Najmniejsza wartość wynosiła 8,1 lat w 2003 roku, po czym stopniowo wzrastała do 12,1 lat (2013). Średni wiek leku rósł względem skali czasu (współczynnik korelacji R wynosi 0,96) co oznacza, że farmaceutyki na liście top 100 powoli się starzeją. Na podstawie obserwowanego trendu można stwierdzić, że przemysł farmaceutyczny nie oferuje takiej liczby innowacyjnych leków, które mogłyby z powodzeniem zastąpić związki biologicznie aktywne starej generacji.

Celem lepszego zrozumienia badanego problemu na wykresie 10.4 przedstawiono linie ilustrujące możliwe scenariusze:

- 0 leków (linia czarna) – scenariusz, który zakłada stagnację. Od 2003 roku aż do 2013 żaden z leków nie opuszczałby listy. W tej sytuacji średnia wieku zwiększyłaby się o 10 lat.
- 1 lek (linia czerwona) – projekcja, w której corocznie, najstarszy z leków zostałby zastąpiony najmłodszym lekiem (0 lat). Grupa w badanym okresie postarzałaby się o 7,2 lat.
- 3 leków (linia niebieska) – przedstawia model, w którym trzy najstarsze leki z grupy zostałyby co roku zastąpione trzema najmłodszymi. Bestselery w przeciągu 10 lat postarzałyby się o 3,9 lat.
- 5 leków (linia zielona) – w tym przypadku pięć najstarszych w liście zostałoby wymienianych na pięć najmłodszych. W wyniku takiej wymiany średnia wieku populacji najlepiej sprzedających się medykamentów od 2003 do 2013 wzrosłaby zaledwie o 1,2 lat.
- 10 leków (linia fioletowa) – analogicznie dziesięć najstarszych ustępowałoby miejsca takiej samej ilości nowych leków. W badanym okresie średnia wieku spadłaby o 3,6 lat.

Dokładna analiza wieku pozwoliła stwierdzić, że w ostatnim dziesięcioleciu nastąpił spadek produktywności farmaceutyków. Potwierdza to również fakt, że od 20 lat obserwujemy stały, systematyczny spadek zarejestrowanych leków (rozdział 3 – wykres 3.11). Niska wydajność spowodowała ekonomiczne problemy w sektorze farmaceutycznym (np. zamknięcia zakładów, zwolnienia kardy), jak również redukcji projektów badawczych. W rozdziale 3.5 pt.: "Rozwój przemysłu farmaceutycznego – koncepcja fast-followers i leki me-too" wspomniano, iż pomimo wielu teorii wyjaśniających przyczynę obecnej sytuacji w przemyśle farmaceutycznym, najbardziej prawdopodobną wydaje się być hipoteza Sams-Dodda [42]. Tradycyjne metody poszukiwania leków w oparciu o właściwości fizyko-chemiczne skupiały się wyłącznie na ogólnym efekcie terapeutycznym. Następnie, rewolucja technologiczna skierowała sektor farmaceutyczny w kierunku badań zależności ligand-receptor. Spowodowało to wzrost kosztów, złożoności oraz trudności nowych projektów badawczych.

Obok popularnej teorii Sams-Dodda istnieje druga hipoteza, zaproponowana przez Pammoliego. Badacz przeanalizował bazę 28000 związków, które obecnie są obiektem zainteresowań sektora R&D. Pammoli stwierdził wzmożoną intensywność poszukiwania potencjalnych leków w miejscach o wyższym ryzyku niepowodzenia. Brak informacji

o mechanizmach działania na nowe cele molekularne oraz wybór ryzykownych strategii innowacyjnych były bezpośrednią przyczyną kryzysu [53].

Z ekonomicznej perspektywy kapitał jest miarą możliwości, natomiast zysk determinuje sukces produktu na rynku. Przemysł farmaceutyczny, obok petrochemii i finansów, jest jednym z najbardziej dochodowych biznesów. Odgrywa też ogromną rolę w regulacjach prawnych, etycznych i podatkowych [60]. Z tego powodu nowe przedsiębiorstwa napotykają wiele przeszkód prawnych i regulacyjnych na drodze do stania się producentem leków. To oznacza, że maksymalizacja zysków na rynku jest głównym celem firm farmaceutycznych. Z tego powodu grupę najlepiej sprzedających się leków można traktować jako zbiór związków wykazujących najbardziej pożądane cechy.

Z drugiej strony powstaje pytanie czy istnieje zróżnicowanie zachowań rynkowych związane z typem leku? Z tego powodu zbiór bestsellerów podzielono na podgrupy ze względu na docelowe miejsce działania. Umożliwiło to dokładniejsze poznanie zachodzących zmian w sektorze farmaceutycznym na przestrzeni ostatnich lat.

Wyodrębniono dziewięć następujących podgrup:

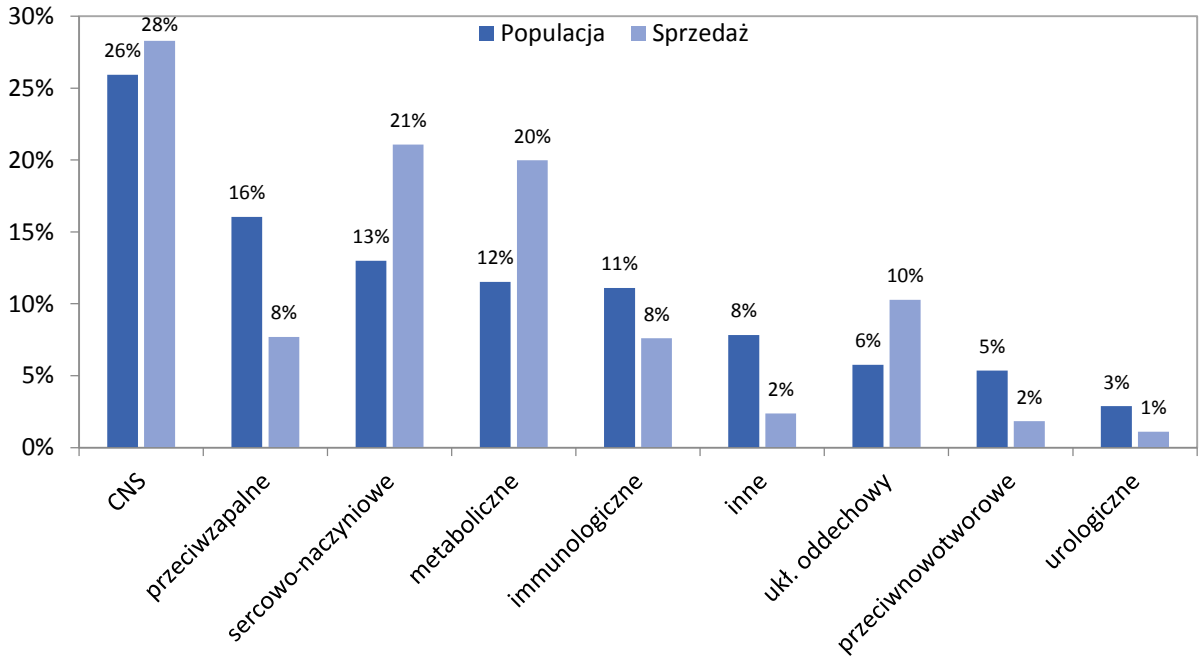
- centralny układ nerwowy (26%, ang. central nervous system, CNS)
- przeciwzapalne (16%, ang. anti-infective)
- sercowo-naczyniowe (13%, ang. cardiovascular)
- metaboliczne (12%, ang. metabolic)
- immunologiczne (11%, ang. immune)
- układ oddechowy (6%, ang. respiratory)
- urologiczne (5%, ang. urologic)
- przeciwnowotworowe (3%, ang. anticancer)
- inne (8%, ang. other).

Procenty w nawiasach informują o liczbie leków w stosunku do całej populacji. Podział uwzględnia również nazwy angielskie.

W kolejnym etapie wyodrębnione grupy sklasyfikowano pod względem woluminu sprzedaży przyjmujące następujące wartości:

- centralny układ nerwowy (28%)
- przeciwzapalne (8%)
- sercowo-naczyniowe (21%)
- metaboliczne (20%)
- immunologiczne (8%)
- układ oddechowy (10%)
- urologiczne (1%)
- przeciwnowotworowe (2%)
- inne (2%).

Otrzymane wyniki przedstawiono na rysunku 10.5. Leki działające na centralny układ nerwowy są zwycięzcami zarówno pod względem liczebności (stanowią 26% całej populacji) jak również zysków (osiągnęły 28% całego woluminu sprzedaży) jakie przynoszą w przeliczeniu na jeden lek [114].



Rysunek 10.5 Procentowy udział podgrupy leków w stosunku do całej populacji oraz ich sprzedaży.

W tabeli 10.3 zestawiono najważniejsze średnie wartości parametrów dla poszczególnych podgrup farmaceutyków w zależności od czasu. Zauważono, że leki CNS są zwycięzcami rankingu liczebności, przychodów oraz charakteryzują się najniższą średnią masą molekularną. Stanowią również jedną z trzech najstarszych populacji (średni wiek tej grupy to 10,6 lat). Zbiór leków wykazujących działanie sercowo-naczyniowe liczy 11,3 lat, natomiast pozostałe niesklasyfikowane farmaceutyki osiągają średni wiek liczący 14,6 lat.

Tabela 10.3 Średnie wartości parametrów (wiek leku, masa molekularna, clog P oraz TPSA) dla odpowiednich klas leków od czasu.

Klasa leku	Parametr	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
CNS (26%)	Wiek	9,1	8,8	9,3	9,4	9,9	10,4	10,8	8,7	13,9	14,4	15,1
	MW	293	286	289	284	279	285	295	301	299	303	303
	clog P	3,2	3,2	3,2	3,1	3,0	3,0	3,0	3,0	3,0	3,1	3,0
	TPSA	48	48	50	51	51	53	57	52	52	53	56
Przeciwzapalne (16%)	Wiek	7,9	8,1	10,0	8,7	8,3	9,1	9,6	12,0	9,6	9,9	11,3
	MW	435	389	451	370	389	388	387	404	481	516	502
	clog P	2,5	2,5	2,3	2,9	2,6	2,9	2,5	2,3	3,0	3,8	3,4
	TPSA	119	118	129	115	123	109	113	125	129	135	144
Sercowo naczyniowe (13%)	Wiek	8,9	9,7	10,5	11,1	10,8	10,1	11,1	10,8	14,6	14,5	13,9
	MW	392	418	416	414	424	454	454	448	441	458	448
	clog P	3,4	3,7	3,7	3,6	3,7	4,0	4,0	4,0	3,9	4,2	3,6
	TPSA	89	93	93	92	96	107	107	104	100	105	110
Metaboliczne (12%)	Wiek	5,0	5,0	6,1	6,6	7,0	8,2	8,6	8,9	10,4	10,4	11,6
	MW	420	387	387	387	389	389	389	395	334	393	399
	clog P	4,2	3,6	3,7	3,7	3,6	3,6	3,6	3,6	3,2	3,6	3,7
	TPSA	95	91	87	87	86	86	86	84	79	82	80
Immunologiczne (11%)	Wiek	6,5	7,4	8,3	9,4	10,4	10,5	10,9	10,7	13,2	12,6	11,0
	MW	443	534	494	537	537	586	740	862	862	677	627
	clog P	4,1	4,4	4,2	3,7	3,7	4,0	4,0	4,5	4,5	4,3	3,8
	TPSA	76	99	90	105	105	122	161	186	186	146	139
Inne (8%)	Wiek	14,2	15,7	16,2	19,2	16,0	13,5	14,5	13,8	9,8	-	10,9
	MW	435	322	322	322	299	355	355	370	272	-	357
	clog P	2,1	1,1	1,1	1,1	1,2	2,7	3,3	4,1	-1,1	-	6,5
	TPSA	132	145	145	145	126	102	87	71	173	-	120
Układu oddechowego (6%)	Wiek	5,5	6,5	5,9	6,2	6,6	7,9	8,0	8,9	9,0	10,2	8,7
	MW	440	506	470	477	430	398	430	392	430	398	360
	clog P	3,4	5,4	4,7	4,6	4,0	3,6	4,0	3,8	4,0	3,6	2,6
	TPSA	110	98	101	103	97	93	97	93	97	93	92
	Wiek	4,5	7,1	9,0	10,0	9,6	10,6	11,8	12,8	10,0	12,0	11,7

Przeciwnowo-tworowe (5%)	MW	474	474	383	383	420	420	386	386	435	430	421
	clog P	6,0	6,0	4,5	4,5	4,5	4,5	4,1	4,1	2,8	2,6	3,7
	TPSA	98	98	88	88	88	88	85	85	115	115	106
Urologiczne (3%)	Wiek	4,9	5,9	6,9	7,9	8,9	10,0	9,0	8,3	7,1	6,7	7,7
	MW	359	359	360	360	360	360	361	361	362	362	362
	clog P	4,3	4,3	4,8	4,8	4,8	4,8	4,4	4,4	3,7	3,7	3,7
	TPSA	61	61	66	66	66	66	55	55	33	33	33

W dalszej części zliczono częstotliwość występowania leków na liście bestselerów według roku akceptacji FDA (dane zestawiono w tabeli 10.4). Zauważono, że molekuły z okresu 1995-2004 wykazują największą liczebność, wynoszącą odpowiednio: 158 (65%) unikatowych farmaceutyków oraz 827 (75%) wielokrotnie występujących struktur na liście top 100.

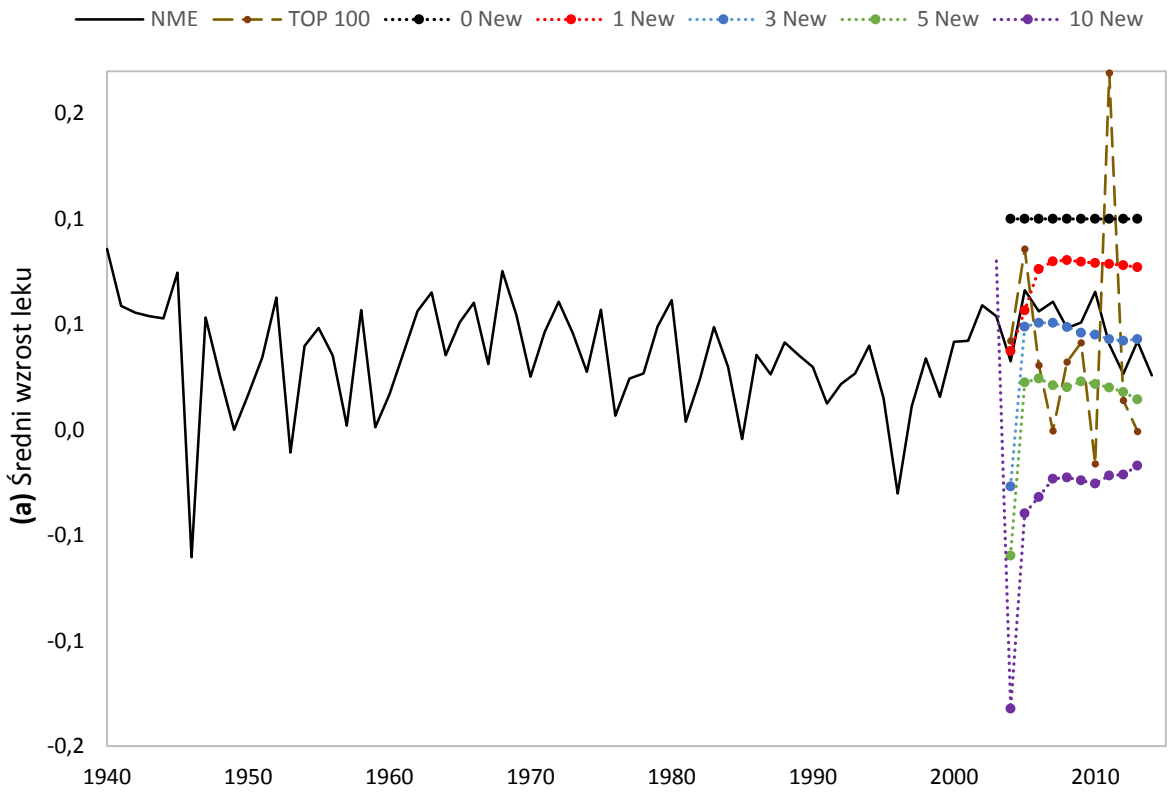
Tabela 10.4 Sprzedaż leków na liście top 100 wg roku rejestracji. Liczby w nawiasach dotyczą danych nieskumulowanych tzn. leków, które występowały tylko raz na liście 2003-2013.

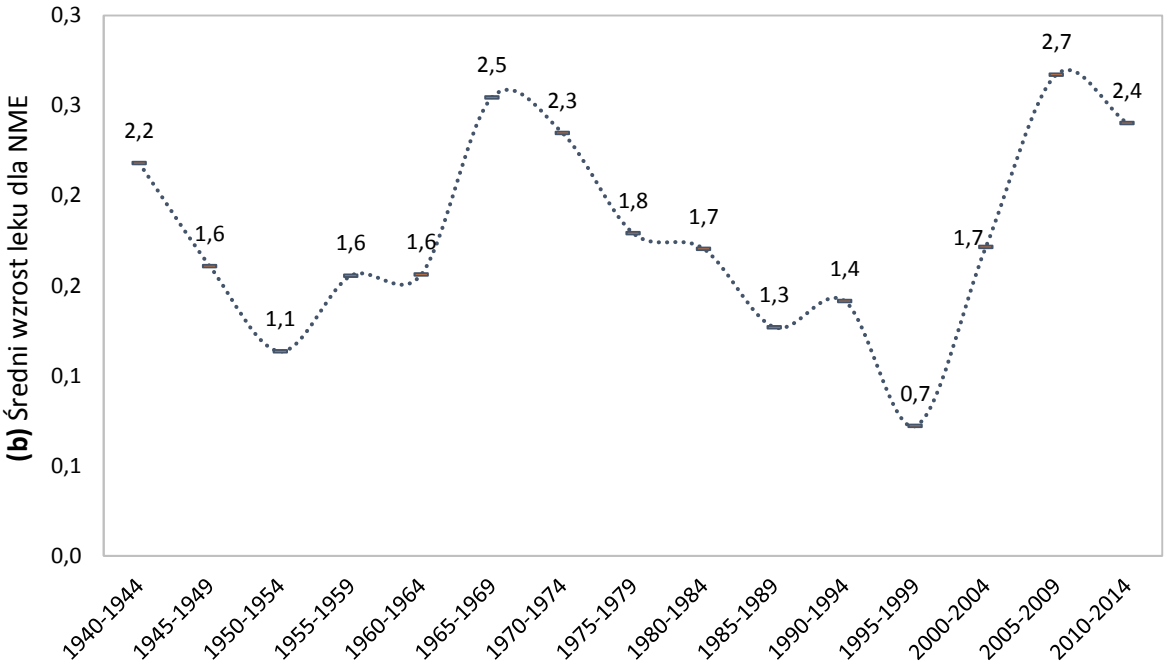
Rok	Liczba leków	Procent	Rok	Liczba leków	Procent
1942	8 (1)	0,7% (0,4%)	**1997**	96 (16)	8,7% (6,6%)
1955	2 (1)	0,2% (0,4%)	**1998**	92 (18)	8,4% (7,4%)
1960	1 (1)	0,1% (0,4%)	**1999**	54 (10)	4,9% (4,1%)
1962	7 (1)	0,6% (0,4%)	**2000**	92 (19)	8,4% (7,8%)
1982	3 (1)	0,3% (0,4%)	**2001**	77 (21)	7,0% (8,6%)
1983	6 (1)	0,5% (0,4%)	**2002**	91 (16)	8,3% (6,6%)
1985	1 (1)	0,1% (0,4%)	**2003**	54 (10)	4,9% (4,1%)
1987	1 (1)	0,1% (0,4%)	**2004**	90 (17)	8,2% (7,0%)
1989	11 (3)	1,0% (1,2%)	**2005**	20 (7)	1,8% (2,9%)
1990	5 (2)	0,5% (0,8%)	**2006**	37 (11)	3,4% (4,5%)
1991	34 (11)	3,1% (4,5%)	**2007**	21 (5)	1,9% (2,1%)
1992	37 (9)	3,34% (3,7%)	**2008**	5 (3)	0,5% (1,2%)
1993	29 (8)	2,6% (3,3%)	**2009**	5 (4)	0,5% (1,6%)
1994	19 (4)	1,7% (1,6%)	**2010**	10 (5)	0,9% (2,1%)
1995	70 (11)	6,4% (4,5%)	**2011**	6 (4)	0,5% (1,6%)
1996	111 (20)	10,1% (8,2%)	**2013**	1 (1)	0,1% (0,4%)

Badanie wieloczynnikowych zmian i trendów w określonych ramach czasowych dla danej populacji jest skomplikowanym procesem. W celu lepszego zrozumienia dynamiki zmian

populacji i subpopulacji bestselerów przeprowadzono badania referencyjne. Korzystając z internetowej bazy FDA "Orange book" [64], zebrano podstawowe informacje na temat całej populacji rejestracji FDA (1939 – 2014) nowych jednostek molekularnych (ang. new molecular entities, NME). W skriningu odrzucono wszystkie wycofane leki (tzw. "withdrawals"). Następnie dla badanej populacji wykonano analogiczną analizę wieku, którą przedstawiono na rysunku 10.6a. Średni wiek leku populacji NME wraz z upływem czasu stopniowo wzrastał. Ponadto średni wiek całej populacji wynoszący 13,7 lat jest zbliżony do średniego wieku bestselerów (12,1 lat).

Zestawienie całej populacji FDA z listą top 100 bestselerów (wykres 10.6a) ułatwiło zidentyfikowanie potencjalnych zależności pomiędzy zbiorami.





Rysunek 10.6 Średni wzrost wieku leku dla NME 1939-2014 (czarna ciągła linia), bestselerów (brązowa przerywana linia) oraz pięciu hipotetycznych scenariuszy listy top 100 (kolorowe kropkowane linie) w rocznym interwale czasowym (a) oraz NME w pięcioletnim interwale (b).

Dodatkowo przeanalizowano dynamikę zmian wieku leku (dynamika leku), która określa średnią zmianę wieku leku w badanym okresie (rocznie – rysunek 10.6a lub 5-letni okres – rysunek 10.6b). Dynamika zmiany wieku została obliczona jako różnica pomiędzy rokiem poprzednim a następnym (dla roku 1939 przyjęto wartość równą zero). Na rysunku 10.6a zaobserwowano podobny średni poziom dynamiki pomiędzy NME (czarna ciągła linia) oraz top 100 (brązowa przerywana linia). Natomiast wyraźne różnice widoczne są w wahaniach dynamiki wieku pomiędzy kolejnymi latami obu grup.

Analizując wykres 10.6a zaobserwowano chaotycznie zmieniającą się dynamikę leku. Sklasteryzowanie danych (ang. data binding) w pięcioletnie okresy ułatwiło interpretację zmian zachodzących na przestrzeni czasu oraz prognozowanie oczekiwanych trendów (rysunek 10.6b). Najdłuższy średni spadek wieku odnotowano w latach 1970-1999. Analogicznie najdłuższy średni wzrost w latach 1950-1970 oraz 2000-2009. Ciekawym wydaje się zmiana tendencji w ostatnim okresie 2010-2014. Czy zatem zmiana trendu w ostatnim okresie jest oznaką innowacyjności w technologii farmaceutycznej [53]? Wartość rynku jest jednym z najważniejszych czynników determinujących kierunki rozwoju farmaceutyków.

Na rysunku 10.6a zestawiono porównanie dynamiki leku listy top 100 (brązowa, przerywana linia) wraz z pięcioma różnymi scenariuszami (linie kropkowane) z całą populacją NME (czarna, ciągła linia). Roczne zmiany dla listy top 100 są bardziej dynamiczne niż całej populacji. Badając trendy najbardziej odpowiadającego rzeczywistości scenariusza (zmiany 3 leków) zaobserwowano wyraźne podobieństwo ze zmianą dynamiki wieku całej populacji FDA. Z drugiej strony przedstawione w pracy badania wskazały, że średni wiek leków na rynku stale rośnie. Zatem wprowadzane na rynek nowe leki nie wykazują dostatecznej efektywności rynkowej by wyprzeć więcej niż 3% starszych farmaceutyków z listy bestsellerów.

W dalszej części niniejszej rozprawy opisano wyniki badań, które dotyczą wpływu parametrów fizyko-chemicznych na produktywność w sektorze farmaceutycznym.

10.2 Model slim pharmy

Projektowanie oparte o strukturę celu działania leku (ang. target-oriented drug discovery) jest głównym paradygmatem badawczym współczesnej chemii leków. Różnica w projektowaniu opartym o budowę topologiczną cząsteczki a strukturą receptora docelowego polega na maksymalizowaniu siły działania oraz selektywności poprzez dopasowanie do celu białkowego. Zmiana sposobu projektowania leków doprowadziła do zwiększenia złożoności struktur molekularnych. Porównując sukces wprowadzenia leku z trendami panującymi w chemii medycznej można wywnioskować, że tak zwana otyłość molekularna (ang. molecular obesity) jest głównym powodem eskalacji kosztów rozwoju farmaceutyków [8,12,136]. Na przykład średnie wartości MW (300-450 Da) i clog P (1,5-4,0) są niższe dla leków istniejących na rynku od potencjalnych kandydatów będących w fazie badawczej [136]. Dlatego tak zwana koncepcja "slim pharmy" opiera się na projektowaniu związków o niskiej złożoności [8]. W rodziale 8 przedstawiono dane, które świadczą o tym, że leki o niskiej złożoności dominują w populacji (liczbowo), natomiast te o większej złożoności zazwyczaj nieco lepiej się sprzedają (w przeliczeniu średniego przychodu na lek). Tak więc leki projektowane w koncepcji "slim pharmy" mają większa szansę rynkowego powodzenia, natomiast leki większe cząsteczkowo potencjalnie mogą przynosić wyższe dochody ze sprzedaży.

Dotychczas przedstawiono analizę trendów ekonomicznych, wieku oraz dynamiki zmiany wieku leku w funkcji czasu. Dodatkowe badanie zmiany masy molekularnej pozwoliło

zaobserwować tendencję tej zmiennej. Zatem w niniejszym rozdziale przedstawiono analizę wpływu zmiany właściwości fizyko-chemicznych ze szczególnym uwzględnieniem parametrów lekopodobieństwa (np. regułę Lipińskiego) leków FDA na sukces rynkowy.

Koncepcja lekopodobieństwa opiera się na bardzo redukcjonistycznych założeniach w porównaniu do metod projektowania leków zorientowanych na cel molekularny. Podstawowe parametry lekopodobieństwa zostały oszacowane na podstawie danych dla leków dostępnych komercyjnie [8,12,137]. Reguła pięciu Lipińskiego (ang. rule of five, Ro5) jest prawdopodobnie najlepiej znanym filtrem, bazująca na czterech deskryptorach molekularnych. Determinują one optymalne właściwości ADMET, takie jak:

- Masa molowa (MW) < 500 Da
- Współczynnik podziału oktanol:woda (log P) < 5
- Liczba donorów wiązania wodorowego (HBD) < 5
- Liczba akceptorów wiązania wodorowego (HBA) < 10

Badając populację leków zauważono, że współczynniki Ro5 są ze sobą wzajemnie skorelowane. W celu eliminacji efektu interkorelacji, Gleeson zdefiniował tak zwany "ADMET Score". Obliczenie wartości ADMET Score jest możliwe za pomocą następującego równania [20]:

$$ADMET\ Score = \frac{|2{,}5 - clog\ P|}{2{,}0} + \frac{|330 - MW^*|}{120}$$

*jeżeli MW < 330, do obliczeń należy przyjąć MW = 0.

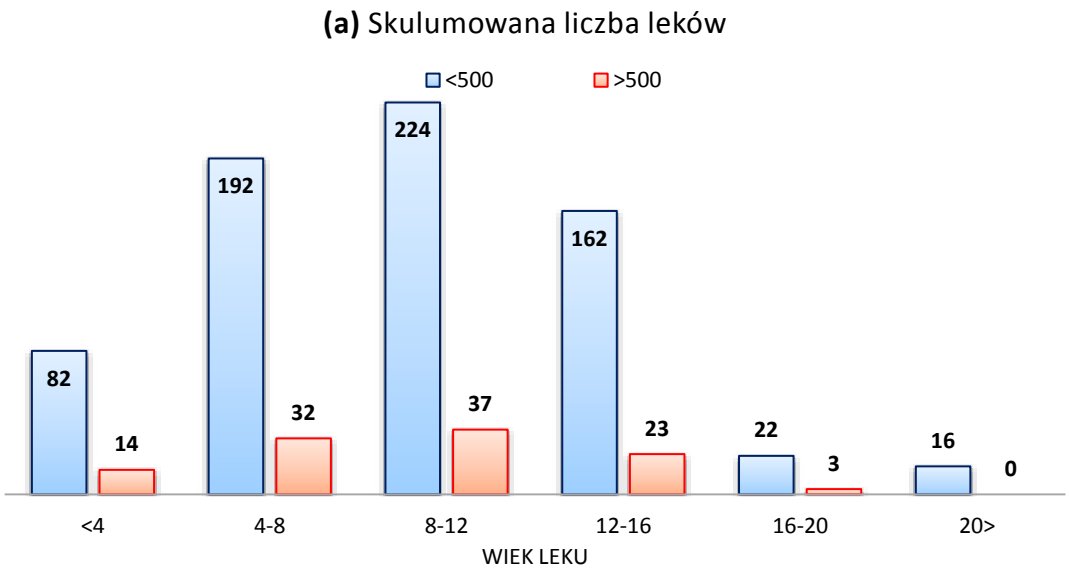
Wielokrotnie próbowano ustalić powiązania pomiędzy właściwościami fizyko-chemicznymi cząstek aktywnych a ich sukcesem terapeutycznym [138]. Określenie optymalnych parametrów (np. Ro5) cząsteczek aktywnych jest przydatne na wczesnym, strategicznym etapie filtrowania bibliotek związków [139]. Różnica pomiędzy projektowaniem zorientowanym na cel a koncepcją lekopodobieństwa polega na zależności, że większe korzyści terapeutyczne *in vitro* (zazwyczaj wiąże się z wyższą złożonością cząsteczki i MW) skutkują słabszymi parametrami ADMET [20]. Podobną zależność między złożonością związku a jego entalpią i entropią obserwuje się w procesach optymalizacji *in silico* [12]. Zatem intuicyjnie farmaceutyki, które osiągnęły sukces kliniczny mogą stać się zbiorem związków o pożądanych

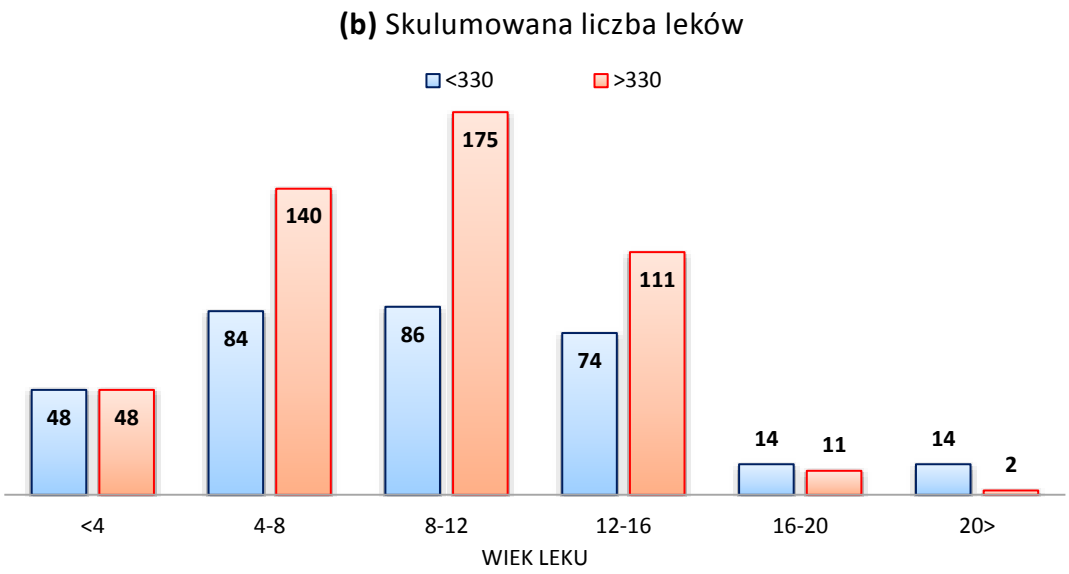
właściwościach. Z drugiej strony zbiór bestsellerów FDA może nieść w sobie informacje, które decydują o najwyższej efektywności i jakości farmakologicznej. W związku z tym w pracy badano współzależność pomiędzy stopniem złożoności struktury substancji chemicznej a jej sukcesem ekonomicznym.

Rysunek 10.7 Diagram porównujący średnią masę molekularną (niebieska linia) z średnimi wartościami clog P (zielona linia) (a) oraz średnimi wartościami ADMET Score (czerwona linia) (b) dla leków FDA (linia ciągła) i listy 100 najlepiej sprzedających się leków (kropki) [140].

Na wykresie 10.7a,b przedstawiono zależność średnich wartości deskryptorów określających parametry lekopodobieństwa (ADMET Score, MW, clog P) w całej populacji cząstek NME (FDA) oraz bestselerów w jednostce czasu. Średnie wartości MW i ADMET wyraźnie różnicują badane zbiory. Parametry te dla bestselerów przyjmują niższe wartości oraz charakteryzują się większą stabilnością w porównaniu do populacji NME. To z kolei świadczy o niższej złożoności oraz lepszych parametrach ADMET związków ze zbioru top 100. Średnia wartość ADMET Score wynosi kolejno 3,6 (NME – FDA) oraz 2,5 (top 100). Średnie wartości MW bestselerów zawierały się w przedziale od 356 Da do 412 Da. Dla zbioru top 100 w okresie od 2003 do 2006 obserwowano spadkową tendencję (z 379 do 356 Da), a w następnych dwóch latach wartość osiągnęła maksimum 410-412 Da.

Ewolucja lipofilowości w badanych populacjach wydaje się być stabilna pod względem dynamiki. W odróżnieniu od MW i ADMET, średnie wartości parametru clog P są na zbliżonym poziomie.





Rysunek 10.8 Skumulowana liczba leków z listy top 100 (a, b) dla dwóch grup sklasyfikowanych na podstawie kryterium Lipińskiego MW=500 Da (a) oraz Gleesona MW=330 Da (b) w zależności od zmieniającego się wieku leku.

Na rysunku 10.8a,b przedstawiono zależność pomiędzy wiekiem (lata) i stopniem złożoności struktury leku (MW) a sukcesem rynkowym zmierzonym liczebnością leków z listy top 100. Diagramy przedstawiają skumulowaną liczbę leków o niższej i wyżej złożoności, zgodnie z kryteriami Lipińskiego (500 Da) i Gleesona (330 Da).

Wykres 10.8a przedstawia wyraźną przewagę sukcesu rynkowego farmaceutyków o MW mniejszej niż 500 Da dla wszystkich klas wiekowych. Obniżenie kryterium MW do założeń Gleesona (330 Da) przedstawionych na rysunku 10.8b różnicuje badaną grupę co pozwala na wizualizację odkrytych informacji.

W przedziale wiekowym pomiędzy 4-16 lat zaobserwowano, że sukces rynkowy jest większy dla związków biologicznie aktywnych z przedziału 330-500 Da. Interesującym jest fakt, iż leki o masie mniejszej niż 330 Da dominują w populacji 16+ lat. Nie mniej jednak uśredniając wszystkie otrzymane wyniki, oszacowano optymalną strefę sukcesu, która znajduje się w przedziale 330-500 Da.

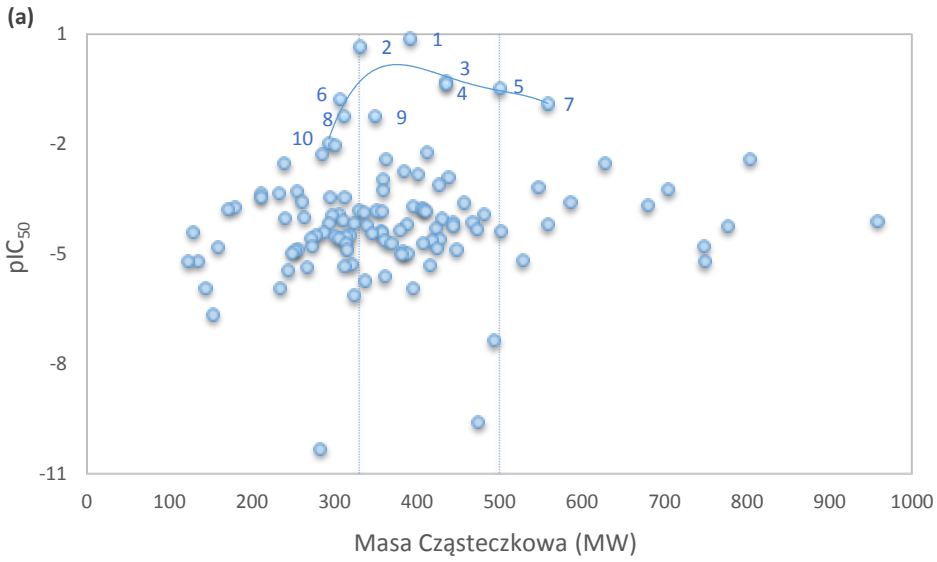
Powstaje jednak pytanie jak mierzyć aktywność związków wobec różnych leków molekularnych (celowanych). IC_{50} jest miarą efektywności czynnika hamującego daną aktywność biologiczną lub funkcje biochemiczne wyrażonego jako 50% wartości

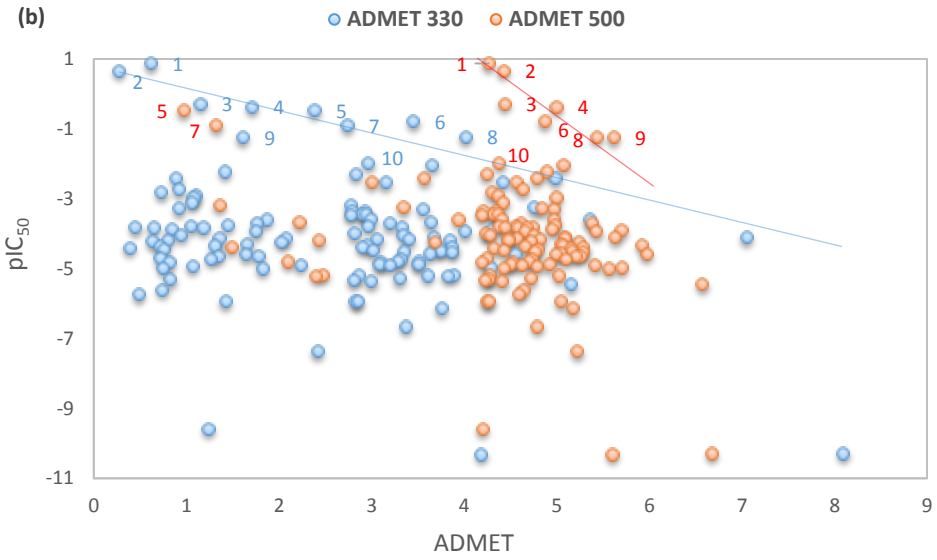
maksymalnego stężenia. W niniejszej pracy wartość IC_{50} wyrażono w skali pIC_{50} zgodnie ze wzorem:

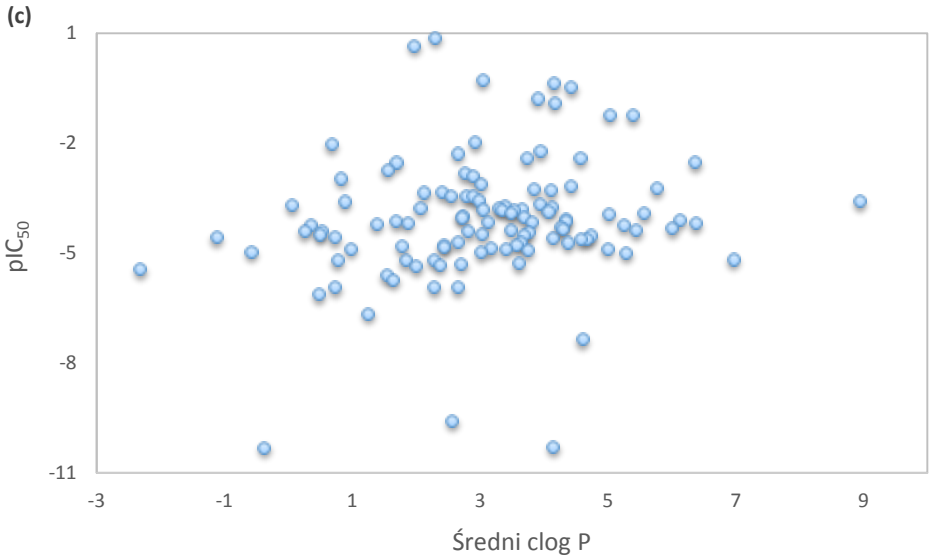
$$pIC_{50} = -log_{10}(IC_{50})$$

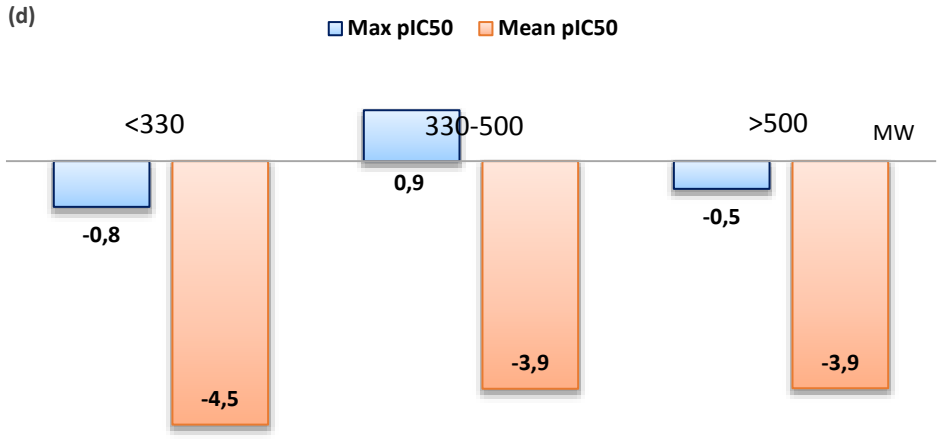
Analiza danych IC_{50} z bazy ChEMBL wykazuje duże wahania IC_{50} wobec tego samego celu działania dla wartości mierzonych w różnych laboratoriach. W niniejszej pracy jako miarę aktywności badanych związków chemicznych przyjęto aktywności polifarmakologiczne, tzn. mierzone względem licznych celów molekularnych. Uzasadnia to fakt, że w przypadku dużej liczby zróżnicowanych danych taka właśnie procedura wydaje się lepiej charakteryzować zbiór molekuł biologicznie aktywnych. Po pierwsze bowiem nie każdy lek wobec indywidualnego celu osiągnąć może optymalne dopasowanie, a po drugie nie każde oznaczenie analityczne *in vitro* jest w równym stopniu precyzyjne. Nie ma też możliwości pojedynczej precyzyjnej kontroli jakości dla tak wielkiej liczby danych. Według badań baza ChEMBL jest wiarygodnym repozytorium danych IC_{50} różnych związków chemicznych [141].

Z bazy wyselekcjonowano wszystkie dane dotyczące IC_{50} dla puli 171 leków, które stanowiły 82% całej populacji oraz uśredniono wszystkie dostępne w bazie wartości IC_{50} dla danej cząsteczki. Wyniki zilustrowano na wykresie 10.9a-d.




(b)
ADMET 330
ADMET 500
pIC50
ADMET


(c)
pIC50
Średni clog P



Rysunek 10.9 Maksymalne i średnie wartości pIC_{50} dla listy top 100 w zależności od masy molekularnej (a), wartości współczynników ADMET 330, ADMET 500 (b) i clog P (c). Porównanie maksymalnych i średnich wartości pIC_{50} poszczególnych grup (d).

Zależność pomiędzy pIC_{50} a MW dla top 100 leków przedstawiono na grafie 10.9a. Wykres uwzględnia podział na trzy grupy zdefiniowane przez próg MW (kryterium Lipinskiego oraz Gleesona). Zauważono, że wartości pIC_{50} maleją wraz ze wzrostem MW w przedziale 330-500 Da. Efekt ten jest szczególnie widoczny dla wartości maksymalnych. Zatem dziesięć związków o najwyższych wartościach pIC_{50} (zestawionych w tabeli 10.5) opisano funkcją wielomianową czwartego stopnia:

$$pIC_{50} = -4 \cdot 10^{-9} \cdot MW^4 + 7 \cdot 10^{-6} \cdot MW^3 - 4.8 \cdot 10^{-3} \cdot MW^2 + 1.5 \cdot MW - 167.8$$

Tabela 10.5 Lista dziesięciu leków o najwyższych wartościach pIC_{50}.

Nr	Nazwa Handlowa	Substancja Aktywna	IC_{50}	pIC_{50}	MW	ADMET 330	ADMET 500
1	Spiriva	Tiotropium bromide	0,13	0,87	392,51	0,62	4,26
2	Cipro	Ciprofloxacin	0,22	0,66	331,34	0,27	4,43
3	Xarelto	Rivaroxaban	1,90	-0,28	435,88	1,15	4,44
4	Diovan	Valsartan	2,39	-0,38	435,52	1,71	5,00
5	Flovent HFA	Fluticasone propionate	3,00	-0,48	500,57	2,39	0,97
6	Gilenya	Fingolimod	6,10	-0,79	307,47	3,45	4,87
7	Benicar	Olmesartan medoxomil	7,97	-0,90	558,59	2,74	1,32

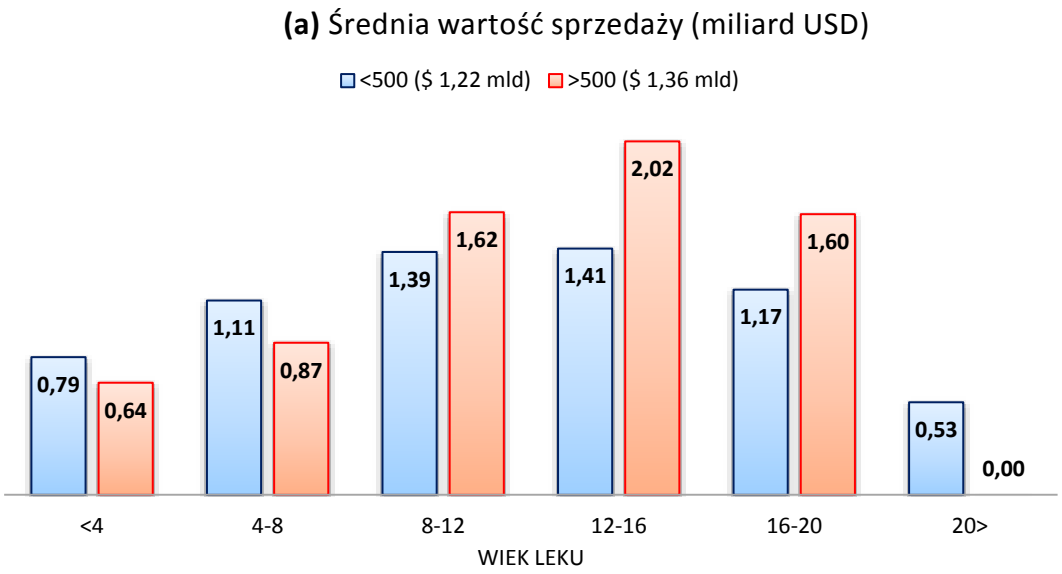
8	Detrol LA	Tolterodine	16,98	-1,23	311,46	4,02	5,43
9	Zytiga	Abiraterone acetate	17,50	-1,24	349,51	1,61	5,62
10	Arimidex	Anastrozole	94,52	-1,98	293,37	2,96	4,38

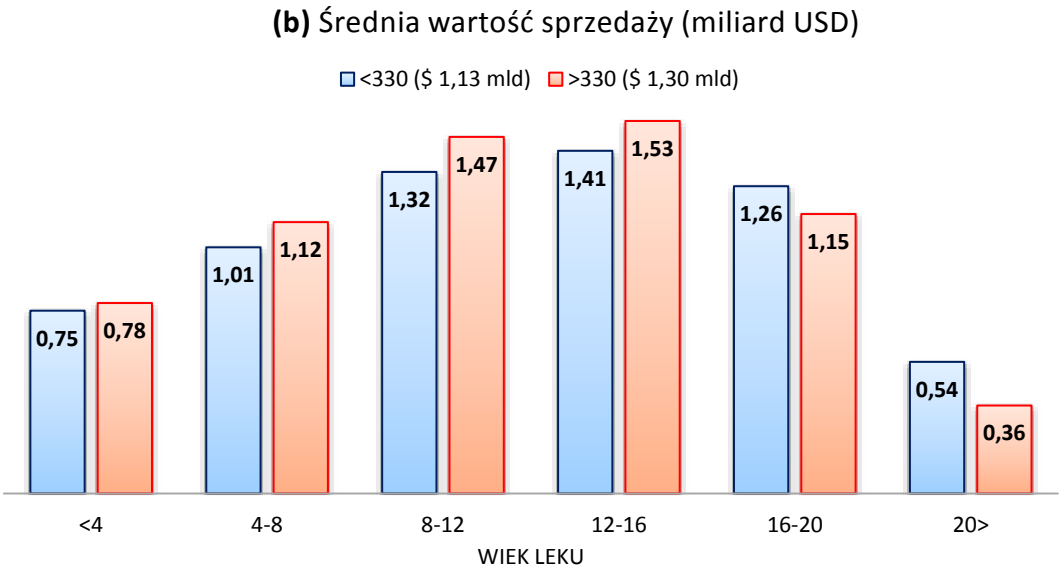
Kolejno dla zgrupowanych leków obliczono średnie i maksymalne wartości pIC_{50}, które zobrazowano na wykresie 10.9d. Odnotowano, że związki o MW<330 Da charakteryzują się najniższymi wartościami pIC_{50} (pIC_{50_MAX} = -0,8 oraz $pIC_{50\_ŚREDNIA}$ = -4,5). W przedziale 330-500 Da zaobserwowano najwyższą maksymalną wartość pIC_{50_MAX} = 0,9. Z drugiej strony średnia wartość parametru pIC_{50} pozostawała na takim samym poziomie jak w przypadku związków o masie przekraczającej 500 Da ($pIC_{50\_ŚREDNIA}$ = -3,9). Analiza parametru lipofilowości (clog P) przedstawiona na ilustracji 10.9c charakteryzuje się zaskakująco podobnym rozkładem jak dla MW (rysunek 10.9a).

Maksymalne wartości pIC_{50} charakteryzują większy potencjał interakcyjny leku z różnymi celami molekularnymi. Statystycznie większą szansę na dopasowanie do receptora wykazują związki o małej złożoności (niskiej masie cząsteczkowej). Analogiczną zależność zaobserwowano pomiędzy rosnącymi wartościami pIC_{50} a malejącym ADMET Score w granicy 330-500 Da (wykres 10.9b).

Omawiany efekt można wyjaśnić w oparciu o prawdopodobieństwo dopasowania fragmentów molekularnych z receptorem [142]. Lek o mniejszej MW ma większe prawdopodobieństwo interakcji z celem molekularnym. Jednakże na rysunku 10.9a widać, że aktywność związków o masie mniejszej niż 330 Da drastycznie spada. Z kolei molekuły większe niż 500 Da mają mniejsze prawdopodobieństwo dopasowania się do przypadkowego miejsca wiążącego ze względu na wyższą selektywność. Analiza ekonomiczna wykazała, że pomimo niskiej wartości pIC_{50} związki o małej masie mają większe szanse odnieść sukces rynkowy. Zależności te przekładają się na wyższe prawdopodobieństwo sukcesu niż bardziej złożone cząsteczki. Tak więc mniejszy sukces zaobserwowano dla leków o skrajnych wartościach MW i logP. Jest to zgodne z modelem tzw. "sweet spot" [12].

Zatem przeanalizowano zależność pomiędzy stopniem złożoności leku (MW) a jego sukcesem ekonomicznym (sprzedaż) i wiekiem leku (rysunek 10.10a,b).



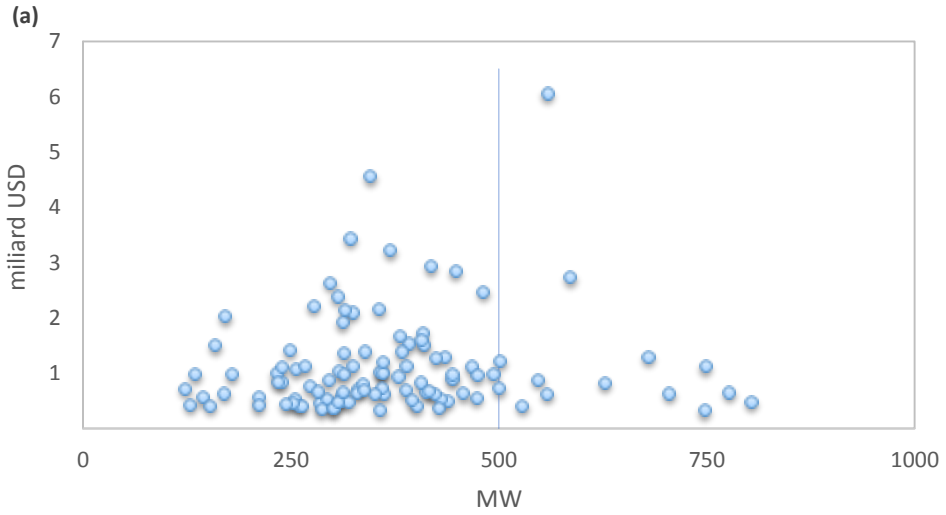


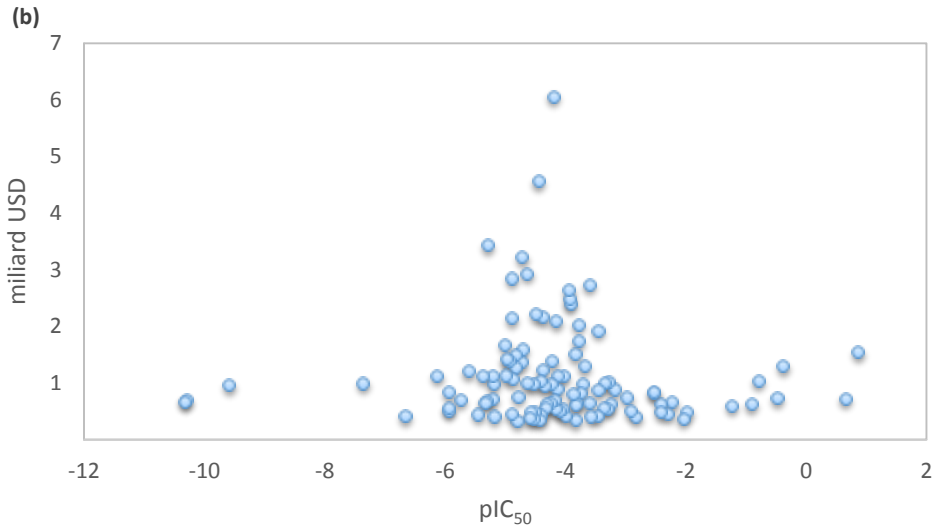
Rysunek 10.10 Średnia wartość sprzedaży (a, b) dwóch grup sklasyfikowanych na podstawie założeń Lipińskiego MW=500 Da (a) oraz Gleesona MW=330 Da (b) w zależności od zmieniającego się wieku leku.

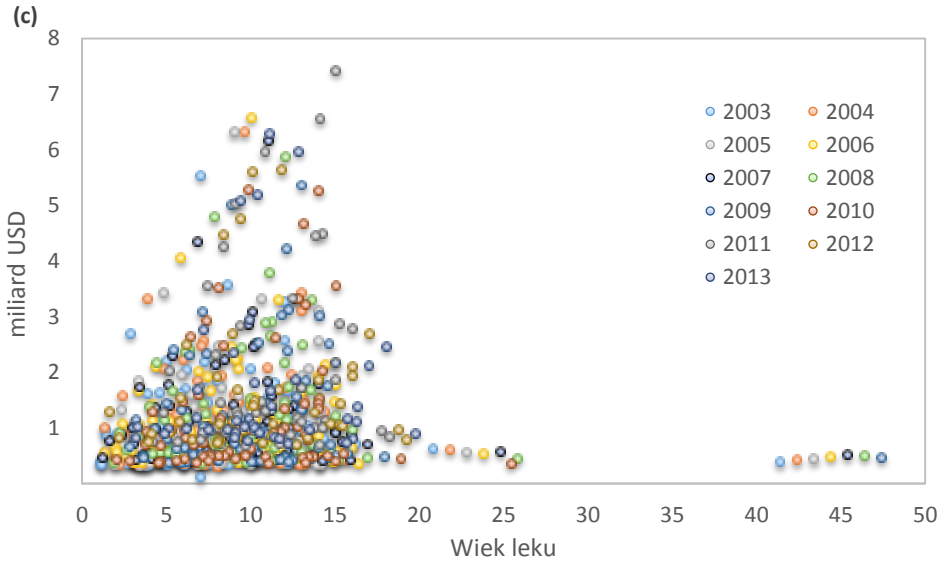
Sukces rynkowy mierzony średnią wartością sprzedaży zmienia się w zależności od wieku leku. Dla badanej populacji leki z przedziału 8-16 lat osiągnęły największe przychody. Na rysunku 10.10b zaobserwowano, że farmaceutyki o masie większej niż 330 Da (wyłączając leki starsze niż 16 lat) odnosiły większy sukces rynkowy. Podobną zależność zanotowano dla związków o masie większej od 500 Da w przedziale 8-20 lat (wykres 10.10a).

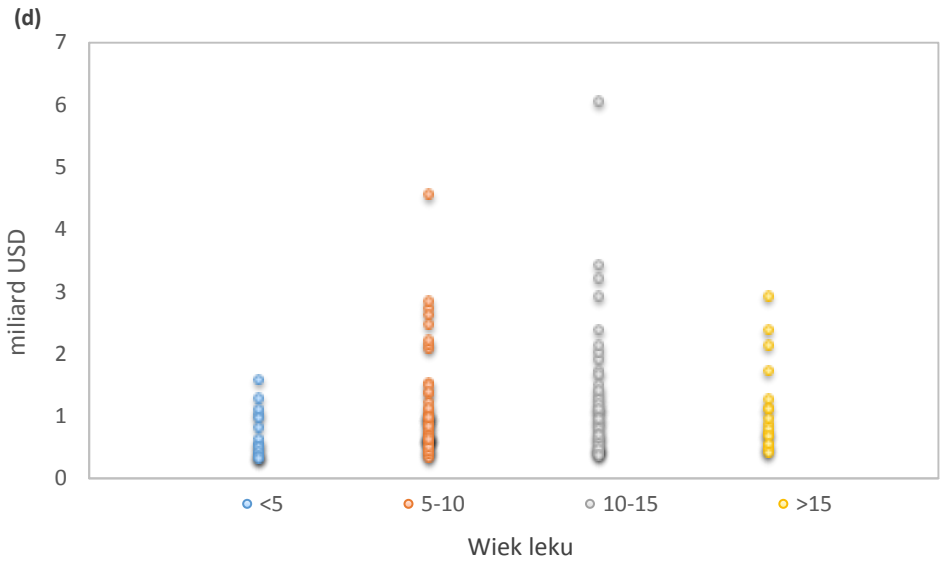
Podczas gdy w populacji młodszych i starszych (poniżej 8 i powyżej 20 lat) leki o MW>500 Da wykazały większą efektywność rynkową.

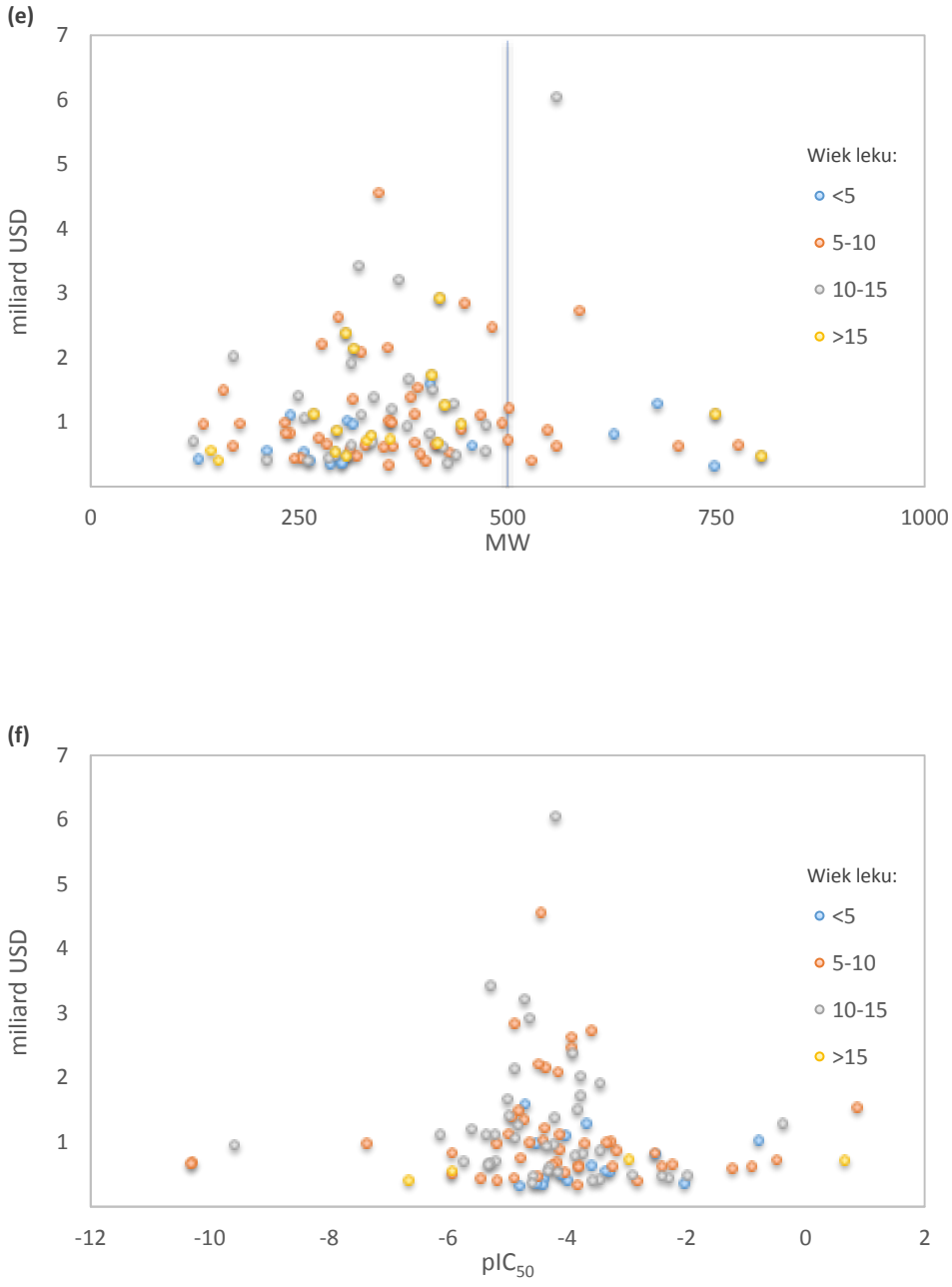
Analizując efektywność rynkową leków zgrupowanych według założeń Lipińskiego (<500 oraz >500 Da), zaobserwowano ekonomiczną przewagę (o 0,14 mld USD) farmaceutyków, których masa przekracza 500 Da (przychód leków o MW<500 i MW>500 Da wyniósł odpowiednio 1,22 i 1,36 mld USD). W zestawieniu parametrów (MW, pIC_{50}, wiek leku) ze średnią wartością sprzedaży (rysunek 10.11a-f) wykazano istnienie optymalnego przedziału, w którym wartość sprzedaży osiąga największe wartości (np. dla MW=330-500 Da, natomiast dla pIC_{50} od -3,5 do -5). Warto zauważyć, że dla grupy związków o MW>500 Da lek o nazwie handlowej "Lipitor" zawyża wyniki sprzedaży (najwyżej położony punkt na wykresie 10.11a-f). Dodatkowo na ilustracji 10.11e,f porównano zależność parametrów (MW, pIC_{50}) ze sprzedażą uwzględniając wiek leku. Opierając się na analizie kowariancji, nie wykazano żadnych zależności pomiędzy indywidualnymi danymi (tabela 10.6a) w odróżnieniu od uśrednionych danych (tabela 10.6b).




(b)
miliard USD
7
6
5
4
3
2
1
-12
-10
-8
-6
-4
-2
0
2
pIC50


(c)
miliard USD
8
7
6
5
4
3
2
1
0
5
10
15
20
25
30
35
40
45
50
Wiek leku
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013


(d)
miliard USD
7
6
5
4
3
2
1
0
<5
5-10
10-15
>15
Wiek leku



Rysunek 10.11 Wartość średniej sprzedaży bestsellerów w zależności od masy molekularnej (a), wartości pIC_{50} (b) oraz wieku leku (c). Dodatkowo, na wykresach d, e, f przedstawiono porównanie tych samych deskryptorów z lekami w odpowiednich grupach wiekowych (<5, 5-10, 10-15, >15).

Tabela 10.6 Zestawienie korelacji (współczynników Pearsona) dla danych:

a) Indywidualnych

Parametr	*Sprzedaż*	*Masa Molekularna*	*pIC_{50}*	*Wiek Leku*
Sprzedaż	1.00			
Masa Molekularna	0.06	1.00		
pIC_{50}	-0.01	0.00	1.00	
Wiek Leku	0.09	-0.07	-0.09	1.00

b) Binowanych w cztery grupy wiekowe (<5, 5-10, 10-15, >15)

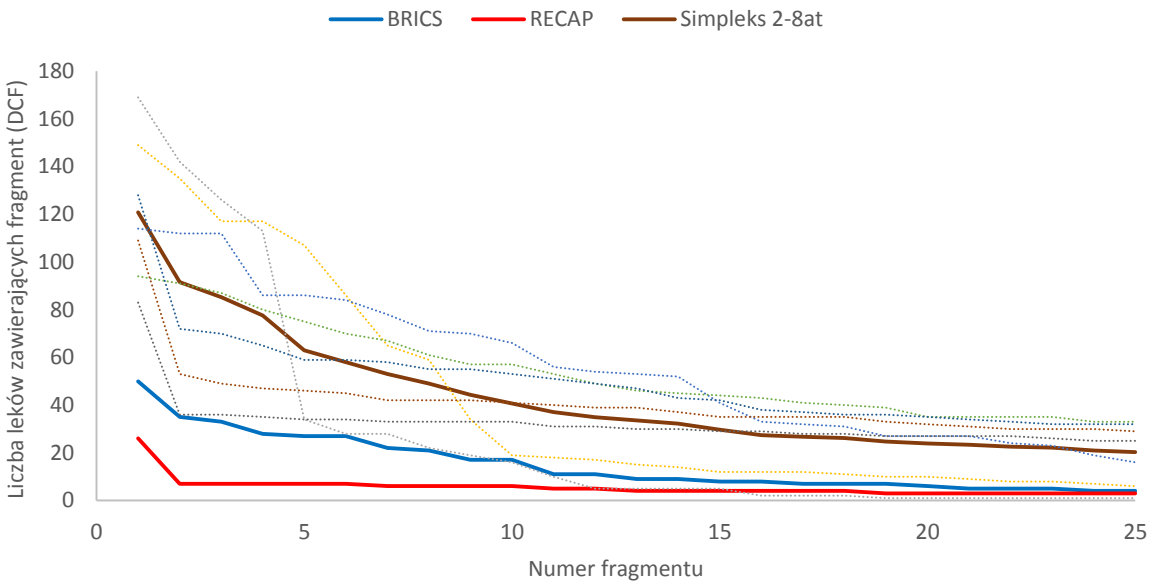
Parametr	*Sprzedaż*	*Masa Molekularna*	*pIC_{50}*	*Wiek Leku*
Sprzedaż	1.00			
Masa Molekularna	-0.47	1.00		
pIC_{50}	-0.44	-0.59	1.00	
Wiek Leku	0.87	-0.84	0.06	1.00

Podsumowując, wpływ złożoności związku molekularnego na sukces rynkowy nie jest do końca jasny. W uproszczeniu można stwierdzić, że leki o mniejszej masie molekularnej mają większą szansę na sukces rynkowy. Chociaż te z większą MW odnotowują wyższe zyski w przeliczeniu na lek. Jedną z przyczyn może być fakt, że strategie poszukiwania leków w przestrzeniach o większym ryzyku niepowodzenia z reguły prowadzą do wzrostu MW. Większe ryzyko może przynieść znaczniejszy zysk.

Przy obecnym stanie zaawansowania technologicznym strategia zorientowana na cel molekularny jest mniej efektywna od klasycznego fenotypowego podejścia projektowania leków. Warto zauważyć, że najlepiej sprzedające się leki są obecne na rynku od około 12 lat. Wydaje się, że obecna wiedza na temat poszukiwania innowacyjnych, selektywnych i skuteczniejszych leków kryje jeszcze wiele niewiadomych, których rozwiązania wciąż poszukujemy.

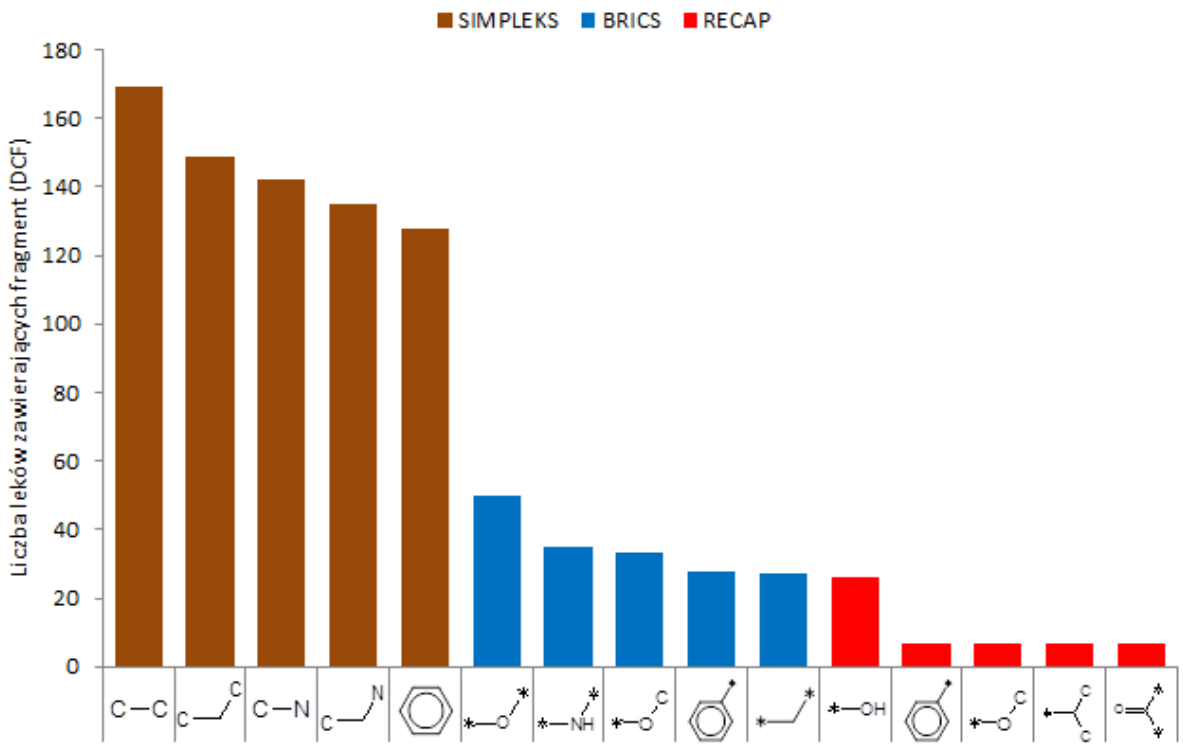
10.3 Badanie struktury topologicznej bestselerów

W pierwszym etapie badań przeprowadzono analizę porównawczą pomiędzy trzema metodami dekompozycji związków chemicznych: RECAP [85], BRICS [86] i SIMPLEX [87,88]. W tym celu zbiór leków z listy top 100 poddano fragmentacji, a następnie zliczono liczbę powtarzających się fragmentów (wykres 10.12).



Rysunek 10.12 Zależność liczby leków (oś y) zawierających dany fragment (oś x) od metody dekompozycji: BRICS (linia niebieska), RECAP (linia czerwona), SIMPLEKS (linie kropkowane odpowiadają fragmentom o długości wiązania kolejno od 2 do 8, natomiast linia brązowa odpowiada liczbie z całego zakresu 2-8) dla leków z listy top 100.

Przedstawiona na wykresach 10.12, 10.13 oraz w tabeli 10.7 analiza liczby leków zawierających fragment (ang. drugs containing fragments, DCF) jednoznacznie wskazuje, że Simpleksy są najbardziej efektywną pod względem zróżnicowania fragmentów, metodą dekompozycji związków chemicznych.



Rysunek 10.13 Liczba leków zawierających pięć najczęściej występujących fragmentów w zależności od sposobu dekompozycji.

Tabela 10.7 Parametr DCF, struktura fragmentu oraz wartość sprzedaży (w przeliczeniu na lek) z uwzględnieniem metody dekompozycji wiązań atomowych.

Metoda	Nr	Fragment	DCF	Sprzedaż (mld USD)	Sprzedaż (%)
SIMPLEKS	1	CC	169	947,185	100%
	2	CCC	149	758,161	80%
	3	CN	142	725,582	77%
	4	CCN	135	669,085	71%
	5	c1ccccc1	128	889,721	94%
BRICS	1	*O*	50	305,921	32%
	2	*N*	35	143,644	15%
	3	*OC	33	217,827	23%
	4	*c1ccccc1	28	144,227	15%
	5	*C*	27	157,073	17%
RECAP	1	*O	26	199,602	21%
	2	*c1ccccc1	7	65,138	7%
	3	*OC	7	55,809	6%
	4	*C(C)C	7	27,910	3%
	5	*C(=O)*	7	20,361	2%

Najliczniejszą klasą bestselerów są leki działające na ośrodkowy układ nerwowy o średnim wieku 14,6 lat (rozdział 10 – tabela 10.3). Tak więc kolejną część badań fragonomiki leków z listy top 100 oparto wyłącznie o zmodyfikowaną metodę simpleksów. Celem badania było sprawdzenie czy grupa związków o podobnych właściwościach posiada również wspólne elementy strukturalne. Analizie poddano dwie najliczniejsze podgrupy: CNS (26% populacji) oraz przeciwzapalne (stanowiących 16% wszystkich leków).

```
import indigo;
from indigo_renderer import *;
import pybel
indigo=Indigo()
molecules=[]

def Simpleksy(limitLeft =range(1,boundsCount+2) ,
limitRight=range(1,boundsCount+2)):
    indigo=Indigo()
    molecules=[]
    file=open('lokalizacja pliku', 'r')
    f=open('lokalizacja pliku','a')
    atomindex = []
    for x in file:
        line=x.strip('\n')
        mol1=indigo.loadMolecule(line)
        mol1.aromatize()
        mol2=mol1.canonicalSmiles()
        mol2=indigo.loadMolecule(mol2)
        molecules.append(mol2)
        boundsCount = mol2.countBonds();
        if limitLeft <= boundsCount:
                for submol in mol2.iterateEdgeSubmolecules(limitLeft,
limitRight):
                        f.write(submol.clone().smiles())
                        f.write('\n')
                        f.closed
        else:
            print 'koniec'
    return

for i in range(3,7):
    Simpleksy(i,i)
```

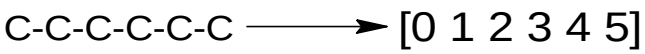
Rysunek 10.14 Fragment podstawowej funkcji służącej do dekompozycji molekuł na simpleksy o zdefiniowanej długości.

Fragmentacji dokonano w środowisku programowania Python. Fragmenty molekularne poddano dekompozycji wg długości wiązania (w zakresie od trzech do siedmiu). Zmodyfikowany algorytm simpleksów nie wykonywał dekompozycji wiązań aromatycznych. Pomimo uproszczenia obliczeń, głównym ograniczeniem wciąż pozostała moc obliczeniowa

jednostki komputerowej. Nie mniej jednak, wybór długich simpleksów odbiega od ich pierwotnej koncepcji i bardziej przypomina analizę MCS. Uwzględniając złożoną problematykę ograniczeń, do metody dobrano najbardziej optymalne warunki:

- długość wygenerowanych simpleksów wahała się w granicach 3-7 wiązań (4-8 atomowe fragmenty)
- zmodyfikowano algorytm, w taki sposób aby wiązania aromatyczne nie ulegały dekompozycji

Otrzymany zbiór simpleksów zapisano w kanonicznym kodzie SMILES. Algorytm dokonał dekompozycji związków metodą kombinacji bez powtórzeń fragmentów o zadanej długości z zachowaniem wiązań występujących między atomami. Jednakże istnieje konieczność ponownego przeszukania zbioru celem eliminacji identycznych struktur topologicznych pochodzących od macierzystej cząsteczki. Dla lepszego zrozumienia, procedurę wyjaśniono na przykładzie cząsteczki heksanu:



Rysunek 10.15 Cząsteczka heksanu zapisana w notacji numerycznej

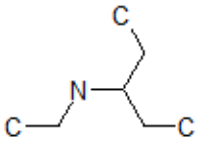
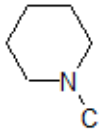
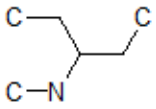
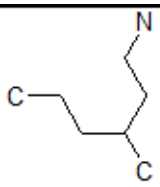
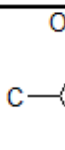
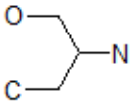
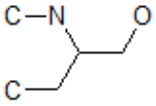
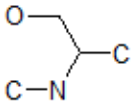
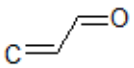
W procesie generowania simpleksów 2D zgodnie z nomenklaturą przedstawioną na rysunku 10.15, heksan ulega dekompozycji na następujące fragmenty:

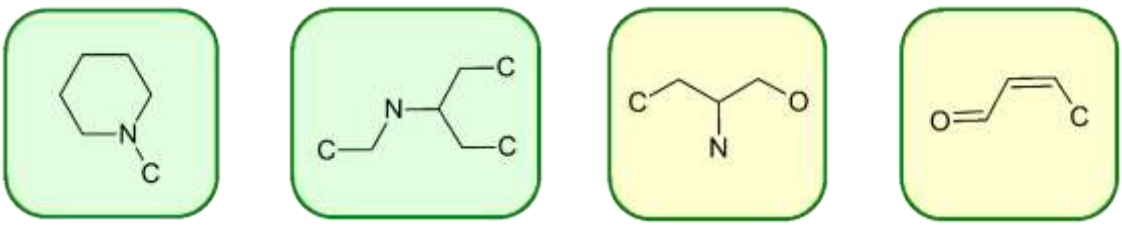
- pięć dwuatomowych fragmentów CC w kombinacji [0 1] [1 2] [2 3] [3 4] [4 5]
- cztery trzyatomowe fragmenty CCC w kombinacji [0 1 2] [1 2 3] [2 3 4] [3 4 5]
- trzy czteroatomowe fragmenty CCCC w kombinacji [0 1 2 3] [1 2 3 4] [2 3 4 5]
- dwa pięcioatomowe fragmenty CCCCC w kombinacji [0 1 2 3 4] [1 2 3 4 5]
- jeden sześcioatomowy fragment CCCCCC, alternatywnie [0 1 2 3 4 5 6]

Każdy wygenerowany fragment heksanu zapisany w kodzie kanonicznym SMILES nie uwzględnia stereochemii. Zatem w kolejnym etapie ze zbioru wygenerowanych simpleksów, usunięto wszystkie powtarzające się struktury 2D (o tym samym kanonicznym kodzie SMILES) pochodzące od tego samego związku. W ten sposób uzyskano zbiór indywidualnych fragmentów molekularnych przypisanych maksymalnie raz dla związków wejściowych. Następnie ze zbioru reprezentatywnych cząsteczek zliczono w obrębie całej

populacji liczbę wszystkich powtarzających się struktur kodu SMILES. W ten sposób oszacowano popularność simpleksów, ponieważ otrzymane wartości informują o liczbie cząsteczek posiadających dany motyw w swojej strukturze. Ostatecznie wyodrębniono pięć najczęściej występujących, indywidualnych dla danej grupy motywów (tabela 10.8 oraz rysunek 10.16).

Tabela 10.8 Różnorodność molekularna najbardziej popularnych klas związków z listy bestsellerów 2003-2013. Fragmenty otrzymane metodą simpleks.

CNS					
	27%	18%	18%	18%	17%
Przeciwzapalne					
	44%	41%	37%	37%	33%
Popularne					
	100%	94%	89%	87%	81%



Rysunek 10.16 Najczęściej występujące struktury, z prawej CNS (27%, 18% własnej populacji) oraz z lewej motywy leków przeciwzapalnych (44%, 33% własnej populacji).

W wyniku przeprowadzonej analizy w każdej z grup udało się wyodrębnić:

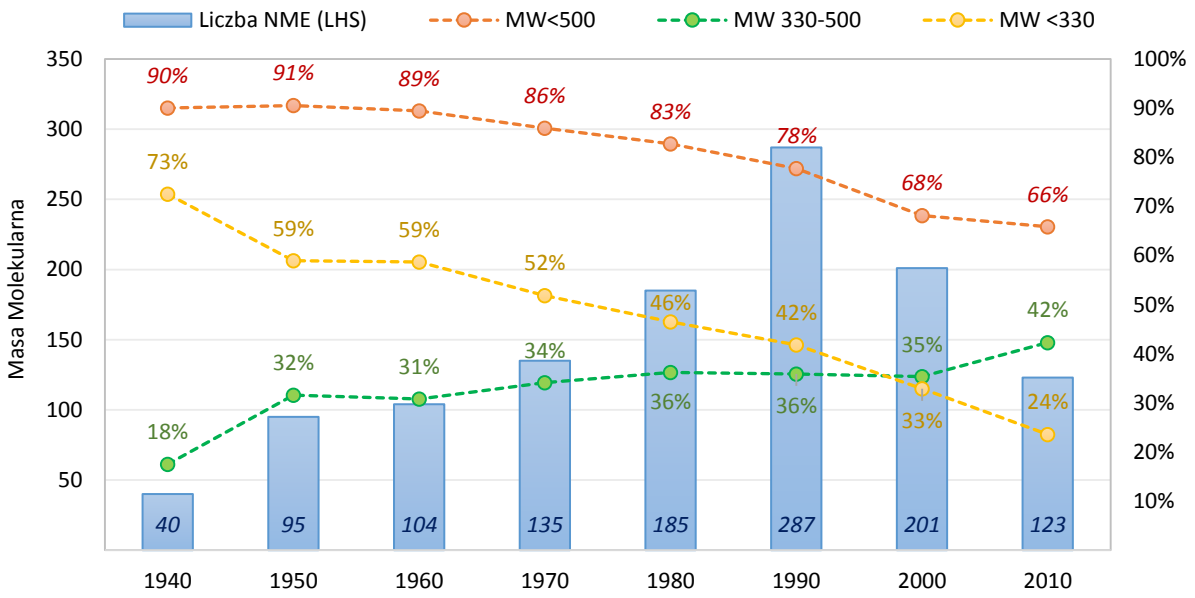
- najczęściej występujące fragmenty w obu podgrupach
- dwa najpopularniejsze simpleksy charakteryzujące wybraną grupę leków

11 Studium fragonomiki leków NME z lat 1939-2014

W wyniku eksploracji danych wyodrębniono populację 1170 nowych jednostek molekularnych (NME) z okresu 1939-2014. Populację podzielono na osiem grup wg roku rejestracji FDA:

- Grupa 1: NME 1939-1949
- Grupa 2: NME 1950-1959
- Grupa 3: NME 1960-1969
- Grupa4: NME 1970-1979
- Grupa5: NME 1980-1989
- Grupa6: NME 1990-1999
- Grupa7: NME 2000-2009
- Grupa8: NME 2010-2014

Następnie wykorzystując możliwości biblioteki Open Babel [143], dla każdej z molekuł obliczono podstawowe parametry lekopodobieństwa (MW, clog P, ADMET Score).



Rysunek 11.1 Zależność procentowego udziału grup związków (podzielonych ze względu na masę cząsteczkową) od roku zatwierdzenia oraz liczby NME (kolumny). Linie przedstawiają procentowy udział danej podgrupy.

Najprostszą miarą złożoności cząsteczki związku chemicznego jest masa molekularna. Zakładając odpowiednie granice MW można dokładniej oszacować różnorodność zbioru i trendów panujących na przestrzeni kolejnych lat. Od ponad siedemdziesięciu lat obserwuje się stopniowy spadek liczebności związków NME o MW<500 Da (rysunek 11.1). Największy spadek liczebności leków wykazały związki o MW<330 Da. Badanie zależności pomiędzy danymi przedstawionymi na wykresie 11.1 oraz w tabeli 11.1 sugeruje poszukiwanie nowych leków wg następujących założeń:

- MW=330-500 Da – jest to najbardziej optymalna i bezpieczna grupa związków. Wykazuje idealne właściwości lekopodobieństwa. Obecnie stanowi 42% NME. Udział procentowy powoli rośnie z roku na rok. Zbiór jest dobrze przebadany. Można łatwo przewidzieć potencjalne właściwości polifarmakologiczne nowych molekuł.
- MW>500 Da – najmniej przebadana grupa molekuł. Wykazuje niekorzystny profil ADMET ale może posiadać lepszą skuteczność terapeutyczną. Z powodu zmiany strategii projektowania leków, udział procentowy w ostatnich latach dynamicznie rośnie (obecnie wynosi 34%). Przyjmuje się, że związki z tej grupy stanowią przyszłość medycyny (leki biologiczne, selektywne, spersonalizowane).

Średni spadek liczebności leków o MW<500 Da w okresie 1940-2010 wyniósł -24 p.p. (punkt procentowy) (z 90% do 66% w stosunku do wszystkich rejestracji FDA). Zauważono, że spadek w ostatnich latach (1990-2010) był zdecydowanie większy i wynosił -12 p.p. (z 78% do 66%) w porównaniu do lat 1940-1980, w który zanotowano spadek o -7 p.p. (z 90% do 83%). W przypadku cząsteczek o MW<330 Da, odnotowano największy procentowy spadek liczby populacji w okresie od 1940 do 2010 (zmiana o -49 p.p. tj. z 73% w 1940 na 24% związków o MW<330 Da w latach 2010-tych).

Interesującym okazuje się fakt, że populacja leków o MW=330-500 Da cały czas rosła (linia zielona). W analizowanym przypadku odnotowano średni 24 p.p. wzrost liczby leków od lat 1940-tych.

Tabela 11.1 Średnie wartości deskryptorów, udziały procentowe podzbiorów NME w zależności od całego okresu rejestracji FDA.

DESKRYPTOR	1940'	1950'	1960'	1970'	1980'	1990'	2000'	2010'
Liczba leków	40	95	104	135	185	287	201	123
MW >400 & clog P >4	8%	9%	8%	10%	12%	21%	28%	38%
MW >500 & clog P >5	5%	2%	0%	4%	2%	7%	15%	16%
>2 niespełnionych warunków Ro5	10%	4%	9%	14%	14%	20%	29%	28%
MW (średnia/mediana)	301 / 261	326 / 304	340 / 304	415 / 327	383 / 340	459 / 367	676 / 409	635 / 450
MW <330	73%	59%	59%	52%	46%	42%	33%	24%
MW <400	85%	80%	76%	72%	66%	59%	49%	36%
MW <500	90%	91%	89%	86%	83%	78%	68%	66%
MW 330-500	18%	32%	31%	34%	36%	36%	35%	42%
clog P (średnia/mediana)	2,70 / 2,62	2,89 / 2,97	2,61 / 2,80	2,11 / 2,33	2,44 / 2,52	2,65 / 3,03	2,97 / 3,27	3,88 / 3,84
clog P <3	63%	53%	52%	63%	59%	49%	44%	33%
clog P <4	70%	77%	78%	80%	74%	70%	63%	54%
clog P <5	85%	94%	97%	89%	89%	86%	79%	71%
ADMET Score (średnia/mediana)	3,33 / 3,30	2,67 / 2,93	2,87 / 3,01	3,49 / 3,04	2,96 / 2,89	3,51 / 3,04	5,16 / 2,99	4,52 / 2,99

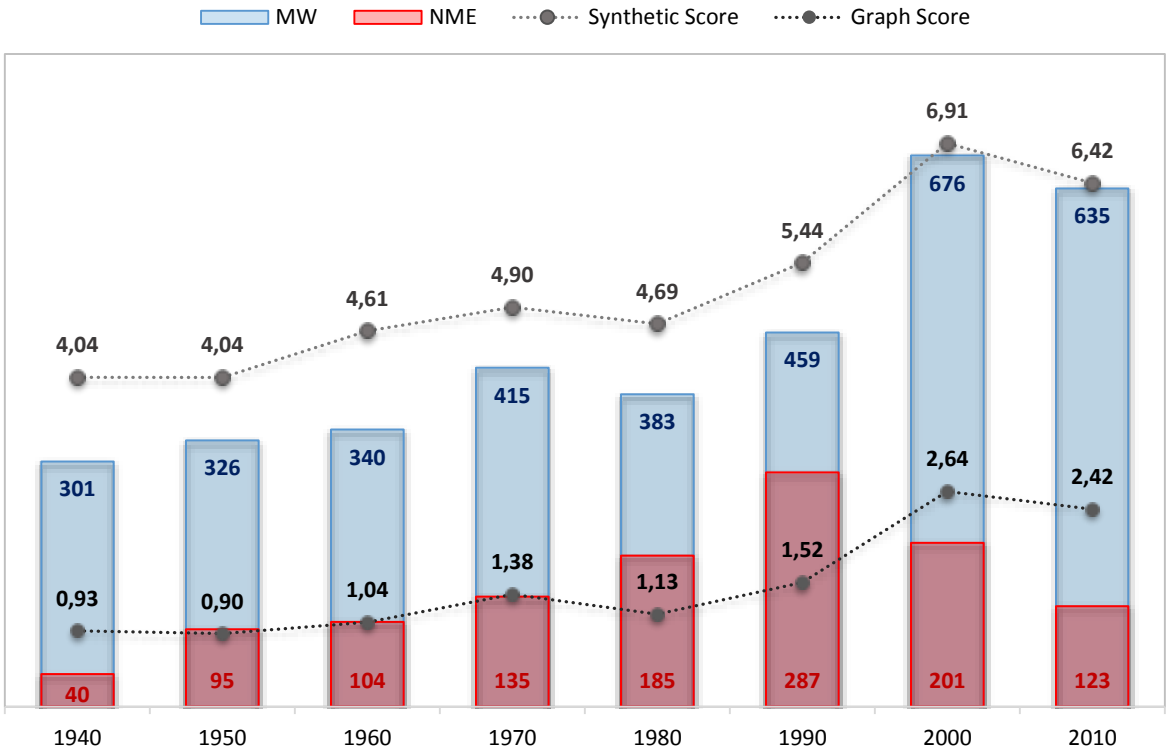
Interesujący problem wiąże się z zależnością rynkowego sukcesu względem problemów syntetycznych. W celu oszacowania syntetycznej/handlowej dostępności związków chemicznych wykorzystano program SYLVIA (Molecular Networks GmbH) [4,144,145].

SYLVIA służy do szacowania syntetycznej dostępności związków chemicznych. Program jest dostępny w wersji komercyjnej oraz demonstracyjnej (po złożeniu odpowiedniego wniosku uzasadniającego jego użycie). SYLVIA jest prosty w obsłudze oraz posiada przyjazny dla użytkownika interfejs GUI.

Dostępność syntetyczną szacowano w oparciu o następujące parametry:

- **Graph Score** – obliczany na podstawie teorii grafów i informacji teoretycznych. Uwzględnia wielkość, symetrię, rozgałęzienie, ilość pierścieni, wiązań wielkokrotnych, heteroatomów itp.
- **Ring Score** – wynik uwzględnia zmostkowane i skondensowane układy pierścieniowe, które mogą utrudniać syntezę związków, a tym samym zwiększyć syntetyczną ocenę dostępności.

- **Stereo Score** – oblicza liczbę tetraedrycznych centrów stereochemicznych w cząsteczce, które wpływają na trudność syntezy związku.
- **BB Score** (ang. Building Blocks/Starting Materials Score) – struktury zawierające złożone motywy mogą być łatwo syntetyzowane pod warunkiem, że fragmenty można otrzymać z łatwo dostępnych substratów. Reagenty są wyszukiwane i porównywane między sobą w specjalnie przygotowanej, zintegrowanej dla programu bazie. Im bardziej powszechne substancje biorące udział w reakcji tym niższy wynik.
- **Reaction Center Score** – związek zostaje poddany analizie retrosyntetycznej, w wyniku której oceniane są etapy syntez oraz struktury związków pośrednich.
- **Synthetic Score** – ogólny wynik syntetycznej dostępności struktury docelowej. Parametr jest obliczany przez zsumowanie pięciu wyżej wymienionych składników (Graph, Ring, Stereo, BB oraz Reaction Center Score). Ocena syntetycznej dostępności opiera się na analizie regresji z uwzględnieniem odpowiednich wag dla każdego ze składników. Skala waha się w granicach od 1 (łatwa synteza, związek łatwo dostępny handlowo) do 10 (trudna synteza, związek trudno dostępny na rynku lub jego brak).



Rysunek 11.2 Ocena dostępności syntetycznej ("Synthetic Score") oraz złożoności strukturalnej ("Graph Score") w zależności od roku zatwierdzenia regulatora FDA. Kolumna czerwona odpowiada liczbie NME, natomiast niebieska średniej masy molekularnej.

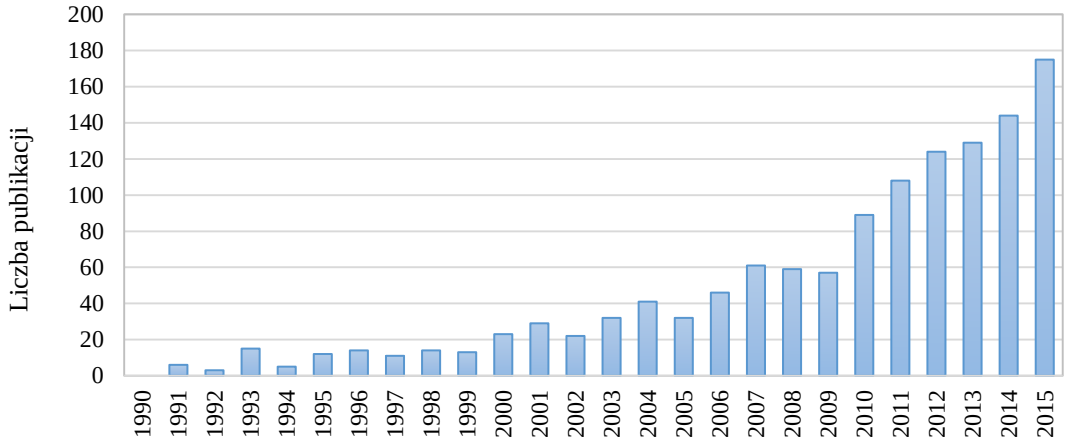
NME poddano kategoryzacji z uwzględnieniem dziesięcioletniego okresu rejestracji FDA. W ten sposób utworzono 8 grup z podziałem na dekady (1940', 1950', 1960', 1970', 1980', 1990', 2000', 2010'). Dla każdego zbioru obliczono średnie arytmetyczne leków. Analizując dane przedstawione na wykresie 11.2 zauważono, że trend linii opisującej dostępność syntetyczną ("Synthetic Score") podąża zgodnie ze zmianą masy molekularnej. Współczynnik korelacji pomiędzy tymi parametrami wynosi $R^2 = 0,98$ (silna, pozytywna korelacja). Podobną zależność zaobserwowano pomiędzy "Graph Score" a MW (współczynnik determinacji R^2 wynosi 0,99). Natomiast korelacja pomiędzy "Synthetic Score" a "Graph Score" R^2 wynosi 0,97. Spośród pozostałych parametrów obliczanych za pomocą programu SYLVIA (tabela 11.2), żadne nie wykazywały istotnych korelacji z masą molekularną.

Tabela 11.2 Średnie wartości parametrów otrzymanych za pomocą programu SYLVIA w zależności od roku rejestracji FDA.

DESKRYPTOR	1940'	1950'	1960'	1970'	1980'	1990'	2000'	2010'
Liczba NME	40	95	104	135	185	287	201	123
Masa Molekularna	300,61	326,35	339,62	415,44	383,05	458,64	675,60	635,28
Synthetic Score	4,04	4,04	4,61	4,90	4,69	5,44	6,91	6,42
BB Score	0,61	0,64	0,84	0,73	0,89	1,12	1,20	1,16
Retrosynthetic Score	0,36	0,37	0,36	0,34	0,35	0,36	0,35	0,35
Graph Score	0,93	0,90	1,04	1,38	1,13	1,52	2,64	2,42
Ring Score	0,78	0,86	0,88	0,77	0,82	0,83	0,86	0,85
Stereo Score	0,69	0,60	0,81	1,02	0,82	0,94	1,19	0,97

Analizując parametry z tabeli 11.2 zauważono, iż wraz ze wzrostem dostępności syntetycznej, stopniem złożoności struktury oraz MW rośnie złożoność struktur szacowana parametrem dostępności substratów ("BB Score"). Wzrost "BB Score" wyniósł 0,55 p. (z 0,61 w 1940' do 1,16 w 2010').

Zaobserwowane zależności mogą świadczyć o wzmożeniu strategii projektowania związków o bardziej złożonej strukturze molekularnej. W ostatnich latach obserwuje się intensywny rozwój oraz większe zainteresowanie środowiska naukowego w zastosowaniu struktur zwanych blokami budulcowymi w farmacji. Ich dostępność ułatwia projektowanie syntez związków o wyższych MW.



Rysunek 11.3 Wzrost liczby publikacji dotyczących badań nad blokami budulcowymi w farmacji (Scopus, data dostępu: 19.07.2016)

11.1 Analiza struktury topologicznej

W kolejnym etapie analizę struktur związków rejestrowanych przez FDA oparto o metody fragmentaryczne oraz koncepcję motywów uprzywilejowanych, takich jak:

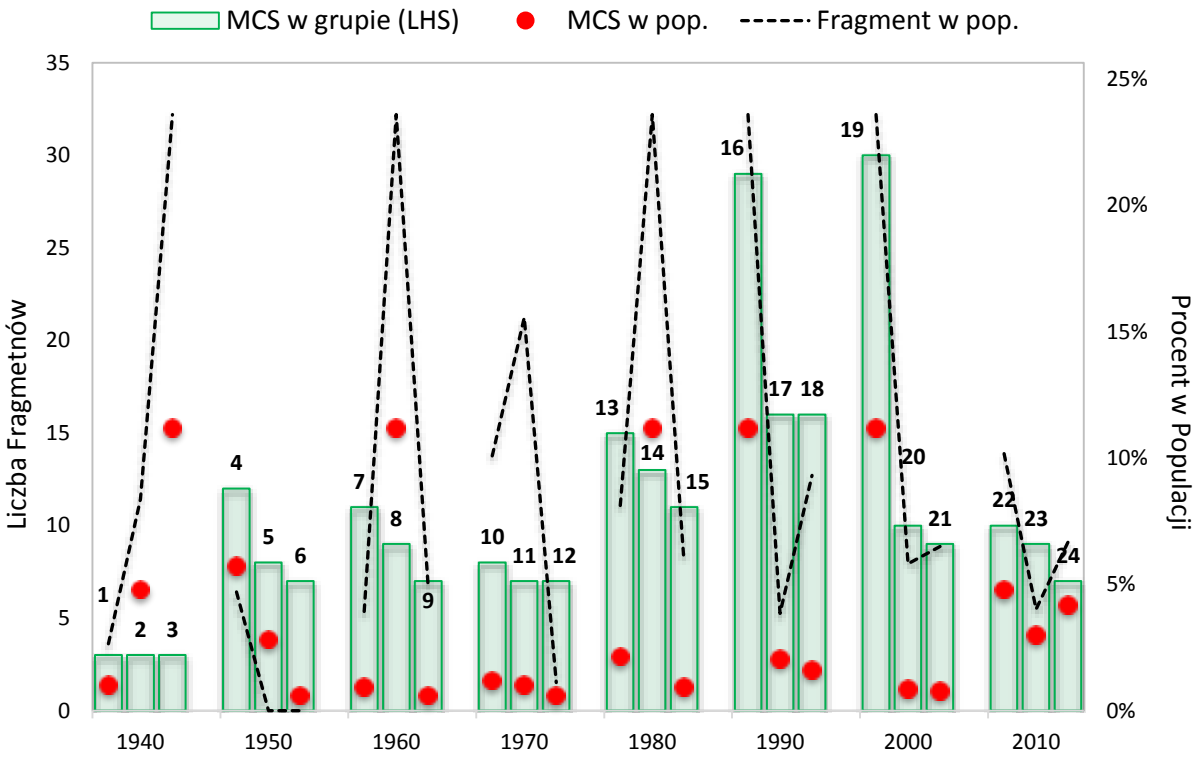
- MCS (ang. maximum common substructure) – poszukiwanie maksymalnej wspólnej podstruktury związków
- Motywy Murcko (ang. Murcko scaffold)
- Simpleksy (ang. simplex)

Formalnie przedstawione wyżej parametry są złożonymi deskryptorami molekularnymi, które oblicza się w oparciu o strukturę cząsteczki chemicznej.

Na początku przebadano fragmenty molekularne celem wyznaczenia maksymalnej wspólnej podstruktury (MCS). Do analizy wykorzystano program "Library MCS", który wchodzi w skład pakietu ChemAxon. Autorem programu jest ChemAxon - amerykańska firma, która w swoim portfolio ma szeroką gamę programów chemoinformatycznych.

Uprzednio przygotowane pliki .smi (smiles) sklasyfikowano względem roku rejestracji FDA. Następnie każdy taki zestaw molekuł (uprzednio zapisanych w notacji liniowej) kolejno wprowadzano jako dane wejściowe do programu. W wyniku obliczeń (algorytm klasteryzacji) uzyskano charakterystyczne drzewa hierarchiczne oraz struktury wraz ze szczegółowymi

informacjami (o pochodzeniu, liczbie powtórzeń, stopniach klasteryzacji). Dla każdego badanego okresu zebrano trzy najczęściej występujące motywy. Następnie obliczono ich częstotliwość występowania. Wykonano również analizę porównawczą, która polegała na wyszukiwaniu MCS w całym zbiorze NME, metodą podstruktur Open Babel/Pybel w języku programowania Python.

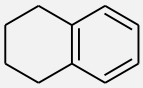
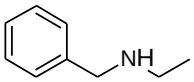
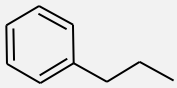
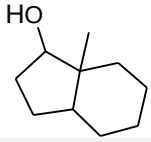
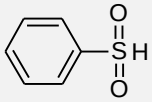
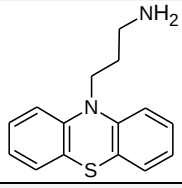
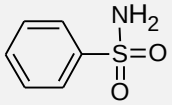
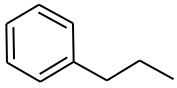
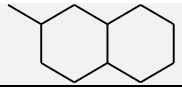
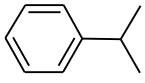
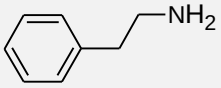


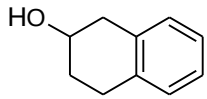
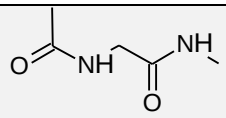
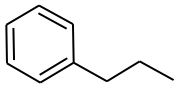
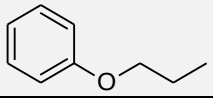
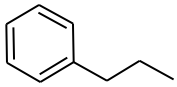
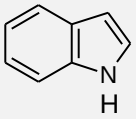
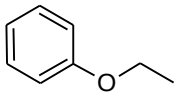
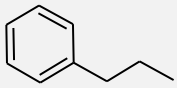
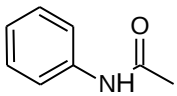
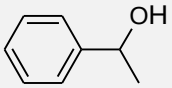
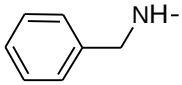
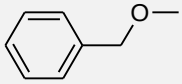
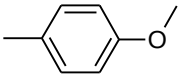
Rysunek 11.4. Udział poszczególnych struktur w zbiorach NME. Kolumny przedstawiają całkowitą liczę molekuł zawierających fragment MCS oznaczony numerem od 1 do 24 (szczegółowe zestawienie w tabeli 11.3). Procentowy udział MCS oznaczono czerwonymi kropkami (wg danych z Library MCS) oraz linią przerywaną (metodą Pybel).

W wyniku przeprowadzonej analizy MCS nie odnotowano zależności pomiędzy częstotliwością występowania wspólnych maksymalnych podstruktur od roku rejestracji FDA. Procentowy udział trzech najbardziej popularnych MCS waha się w granicy od 5,4% (1970') do 9,5% (1950'), natomiast średnia wyniosła 7,5%. Najniższy średni procentowy udział MCS odnotowują struktury z lat 1970-tych (0,9%), natomiast najwyższy z lat 1940-tych (5,7%). Wykonując analogiczne badania w środowisku programowania Python otrzymano wyższe

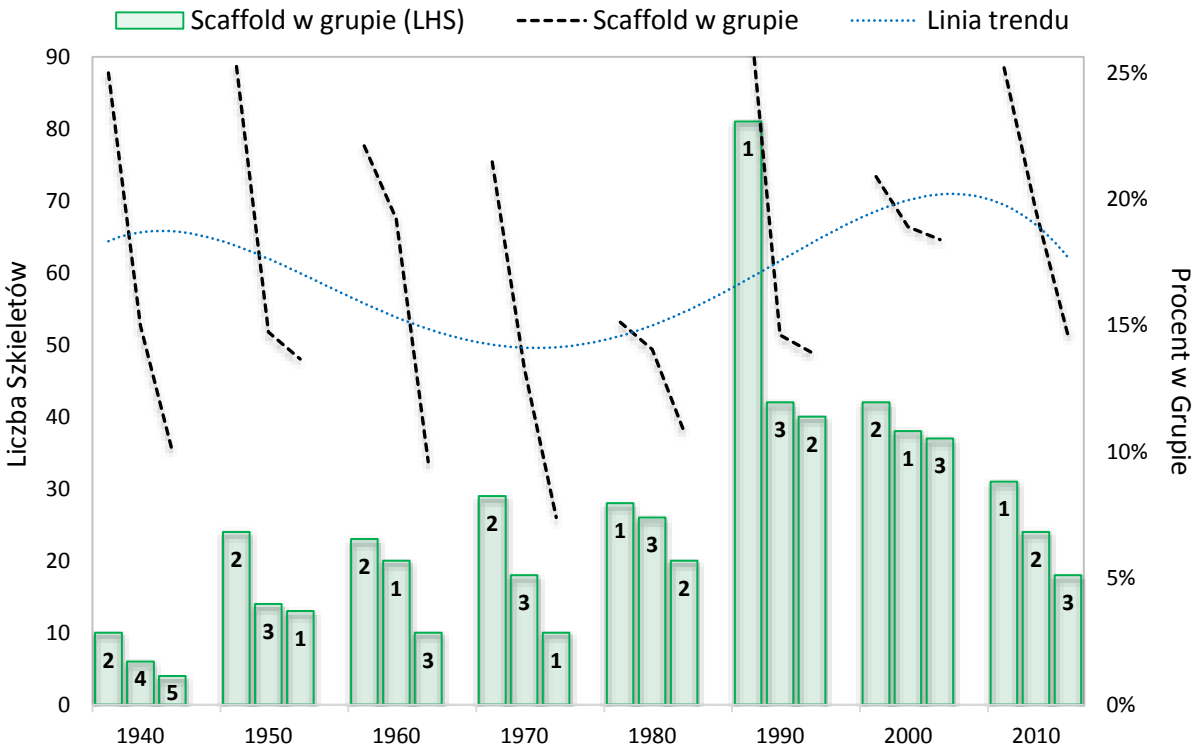
wartości udziału MCS w populacji, przy jednoczesnym zachowaniu proporcji dla indywidualnych podstruktur. W przypadku biblioteki Open Babel/Pybel, algorytm zlicza każdą cząsteczkę, w której występuje poszukiwany fragment. Program Library MCS stosuje metodę klasteryzacji, w wyniku której następuje optymalizacja pomiędzy parametrami: podobieństwa molekuł, liczbą drzew hierarchicznych, a złożonością podstruktury. Efektem metody ChemAxon może być sklasyfikowanie cząsteczek o wspólnej podstrukturze do innych zbiorów celem redukcji poziomów dekompozycji (ilości drzew hierarchicznych).

Tabela 11.3. Zestawienie struktur MCS wraz z odpowiadającą numeracją, zbiorem oraz procentowym udziałem obliczonym na podstawie danych z Library MCS (Ax) oraz Pybel (Py).

Zbiór	*Nr związku*	*Struktura*	*% Ax w zbiorze*	*% Ax w całości*	*% Py w całości*
1939-1949	1		8%	1%	3%
	2		8%	5%	8%
	3		8%	11%	24%
1950-1959	4		13%	6%	5%
	5		8%	3%	0%
	6		7%	1%	0%
1960-1969	7		11%	1%	4%
	8		9%	11%	24%
	9		7%	1%	5%
1970-1979	10		6%	1%	10%
	11		5%	1%	16%

	12		5%	1%	1%
1980-1989	13		8%	2%	8%
	14		7%	11%	24%
	15		6%	1%	6%
1990-1999	16		10%	11%	24%
	17		6%	2%	4%
	18		6%	2%	9%
2000-2009	19		15%	11%	24%
	20		5%	1%	6%
	21		4%	1%	6%
2010-2014	22		8%	5%	10%
	23		7%	3%	4%
	24		6%	4%	7%

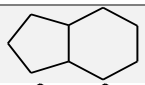
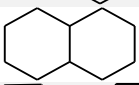
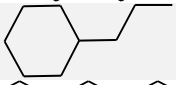
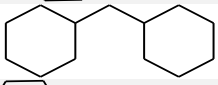
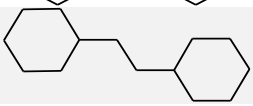
W kolejnym etapie studium fragonomiki, przeprowadzono analizę fragmentów Murcko. Wykorzystując funkcję Python oraz bibliotekę RDKit, molekuły przekonwertowano do fragmentów Murcko wykorzystując notację liniową SMILES. Analogicznie jak w przypadku analizy MCS, dla każdej z 8 grup zliczono i wyodrębniono trzy najczęściej występujące struktury. Dane przedstawiono na wykresie 11.5 oraz zebrano w tabeli 11.4.



Rysunek 11.5. Kolumny przedstawiają liczbę szkieletów Murcko w grupie. Szkielety oznaczono odpowiednimi liczbami przypisanymi dla struktur z tabeli 11.4. Linie przerywane określają udział procentowy, natomiast niebieska linia przedstawia trend występowania motywów na przestrzeni lat.

W badanym zbiorze NME zidentyfikowano pięć najczęściej występujących fragmentów Murcko (tabela 11.4). Wyłączając grupę 1940' najpopularniejszymi strukturami pozostają tylko trzy struktury. Do lat 1980-tych najliczniejszą grupę stanowiła struktura nr 2, aby w kolejnych latach (oprócz 2000') ustąpić miejsca strukturze nr 1, która w zasadzie jest analogiem motywu nr 2. Występowanie popularnych motywów Murcko na poziomie 15-20% (linia trendu przedstawiona na rysunku 11.5) może świadczyć o istnieniu wspólnych cech, które odróżniają leki od związków biologicznie nieaktywnych.

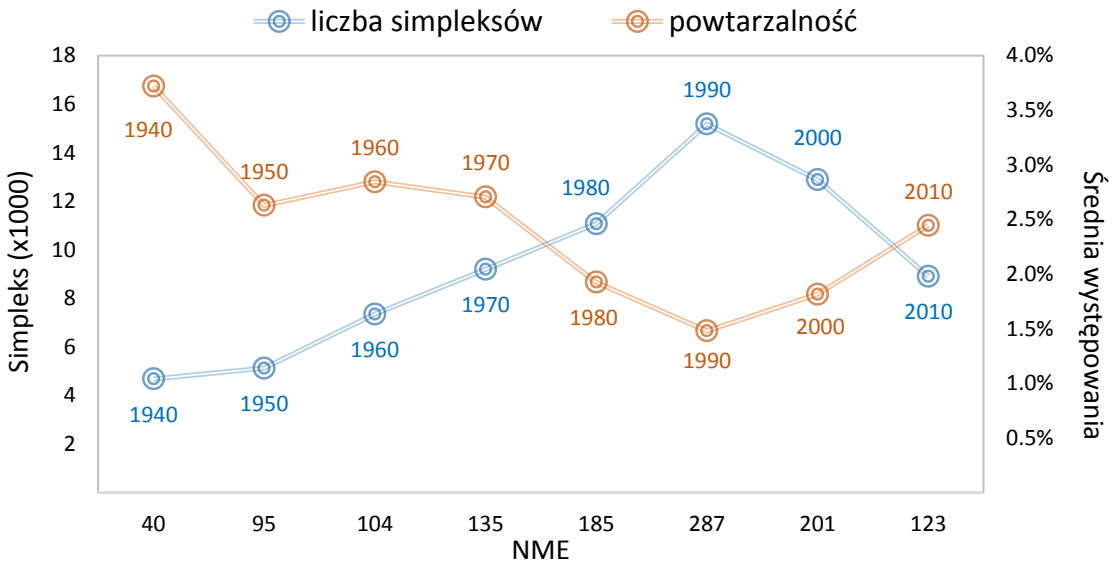
Tabela 11.4. Struktury najczęściej spotykanych fragmentów Murcko wraz z odpowiadającą im numeracją.

Lp.	*Struktura Szkieletu (Scaffold)*
1	
2	
3	
4	
5	

Ostatni etap analizy fragonomicznej leków FDA wykonano w oparciu o zmodyfikowaną metodę simpleksów.

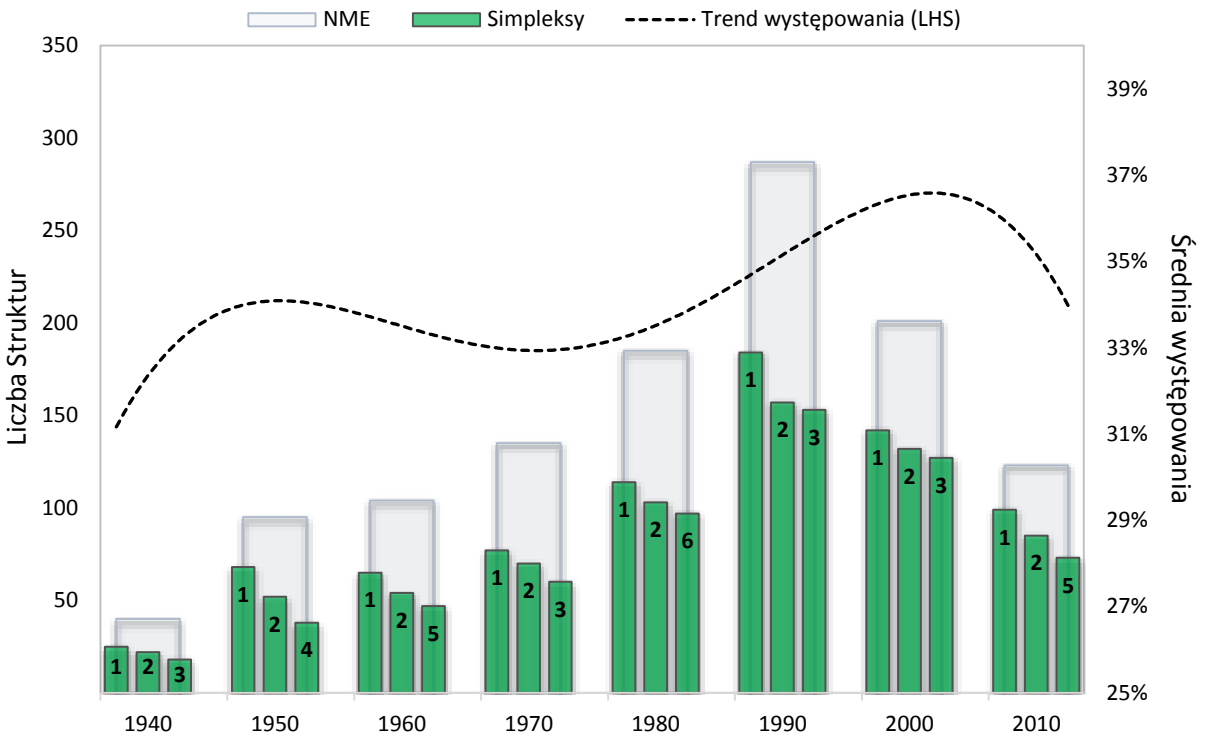
W oparciu o skrypt przedstawiony na rysunku 10.14 (rozdział 10), dokonano fragmentaryzacji wszystkich molekuł NME wg następującego schematu:

- Wygenerowano simpleksy o długości wiązania 3-7
- Każdy fragment przypisano do cząsteczki macierzystej
- Fragmenty (zapisane w generycznym kodzie SMILES) przekonwertowano do postaci kanonicznej kodu SMILES
- Usunięto powtarzające się struktury w obrębie danej cząsteczki (wyjaśnienie na przykładzie heksanu – rozdział 10)
- Struktury sklasyfikowano według roku rejestracji FDA (data NDA)
- Zliczono powtarzające się struktury w obrębie danej grupy wiekowej
- Dane poddano analizie statystycznej



Rysunek 11.6. Korelacja pomiędzy liczbą cząsteczek NME w zbiorze a liczbą wygenerowanych simpleksów (linia pomarańczowa, R^2 = -0,89), oraz częstotliwością fragmentu (linia niebieska, R^2 = 0,94).

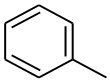
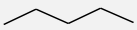
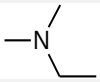
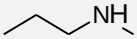
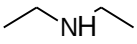
W odróżnieniu od wyników uzyskanych przez analizę struktur MCS oraz motywów Murcko, w przypadku analizy simpleksów zauważono występowanie pewnych zależności. Wykres 11.6 przedstawia korelację pomiędzy liczebnością zbioru NME, liczbą wygenerowanych simpleksów a średnią powtarzalnością fragmentów w zbiorze. Obliczony współczynnik Pearsona wykazuje silną zależność pomiędzy badanymi parametrami (dodatnią - pomiędzy liczbą NME a liczbą simpleksów lub ujemną – między liczbą NME a średnią występowania fragmentu w zbiorze).



Rysunek 11.7. Liczba NME (kolumna niebieska) wraz z liczbą cząsteczek zawierających w swojej strukturze odpowiednie simpleksy (kolumny zielone z etykietą – numer przypisany do odpowiedniej struktury).

Badając dystrybucję simpleksów zauważono, że trzy najczęściej występujące fragmenty stanowią średnio 58,15% populacji całego zbioru. Największą wartość 69,65% obserwowano w latach 2010-nych, a najniższą 52,27% w 1970-tych. Przedstawiona na wykresie 11.7 linia trendu obliczona została na podstawie równania wielomianowego czwartego stopnia. Źródłem danych była liczba występowania simpleksów powtarzających się co najmniej 25% w danej populacji. Na wykresie 11.7 widoczna jest rozbieżność pomiędzy kolumnami zielonymi a linią trendu. W tym przypadku, najniższą wartość występowania 32,23% odnotowano w latach 1940-tych, a najwyższą 35,91% w latach 2000-nych (analogicznie do struktur Murcko).

Tabela 11.5. Struktury wygenerowanych fragmentów wraz z odpowiadającą im numeracją.

Lp.	*Metoda Simpleksów*
1	
2	
3	
4	
5	
6	

Podsumowując, w niniejszym rozdziale opisano studium fragonomiki leków NME w oparciu o metodę MCS, szkielety Murcko oraz metodę simpleksów. Największą liczbę fragmentów uzyskano w zmodyfikowanej metodzie Kuz'mina. Spośród najpopularniejszych simpleksów wyróżniono benzen, toluen oraz cztery fragmenty alifatyczne. Wszystkie szkielety wygenerowane metodą Murcko posiadały sześcioczłonowy cykliczny pierścień, natomiast trzy z pięciu (60%) motyw toluenu. W podgrupie MCS zidentyfikowano 21 (87,50%) cząsteczek zawierających benzen, 12 (50%) z motywem toluenu oraz 3 fragmenty alifatyczne (1 łańcuchowy a 2 pierścieniowe).

12 Podsumowanie

Poszukiwanie nowych leków, odkrywanie zależności warunkujących ich aktywność leży w centrum zainteresowania chemii i przemysłu farmaceutycznego. Organizacja przemysłu farmaceutycznego decyduje o fakcie, że istnieje potrzeba poszukiwania innowacyjnych technik pozwalających zwiększyć efektywność projektowania leków. Pomimo ciągłego postępu technologicznego praca nad nowym lekiem jest procesem długim a przede wszystkim wymagającym dużego nakładu finansowego. W literaturze opisano wiele technik pozwalających zoptymalizować lub przyśpieszyć ten proces, z których coraz większy odsetek stanowią komputerowe metody *in silico*. Modelowanie i symulacje są wykorzystywane na każdym etapie badań nad farmaceutykiem.

W niniejszej pracy przeprowadzono szereg analiz z wykorzystaniem metod programowania i obliczeń w oparciu o dane dostępne w bazach (np. World Bank, Eurostat, Drugs.com, FDA, PubChem, ChEMBL). Badania na które składa się niniejsza rozprawa doktorska przebiegały według następującej kolejności:

1. **Określono efektywność publikacyjną jednostek badawczych.**
   W szczególności przeanalizowano zależność pomiędzy liczbą publikacji w Nature a wydatkami na naukę przez indywidualne państwa i uniwersytety. Wykorzystano parametry ekonomiczne (finansowanie) jako wskaźnik naukowego sukcesu i prestiżu.
   - Załącznik 2
     Bogocz J, Bak A, Polanski J. (2014) No free lunches in nature? An analysis of the regional distribution of the affiliations of Nature publications. *Scientometrics*. 101:547-568. (IF=2,084; pkt KEJN=35)

2. **Zaproponowano nową metodę miary wieku leku w oparciu o datę rejestracji regulatora rynku amerykańskiego (agencji FDA).**
   Dokonano eksploracji baz molekularno-strukturalnych w tym internetowego zestawienia najlepiej sprzedających się leków na rynku amerykańskim. Opracowano nową metodę pomiaru wieku leku. Wykorzystując metodę odkrywania wiedzy (na podstawie listy bestselerów oraz wieku leku), zauważono powiązanie spadku produktywności przemysłu farmaceutycznego z problemem starzenia się leków.

- Załącznik 3

  Polanski J, Bogocz J, Tkocz A. (2015) Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. *Drug Discovery Today*. 20:1300-1304. (IF=5,625; pkt KEJN=45)

3. **Wykonano wieloczynnikową analizę rynkowego sukcesu leków**

   W oparciu o kryteria lekopodobieństwa oraz parametry ekonomiczne oszacowano produktywność oraz konkurencyjność na rynku farmaceutyków. Powiązano koncepcję "slim pharmy" oraz "sweet spot" z sukcesem rynkowym. Porównano listę stu najlepiej sprzedających się leków z całą populacją rejestracji regulatora FDA, w tym celu również zaproponowano nowy parametr zwany zmianą wieku leku. Przeanalizowano wpływ pIC_{50} na sukces rynkowy leku.

   - Załącznik 4

     Polanski J, Bogocz J, Tkocz A. (2016) The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *Journal of Computer-Aided Molecular Design.* 30:381-389. (IF=3,199; pkt KEJN=30)

4. **Studium fragonomiki leków zarejestrowanych przez regulatora FDA**

   Bazując na strukturach znanych leków z lat 1939-2014 zaproponowano wykorzystanie metod fragmentarycznych w analizach architektury aktywnych związków. Badania przeprowadzono w oparciu o metody: MCS, Murcko, Simpleks.

Zaobserwowano wiele ciekawych zależności i trendów m.in. zmian jakie zaszły w projektowaniu farmaceutyków na przestrzeni ostatnich lat. Przeprowadzone badania rzuciły nowe spojrzenie na znane procedury projektowania struktur potencjalnych bioefektorów.

13 Dorobek naukowy

13.1 Spis publikacji

1. Bogocz J, Bak A, Polanski J. (2014) No free lunches in nature? An analysis of the regional distribution of the affiliations of Nature publications. *Scientometrics*. 101:547-568.

2. Polanski J, Bogocz J, Tkocz A. (2015) Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. *Drug Discovery Today*. 20:1300-1304.

3. Polanski J, Bogocz J, Tkocz A. (2016) The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *Journal of Computer-Aided Molecular Design.* 30:381-389.

4. J. Bogocz, A. Tkocz, R. Rzycka, et al. (2015) Analiza profilu leków dopuszczonych przez FDA w latach 2003-2012. *Nauka i wiedza kluczem do poznania świata*. Słupsk, Polska. ISBN 978-83-63216-01-6.

5. J. Bogocz, J. Polanski. (2013) Polifarmakologia i lekotypia pochodnych tiosemikarbazonu. *Dokonania naukowe doktorantów*. Kraków, Polska. ISBN 978-83-63058-34-0.

13.2 Spis prezentacji konferencyjnych

1. J. Bogocz. (2015) Innowacyjne metody eksploracji baz danych w poszukiwaniu nowych reguł projektowania leków. *IV ogólnopolska konferencja naukowa pt. Pomiędzy Naukami – zjazd fizyków i chemików*. Chorzów, Polska.

2. J. Bogocz, A. Tkocz, J. Polanski. (2015) Top 100 drug bestsellers are getting older. *The XXXVIIIth Symposium: Chromatographic methods of investigating the organic compounds*. Szczyrk, Polska.

3. J. Bogocz, A. Tkocz, J. Polanski. (2015) Simpleksowa reprezentacja struktury molekularnej. *Nauka i biznes, czyli dwa przenikające się światy*. Katowice, Polska.

4. J. Bogocz, A. Tkocz, J. Polanski. (2014) Studium fragonomiki w oparciu o metodę simpleksów. *Nauka i wiedza kluczem do poznania świata*. Kraków, Polska.

5. J. Bogocz, A. Tkocz, J. Polanski. (2014) The simplex method of dolutegravir. *18th Gliwice Scientific Meetings*. Gliwice, Polska.

6. J. Bogocz, A. Tkocz, J. Polanski. (2014) Application of simplex method in drug design. *7th Central Europe Conference, Chemistry towards Biology*. Katowice, Polska.

7. J. Bogocz, A. Tkocz, J. Polanski. (2014) Studium fragonomiki w oparciu o metodę simpleksów. *III ogólnopolska konferencja naukowa pt. Pomiędzy Naukami – zjazd fizyków i chemików*. Chorzów, Polska.

8. J. Bogocz, J. Polanski. (2013) Mining chemical database for the analysis of fragmental drug-likeness topology. Application for thiosemicarbazone derivatives. *YoungChem2013 International Congress of Young Chemists*. Poznań, Polska.

9. J. Bogocz, J. Polanski. (2013) Structure activity relationship of thiosemicarbazone derivatives. *The XXXVIth Symposium: Chromatographic methods of investigating the organic compounds*. Szczyrk, Polska.

10. A. Tkocz, J. Bogocz, J. Polanski. (2015) An analysis of fragmental drug-likeness topology. *The XXXVIIIth Symposium: Chromatographic methods of investigating the organic compounds*. Szczyrk, Polska.

14 Spis ilustracji

15 Spis tabel

16 Literatura

1. Brown FK. (1998) Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery. Annual Reports in Medicinal Chemistry. James AB, Editor. *Academic Press*. 375-384.

2. Polanski J, Bak A. (2010) Podstawy chemoinformatyki lekow. *Uniwersytet Slaski*.

3. Ertl P, Schuffenhauer A. (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*. 1:1-11.

4. Polanski J, et al. (2016) Molecular descriptor data explain market prices of a large commercial chemical compound library. *Scientific Reports*. 6:(28521):1-10.

5. Zoete V, Grosdidier A, Michielin O. (2009) Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of Cellular and Molecular Medicine*. 13:238-248.

6. Shoichet BK. (2004) Virtual screening of chemical libraries. *Nature*. 432:862-865.

7. Walters WP, Stahl MT, Murcko MA. (1998) Virtual screening - an overview. *Drug Discovery Today*. 3:160-178.

8. Hann MM. (2011) Molecular obesity, potency and other addictions in drug discovery. *Medicinal Chemical Communications*. 2:349-355.

9. Sliwoski G, et al. (2014) Computational Methods in Drug Discovery. *Pharmacological Reviews*. 66:334-395.

10. White RE. (1998) Short-and long-term projections about the use of drug metabolism in drug discovery and development. *Drug Metabolism and Disposition*. 26:1213-1216.

11. Dickson M, Gagnon JP. (2004) The cost of new drug discovery and development. *Discovery Medicine*. 4:172-179.

12. Hann MM, Keseru GM. (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nature Reviews Drug Discovery*. 11:355-365.

13. Collier R. (2009) Drug development cost estimates hard to swallow. *Canadian Medical Association Journal*. 180:279-280.

14. van de Waterbeemd H, Gifford E. (2003) ADMET in silico modelling: Towards prediction paradise? *Nature Reviews Drug Discovery*. 2:192-204.

15. Lipinski CA. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*. 44:235-249.

16. Lipinski CA, et al. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*. 23:3-25.

17. DiMasi JA. (2001) New drug development in the United States from 1963 to 1999. *Clinical Pharmacology & Therapeutics*. 69:286-296.

18. Ciociola AA, Cohen LB, Kulkarni P. (2014) How drugs are developed and approved by the FDA: current process and future directions. *The American Journal of Gastroenterology*. 109:620.

19. Li AP. (2001) Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today*. 6:357-366.

20. Gleeson MP, et al. (2011) Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nature Reviews Drug Discovery*. 10:197-208.

21. EUPATI. (2015) Key principles of pharmacology. Dostępne na: www.eupati.eu

22. Testa B, et al. (2001) Pharmacokinetic optimization in drug research. *Wiley-VCH*.

23. Bianconi E, et al. (2013) An estimation of the number of cells in the human body. *Annals of Human Biology*. 40:463-471.

24. Palm K, et al. (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research*. 14:568-571.

25. Stenberg P, et al. (2002) Theoretical Predictions of Drug Absorption in Drug Discovery and Development. *Clinical Pharmacokinetics*. 41:877-899.

26. Kawabata Y, et al. (2011) Formulation design for poorly water-soluble drugs based on biopharmaceutics classification system: basic approaches and practical applications. *International Journal of Pharmaceutics*. 420:1-10.

27. Bartzatt R, Malesa C. (2003) Synthesis, structural analysis and antibacterial activity of a butyl ester derivative of ampicillin. *Chemotherapy*. 49:213-221.

28. World Drug Index. *Thomson Reuters.* Dostępne na: www.thomsonreuters.com

29. Stenberg P. (2001) Computational models for the prediction of intestinal membrane permeability.

30. Molinspiration - Cheminformatics on the Web. *Molinspiration Cheminformatics.* Dostępne na: www.molinspiration.com

31. Ertl P. (2010) Molecular structure input on the web. *Journal of Cheminformatics*. 2:1-9.

32. Abraham MH, et al. (2002) Application of hydrogen bonding calculations in property based drug design. *Drug Discovery Today*. 7:1056-1063.

33. Aihara J-i. (1999) Reduced HOMO-LUMO gap as an index of kinetic stability for polycyclic aromatic hydrocarbons. *The Journal of Physical Chemistry A*. 103:7487-7495.

34. Oprea TI, et al. (2001) Is There a Difference between Leads and Drugs? A Historical Perspective. *Journal of Chemical Information and Computer Sciences*. 41:1308-1315.

35. Patrick GL. (2013) An introduction to medicinal chemistry. *Oxford University Press*.

36. BIO-HPC. (2016) Prediction studies of the biological activity of chemical substances by QSAR methods. *Bioinformatics and High Performance Computing Research Group.* Dostępne na: www.bio-hpc.eu

37. Nikolova N, Jaworska J. (2005) Review of the methods for assessing the applicability domains of SARS and QSARS. *Institute for Health & Consumer Protection - ECVAM*:1-38.

38. Bielenica A, E Koziol A, Struga M. (2013) Computational Methods in Determination of Pharmacophore Models of 5-HT1A 5-HT2A and 5-HT7 Receptors. *Mini-Reviews in Medicinal Chemistry*. 13:933-951.

39. Milne G, Nicklaus M, Wang S. (1998) Pharmacophores in drug design and discovery. *SAR and QSAR in Environmental Research*. 9:23-38.

40. Al-Balas QA, et al. (2013) Virtual lead identification of farnesyltransferase inhibitors based on ligand and structure-based pharmacophore techniques. *Pharmaceuticals*. 6:700-715.

41. Micheel CM, Nass SJ, Omenn GS. (2012) Evolution of translational omics: lessons learned and the path forward. *National Academies Press*.

42. Sams-Dodd F. (2013) Is poor research the cause of the declining productivity of the pharmaceutical industry? An industry in need of a paradigm shift. *Drug Discovery Today*. 18:211-217.

43. Kalyaanamoorthy S, Chen Y-PP. (2011) Structure-based drug design to augment hit discovery. *Drug Discovery Today*. 16:831-839.

44. Drwal MN, Griffith R. (2013) Combination of ligand-and structure-based methods in virtual screening. *Drug Discovery Today*. 10:395-401.

45. Hopkins AL. (2008) Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*. 4:682-690.

46. Sams-Dodd F. (2005) Target-based drug discovery: is something wrong? *Drug Discovery Today*. 10:139-147.

47. Arrowsmith J. (2011) Phase II failures: 2008-2010. *Nature Reviews Drug Discovery*. 10:328-329.

48. Swinney DC, Anthony J. (2011) How were new medicines discovered? *Nature Reviews Drug Discovery*. 10:507-519.

49. Bohacek RS, McMartin C, Guida WC. (1996) The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*. 16:3-50.

50. Ertl P. (2003) Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *Journal of Chemical Information and Modeling*. 43:374-380.

51. Reymond J-L. (2015) The chemical space project. *Accounts of Chemical Research*. 48:722-730.

52. Gaulton A, et al. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*. 40:1100-1107.

53. Pammolli F, Magazzini L, Riccaboni M. (2011) The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*. 10:428-438.

54. Miller BA, Arcolano N, Bliss NT. (2013) Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data. *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*. IEEE.

55. Chen H, Chiang RH, Storey VC. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*. 36:1165-1188.

56. Fayyad U, Piatetsky-Shapiro G, Smyth P. (1996) From data mining to knowledge discovery in databases. *AI Magazine*. 17:37-54.

57. Morphy R, Rankovic Z. (2007) Fragments, network biology and designing multiple ligands. *Drug Discovery Today*. 12:156-160.

58. Keiser MJ, et al. (2009) Predicting new molecular targets for known drugs. *Nature*. 462:175-181.

59. Perez-Nueno VI, et al. (2011) Predicting drug polypharmacology using a novel surface property similarity-based approach. *Journal of Cheminformatics*. 3:O19.

60. Cockburn IM. (2007) Is the pharmaceutical industry in a productivity crisis? Innovation Policy and the Economy. *MIT Press*. 1-32.

61. Fuoco D. (2015) Hypothesis for changing models: current pharmaceutical paradigms, trends and approaches in drug discovery. *PeerJ PrePrints*:1-9.

62. Sweet BV, Schwemm AK, Parsons DM. (2011) Review of the processes for FDA oversight of drugs, medical devices, and combination products. *Journal of Managed Care Pharmacy*. 17:40-50.

63. Kinch MS, et al. (2014) An overview of FDA-approved new molecular entities: 1827–2013. *Drug Discovery Today*. 19:1033-1039.

64. Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations. *US Food and Drug Administration.* Dostępne na: www.fda.gov

65. Ma'ayan A, et al. (2007) Network analysis of FDA approved drugs and their targets. *Mount Sinai Journal of Medicine*. 74:27-32.

66. Kinch MS, Hoyer D. (2015) A history of drug development in four acts. *Drug Discovery Today*. 20:1163-1168.

67. Kinch MS, et al. (2014) An analysis of FDA-approved drugs for infectious disease: antibacterial agents. *Drug Discovery Today*. 19:1283-1287.

68. BMJ. (1952) The History of Pneumonia. *British Medical Journal*. 1:156-158.

69. Ballentine C. (1981) Sulfanilamide Disaster: Taste of Raspberries, Taste of Death. The 1937 Elixir Sulfanilamide Incident. *FDA Consumer Magazine.* Dostępne na: www.fda.gov

70. Pekarsky B. (2010) Should Financial Incentives be Used to Differentially Reward 'Me-Too' and Innovative Drugs? *PharmacoEconomics*. 28:1-17.

71. Cohen J, Kaitin K. (2008) Follow-On Drugs and Indications: The Importance of Incremental Innovation to Medical Practice. *American Journal of Therapeutics*. 15:89-91.

72. Gagne JJ, Choudhry NK. (2011) How many "me-too" drugs is too many? *The Journal of the American Medical Association*. 305:711-712.

73. Frothingham R, Relman AS, Lee TH. (2004) "Me-Too" Products - Friend or Foe? [4] (multiple letters). *New England Journal of Medicine*. 350:2100-2101.

74. DiMasi JA, Paquette C. (2004) The economics of follow-on drug research and development. *PharmacoEconomics*. 22:1-14.

75. Davit BM, et al. (2009) Comparing generic and innovator drugs: a review of 12 years of bioequivalence data from the United States Food and Drug Administration. *Annals of Pharmacotherapy*. 43:1583-1597.

76. Kesselheim AS, et al. (2008) Clinical equivalence of generic and brand-name drugs used in cardiovascular disease: a systematic review and meta-analysis. *The Journal of the American Medical Association*. 300:2514-2526.

77. Moore GE. (1965) Cramming More Components onto Integrated Circuits. *Electronics*. 38:114-117.

78. Maruszak B. (2016) Prawo Moore'a. *Encyklopedia Zarządzania "M-files"*. Dostępne na: www.mfiles.pl

79. Scannell JW, et al. (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*. 11:191-200.

80. Agyeman AA, Ofori-Asenso R. (2015) Perspective: Does personalized medicine hold the future for medicine? *Journal of Pharmacy And Bioallied Sciences*. 7:239-244.

81. Roth BL, Sheffler DJ, Kroeze WK. (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery*. 3:353-359.

82. Goldblatt EM, Lee W-H. (2010) From bench to bedside: the growing use of translational research in cancer medicine. *American Journal of Translational Research*. 2:1-18.

83. Fishburn CS. (2013) Translational research: the changing landscape of drug discovery. *Drug Discov Today*. 18:487-494.

84. Reutlinger M, et al. (2014) Multi-Objective Molecular De Novo Design by Adaptive Fragment Prioritization. *Angewandte Chemie International Edition*. 53:4244-4248.

85. Lewell XQ, et al. (1998) RECAP retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*. 38:511-522.

86. Degen J, et al. (2008) On the Art of Compiling and Using'Drug-Like'Chemical Fragment Spaces. *ChemMedChem*. 3:1503-1507.

87. Kuz'min V, Artemenko AG, Muratov EN. (2008) Hierarchical QSAR technology based on the Simplex representation of molecular structure. *Journal of Computer-Aided Molecular Design*. 22:403-421.

88. Kuz'min VE, et al. (2005) Hierarchic system of QSAR models (1D–4D) on the base of simplex representation of molecular structure. *Journal of Molecular Modeling*. 11:457-467.

89. Severance C. (2015) Guido van Rossum: The Early Years of Python. *Computer*. 48:7-9.

90. Lutz M, Ascher D. (2002) Python. Wprowadzenie. *Helion*: Gliwice.

91. Martelli A. (2006) Python in a Nutshell. *O'Reilly Media, Inc.*

92. Van Rossum G. (2007) Python Programming Language. *USENIX Annual Technical Conference*.

93. TIOBE. (2016) TIOBE Programming Community Index. Dostępne na: www.tiobe.com

94. Timmerman H, et al. (2002) Handbook of Molecular Descriptors. *Weinheim: Wiley-VCH*.

95. Gasteiger J. (2003) Handbook of chemoinformatics. *Wiley Online Library* 1.

96. James CA, Weininger D, Delany J. (1995) Daylight theory manual. *Daylight Chemical Information Systems*. 3951.

97. Marburger JH. (2005) Wanted: better benchmarks. *Science*. 308:1087.

98. Bogocz J, Bak A, Polanski J. (2014) No free lunches in nature? An analysis of the regional distribution of the affiliations of Nature publications. *Scientometrics*. 101:547-568.

99. Leydesdorff L, Wagner C. (2009) Macro-level indicators of the relations between research funding and research output. *Journal of Informetrics*. 3:353-362.

100. de Solla Price DJ. (1963) Little science, big science. *New York: Columbia University Press*.

101. NALIMOV V. (1969) Naukometriya: Izuhechinie nauki kak informatsinnogo protessa. Moscú. *Nauka editores*.

102. Granovsky YV. (2001) Is it possible to measure science? VV Nalimov's research in scientometrics. *Scientometrics*. 52:127-150.

103. Garfield E. (2009) From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics*. 3:173-179.

104. Hood W, Wilson C. (2001) The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*. 52:291-314.

105. Garfield E. (2006) Citation indexes for science. A new dimension in documentation through association of ideas. *International Journal of Epidemiology*. 35:1123-1127.

106. Garfield E. (1964) Science Citation Index-A new dimension in indexing. *Science*. 144:649-654.

107. Reuters T. Web of Science. Dostępne na: www.webofscience.com

108. Elsevier S. (2016) Scopus Content Coverage Guide. *Elsevier*. Dostępne na: www.elsevier.com

109. Nature. (2013) Beware the impact factor. *Nature Materials*. 12:89-89.

110. Hirsch JE. (2005) An index to quantify an individual's scientific research output. *PNAS*:16569-16572.

111. Kim S, et al. (2015) PubChem substance and compound databases. *Nucleic Acids Research*:gkv951.

112. Coordinators NR. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 43:8-20.

113. Bento AP, et al. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Research*. 42:1083-1090.

114. Polanski J, Bogocz J, Tkocz A. (2015) Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. *Drug Discovery Today*. 20:1300-1304.

115. Venables W, Ripley B, Venables W. (2002) Modern applied statistics with S. *Springer*.

116. Vanderelst D, Speybroeck N. (2013) Scientometrics reveals funding priorities in medical research policy. *Journal of Informetrics*. 7:240-247.

117. Moran M, et al. (2009) Neglected disease research and development: how much are we really spending? *PLoS Medicine*. 6:e30.

118. Gross CP, Anderson GF, Powe NR. (1999) The Relation between Funding by the National Institutes of Health and the Burden of Disease. *New England Journal of Medicine*. 340:1881-1887.

119. NIH. (2016) Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC). *National Institutes of Health.* Dostępne na: www.report.nih.gov

120. Frame JD, Narin F. (1976) NIH funding and biomedical publication output. *Federation proceedings*.

121. Kaneiwa K, et al. (1988) A comparison between the journals Nature and Science. *Scientometrics*. 13:125-133.

122. Symonds MR. (2004) Nature and Science know best. *Trends in Ecology & Evolution*. 19:564.

123. The World Bank Group. Dostępne na: www.worldbank.org

124. Eurostat (European Commision Database). Dostępne na: www.europa.eu

125. The Center for Measuring University Performance (MUP). Dostępne na: www.mup.asu.edu

126. Nature Publishing Index (NPI).

127. National Science Board (NSB) - NSF. Dostępne na: www.nsf.gov/nsb

128. Paul SM, et al. (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*. 9:203-214.

129. Feldman AM. (2015) Bench-to-Bedside; Clinical and Translational Research; Personalized Medicine; Precision Medicine-What's in a Name? *Clinical and Translational Science*. 8:171-173.

130. Chan Kim W, Mauborgne R. (2005) Blue Ocean Strategy Harvard Business School Press.

131. Drugsite Trust (drugs.com). Dostępne na: www.drugs.com

132. Harrison C. (2013) Patent watch. *Nature Reviews Drug Discovery*. 12:14-15.

133. Law V, et al. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*. 42:1091-1097.

134. Kinch MS, Patridge E. (2014) An analysis of FDA-approved drugs for infectious disease: HIV/AIDS drugs. *Drug Discovery Today*. 19:1510-1513.

135. Kinch MS. (2014) An analysis of FDA-approved drugs for oncology. *Drug Discovery Today*. 19:1831-1835.

136. Walters WP, et al. (2011) What do medicinal chemists actually make? A 50-year retrospective. *Journal of Medicinal Chemistry*. 54:6405-6416.

137. Lipinski C, Hopkins A. (2004) Navigating chemical space for biology and medicine. *Nature*. 432:855-861.

138. Ferenczy GG, Keseru GM. (2013) How are fragments optimized? A retrospective analysis of 145 fragment optimizations. *Journal of Medicinal Chemistry*. 56:2478-2486.

139. Leeson PD, Springthorpe B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*. 6:881-890.

140. Polanski J, Bogocz J, Tkocz A. (2016) The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *Journal of Computer-Aided Molecular Design*. 30:381-389.

141. Kalliokoski T, et al. (2013) Comparability of Mixed IC(50) Data – A Statistical Analysis. *PLoS ONE*. 8:e61007.

142. Zartler ER, Shapiro MJ. (2005) Fragonomics: fragment-based drug discovery. *Current Opinion in Chemical Biology*. 9:366-370.

143. O'Boyle NM, et al. (2011) Open Babel: An open chemical toolbox. *Journal of Cheminformatics*. 3:1-14.

144. Boda K, Seidel T, Gasteiger J. (2007) Structure and reaction based evaluation of synthetic accessibility. *Journal of Computer-Aided Molecular Design*. 21:311-325.

145. Gasteiger J. (2015) Cheminformatics: Computing target complexity. *Nature Chemistry*. 7:619-620.

17 Załączniki

17.1 Załącznik 1

Tabela 17.1 Wybrane parametry ekspertyzy dla indywidualnych leków.

Rok	*Nazwa leku*	*Translacyjność*	*Personalizacja*	*Target*	*Innowacyjność "naukowa"*	*Innowacyjność "blue ocean"*	*Innowacyjność "red ocean"*
2003	Fuzeon	3	-2	2	2	2	1
2003	Somavert	3	-2	2	2	2	1
2003	Emend	2	-2	1	1	1	1
2003	Factive	1	-2	2	-1	-2	-1
2003	Iressa	3	0	2	2	2	1
2003	Velcade	3	-2	2	2	2	1
2003	Boniva	1	-2	1	0	-2	0
2003	Uroxatral	1	-2	1	-1	-2	0
2003	Reyataz	2	-2	2	0	-2	1
2003	Emtriva	1	-2	2	-1	-2	1
2003	Aloxi	1	-2	1	0	-2	2
2003	Zavesca	2	0	2	2	2	1
2003	Crestor	1	-2	1	-2	-2	1
2003	Levitra	1	-2	2	0	-2	0
2003	Cubicin	2	-2	2	1	2	1
2003	Radiogardase	0	-2	1	0	0	0
2003	Namenda	1	-2	0	1	2	1
2003	Elestat	1	-2	2	0	0	0
2003	Cialis	1	-2	2	0	-2	2
2003	Plenaxis	2	-2	1	1	0	1
2003	Ertaczo	0	-2	0	0	-2	0
2003	Amevive	2	-2	2	2	2	1
2003	Fabrazyme	3	-2	2	2	2	1
2003	Aldurazyme	3	-2	2	2	2	1
2003	Xolair	3	-2	2	2	0	2
2003	Bexxar	2	-2	1	1	0	1
2003	Raptiva	3	-2	1	1	-2	2
2004	Spiriva	2	-2	2	1	-2	1
2004	Alimta	1	-2	1	0	0	2
2004	Sensipar	3	-2	2	2	0	0
2004	Ketek	3	-2	2	1	-2	2
2004	Chirhostim	2	-2	1	1	-2	0
2004	Apidra	1	-2	2	1	-2	2
2004	Apokyn	0	-2	0	0	-2	0
2004	Vitrase	0	-2	0	0	-2	0
2004	Tindamax	0	-2	0	-2	-2	0
2004	Vidaza	2	-2	1	1	2	1

2004	Xifaxan	1	-2	2	1	-2	0
2004	Sanctura	1	-2	2	0	-2	0
2004	Nutrestore	0	-2	-2	0	0	0
2004	Campral	1	-2	1	1	-2	1
2004	Cymbalta	1	-2	1	0	-2	1
2004	Pentetate calcium trisodium	2	-2	2	0	0	0
2004	Pentetate zinc trisodium	2	-2	2	0	0	0
2004	Fosrenol	2	-2	2	1	1	0
2004	Amphadase	0	-2	2	0	-2	0
2004	Omacor	1	-2	-2	0	0	0
2004	Tarceva	2	0	1	1	-2	0
2004	Vesicare	1	-2	2	0	-2	0
2004	Multihance	1	-2	0	0	0	0
2004	Lunesta	0	-2	1	0	-2	-2
2004	Vision blue	0	-2	0	-2	-2	-2
2004	Macugen	3	-2	2	2	2	1
2004	Enablex	1	-2	2	0	0	2
2004	Prialt	2	-2	1	2	2	2
2004	Clolar	3	-2	1	1	-1	2
2004	Ventavis	2	-2	2	0	-2	0
2004	Lyrica	1	-2	0	1	-2	1
2004	Erbitux	3	2	2	1	1	1
2004	Avastin	3	2	2	2	2	1
2004	Neutrospec	1	-2	1	1	1	0
2004	Tysabri	3	-2	1	2	2	1
2004	Kepivance	3	-2	2	2	2	1
2005	Symlin	2	-2	2	2	2	1
2005	Mycamine	2	-2	2	1	-2	2
2005	Baraclude	2	-2	2	0	0	2
2005	Byetta	2	-2	2	2	2	1
2005	Tygacil	1	-2	1	0	-2	2
2005	Levemir	2	-2	2	1	-2	2
2005	Aptivus	3	-2	2	1	0	1
2005	Rozerem	3	-2	2	2	2	1
2005	Nevanac	1	-2	1	0	1	1
2005	Increlex	2	-2	2	0	1	2
2005	Hydase	0	-2	2	0	-2	0
2005	Arranon	2	-2	1	1	-2	2
2005	Exjade	1	-2	1	0	-2	2
2005	Hylenex	0	-2	2	0	-2	0
2005	Iplex	2	-2	2	0	1	1
2005	Nexavar	3	-2	2	2	2	1
2005	Revlimid	2	-2	0	1	-2	2
2005	Vaprisol	3	-2	-2	2	2	1
2005	Naglazyme	3	-2	2	2	2	1
2005	Orencia	3	-2	2	2	1	1
2006	Sutent	3	-2	2	2	2	1
2006	Ranexa	2	-2	0	1	1	2

2006	Amitiza	3	-2	2	2	2	1
2006	Eraxis	2	-2	1	1	-2	2
2006	Dacogen	2	-2	1	1	1	2
2006	Chantix	3	-2	2	2	2	1
2006	Azilect	2	-2	2	0	0	2
2006	Prezista	3	-2	2	1	1	2
2006	Sprycel	3	0	2	1	0	2
2006	Anthelios sx	0	-2	0	0	-2	2
2006	Noxafil	1	-2	1	0	-2	1
2006	Pylera	0	-2	2	0	-2	0
2006	Zolinza	3	-2	2	2	1	2
2006	Januvia	3	-2	2	2	1	2
2006	Omnaris	2	-2	1	1	-2	2
2006	Tyzeka	3	-2	2	1	0	2
2006	Veregen	0	-2	0	0	-2	0
2006	Invega	1	-2	0	0	0	2
2006	Myozyme	3	-2	2	2	2	1
2006	Lucentis	2	-2	2	1	-1	2
2006	Elaprase	3	-2	2	2	2	1
2006	Vectibix	2	0	2	1	2	1
2007	Vyvanse	1	-2	1	0	0	1
2007	Tekturna	2	-2	2	1	1	2
2007	Tykerb	3	2	2	1	2	1
2007	Altabax	2	-2	2	1	2	1
2007	Neupro	1	-2	1	0	-1	1
2007	Torisel	3	-2	2	1	1	2
2007	Letairis	1	-2	1	1	-2	2
2007	Selzentry	3	0	2	1	2	1
2007	Ammonia n13	1	0	0	0	0	0
2007	Somatuline depot	1	-2	2	0	-2	1
2007	Doribax	1	-2	1	0	-2	1
2007	Isentress	2	-2	2	1	2	1
2007	Ixempra	1	-2	1	-1	-2	0
2007	Tasigna	3	1	2	1	0	2
2007	Kuvan	3	2	2	2	2	1
2007	Bystolic	1	-2	1	0	-2	2
2007	Soliris	3	0	2	2	2	1
2007	Mircera	2	-2	2	0	-2	1
2008	Intelence	3	-2	2	0	-2	0
2008	Arcalyst	2	-2	2	2	2	2
2008	Pristiq	0	-2	0	-2	-2	0
2008	Treanda	0	-2	0	-2	-2	1
2008	Lexiscan	1	-2	2	1	0	2
2008	Cimzia	1	-2	1	0	-2	1
2008	Relistor	3	-2	2	2	2	1
2008	Entereg	3	-2	2	2	2	1
2008	Durezol	2	-2	2	1	0	2
2008	Eovist	1	-2	1	1	1	1
2008	Cleviprex	2	0	2	0	-1	2
2008	Xenazine	0	-2	0	0	-2	0

2008	Nplate	3	-2	2	2	2	1
2008	Adreview	2	0	1	0	-1	1
2008	Rapaflo	2	-2	2	1	2	1
2008	Vimpat	1	-2	1	1	-2	1
2008	Toviaz	1	-2	2	-1	-2	0
2008	Banzel	1	-2	0	1	-1	2
2008	Promacta	2	-2	2	1	1	2
2008	Nucynta	2	-2	1	1	-2	2
2008	Lusedra	1	-2	0	0	0	1
2008	Mozobil	3	-2	2	2	2	1
2008	Ablavar	2	-2	1	1	0	2
2008	Firmagon	1	-2	2	0	0	0
2009	Savella	1	-2	1	0	0	0
2009	Uloric	2	-2	2	2	-1	1
2009	Afinitor	3	-2	2	2	1	2
2009	Coartem	1	-2	0	0	-2	2
2009	Ulesfia	2	-2	1	1	0	1
2009	Fanapt	1	-2	0	0	-2	-2
2009	Samsca	3	-2	2	2	0	2
2009	Besivance	1	-2	1	0	-2	1
2009	Multaq	2	-2	0	1	-2	2
2009	Effient	2	-2	2	1	-1	2
2009	Onglyza	3	-2	2	2	0	2
2009	Livalo	2	-2	2	0	0	0
2009	Saphris	0	-2	0	-2	-2	-1
2009	Sabril	0	-2	0	-2	-2	0
2009	Bepreve	0	-2	-2	2	-2	0
2009	Vibativ	2	-2	2	1	1	2
2009	Folotyn	2	-2	1	1	0	1
2009	Votrient	2	-2	1	1	0	1
2009	Istodax	1	-2	1	1	-1	2
2009	Qutenza	2	-2	2	1	2	1
2009	Simponi	1	-2	1	0	-2	1
2009	Dysport	1	-2	2	0	-2	1
2009	Ilaris	2	-2	1	1	-2	2
2009	Stelara	3	-2	2	2	2	1
2009	Arzerra	2	-2	2	2	0	1
2009	Kalbitor	3	-2	2	2	2	1
2010	Ampyra	1	-2	1	0	0	0
2010	Victoza	2	-2	2	1	0	2
2010	Vpriv	2	0	2	1	-2	2
2010	Carbaglu	3	0	2	2	2	1
2010	Asclera	2	-2	1	0	-2	1
2010	Natazia	2	-2	2	0	-1	1
2010	Jevtana	3	0	2	1	-2	2
2010	Lastacaft	1	-2	2	-1	-2	0
2010	Ella	2	-2	2	1	1	1
2010	Gilenya	3	-2	2	2	2	1
2010	Pradaxa	3	0	2	2	0	2
2010	Latuda	1	-2	0	0	-2	1

2010	Teflaro	3	0	2	1	-2	2
2010	Egrifta	3	0	2	2	2	1
2010	Halaven	1	-2	1	0	-2	2
2010	Actemra	3	0	2	2	2	1
2010	Xiaflex	2	0	2	1	1	1
2010	Lumizyme	0	0	2	0	-2	1
2010	Prolia	3	0	2	2	2	1
2010	Xeomin	1	0	2	0	-2	2
2010	Krystexxa	3	-2	2	2	2	1
2011	Datscan	1	-2	1	0	-2	-1
2011	Natroba	1	-2	1	0	0	0
2011	Viibryd	1	-2	1	0	-2	0
2011	Edarbi	1	-2	2	-1	-2	1
2011	Daxas	2	0	2	2	1	2
2011	Benlysta	3	-2	2	2	2	1
2011	Gadavist	0	-2	0	-1	-2	0
2011	Yervoy	2	0	2	2	2	1
2011	Horizant	1	-2	-1	0	-2	1
2011	Caprelsa	3	0	2	2	2	1
2011	Zytiga	3	0	2	2	2	1
2011	Tradjenta	2	0	2	0	0	2
2011	Edurant		0	2	1	0	2
2011	Victrelis	3	1	2	2	2	1
2011	Incivek	3	1	2	2	2	1
2011	Dificid	3	0	2	2	2	1
2011	Potiga	2	-2	2	2	2	1
2011	Nulojix	3	0	2	1	0	1
2011	Arcapta	1	-2	2	-1	-2	1
2011	Xarelto	3	-2	2	2	2	1
2011	Brilinta	3	-2	2	2	2	1
2011	Zelboraf	3	2	2	2	2	1
2011	Adcetris	3	0	2	2	2	1
2011	Firazyr	3	-2	2	1	2	1
2011	Xalkori	3	2	2	2	2	1
2011	Ferriprox	1	-2	1	0	0	1
2011	Onfil	1	-2	0	1	-2	-2
2011	Jakafi	3	0	2	2	2	1
2011	Erwinaze	1	0	1	-1	-2	1
2011	Eylea	3	-2	2	0	-2	2
2012	Amyvid	0	0	1	1	1	1
2012	Choline c11	1	1	0	0	1	1
2012	Erivedge	3	1	2	2	2	1
2012	Kalydeco	3	2	2	2	2	1
2012	Perjeta	3	2	2	2	2	1
2012	Stivarga	1	0	1	0	-2	1
2012	Voraxaze	2	-1	1	1	1	1
2012	Xtandi	2	-1	2	0	0	2
2012	Zaltrap	2	0	2	0	-2	1
2012	Belviq	2	-2	2	1	1	2
2012	Inlyta	2	0	1	1	-2	1

2012	Kyprolis	2	0	2	1	0	2
2012	Linzess	3	0	2	2	2	1
2012	Myrbertiq	1	-2	2	1	0	2
2012	Neutroval	0	-2	1	-2	-2	2
2012	Omontys	0	-2	2	-2	-2	2
2012	Picato	0	-2	-2	0	0	0
2012	Prepopik	0	-2	0	-2	-2	-2
2012	Stendra	1	-2	2	0	-2	0
2012	Stribild	0	-2	2	-2	-2	-2
2012	Surfaxin	0	-2	0	-1	-1	-1
2012	Tudorza	1	-2	2	-1	-2	1
2012	Zioptan	1	-2	1	-1	-2	0
2012	Elelyso	0	0	2	1	-2	2
2012	Bosulif	1	0	2	1	-1	1
2012	Aubagio	0	0	0	0	0	1
2012	Jetrea	1	-2	1	2	2	1
2012	Fycompa	3	-2	2	2	2	1
2012	Synribo	0	-2	-2	-2	-2	-2
2012	Xeljanz	2	0	1	2	-1	1
2012	Cometriq	2	0	1	1	0	0
2012	Iclusig	3	1	2	1	-1	2
2012	Signifor	3	0	2	2	0	1
2012	Raxibacumab	2	0	2	2	2	2
2012	Gattex	3	0	2	2	2	1
2012	Juxtapid	2	0	0	1	1	1
2012	Eliquis	2	1	2	1	0	1
2012	Sirturo	3	0	2	2	2	1
2012	Fulyzaq	2	0	2	1	1	2

17.2 Załącznik 2

Kopia publikacji naukowej:

Bogocz J, Bak A, Polanski J. (2014) No free lunches in nature? An analysis of the regional distribution of the affiliations of Nature publications. *Scientometrics*. 101:547-568. (IF=2,084; pkt KEJN=35)

Scientometrics (2014) 101:547–568
DOI 10.1007/s11192-014-1252-4

No free lunches in nature? An analysis of the regional distribution of the affiliations of Nature publications

Jacek Bogocz · Andrzej Bak · Jaroslaw Polanski

Received: 1 October 2013 / Published online: 25 February 2014

Abstract Nature is among the world's most highly cited multidisciplinary science journals with one of the highest impact factors of 38.597 (Nature Publishing Group (NPG) 2013), which is used relatively often in many scientific rankings. When analysing the regional distribution of Nature publications, we found a high correlation between the expenditures and the number of local affiliations that are counted on a national basis. The same regularity can be observed for the world's top 30 and the US's top 50 universities; however, the correlation is now skewed by the so-called *cumulative advantage* or the *Matthew Effect*, which evidently rewards those that are ranked at the top of the Academic Ranking of World Universities. *The rich get richer and the poor get poorer*. Surprisingly, the amount of the endowment better determines the number of Nature publications for universities than the total research expenditure.

Keywords Nature publications · Academic Ranking of World Universities · Matthew Effect · Higher education expenditures

Introduction

There has been an enormous interest in the performance metrics for scientific output in recent years. The quantitative aspects of "the science of science and science policy" are the ultimate target. How efficient is science and how much should we invest in it (ITG 2008)? Is it possible to quantify the scientific performance and draw an exact borderline between 'good' and 'outstanding' science? Along with the availability of scientific

Electronic supplementary material The online version of this article (doi:10.1007/s11192-014-1252-4) contains supplementary material, which is available to authorized users.

J. Bogocz · A. Bak · J. Polanski (✉)
Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland
e-mail: polanski@us.edu.pl

databases that provide easy access to the ever-increasing number of publications, scientometrics was developed with the aim of analysing and evaluating them. Although scientometrics is sometimes disregarded by scientists, it can determine the fate of universities and faculties as well as the careers of individuals. 'Publish-or-perish' has been declared a strategy, a requirement and a golden rule of academia that pushes scientists to publish in the most widely known/cited periodicals.

Quantifying scientific output is a serious and controversial problem. The Hirsch Index and the Academic Ranking of World Universities (ARWU) are examples of the classifications that resemble those that are used in sport rankings. The early scientometrics was used specifically for descriptive purposes to provide the parameters for the statistical monitoring of science (Leydesdorff and Wagner 2009a, b). However, the globalisation of science, interactions between business and science, the increasing number of scientific journals, the availability of scientific databases and strong competition have also brought about an active manipulation that is aimed at improving the individual position in rankings, for which citation abuses are probably the best known example. Consequently, a higher position in the rankings attracts investment and also provides better possibilities for further research. Thus, the normative aspect of scientometrics has gained in importance.

We are well aware that financial support is important in science; however, can we quantify this in terms of scientific 'productivity'? How important is to map funding priorities and to tailor them reasonably to gain growth in science? American universities are the traditional winners of Nobel prizes, Nature publications, innovations, etc. The US federal budget for science is more than $150b a year. The resources of US science are enormously high, e.g., a single Stanford University budget is more than $4b a year and it had an endowment of $17b in 2012. This enabled Stanford alone to obtain 17 Nobel prizes. Moreover, there are at least several other strong players in the US team that gambles in scientific roulette. Asia along with China has just come to the table. On the other hand, research funding cliffs endanger science in the US (Garrison 2013) as well as young scientists in Europe (Broberger and Sjostrom Douagi 2012). The number of papers published in mainstream journals (indexed by SCI) is still regarded as the one of the most unambiguous indicators of the research outcome (Guan and Wang 2004). The correlation coefficient between the total number of publications and the level of funding ranged from 0.7 to 0.9 (Venables and Ripley 2002). Similar high correlations were revealed by Vanderelst and Speybroeck (2013), Moran et al. (2009), Gross et al. (1999) and the National Institutes of Health (NIH) (2011), respectively. Vanderelst and Speybroeck (2013) concluded that the burden of a disease was a good predictor for its associated research funding in medical research policy. On the other hand, the number of papers published on a given disease was a measure of the level of funding spent on a specific disease. Frame and Narin (1976) indicated a high correlation coefficient of 0.9 between the number of papers and the level of funding of individual institutes or hospitals. Institutions that receive more funding publish more papers.

In this study, we analysed the relationship between the number of publications in Nature with the expenditures on research by individual countries and universities. Nature is among the world's most highly cited multidisciplinary science journals with one of the highest impact factors of 38.597 (NPG 2013). Kaneiwa et al. (1988) provided a comparison study between Nature and Science. The former covers science developments in the US and Europe, however the latter is more provincial. In this study Nature was used as a measure of a country's research production for the previously mentioned continents. Although recently we are much more cautious about the possible misinterpretation of the journal impact factors that measure the impact of the journal and not the individual contribution,

the number of publications in Nature has been used as an indicator of scientific ranking and as an assessment of outstanding research performance. For many *getting work published in Nature or Science is akin to finding the Holy Grail*, which undoubtedly increases prospects of promotion and funding (Symonds 2004). By analysing the regional distribution of Nature publications, i.e., the number of publications in Nature that are affiliated with a given country or university (Nature), we found a high correlation between the expenditures in higher education and the number of local affiliations counted on a national basis. The same regularity can be observed for the Nature value of the world's top 30 and the US's top 50 universities; however, the correlation is now skewed and evidently rewards those ranked at the top of the ARWU.

Surprisingly, the amount of the endowment and not the total research expenditure better determines the number of Nature publications for universities. It seems that the so-called cumulative advantage or the Matthew Effect can cause this phenomenon. It was shown previously that those who publish in Nature are likely to attract more funds thus considerably increasing their productivity and, as a result, the research output (Symonds 2004).

On the other hand, it is questionable whether to rely solely on a single journal outcome; however, it can serve as a heuristics showing general trends in science (Arkhipov 1999). Basically, the publication practices seem to be discipline-specific (Leydesdorff and Wagner 2009a, b) and can differ considerably among various fields of science and therefore research performance—irrespective of the discipline—cannot be easily comparable. There are at least several other important problems in our analyses. First are limitations in the application of the Nature affiliations as a general descriptor of scientific output. Nature has a focus on certain disciplines, for example, paying less attention to social science and having a bias in favor of medical approaches, which prefers certain universities. Moreover, mapping affiliations to the universities and/or countries is ambiguous because of *double counting* in multiple co-authored publications. This defines the obvious pros and cons of the Nature indicator. On the other hand, there are many problems in collecting the financial data both regarding the availability and comparability, i.e., in the way the statistics on university endowments, annual giving and total research expenditure are arranged.

Data collection

For the analyses that were performed for this article, we queried various types of data, such as: education, research and development expenditures, gross domestic product (GDP), the number of publications in Nature, the rankings of universities, their budgets, endowments and annual gifts, respectively. Economic-based data were collected from the World Bank and Eurostat databases. The World Bank allows access to data across time, country and indicator in order to get a subset of data. The DataBank provides download access to over 8,000 various indicators and over 200 countries and economic profiles. Eurostat publishes statistical information at the European level. Both of these databases contain easily accessible information that can be downloaded in many file formats. Expenditure indicators for universities such as: endowments, annual giving and total research datasets were obtained from the Annual Report on Research Universities in the US (ARRU), which is available at the Center for Measuring University Performance (MUP). The Center for MUP is focused on the competitive national context for major research universities. The ARRU provides classification into groups based on nine quality indicators and collects data on expenditures for research and experimental development (R&D) from over 200 US

academic institutions. gross expenditure on R&D (GERD), higher education expenditure on R&D (HERD) and other expenditure indicators were determined for each country or university by determining its average over the available period of 2001–2011 (HERD and GERD), for endowment and annual giving over the period of 2001–2010 and total research over the period of 2001–2009, respectively. All of the expenditure indicators were converted into US dollars. In our analyses of the performance of a university, we used three expenditure indicators: $\text{endowment}_{2001-2010}$, $\text{annual giving}_{2001-2010}$ and $\text{total research}_{2001-2009}$ as shown in Table 1. We used the Web of Knowledge database (Thomson Reuters) for screening publications in the journal Nature and for the analysis of their regional distribution as well. We probed the time period between 2001 and 2011 estimating both the number of publications in individual years and the cumulative index for 2001–2011 ($\text{Nature}_{2001-2011}$). The Web of Knowledge is a comprehensive interdisciplinary, bibliographic repository with article references from journals, books, etc. and its user friendly search engine allows for effective data mining. We did not differentiate this parameter according to the authors' share within the affiliation in publications and avoided the institutional analysis as well. Therefore, a single affiliation counts for one the same as all of the other affiliation shares, which also count for one; however, we are aware that some top institutions have hospitals and research centres affiliated with them. As a matter of fact the Nature publications were not differentiated according to the specific type of documents, for instance full articles and short correspondences were combined and counted together. Additionally, we used the Nature Publishing Index (NPI) where the total number of publications in Nature are recorded for each of the universities. The NPI ranks institutions according to the number of publications in the Nature journal. The Global Top 200 NPI ranking counts the primary research published in the articles, letters and brief communications sections of Nature by any institution worldwide. The information for the top 200 world's institutions is available online; however, in our case only that for the period 2008–2012 is available ($\text{NPI}_{2008-2012}$).

We included 100 countries from all continents. Countries that did not have publications in Nature were not taken into account. The lack of information in the databases, e.g., public spending on education for China, Jordan, Gabon or Libya is the reason that these countries are not included in some statistics. The data collected concerned the period of time from 2001 to 2011 due to the time delay in data acquisition. Data were collected in May–July 2013. The Pearson correlation coefficient r was used to indicate the relationship between the variables in our analysis.

For the analysis of the performance of a university, we probed the data for the world's top 30 and the US's top 50 universities, which were selected on the basis of the ARWU, a university rank that is available on the internet (http://www.shanghairanking.com/) and the ARRU. ARRU specifies, for example, federal and total research expenditures, annual funds, endowment assets and faculty awards data. ARWU and ARRU were used further to provide an additional benchmark for the analysis of the university affiliations. In order to eliminate the direct contribution of Nature publications in our study; we constructed another rank, AWARD, based on the 2003–2012 period. AWARDs include the total number of the staff of an institution that won Nobel Prizes and Fields Medals, but does not include a number of publications in Nature. The highest score of an institution is 100 and other institutions are calculated as a percentage of the top score—the $\text{AWARD}_{2003-2012}$ data was collected from the ARWU website.

Table 1 Top 50 American research universities with six different measures

L.p.	University name	$ARWU_{2012}$	$AWARD_{2003-2012}$	$Nature_{2003-2012}$	$Endowment_{2003-2010}$ (x $1000)	Annual $giving_{2003-2010}$ (x $1000)	Total $research_{2003-2009}$ (x $1000)
1.	Harvard University	1	1	1395	25,491,382	594,702	433,725
2.	Stanford University	2	6	633	12,153,768	630,549	641,074
3.	Massachusetts Institute of Technology (MIT)	3	3	629	7,392,201	261,458	567,007
4.	University of California Berkeley	4	5	599	2,256,400	233,428	539,000
5.	California Institute of Technology	6	7	439	1,462,057	128,482	263,701
6.	Princeton University	7	4	263	11,873,059	197,380	184,084
7.	Columbia University in the City of New York	8	8	351	5,537,697	365,748	490,633
8.	University of Chicago	9	2	258	4,618,729	217,543	288,421
9.	Yale University	11	11	389	15,663,975	343,282	424,814
10.	University of California Los Angeles	12	12	318	1,532,746	324,381	809,444
11.	Cornell University	13	10	336	3,416,463	288,235	439,059
12.	University of Pennsylvania	14	14	270	4,771,061	383,545	618,654
13.	University of California San Diego	15	17	395	346,806	122,003	721,424
14.	University of Washington	16	18	386	1,810,623	276,010	711,258
15.	Johns Hopkins University	17	21	347	2,133,738	374,247	1,421,522
16.	University of California San Francisco	18	13	373	779,075	260,352	749,780
17.	University of Wisconsin-Madison	19	15	230	1,411,435	345,975	783,992
18.	University of Michigan	22	39	291	5,207,911	240,771	1,791,680
19.	New York University	27	24	186	1,741,400	273,335	265,821
20.	University of Minnesota	29	36	154	2,022,002	256,873	574,604
21.	Northwestern University	30	27	153	4,749,231	188,964	385,382
22.	Washington University in St. Louis	31	23	264	4,334,934	141,479	514,676
23.	Rockefeller University	32	9	275	1,590,424	71,519	204,953
24.	University of Colorado Boulder	33	19	208	292,252	55,229	243,339

Table 1 continued

L.p.	University name	$ARWU_{2012}$	$AWARD_{2003–2012}$	$Nature_{2001–2012}$	$Endowment_{2001–2010}$ (x $1000)	Annual $giving_{2001–2010}$ (x $1000)	Total $research_{2001–2009}$ (x $1000)
25.	University of Texas at Austin	35	29	141	4,043,301	221,134	399,180
26.	Duke University	36	42	248	4,201,218	310,781	611,037
27.	University of Maryland	38	30	162	344,438	85,541	344,088
27.	University of North Carolina at Chapel Hill	41	34	144	1,556,846	218,693	446,173
29.	University of Southern California	46	22	110	2,746,674	393,794	445,521
30.	University of California Davis	47	41	172	435,729	87,576	547,508
31.	University of Texas SW Medical Center - Dallas	48	16	135	995,624	124,175	318,332
32.	Pennsylvania State University	49	40	234	899,281	115,191	541,417
33.	Vanderbilt University	50	20	87	2,689,401	135,448	329,999
34.	Purdue University	56	28	66	1,398,777	158,008	361,428
35.	University of Pittsburgh	58	42	87	1,656,868	103,118	493,224
36.	Rutgers - The State University of New Jersey	61	26	163	471,809	79,714	267,241
37.	Ohio State University	65	42	88	1,967,002	134,914	136,120
38.	Brown University	65	31	94	1,647,059	212,129	582,008
39.	Boston University	71	38	139	838,112	87,830	235,289
40.	Arizona State University	79	32	129	337,447	100,111	185,984
41.	University of Utah	82	37	97	464,714	129,023	244,543
42.	University of Rochester	86	35	57	1,377,686	69,462	327,750
43.	Rice University	91	25	73	3,669,987	71,682	61,836
44.	Case Western Reserve University	99	33	72	1,509,873	88,854	310,738
45.	Baylor College of Medicine	101	42	149	964,409	66,416	443,212
46.	Emory University	104	42	95	4,672,648	154,365	345,758
47.	University of Virginia	197	42	96	3,120,671	224,699	221,541

Table 1 continued

L-p.	University name	$ARWU_{2012}$	$AWARD_{2003-2012}$	$Nature_{2001-2012}$	$Endowment_{2001-2010}$ (x $1000)	Annual $giving_{2001-2010}$ (x $1000)	Total $research_{2001-2009}$ (x $1000)
48.	Mount Sinai School of Medicine	207	42	76	536,415	118,417	241,370
49.	Oregon Health & Science University	210	42	58	332,511	73,680	242,722
50.	University of Colorado Denver	359	42	38	197,409	42,724	241,652

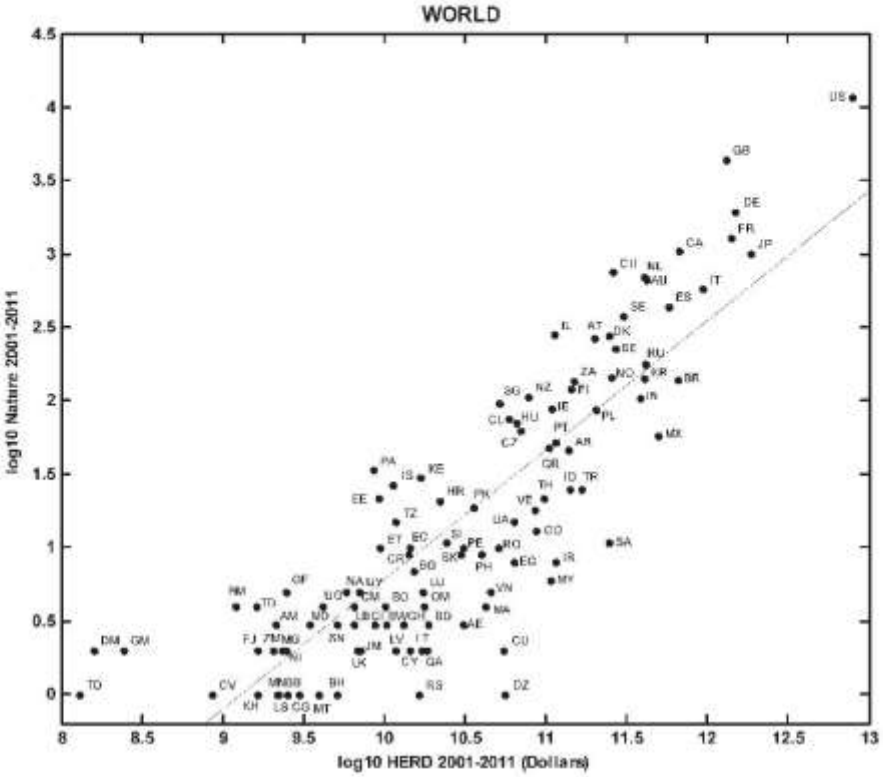


Fig. 1 Relationship between the total number of publications in Nature affiliated by individual countries $\text{Nature}_{2001-2011}$ and expenditures in higher education $\text{HERD}_{2001-2011}$ (average spending during 2001–2011), illustrated in a double logarithmic scale, $n = 105$, $r = 0.96$ (correlation calculated for logarithm untransformed data). *Sources* World Bank Database, Web of Knowledge

Expenditure indicators

We used several indicators to express the amount of expenses. GDP, which is the market value of all officially recognised final goods and services that are produced within a country and GERD, which is the total intramural expenditure on R&D that is performed on the national territory. GERD includes research and development that is performed within a country and funded from abroad but excludes payments for R&D performed abroad, public spending on education as a percentage of GDP in a given year and includes direct public funding for educational institutions and transfers to households and enterprises. Each factor was analysed separately, which made it possible to determine the relationship between variables and to analyse and discuss local differences.

The higher education sector includes all research institutes (universities, colleges, etc.). HERD is an indicator of the total higher education expenditures on research and development that correlates well with the level of education of a given country. High investments in the higher education R&D sector are an indicator of an increase in the production of knowledge. Generally, HERD is expressed as a percentage of GDP. As Nature is an extensive type of parameter, i.e., an additive absolute value quantity related to the whole

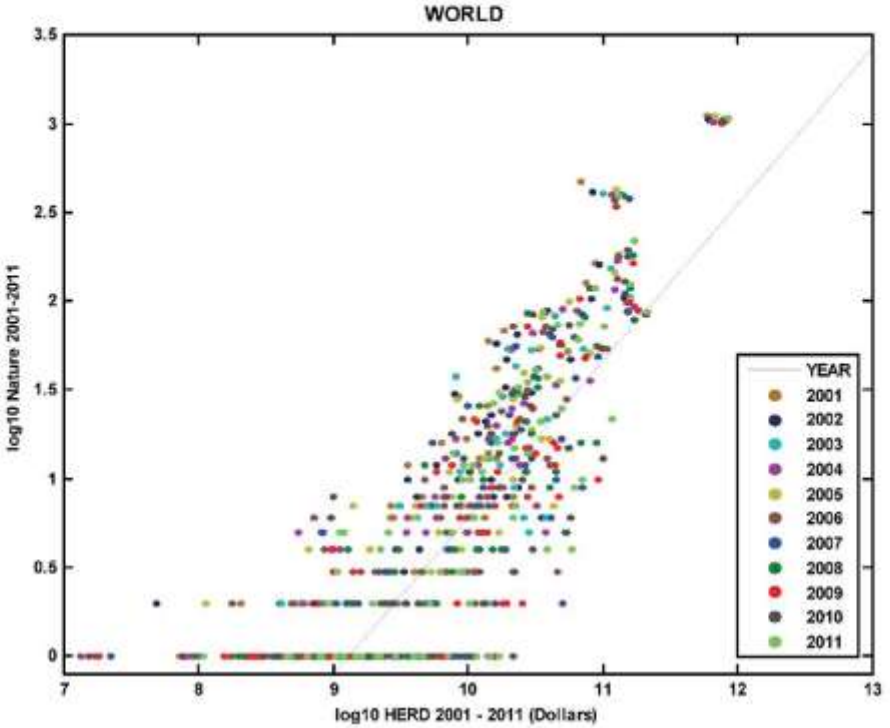


Fig. 2 Relationship between the number of publications in Nature affiliated by individual countries $\text{Nature}_{2001-2011}$ and expenditures in higher education $\text{HERD}_{2001-2011}$, illustrated in a double logarithmic scale; *different colors* indicate individual years, $n = 105$, $r = 0.95$ (correlation calculated for logarithm untransformed data). *Sources* World Bank Database, Web of Knowledge

country or university, we also recalculated the HERD into the absolute value of expenditures in US dollars. If given as a percentage of the GDP, this will provide completely different levels of financial involvement for different countries.

Several indicators can be used to describe the financial support obtained by universities. However, unlike national data not all of these are easily available on a global scale. The total research expenditure is an indicator of the scale of the research. Total research includes all of the expenditures reported by a university to the National Science Board (NSB 2012). Annual giving provides an indicator of the current level of an institution's private contributions both to current expenses and towards increased endowment. Annual gifts are important for universities, because they are a significant factor in balancing the operating budget. Fundraising programs allow an institution to invest in innovative curricular initiatives, attract top students and offer financial aid and scholarships as well. The annual giving data includes all of the contributions that are actually received during the institution's fiscal year in the form of cash, securities, company products and other property from alumni, non-alumni individuals, corporations, foundations, religious organisations and other groups. Not included in the total are public funds, earnings on investments held by the institution and unfulfilled pledges (Council for Aid to Education's VSE Survey).

Fig. 4 Total number of publications Nature$_{2001–2009}$ or Nature$_{2001–2010}$ of top 50 US universities plotted▶ against: endowment$_{2001–2010}$, $r = 0.74$ (**a**), annual giving$_{2001–2010}$, $r = 0.66$ (**b**), total research$_{2001–2009}$, $r = 0.30$; $r = 0.46$, without Harvard No. 1, MIT No. 2 and Stanford No. 3 (**c**). Total research$_{2001–2009}$, shown in each year separately (**d**). *Plus* and *square labels* indicated the university located in the upper and lower half of the top50 AWARD$_{2003–2012}$, respectively, while the *numbers* in labels according to the ARWU$_{2012}$ rank list. *Plus labels* and *lower numbers* tend to gather above the model *straight line*. *Sources* Web of Knowledge, ARRU, ARWU

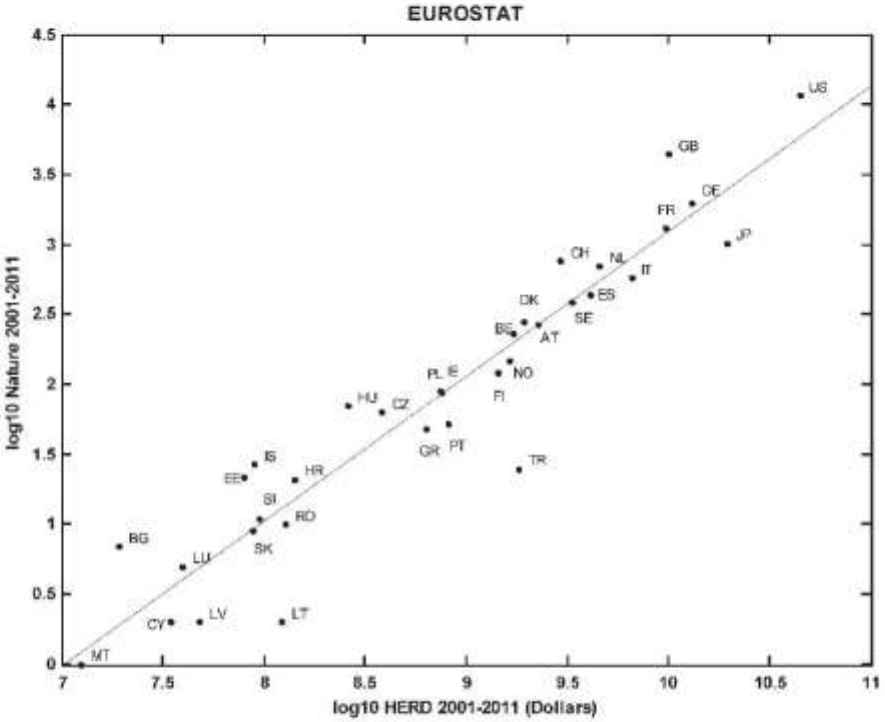
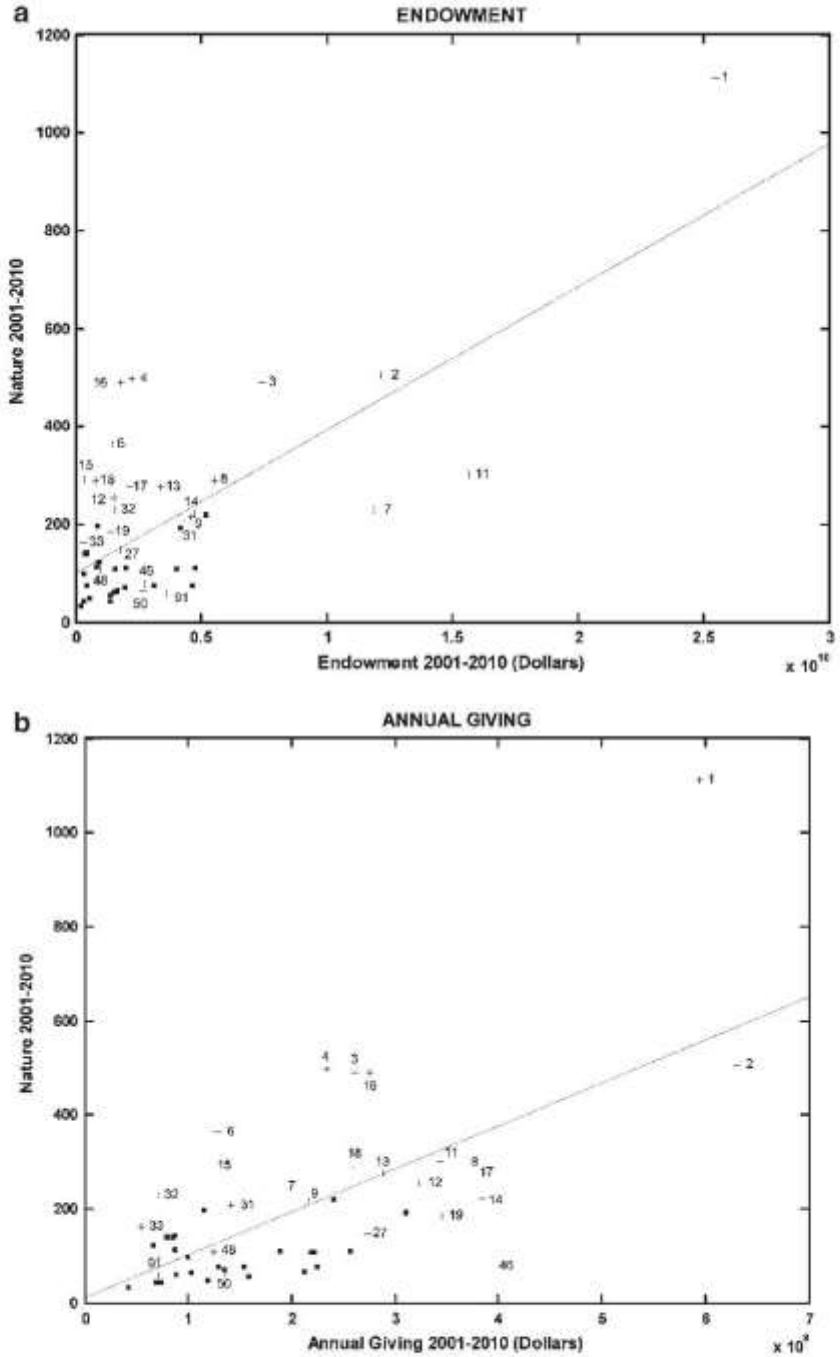


Fig. 3 Relationship between Nature$_{2001–2011}$ and HERD$_{2001–2011}$ for Eurostat covered countries, illustrated in a double logarithmic scale, $n = 32$, $r = 0.92$ (correlation calculated for logarithm untransformed data). *Sources* Eurostat, Web of Knowledge

Endowments reflect the long-term strength of accumulated private support and in some cases institutional savings that deliver an income that can be used for important purposes each year. Generally, an endowment is quite large and an institution often receives multiple endowments that are pooled in a common fund. Typically, the principal of the endowment is invested and the interest is used to fund projects. The practice of investing the principal rather than spending it allows an endowment to grow over time, rather than diminishing by being spent all at once. For universities, this practice especially allows the university to accrue large amounts of wealth, which can then be used to keep pace with other, competitive universities. Moreover, an endowment is used to cover operating expenses, often by taking far more than the normal distribution of 4.5–5 %, it provides stability against downturns in the economy, federal budget cuts and other changes. Endowment assets provide the type of permanence that creates the space for and fuels innovations. Thus,


a
ENDOWMENT
Nature 2001-2010
Endowment 2001-2010 (Dollars)
x 10^10
b
ANNUAL GIVING
Nature 2001-2010
Annual Giving 2001-2010 (Dollars)
x 10^8

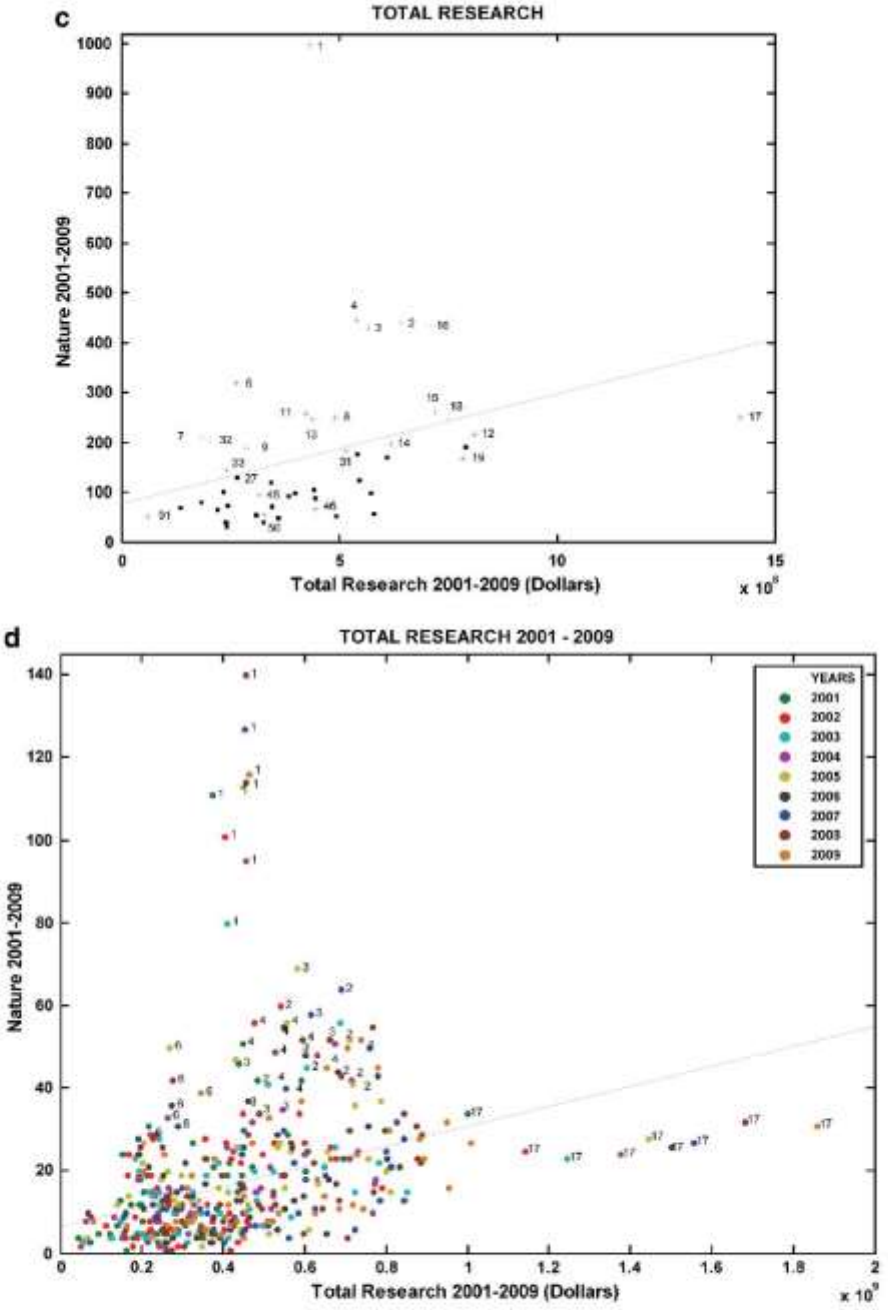
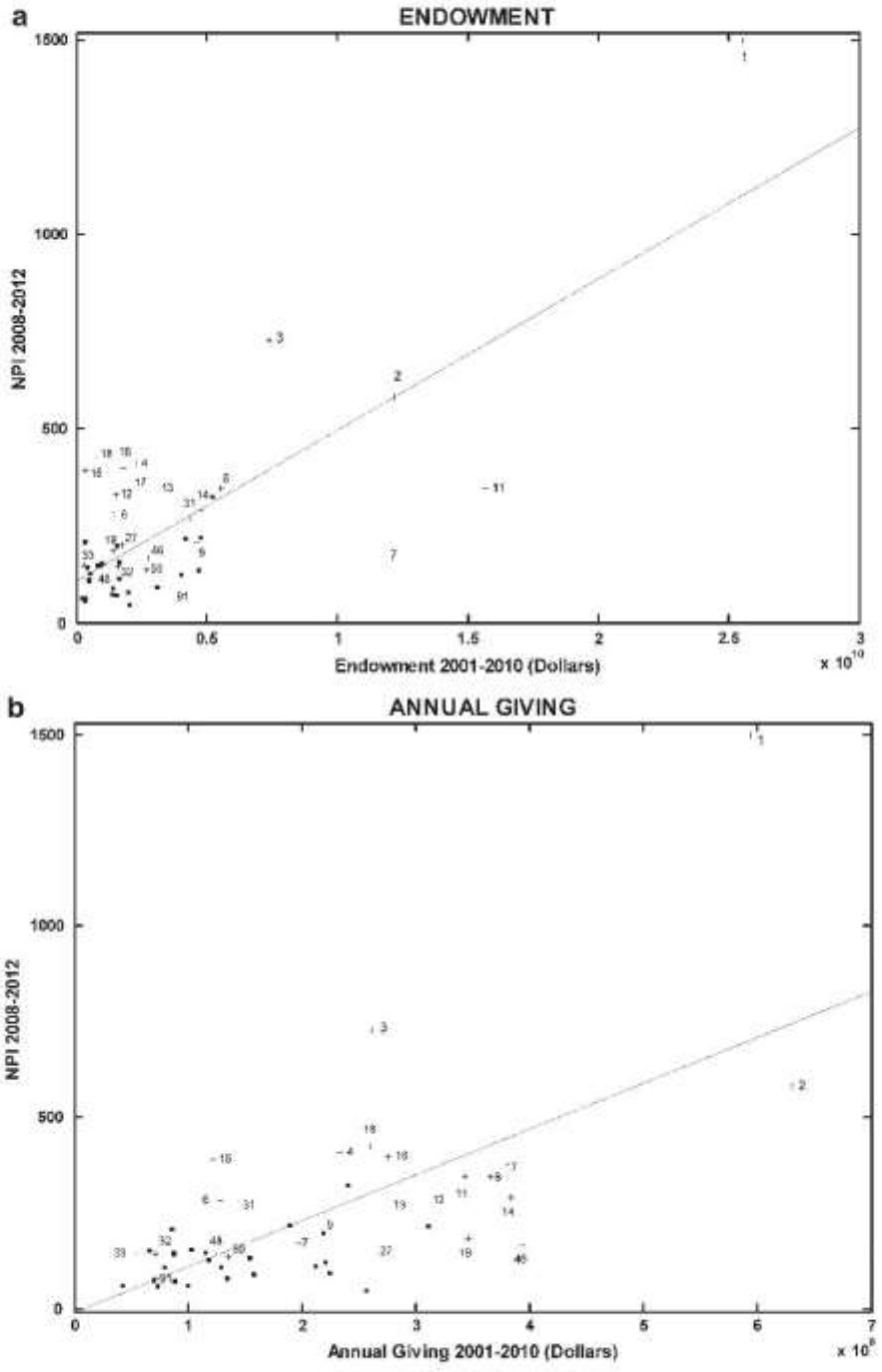


Fig. 4 continued

universities with large endowments are able to pursue critical initiatives that enhance their ability to be leaders in research and higher education. Often an endowment can have a restricted form that supports specific areas within the institution. Endowed professorships or chairs are the best examples illustrating these types of funds. In the United States, endowments are believed to be essential for the financial vigor of educational institutions. Alumni or friends of institution sometimes contribute capital to the endowment. The endowment funding culture is strong in the United States and Canada but is less pronounced overseas with the exceptions of Cambridge and Oxford universities. Endowment funds have also been created to support secondary and elementary school districts in several states in the United States.

Results and discussion

Figure 1 plots parameter $\text{Nature}_{2001-2011}$ by individual countries as a function of $\text{HERD}_{2001-2011}$, expenditures in higher education between 2001 and 2011 or $\text{GERD}_{2001-2011}$, expenditures in R&D between 2001 and 2011 (Fig. 1, supplementary materials) based on the World Bank Database. In Fig. 2 we present a plot that analyses a similar relationship, but for each year separately. The 'comet-like' distribution indicates that a higher level of funding is in line with the number of publications in Nature, which is true for the best countries as shown in the 'comet head'. On the other hand, the 'tail of the comet' depicts players that have a quite different Nature outcome irrespective of the subsidy obtained, which is illustrated by the points scattered across the plot. This also indicated that for a stable Nature output high investment is needed, while a lower subsidy is not a guarantee for a longstanding Nature performance. In Fig. 3 we plot the data according to Eurostat. Eurostat data are highly reliable and uniform for the countries included; however, only a limited number of countries (32) are covered by these statistics. Basically, linear relationships can be observed in Figs. 1–3. In our analyses the regression lines as well as the Pearson correlation values were calculated before applying the logarithmic scale used only for illustrative purposes. In other words, no log transformation was carried out in the protocol of r specification. Logarithmic transformation is one the most frequently used yet controversial procedures in data analysis, which makes the interpretation of results a more challenging process. As 'conventional wisdom' dictates the data should be analysed untransformed, the decisions on transformation are basically subjective. However, a log transformation is sometimes 'theoretically justified' and even explicitly recommended when the standard deviation is proportional to the mean data value—the increment of standard deviation is combined with the increase of the mean value (Keene 1995). On the other hand, a 'golden rule' in statistics is an approach called 'let the data decide' where unnecessary data transformation should be avoided or a rationale for the log preprocessing should be provided, i.e., variables that describe the biochemical measurements typically exhibit a skewed distribution where small responses are more clinically valid than changes to large responses, respectively. Moreover, interpreting the coefficient parameters in a linear regression with log-transformed variables is also not straightforward. There is a positive correlation between local expenditures and the number of local affiliations in Nature publications. The data are correlated better when taken for a longer time period $\text{Nature}_{2001-2011}$ (Fig. 1), than when analysed year by year (Fig. 2). The correlation coefficients range from 0.96 (HERD, Fig. 1) to 0.91 (GERD; Fig. 1, supplementary materials). The correlation for the Eurostat data from $r = 0.92$ (HERD, Fig. 3) to $r = 0.91$


a
ENDOWMENT
NPI 2008-2012
Endowment 2001-2010 (Dollars)
x 10^10
b
ANNUAL GIVING
NPI 2008-2012
Annual Giving 2001-2010 (Dollars)
x 10^8

◀ **Fig. 5** Plot represents correlation between Nature publications $NPI_{2008-2012}$ and: $endowment_{2001-2010}$, $r = 0.76$ (**a**), $annual\ giving_{2001-2010}$, $r = 0.68$ (**b**) and $total\ research_{2001-2009}$, $r = 0.34$; $r = 0.64$, without Harvard No. 1, MIT No. 2 and Stanford No. 3 (**c**). *Plus* and *square labels* indicated the university located in the upper and lower half of the top50 $AWARD_{2003-2012}$, respectively, while the numbers in labels according to the $ARWU_{2012}$ rank list. *Sources* Nature Publishing Index 2012, ARRU, ARWU

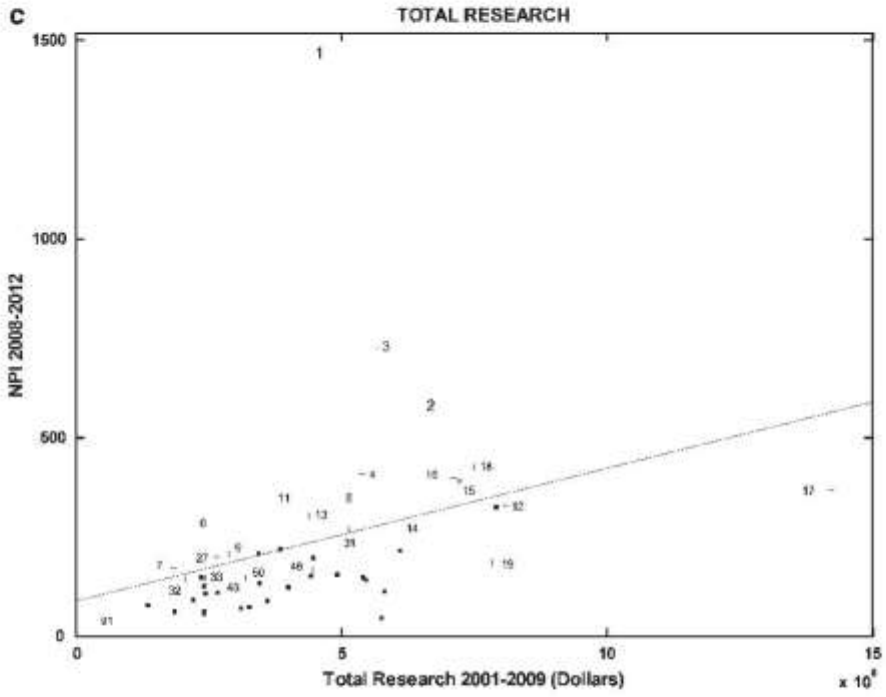


Fig. 5 continued

(GERD; Fig. 2 supplementary materials) is more or less similar to the statistics of the whole world.

The United States is the country that contributes the largest number of affiliations to $Nature_{2001-2011}$ (11,648 counts). This also complies with the largest expenditures in the higher education sector. The United Kingdom, Germany and France are the collective 'triad' of the European countries that maintain dominance and invest the most in the education sector, which is also reflected by their value of $Nature_{2001-2011}$. An interesting case is that the European Union taken as a whole is superior to all of the other continents. During 2001–2011 EU countries published 12,306 scientific articles in Nature. However, this still compares advantageously to the US, when we compare the overall HERD of the EU with $60b and the US $44b, respectively. On the other hand, Eastern European countries are those which contribute a significantly lower number of Nature publications. Countries in old EU countries affiliate 11,829 articles in the Nature journal, which is 96.1 % of total European articles, while the countries of the former Eastern Bloc affiliated 477 papers (3.9 %) during the same period of time. Japan leads in Asia and then fast-growing South Korea and China come second and third, respectively. It can be clearly

noticed that in addition to investment in education/research and development these countries are the economic and high-technology leaders of the global world. Moreover, an inflation effect appears, i.e., quite obviously the number of publications in Nature remained almost at a constant level (2658 in 2001 vs. 2933 articles in 2011), unlike the expenditure which significantly increased during 2001 and 2011. According to the overview of the National Science Board's (NSB 2012), global R&D expenditures over the past decade have grown faster than global GDP. Their analysis found that funding rose from an estimated $522b in 1996 to approximately $1.3t in 2009. Currently, 42.8 % of the world's scientific activity is concentrated in Europe, 40.6 % in the US and 3.6 % in Japan, respectively.

Country counts in $\text{Nature}_{2001-2011}$ or Nature_{year} for any single year are based on the institutional affiliation addresses that are ascribed to a country when a published paper carries at least one address associated with that country. If a country appears more than once in a single paper, it is still counted as a single record for that country. On the other hand, a single Nature publication can count for several countries that are included within the affiliation. This means that the sum of $\text{Nature}_{2001-2011}$ for individual countries does not amount to the total sum of Nature publications in 2001–2011.

The NPI ranks institutions according to the number of publications in the Nature journal. Scientists associate themselves with universities that are first order affiliations for publications. Universities are knowledge generators and drivers of innovation. The problem is the availability of reliable data on the budgets for these institutions worldwide. Conversely, scientific efficiency indicators are easily available. Six indicators are used to construct the ARWU ranking, including the number of alumni and Nobel Prize and Fields Medal winners, the number of highly cited researchers (Thomson Scientific), the number of articles published in the journals Nature or Science, the number of articles indexed in the Science Citation Index and the per capita performance with respect to the size of the institution.

Most of the top ARWU universities are located in America or Europe; among the top 500, 151 (30.2 %) are located in the US, 40.8 % in the EU, 36.8 % in North and South America, 21.6 % in Asia & Pacific and 0.8 % in Africa, respectively. China along with Japan has 58 universities on the list, which is 11.6 % of the total or 53.7 % of Asia and Pacific region. African institutes remain largely absent and are represented by only two countries that are on the list, South Africa with three universities and Egypt with one, respectively.

Harvard University, which is placed at the top of the ARWU statistics, had an average total research fund of $434 m during the period of 2001–2009, had an average $25b of $\text{endowment}_{2001-2010}$ and reached 1111 in $\text{Nature}_{2001-2010}$, respectively. Harvard is at the top of the endowment list and second on the annual giving list, but surprisingly only in 24th place in relation to total research funds (Table 1 columns $\text{endowment}_{2001-2010}$, $\text{annual giving}_{2001-2010}$ and $\text{total research}_{2001-2009}$). Stanford, which is ranked second in the ARWU, is third on the $\text{endowment}_{2001-2010}$ list and first on the $\text{annual giving}_{2001-2010}$ list. The average Stanford amount that spent on total research during the period of 2001–2009 was $641 m (8th on the $\text{total research}_{2001-2009}$, Table 1). MIT (No. 3, ARWU) is 5th on the endowment list but 13th on the total research list and 14th on the annual giving list. The data changes with the individual years. Yale University (received an average of $425 m of $\text{total research}_{2001-2009}$ and $343 m of $\text{annual giving}_{2001-2010}$ and is No. 11 in the ARWU rank list) had a $16b $\text{endowment}_{2001-2010}$, which is approximately $4b more than Stanford University (No. 2, ARWU). On the other hand, it had two times fewer counts in $\text{Nature}_{2001-2010}$ (302 and 504 papers, respectively). Similar values of $\text{Nature}_{2001-2012}$ and their endowments in 2012 can be observed for two British Universities, Cambridge (No. 5 in

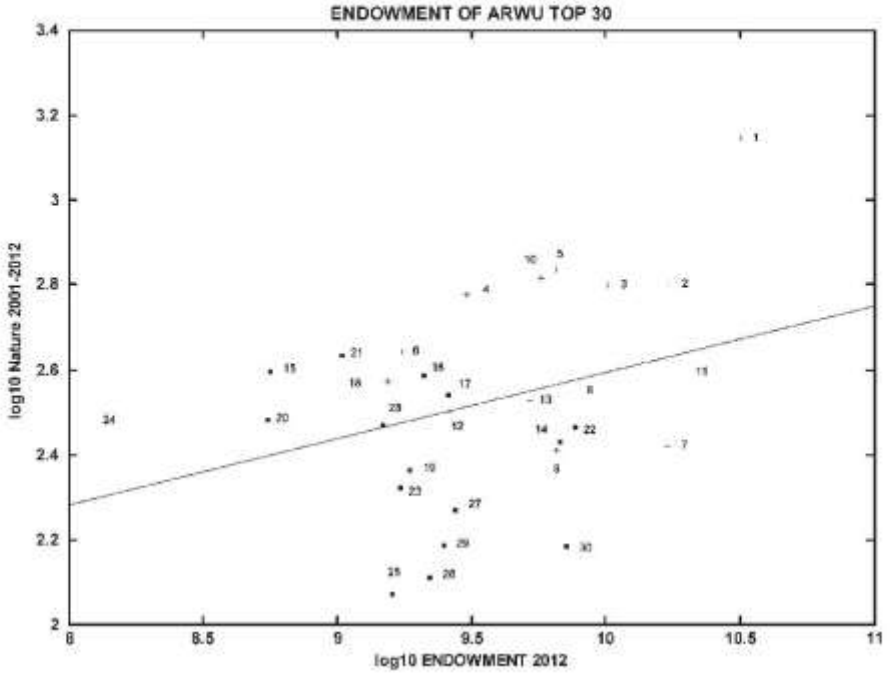


Fig. 6 $Nature_{2001–2012}$ of top 30 world universities plotted against $endowment_{2012}$. Labels are ascribed according the rank in $ARWU_{2012}$ rank, $r = 0.67$ (correlation calculated for logarithm untransformed data). Lower labels tend to be shifted right; labeled by higher numbers—*left*. *Plus* and *square labels* indicated the university located in the upper and lower half of the top50 $AWARD_{2003–2012}$. *Sources* ARWU, Web of Knowledge, Endowment FY 2012

ARWU) received \$6.6b and had 683 counts and Oxford (No. 10 in ARWU), which had \$5.8b had 655 counts in $Nature_{2001–2012}$, respectively. The University of Tokyo (No. 20) received just \$553 m of endowments and published 304 papers.

The scientific potential, research funding policy, the endowment stream, the importance of science for societies and economies across the globe depend upon culture and traditions. Thus, an analysis of a more homogeneous system can provide us with more reliable results. Therefore, we focused on the top 50 US universities. In Figs. 4 and 5 we plot the relationship between the $Nature_{2001–2009(2010)}$ and, additionally, the $NPI_{2008–2012}$ for the selected economic indicators for the top 50 US research universities, according to $endowment_{2001–2010}$, annual $giving_{2001–2010}$ and total $research_{2001–2009}$ from the ARRU. Quite surprisingly, $endowment_{2001–2010}$ ($r = 0.74$; Fig. 4a versus $r = 0.76$; Fig. 5a) and annual $giving_{2001–2010}$ ($r = 0.66$; Fig. 4b versus $r = 0.68$; Fig. 5b) appeared to correlate much better to $Nature_{2001–2010}$ and $NPI_{2008–2012}$ than the total $research_{2001–2009}$ expenditures ($r = 0.30$; Fig. 4c, d versus $r = 0.34$; Fig. 5c). The replacement of Nature for $NPI_{2008–2012}$ parameter (Fig. 5a–c) does not change these regularities; however, Nature and NPI are not always completely consistent. A similar correlation was found for the relationship between $Nature_{2001–2012}$ and the $endowment_{2012}$ of 30 top world ARWU universities ($r = 0.67$; Fig. 6).

The correlation of $Nature_{2001–2010}$ and $NPI_{2008–2012}$ to $endowment_{2001–2010}$ seems to be especially interesting ($r = 0.74$; $r = 0.76$, respectively). Generally, endowments are

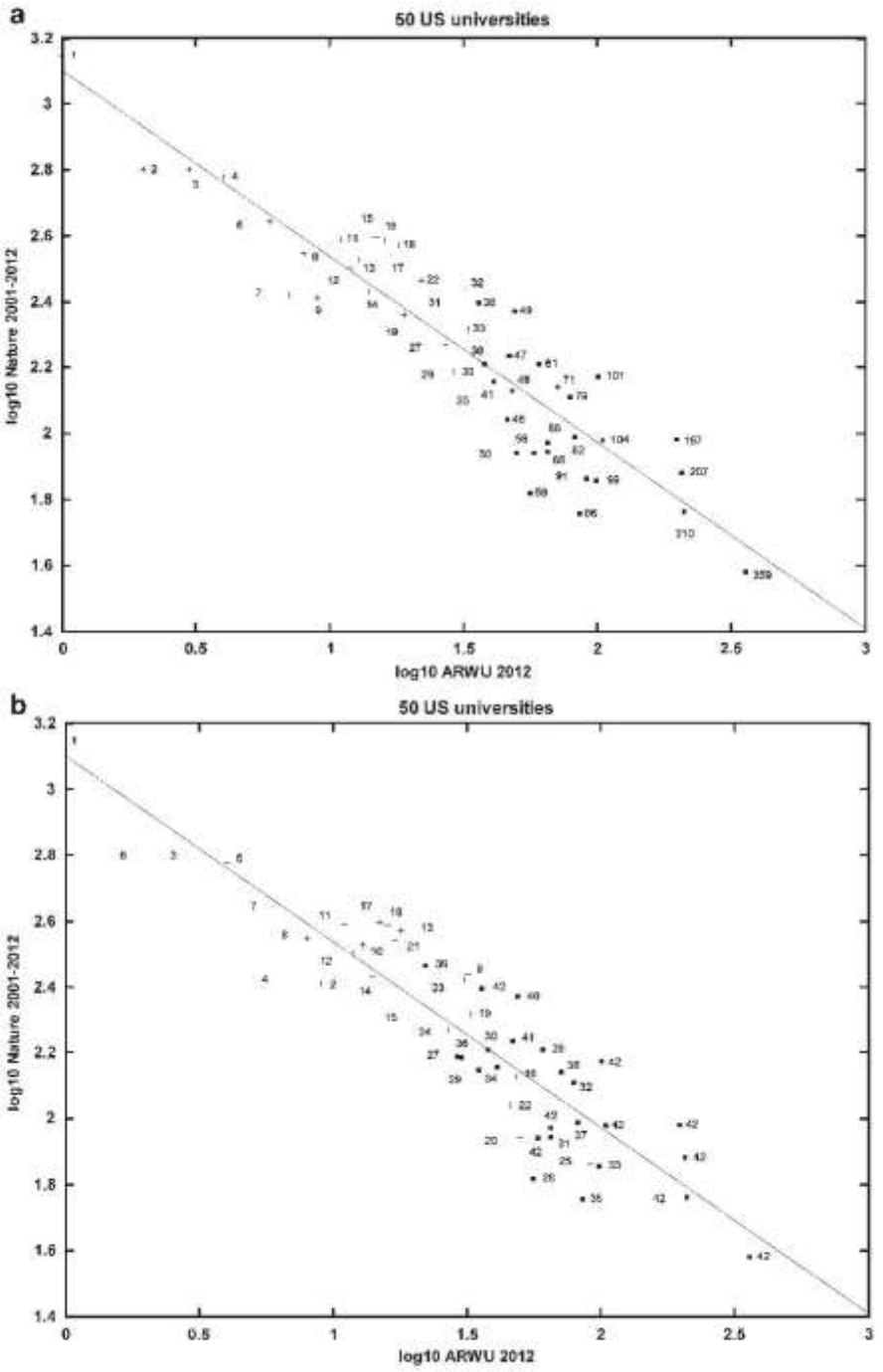
Fig. 7 The plots show a relationship between $Nature_{2001-2012}$ and Academic Ranking of World Universities ($ARWU_{2012}$) rank, $r = -0.48$ (correlation calculated for logarithm untransformed data). Labels according to the $ARWU_{2012}$ (**a**) and $AWARD_{2003-2012}$ (**b**) ranks. *Sources* Web of Knowledge, ARWU

pooled in a common fund. Typically, the principal of the endowment is invested and the interest is used to fund projects. The practice of investing the principal, rather than spending it, allows an endowment to grow over time, rather than diminishing by being spent all at once. For universities especially, this practice allows them to accrue resources thereby providing the possibility to keep pace with other competitive universities (McMahon 2013). Universities live on the resources generated from many sources, but critical to their success are the sizes of their endowments and annual giving. Thus, universities with high endowments are able to pursue critical initiatives that enhance their ability to be leaders in research and higher education.

However, a careful inspection of the relationship of $Nature_{2001-2009}$ or $NPI_{2008-2012}$ versus total $research_{2001-2009}$ expenditures (Figs. 4c, d, 5c) reveals another effect. Although the correlation for the top 50 US universities is low (0.30; 0.34), the elimination of the three top outlying ARWU universities of Harvard, Stanford and MIT, i.e., those with the first three places in ARWU ranking, increases r to a value of 0.47 for $Nature_{2001-2009}$ and 0.64 for $NPI_{2008-2012}$. At the same time, $Nature_{2001-2009}$ for these three universities outperforms the parameter value predicted by the $Nature_{2001-2009}$ versus total $research_{2001-2009}$ expenditures model, i.e., the observables are high above the model straight line. This suggests the importance of the reputation of a university for a Nature value. Due to the fact that the status and reputation of a university are also decisive for annual giving and endowment, the observed regularity indicates that the university logo is important for its success. However, when we tested the direct correlation between the ARWU 2012 rank and $Nature_{2001-2012}$ for the top 50 US universities, we obtained a rather low value of $r = -0.48$ (Fig. 7a, b).

The question now arises as to what the nature of the connection between a university's reputation and Nature is. What are the origins of the extreme success of the top universities in Nature? Several explanations can be hypothesised. First of all, the better organisation of scientific research can be suggested. At the least, excellent science has brought these universities to the top of the rankings. However, what do we mean by better science organisation? Higher endowments and the availability of endowed chairs and professorships attract the best possible scientists. Today, this rule applies on a global scale. This implies that both the prospects of scientific success and the motivation of faculty members may overwhelm that of other universities that do not have the ability to attract the best. However, can human resources on their own win the competition even when there are lower total research funds? It seems that the so-called Matthew Effect in Science works on the advantage of the eminent universities (Merton 1968; Shermer 2008).

The Matthew Effect is an example of the most general effect known as the so-called *cumulative advantage*. "*Them as has, gits*" or as it says in the Bible "*the rich get richer and the poor get poorer*" (Matthew 25:29, http://lds.org/scriptures/nt/matt/25?lang=eng). Once a product gets the lead on the market among consumers, it is desired by the consumers (Shermer 2008). *Eminent scientists get disproportionately great credit for their contributions to science while relatively unknown scientists tend to get disproportionately little credit for comparable contributions* (Merton 1968). Tol proved that this effect controls the number of citations in economics (Tol 2013, 2009). He also provides key references for recent publications in this area. To prove the importance of the Matthew Effect on an anecdotal basis, this effect is used to explain the orientation of the addition


a
50 US universities
log10 Nature 2001-2012
log10 ARWU 2012
b
50 US universities
log10 Nature 2001-2012
log10 ARWU 2012

reaction to a double carbon–carbon bond. If you as a carbon atom have more hydrogen atoms, you will obtain the next ones (Morrison and Boyd 1973).

It is the name of a famous university or a famous scientist, which when present among others, that *sticks* to the publication. Nature is one of the leading journals of science. Harvard, Stanford and MIT are first three universities on the ARWU ranking. The affiliation of such a university is a clear advantage for the journal. The risk of scientific roulette is lower now. A publication attracts the attention of the scientific audience from the very beginning, the affiliation. This also often provides higher citation expectations, thus increasing the IF value and consequently guaranteeing the future supply of valuable contributions to the journal. On the other hand, as remarked by the anonymous reviewer of this contribution during review '*good researchers are attracted to institutions with a good reputation not so much because these institutions have a lot of money but simply because of the reputation of these institutions. Following this hypothesis, there would be a Matthew Effect in which the presence of good researchers at an institution attracts even more good researchers. So good researchers like to be together at the same place. So perhaps the Matthew Effect does not necessarily need to be related to the journal publishing system*'. This ambiguity can be illustrated better if we interpret the scientific institutions and journals according to the economic categories of the competitive market where demand for better publications is higher at the best universities, due to better staff and at the same time the demand is also higher on the publisher's side due to the better university logo. Apparently, this strong positive feedback loop is responsible for the observed output imbalance *that* can be observed, which is consistently increasing its magnitude.

The hypothesis of the Matthew Effect can be strengthened if we now come back to the $\text{Nature}_{2001-2010}$ versus $\text{endowment}_{2001-2010}$ relationship for the US top 50 universities (Fig. 4a). We can observe that the lower university labels, which have a higher ARWU rank, are located over the modelled straight line, while the higher labels are those universities below this line. This means that higher ARWU ranked universities submitted relatively more to $\text{Nature}_{2001-2010}$. Similarly, when analysing the position of the world's top 30 ARWU universities within the $\text{Nature}_{2001-2012}$ versus endowment_{2012} plot (Fig. 6), we can find that now those ranked at the top are shifted to the upper right quarter of the plot. Because the ARWU rank uses the number of Nature publications as one of the efficiency estimators of the university rank, it is not clear whether separation according to ARWU rank is affected by this to such an extent so as to be predominant in the $\text{Nature}_{2001-2012}$ versus endowment_{2012} plots or whether this is a real illustration of the Matthew Effect. Thus, in Figs. 5–7 we also used the plus and square labels to indicate the university located in the upper and lower half of the top50 $\text{AWARD}_{2003-2012}$. Generally, similar to ARWU the AWARD labels also privileged those at the top of the rank.

The Matthew Effect provides a positive feedback loop for the performance in science (Tol 2009). We can clearly observe this in the current analysis. Thus, Nature is limited by financial support, which in turn depends upon the ARWU ranking, which is estimated by also counting, among others, the Nature indicator. This means that all of the indicators are highly inter-correlated, which brings uncertainty to the statistics. Regardless of this uncertainty, a higher ARWU rank also means a lower cost for a single Nature count.

Conclusions

In today's world, scientific activity plays a significant role in technological and economic development. This has brought an enormous interest in the performance metrics for

scientific output, which is still a controversial and serious problem. Nature is among the world's most highly cited multidisciplinary science journals with one of the highest impact factors of 38.597, which is relatively often used in scientific rankings and as an assessment of outstanding research performance. Research output can be measured in terms of patents and publications, especially in the highly cited mainstream journals. By analysing the regional distribution of Nature publications, we found a high correlation between the expenditures and the number of local affiliations in Nature when counted on a national basis.

The correlations between a Nature descriptor and financial data describing the involvement of the individual countries and universities undoubtedly indicate that the country level provides much better models than the university one. The reason for that is quite obvious. Individual universities have quite different specialisations that are a bettor or worse fit for Nature's specificity. Data ambiguity and deviations for individual universities are also higher. On the other hand, the statistics at the country level provides a model of the average university.

The regularities observed for the world's top 30 and the US's top 50 universities indicate that the correlation is skewed and evidently rewards those ranked at the top of the ARWU. Surprisingly, the amount of the endowment better determines the number of Nature publications for universities than the total research expenditure. This illustrates that the status and reputation of a university is important for success in Nature. We believe that the so-called cumulative advantage or the Matthew Effect created this situation.

Acknowledgements J. Bogocz appreciates the support of the Doktoris fellowship.

References

Arkhipov, D. B. (1999). Scientometic analysis of Nature, the journal. *Scientometrics, 46*(1), 51–72.

Broberger, C., & Sjostrom Douagi, A. (2012). Funding model: Cuts endanger young scientists in Europe. *Nature, 491*, 672.

Frame, J. D., & Narin, F. (1976). NIH funding and biomedical publication output. *Federal Proceedings, 35*, 2529–2532.

Garrison, H. H. (2013). Research funding: Fiscal cliff is bad news for US science. *Nature, 493*, 163.

Gross, C., Anderson, G., & Powe, N. (1999). The relation between funding by the national institutes of health and the burden of disease? *New England Journal of Medicine, 340*(24), 1881–1887.

Guan, J., & Wang, J. (2004). Evaluation and interpretation of knowledge production efficiency. *Scientometrics, 59*(1), 131–155.

ITG. (2008). *The Science of Science Policy: A Federal Research Roadmap*. Washington, DC: National Science and Technology Council and the Office of Science and Technology Policy.

Kaneiwa, K., Adachi, J., Aoki, M., Masuda, T., Midorikawa, N., Tanimura, A., et al. (1988). A comparison between the journals Nature and Science. *Scientometrics, 13*(3–4), 125–133.

Keene, O. N. (1995). The log transformation is special. *Statistical Medicine, 14*, 811–819.

Leydesdorff, L., & Wagner, C. (2009a). Is the United States losing ground in science? A global perspective on the world science system. *Scientometrics, 78*(1), 23–36.

Leydesdorff, L., & Wagner, C. (2009b). Macro-level indicators of the relations between research funding and research output. *Journal of Informetrics, 3*, 353–362.

Matthew 25:29. *The Holy Bible*. King James Version. Retrieved from http://lds.org/scriptures/nt/matt/25?lang=eng.

McMahon, M. (2013). *What is an endowment ?* wiseGEEK, Conjecture Corp. Retrieved from http://wisegeek.com/what-is-an-endowment.htm. Accessed July 2013.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63.

Moran, M., Guzman, J., Ropars, A. L., McDonald, A., Jameson, N., Omune, B., et al. (2009). Neglected disease research and development: How much are we really spending? *PLoS Medicine, 6*(2), e1000030.

Morrison, R. T., & Boyd, R. N. (1973). *Organic Chemistry* (3rd ed., p. 188). Boston: Allyn & Bacon Inc.

National Institutes of Health (NIH). (2011). *Estimates of funding for various diseases, conditions, research areas*. National Institutes of Health, Department of Health and Human Services. Retrieved from http://report.nih.gov/categorical_spending.aspx. Accessed May 2013.

National Science Board (NSB). (2012). *Science and Engineering Indicators overview*. Washington, DC: National Science Foundation. Retrieved from http://nsf.gov/statistics/seind12/c0/c0i.htm. Accessed July 2013.

Nature Publishing Group (NPG). (2013). *Nature Publishing Index 2012 Global*. London: Macmillan Publisher Limited. Retrieved from http://natureasia.com/en/publishing-index. Accessed May–July 2013.

Shermer, M. (2008). *The mind of the market: Compassionate apes, competitive humans and other tales from evolutionary economics*. New York: Henry Holt and Company.

Symonds, M. (2004). Nature and Science know best. *Trends in Ecology and Evolution, 19*(11), 564.

Tol, R. S. J. (2009). The Matthew effect defined and tested for the 100 most prolific economists. *Journal of the American Society for Information Science and Technology, 60*(2), 420–426.

Tol, R. S. J. (2013). The Matthew effect for cohorts of economists. *Journal of Informetrics, 7*, 522–527.

Vanderelst, D., & Speybroeck, N. (2013). Scientometrics reveals funding priorities in medical research policy. *Journal of Informetrics, 7*, 240–247.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer. ISBN 0-387-95457-0.

17.3 Załącznik 3

Kopia publikacji naukowej:

Polanski J, Bogocz J, Tkocz A. (2015) Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. *Drug Discovery Today*. 20:1300-1304. (IF=5,625; pkt KEJN=45)

ELSEVIER

feature

Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator

Jaroslaw Polanski*, jaroslaw.polanski@us.edu.pl, Jacek Bogocz and Aleksandra Tkocz

We analyzed a list of the top 100 bestselling drugs as a struggling market for new FDA approvals. Using the time from drug approval by the FDA as a measure of drug age, our analysis showed that the top 100 bestselling drugs are getting older. This reflects the stalled launch of new drugs into the market during recent years.

Introduction

Drug design and development is an extremely complex technological and economic challenge. The productivity of the pharmaceutical industry appears to be declining; thus, over the past 20 years, pharmaceutical R&D has been unable to develop a sufficient number of drugs to supplant those with expiring patents. Lower profits in recent years have resulted in massive economic problems (e.g. site closures and redundancies), as well as a reduction in the number of discovery and development projects [1].

The reason for the productivity gap is controversial. Sams-Dodd hypothesized that the current drug design paradigm caused the decline in productivity [1]. Of the two essential drug development approaches, namely physiology-based or phenotype and target-based approaches, it is the latter that has dominated contemporary pharma practice [2]. This was the result of large investments in screening equipment and the reorganization of research departments, and it means that disease-focused designs have become much less valuable in the corporate view. Target identification, mode of action, target validation and the screening cascade distinguish this design dogma. Although this seems reliable, in reality it represents a huge simplification of the biological system at the level of the whole organism, with all of its complexity and variability. At the same time, even at this level of reductionism, we still cannot make the investigations precise enough to include multiple targeting (polypharmacology) or multidrug interference (synergy). According to Sams-Dodd's hypothesis, a fully rational target-based design is too optimistic given the current state of drug design, and it is becoming increasingly apparent that understanding the mode of action of a drug is not a prerequisite for successful molecular design [1]. In fact, target-oriented approaches are more likely to fail during development than those that are based on physiology [3].

A completely different explanation of the current trends in the pharmaceutical industry was offered by Pammolli *et al.*, who performed a pharmacoeconomic analysis of a large database of 28 000 compounds that had been investigated by pharmaceutical R&D. The authors claimed that the decline in R&D productivity is associated with the 'increasing concentration of the R&D investment in the areas in which the risk of failure is high, which correspond to unmet therapeutic needs and unexploited biological mechanisms'. Thus, pharma is currently looking for completely novel targets and breakthrough innovations. The main conclusion from this study was that there is no productivity gap [4].

In economics, money is used to measure the behavior and fate of a product on the market. Although medicine and pharmacy are lucrative businesses, they also have missions that are steered by ethical imperatives and bans. Cockburn stated: 'If pharmaceutical productivity crisis is a fact, it presents policy makers with some

http://dx.doi.org/10.1016/j.drudis.2015.06.015

difficult questions. Taxpayers around the world support well over $25 bn per year of biomedical research' [5]. This means that, for the pharmaceutical industry, profit is more than a measuring stick. To what extent can we probe this market by investigating bestselling drugs? Intuitively, medications that win the market competition can be considered a standard for the ideal drug, and a list of the bestselling drugs could be used to probe the whole drug population at the highest quality level. The purpose of this paper is to analyze whether the list of bestselling drugs, which model the ideal medications, can reflect or shed light on current trends in pharma. We used the top 100 bestselling prescribed drugs in the US market (http://www.drugs.com/stats) as a model, which enabled us to compare the data with the FDA approval statistics.

The top 100 drugs are getting older

An overview of new molecular entities (NMEs) that were approved by the FDA from 1827 to 2013 was recently published by Kinch *et al.* [6]. The number of FDA approvals during this period was 1453, although the rate of approvals for NMEs was an average of 15 per year in the 1950s and remained at this rate through the 1970s. It increased to 25–30 NMEs per year in the 1980s, and this is where it remains today. The maximum annual approval rate was observed sometime during the mid-1990s, when 55 NMEs were approved in a single year. This analysis demonstrates that FDA approvals can be used to indicate basic trends in the pharmaceutical industry. In particular, it highlights the emergence of a handful of companies that control two-thirds of NMEs. Moreover, the growth in the number of NMEs that are controlled by marketing organizations that have little or no internal drug discovery or development activities can also be observed. The entire batch of FDA approvals was further divided into subpopulations of specific drug classes; namely HIV/AIDS and other infectious disease drugs [7], drugs for oncology [8] and antibacterial agents [9].

In this study, we analyzed another subpopulation of FDA approvals, the top 100 bestselling list (i.e. the drugs that appeared to win the market competition). First, we tried to determine whether the productivity trends could be probed and were reflected by the competition among bestselling drugs. If newer drugs penetrated the bestseller list much faster than previous drugs we could claim a positive increase in R&D productivity. Similarly, if the new R&D products were not competitive enough to prevail over the older drugs then we would observe the reverse trend. But how can the age of drugs be measured? This is a complex problem that can be analyzed by identifying the first patents or publications. These statistics can, however, be skewed because a company's strategy determines the timing for patenting and publishing, and companies try to keep their projects secret for as long as possible. Therefore, the availability of reliable data is not entirely straightforward. Because of this, we used an open and unambiguous parameter to define drug age. Each new drug can appear on the market immediately after FDA approval; therefore we defined the length of time after approval as the measure of drug age. Accordingly, we analyzed how quickly FDA-approved drugs penetrate the top 100 bestseller market (Fig. 1). Surprisingly, to the best of our knowledge, no one has previously attempted a similar analysis.

We found that 49.6% of the top 100 drugs are FDA approvals that originated before 2000, whereas 50.4% of them appeared on the market during the 2000s. The average drug age is approximately 10.2 years and increased from 8.1 years in 2003 to 12.1 years in 2013. The average age, which is plotted year by year in Fig. 1, not only indicates a clear trend but also a high positive correlation ($R^2 = 0.93$), implying that the top 100 bestselling drugs are steadily getting older. The data seem to indicate that we do not

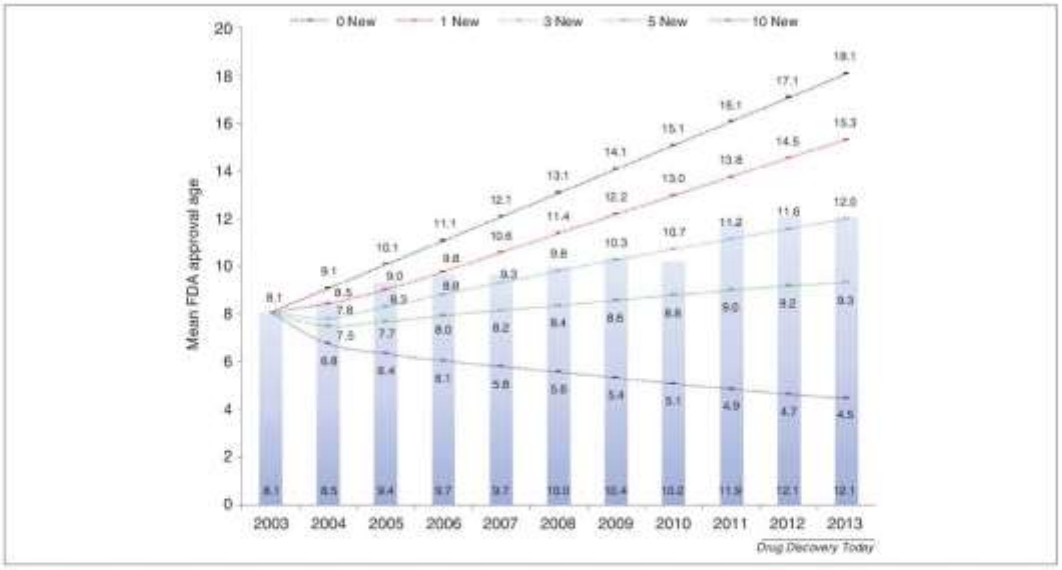


FIGURE 1

Mean FDA approval age of drugs in the top 100 bestselling list compared with simulated aging scenarios. Black line indicates a hypothetical full stagnation situation (i.e. if the situation in 2003 was fully preserved for the entire decade). Violet line at the lower edge of the plot shows a situation in which the ten oldest drugs are replaced by fresh FDA approvals each year. Similarly, the red, blue and green lines illustrate situations with one, three and five replacements, respectively.

have FDA approvals (i.e. new drugs) that are strong enough to win the top 100 tournament at a younger age than earlier candidates. In order to understand Fig. 1 better, we plotted 'what if' lines for simulated scenarios. Thus, a black line indicates a hypothetical full stagnation situation (i.e. if the situation in 2003 was fully preserved for the entire decade). The top 100 population is getting older naturally (i.e. each drug is getting 1 year older each year and reached a mean age of 18.1 years in 2013). By contrast, a violet line at the lower edge of the plot shows a situation in which the oldest ten drugs are replaced by fresh FDA approvals each year. Similarly, the red, blue and green lines illustrate situations with one, three and five replacements, respectively. Best-fit analysis revealed that the difference between the real situation, indicated by the bar plot and the simulated scenarios, occurred with the three-replacement scenario.

The simulated developments indicate that, although a high correlation can result from a high retention level and the effect of natural aging, the increase in drug age is clear. To study the structure of the top 100 drugs on the market further, we classified the list according to specific disease areas: central nervous system (CNS) (26%), anti-infective (16%), cardiovascular (13%), metabolic (12%), immune (11%), respiratory (6%), urologic (5%), anticancer (3%) and other (8%). The percentages refer to the number of drugs, and this transforms to the following values describing market value share: CNS (28%), anti-infective (8%), cardiovascular (21%), metabolic (20%), immune (8%), respiratory (10%), urologic (1%), anticancer (2%) and other (2%). Accordingly, CNS drugs, which have a share of 26% by drug units or 28% by market volume, are the winners on the list. At the same time, a comparison of the drug-aging dynamics according to the drug classes clearly indicates that CNS drugs (10.6 years) are among the oldest populations, after other (14.6 years) and cardiovascular (11.3 years). In particular, CNS drugs were the oldest population in 2013. At the same time, urologics (7.6 years) were the youngest medications, well below the average drug age (Table 1).

In Table 2 we additionally provide data on the structure of the top 100 list when it is dissected by year of approval. This analysis characterizes the frequency (multiple appearances were counted) of the appearance of approvals within the top 100 list. The drugs originating in 1995–2004 are the winners of this competition, with a total of 158 unique drugs (65%) or 827 appearances (~75%).

Measuring short- and long-term changes in the size and age composition of populations is a complex problem. Accordingly, probing the dynamics of FDA approval age in drug

TABLE 1

Mean drug age in years for different drug classes.

Disease area	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
CNS (26%)	9.1	8.8	9.3	9.4	9.9	10.4	10.8	8.7	13.9	14.4	15.1
Anti-infective (16%)	7.9	8.1	10.0	8.7	8.3	9.1	9.6	12.0	9.6	9.9	11.3
Cardiovascular (13%)	8.9	9.7	10.5	11.1	10.8	10.1	11.1	10.8	14.6	14.5	13.9
Metabolic (12%)	5.0	5.0	6.1	6.6	7.0	8.2	8.6	8.9	10.4	10.4	11.6
Immune (11%)	6.5	7.4	8.3	9.4	10.4	10.5	10.9	10.7	13.2	12.6	11.0
Respiratory (6%)	5.5	6.5	5.9	6.2	6.6	7.9	8.0	8.9	9.0	10.2	8.7
Anticancer (5%)	4.5	7.1	9.0	10.0	9.6	10.6	11.8	12.8	10.0	12.0	11.7
Urologic (3%)	4.9	5.9	6.9	7.9	8.9	10.0	9.0	8.3	7.1	6.7	7.7
Other (8%)	14.2	15.7	16.2	19.2	16.0	13.5	14.5	13.8	9.8	–	10.9

TABLE 2

Multiple or unique drug appearances in the top 100 list by year of approval.

Year	Appearances (unique drugs)	Percentage	Year	Appearances (unique drugs)	Percentage
1942	8 (1)	0.7% (0.4%)	1997	96 (16)	8.7% (6.6%)
1955	2 (1)	0.2% (0.4%)	1998	92 (18)	8.4% (7.4%)
1960	1 (1)	0.1% (0.4%)	1999	54 (10)	4.9% (4.1%)
1962	7 (1)	0.6% (0.4%)	2000	92 (19)	8.4% (7.8%)
1982	3 (1)	0.3% (0.4%)	2001	77 (21)	7.0% (8.6%)
1983	6 (1)	0.5% (0.4%)	2002	91 (16)	8.3% (6.6%)
1985	1 (1)	0.1% (0.4%)	2003	54 (10)	4.9% (4.1%)
1987	1 (1)	0.1% (0.4%)	2004	90 (17)	8.2% (7.0%)
1989	11 (3)	1.0% (1.2%)	2005	20 (7)	1.8% (2.9%)
1990	5 (2)	0.5% (0.8%)	2006	37 (11)	3.4% (4.5%)
1991	34 (11)	3.1% (4.5%)	2007	21 (5)	1.9% (2.1%)
1992	37 (9)	3.34% (3.7%)	2008	5 (3)	0.5% (1.2%)
1993	29 (8)	2.6% (3.3%)	2009	5 (4)	0.5% (1.6%)
1994	19 (4)	1.7% (1.6%)	2010	10 (5)	0.9% (2.1%)
1995	70 (11)	6.4% (4.5%)	2011	6 (4)	0.5% (1.6%)
1996	111 (20)	10.1% (8.2%)	2013	1 (1)	0.1% (0.4%)

populations or subpopulations results in a complex metric that combines discovery and development with regulatory and marketing aspects. To understand how this method works, we probed the entire FDA (NME) approvals population (http://www.accessdata.fda.gov/scripts/cder/drugsatfda) that was launched on the market from 1939 to 2014 (Fig. 2). Using this method, unlike in the top 100 list, now all new approvals entering the tested population just after approval remain there, with the rare exceptions of withdrawals. As expected, the mean drug age is steadily increasing in this population, yielding a value of 13.7 years for the entire period of 1939–2014. This value is surprisingly close to the mean age of the top 100 list (12.1 years in 2013).

The question now arises as to whether there is some more interesting information encoded beyond these data. The solid line in Fig. 2a illustrates a differential dissection of the drug age dynamics within this drug population, plotting the mean age growth rate (i.e. the difference in the mean drug age in a given year and the year preceding this year – with this definition the data from the preceding year are required and thus the value cannot be calculated precisely for the first year in the series). Within the chaotic short-term year-to-year changes, we can recognize more-stable long-term trends. These trends are revealed more clearly if we calculate the mean age growth rate in 5-year intervals (Fig. 2b). With some minor exceptions, mean age growth rate decreased in 1970–1999 and increased in 1950–1970 and 2000–2009. Interestingly, the latest period of 2010–2014 reverses the previous increasing trend in mean NME approval age growth rate. Does this indicate an innovation breakthrough in novel technologies and a breakthrough in the 'productivity crisis', as suggested by Pammolli [4]? As is usually the case with economic data, the answer is far from simple. We should remember, however, that, unlike the top 100 list analysis, which includes market performance of the

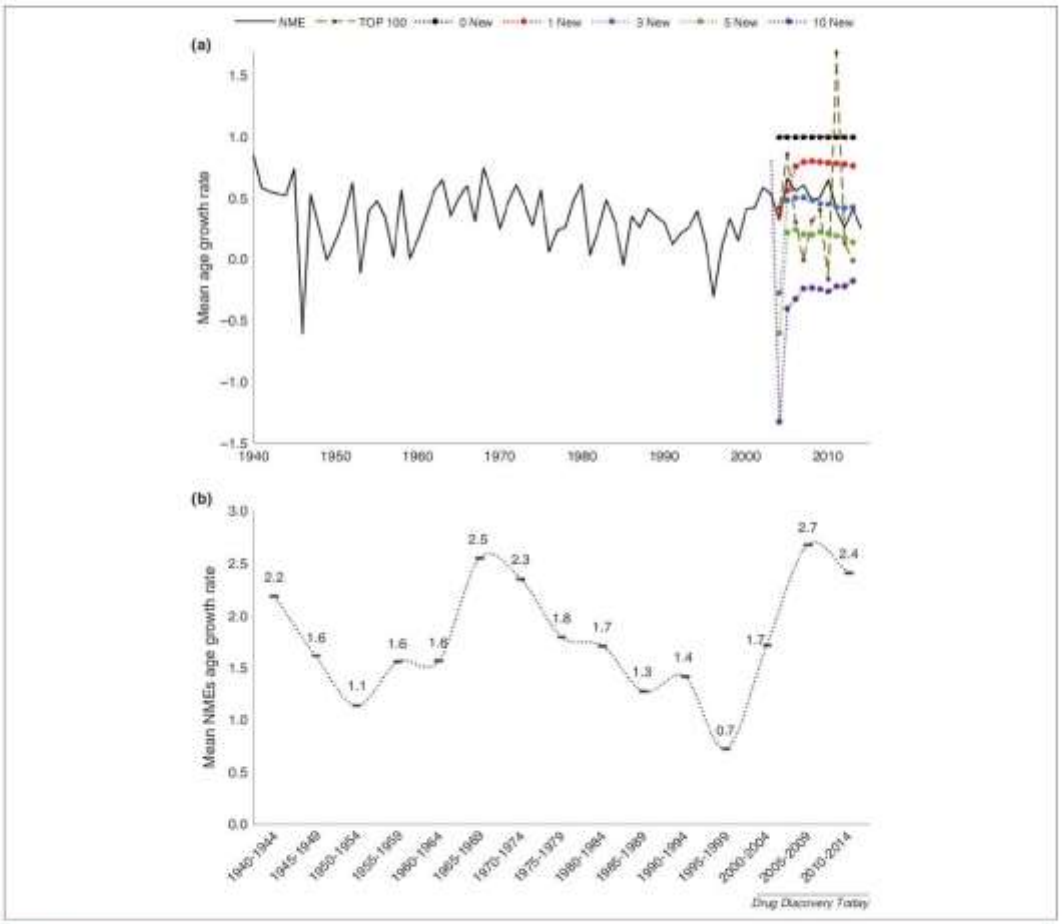


FIGURE 2

Mean age growth rate for the entire population of FDA approvals in 1939–2014 (solid line) compared with that of the top 100 list (dashed line) and the top 100 list with simulated replacement data (colored dots) **(a)** and mean age growth rate calculated in 5-year intervals **(b)**. Compared with 1939 for the first time period.

drugs, in this analysis we are not testing anything other than FDA approval dynamics. Thus, in Fig. 2a we compared the mean age growth rate dynamics within all drug approvals with those of the top 100 list (dashed line) and top 100 simulated data (colored dots). The list changes each year; thus, fluctuations of the mean age growth rate are also higher than those for the whole population of drug approvals. However, the plots for the list and all approvals do not differ dramatically, as might be expected if highly innovative new approvals were aggressive enough to penetrate the top 100 list preferentially.

Concluding remarks

The market is an important performance metric controlling drug development. Our analysis showed that the time from drug approval (FDA), if used as a measure of drug age for probing the top 100 bestselling drug data, indicates that drugs are getting older. Our simulated developments indicate that, although a high correlation can result from high retention and the effect of natural aging, the increase in drug age is clear. Moreover, a comparison of the drug age growth rate dynamics for the entire population of FDA approvals and those winning the market competition (top 100 list) does not reveal the large differences that would be expected for highly innovative development when aggressive newcomers are entering the list. This reflects the stalled launch of new drugs into the market in recent years, which is in concordance with other analyses in this area.

Acknowledgments

The research was co-financed by the National Research and Development Center (NCBiR) under Grant ORGANOMET No: PBS2/A5/40/2014.

References

1 Sams-Dodd, F. (2013) Is poor research the cause of the declining productivity of the pharmaceutical industry? An industry in need of a paradigm shift. *Drug Discov. Today* 18, 211–217
2 Swinney, D.C. and Anthony, J. (2011) How were new medicines discovered? *Nat. Rev. Drug Discov.* 10, 507–519
3 Arrowsmith, J. (2011) Trial watch: Phase II failures: 2008–2010. *Nat. Rev. Drug Discov.* 10, 328–329
4 Pammolli, F. *et al.* (2011) The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* 10, 428–438
5 Cockburn, I.M. (2007) Is the pharmaceutical industry in a productivity crisis? *In Innovation Policy and the Economy*, (vol. 7) (Lerner, J. and Stern, S., eds) pp. 1–32, MIT Press
6 Kinch, M.S. *et al.* (2014) An overview of FDA-approved new molecular entities: 1827–2013. *Drug Discov. Today* 19, 1033–1039
7 Kinch, M.S. and Patridge, E. (2014) An analysis of FDA-approved drugs for infectious disease: HIV/AIDS drugs. *Drug Discov. Today* 19, 1510–1513
8 Kinch, M.S. (2014) An analysis of FDA-approved drugs for oncology. *Drug Discov. Today* 12, 1831–1835
9 Kinch, M.S. *et al.* (2014) An analysis of FDA-approved drugs for infectious disease: antibacterial agents. *Drug Discov. Today* 19, 1283–1287

*Jarosław Polański**
Jacek Bogocz
Aleksandra Tkocz
Institute of Chemistry, University of Silesia, Szkolna 9, 40-006 Katowice, Poland

17.4 Załącznik 4

Kopia publikacji naukowej:

Polanski J, Bogocz J, Tkocz A. (2016) The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *Journal of Computer-Aided Molecular Design.* 30:381-389. (IF=3,199; pkt KEJN=30)

J Comput Aided Mol Des (2016) 30:381–389
DOI 10.1007/s10822-016-9912-5

The analysis of the market success of FDA approvals by probing top 100 bestselling drugs

Jarosław Polanski[1] · Jacek Bogocz[1] · Aleksandra Tkocz[1]

Received: 21 December 2015 / Accepted: 21 April 2016 / Published online: 28 April 2016

Abstract Target-oriented drug discovery is the main research paradigm of contemporary drug discovery. In target-oriented approaches, we attempt to maximize in vitro drug potency by finding the optimal fit to the target. This can result in a higher molecular complexity, in particular, the higher molecular weight (MW) of the drugs. However, a comparison of the successful developments of pharmaceuticals with the general trends that can be observed in medicinal chemistry resulted in the conclusion that the so-called molecular obesity is an important reason for the attrition rate of drugs. When analyzing the list of top 100 drug bestsellers versus all of the FDA approvals, we discovered that on average lower-complexity (MW, ADMET score) drugs are winners of the top 100 list in terms of numbers but that, especially, up to some optimal MW value, a higher molecular complexity can pay off with higher incomes. This indicates that slim drugs are doing better but that fat drugs are bigger fishes to catch.

Keywords Drug design · Drug-like properties · Log P · ADMET · Database

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9912-5) contains supplementary material, which is available to authorized users.

✉ Jarosław Polanski
jaroslaw.polanski@us.edu.pl

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

Introduction

Drug design and development is an extremely complex technological and economic problem. We can illustrate this by the recent controversy over the efficiency in pharmaceutical R&D. There is sufficient argumentation and evidence to provide support for the thesis that R&D productivity has been steadily declining for decades [1]. This decline is so important that, for example, new means were suggested in the "design of Phase II trials to increase the productivity" gap [2]. Conversely, other arguments proved that there is no productivity crisis [3] or at least that "productivity rides again" [4]. Economically, the declining efficiency was demonstrated by the number of new drugs that have been approved per billion US dollars spent on R&D, which has halved roughly every nine years since 1950 [1]. This issue evidently needs a more systematic analysis.

In economics income is a measure of performance and, therefore, the market controlled by money provides us with the performance appraisals that determines the selection of the best products. Each new drug appears on the market immediately after the approval of the drug administration to begin a struggle for popularity and income and the list of drug bestsellers indicates the winners. We have recently used the length of time after approval as the measure of drug age in order to test drug populations (FDA approvals) to analyze overall trends in pharmaceutical R&D, in particular, drug bestsellers [5]. Here, we ask the question to what extent we can explain pharma R&D efficiency exploring drug bestsellers in particular the sales statistics as a function of drug-likeness. The commercial success of any drug is a complex problem that is limited by the compound quality, marketing efforts but also the unmet medical need impacts. In fact, the direct analysis of all these issues is

practically impossible. Instead, the analysis which we followed in this study mimics the viewpoint of natural selection in which drugs bestsellers are the winners among all FDA approvals. All the above mentioned impacts are superposed within reaching the top 100 list.

Target-oriented drug discovery is the main research paradigm of contemporary drug discovery. This has resulted in large investments in screening equipment and the reorganization of research departments. It is the target, mode-of action, target validation and screening cascade that identifies the design dogma [6]. In target-oriented approaches, we attempt to maximize in vitro drug potency by finding the optimal fit to the target. Usually, this results in a higher molecular complexity, in particular, the higher molecular weight (MW) of the drugs [7]. However, a comparison of the successful developments of pharmaceuticals with the general trends that can be observed in medicinal chemistry resulted in the conclusion that this trend or the so-called molecular obesity is an important reason for the attrition rate of drug candidates [7–9]. For example, the average MW and clog *P* of the drugs that are being marketed are lower than those for all of the drug candidates taking a value of 300–450 Da (MW) and 1.5–4.0 (clog P) for the marketed drugs, respectively [7]. Accordingly, the concept of the so-called slim pharma endorses lower molecular complexity drug projects [9].

In comparison with the target-oriented approach, drug-likeness, which has been unexpectedly successful in drug development, appears to be based on a highly reductionist view. The concept of drug likeness relates the success of drug candidates to the simple molecular properties that are typical for the commercial drug population [8–10]. The Lipinski Rule of Five (Ro5), which is probably the best-known filter, is based on the four molecular descriptors that determine good oral absorption and/or ADMET properties, i.e. the logarithm of the octanol:water partition coefficient (log P) <5; MW < 500 Da; the number of H-bond donors (HBDs) < 5 and the number of H-bond acceptors (HBAs) < 10.

It is worth noting that the Ro5 parameters in drug population are highly inter-correlated, and Gleeson defined the so-called ADMET score as the one value ADMET measure to be used when analyzing the oral drug population [11].

$$ADMET\ Score = \frac{|2.5 - c\log P|}{2.0} + \frac{|330 - molecular\ mass^{*}|}{120}$$

* if molecular mass <330 (a mean value of MW for oral drugs), use molecular mass = 0 for calculations

A number of recent investigations have attempted to estimate the physico-chemical properties and ligand efficiencies in the successful designs of drug candidates or their fragments [12] or to evaluate the influence of drug-like concepts on decision-making in medicinal chemistry [13]. However, the direct influence of drug complexity on market behavior of drugs has never been analyzed.

Probing market success versus drug complexity in the top 100 drug bestsellers

Essentially, a disagreement of the target oriented drug design versus drug-likeness is that "the benefit of high in vitro potency" (target based strategy usually resulted in the increase of drug complexity and MW) "may be negated by poorer ADMET properties" [11]. Similar, dichotomy limited by drug complexity can be observed for the enthalpy versus entropy controlled effects in drug optimization. Although these problem is only now being revealed at a structural level, the opportunity for enthalpy-driven optimization decreases with increasing molecular size [8].

Intuitively, medicines winning market competition could form a standard of the ideal drug, while a list of bestselling drugs could be a probe of the whole drug population at the highest quality level. A question arises if drug's success on the market can be related to their *complexity*. The purpose of this study is to analyze to what extent the list of bestselling drugs, modeling the ideal medicines, can reflect and/or clear current trends in pharma. We used the US market as with top 100 bestselling drugs (see: http://www.drugs.com/stats) a model, which allows to compare the data with FDA approvals' statistics.

In Fig. 1 we compared the mean ADMET score, clog *P* (Fig. 1a) and MW (Fig. 1b) in the populations of the top 100 list versus all of the FDA approvals versus timing. The mean ADMET score amounted to 3.7 for the FDA versus 2.5 for the top 100 list. MW is another measure of molecular complexity. Both parameters clearly distinguish the tested populations, i.e. the MW and ADMET score are higher for the whole FDA approval population than for those on the top 100 list on average. This revealed a lower degree of complexity and a more drug-like ADMET profile for those on the top 100 list. In particular, for the top 100, the mean MW ranged from 356 to 412 Da, which beginning at 379 Da (2003) decreased slightly to the lowest level of 356 Da (2006) and then steadily increased to 412 Da (2013). The high mean value of MW for FDA approvals amounting at its maximum to 372 Da (2005). On the other hand, the evolution of the clog *P* values in the population of FDA approvals and the top 100 list was compared in Fig. 1a. It can be observed that a mean clog *P* is very stable, while its value does not clearly differentiate these populations.

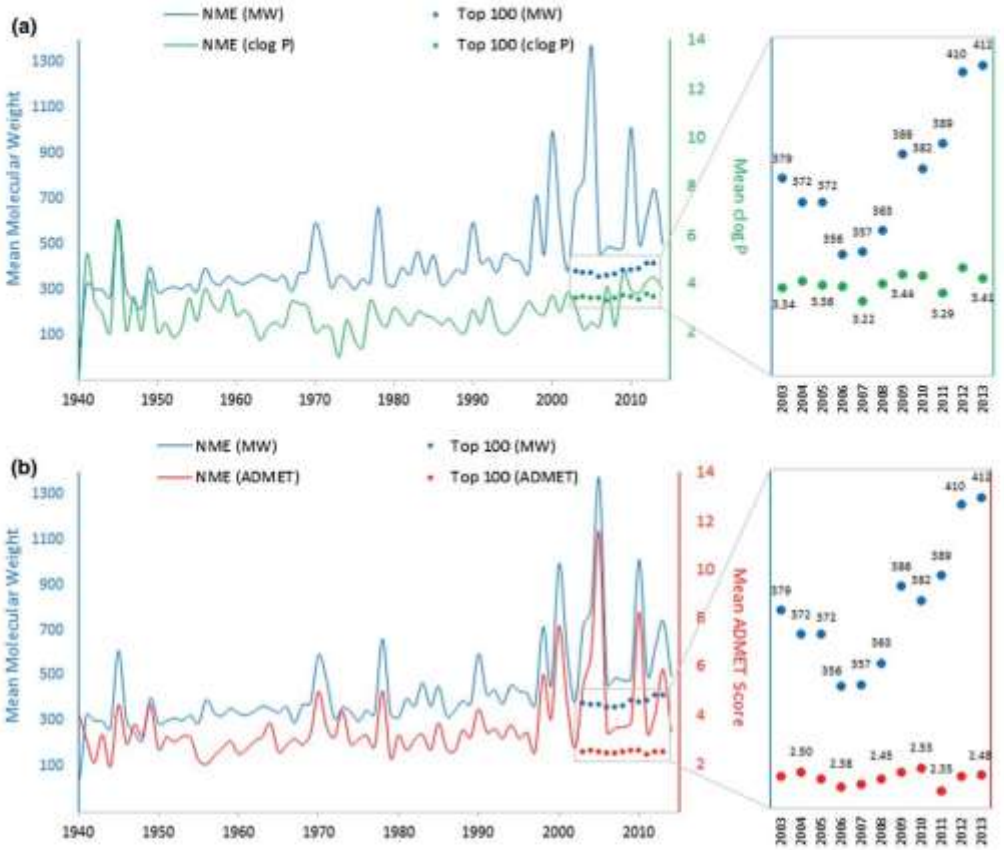


Fig. 1 A comparison of mean MW in *blue* with mean clog *P* in *green* (**a**) and mean ADMET score in *red* (**b**) for FDA approvals (*line*) and the *top* 100 bestselling drugs (*dots*). *NME* new molecular entities were included only

In Fig. 2a, b we analyzed the influence of drugs' complexity (MW) on their market success measured by the presence on the top 100 bestsellers list. Accordingly, we plotted here the cumulative number of drugs of the lower and higher complexity in the top 100 list versus the binned drug age. A value of MW = 500 Da (Lipinski rule) or 330 Da (Gleeson criterion) was used to form two categories of molecular complexity, respectively.

The analysis of Fig. 2 indicates that the Lipinski MW rule, i.e., MW < 500 indicates a clear threshold for success measured by the appearance on the top 100 list. The cumulative number of drugs of MW below the threshold value is much higher for the drugs independent of the drug (MW < 500, Fig. 2a). The decrease of the MW threshold to the Gleeson criterion (MW < 330) reverses this trend for the drugs of the age between 4 and 16 years. Interestingly, for the older drugs (over 16 years old) lower complexity molecules still win the competition (Fig. 2b). These results indicate that, more or less, the MW between 330 and 500 Da is an optimum which assures winning behavior on the market. Since binning data, i.e., plotting mean values instead of the individual ones can skew correlations [14] we analyzed if actually the MW values can influence the pIC_{50} values of drugs plotting the individual vales of the observables. The comparison of the IC_{50} values even for a single compound is risky if experiments are performed in different laboratories. Values could significantly differ. Moreover, we are comparing here different compounds and different assays designed for various targets. Recent study reports that ChEMBL database (see: http://www.ebi.ac.uk/chembl) is a

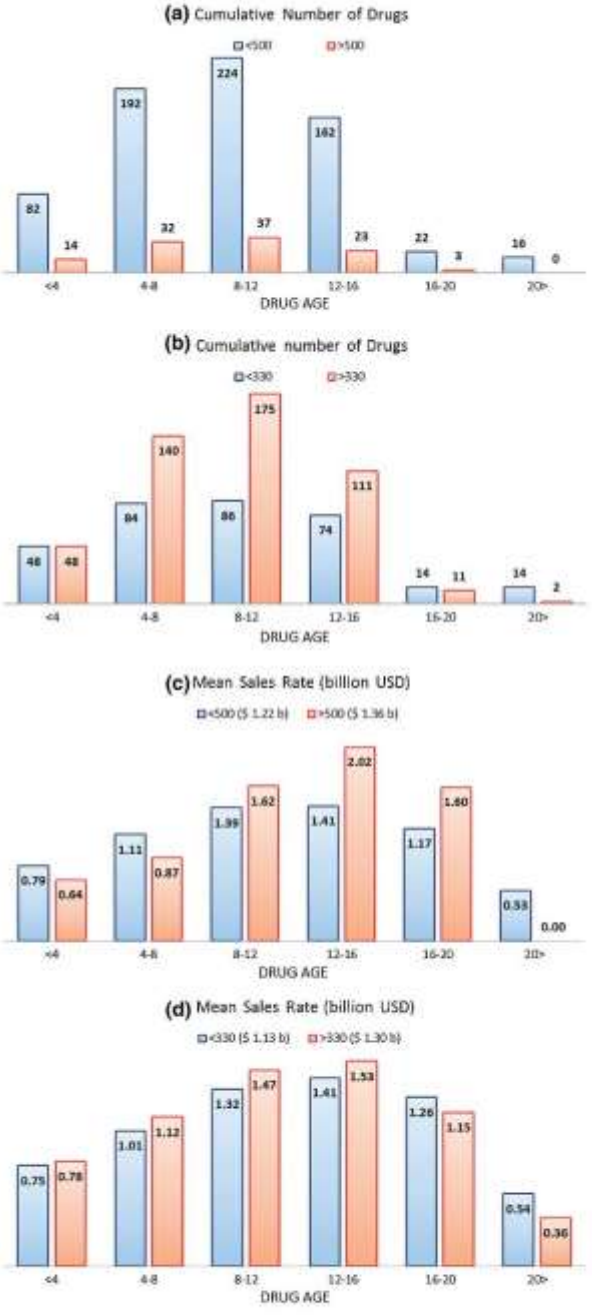


Fig. 2 Cumulative number of drugs in the top 100 list (**a**, **b**) and the corresponding mean sales values (**c**, **d**) for the two drug classes of the different complexity defined by the Lipinski MW = 500 Da (**a**, **c**) or the Gleeson MW = 330 Da criterion (**b**, **d**) versus drug age i.e. time after FDA approval. *NME* new molecular entities were included only

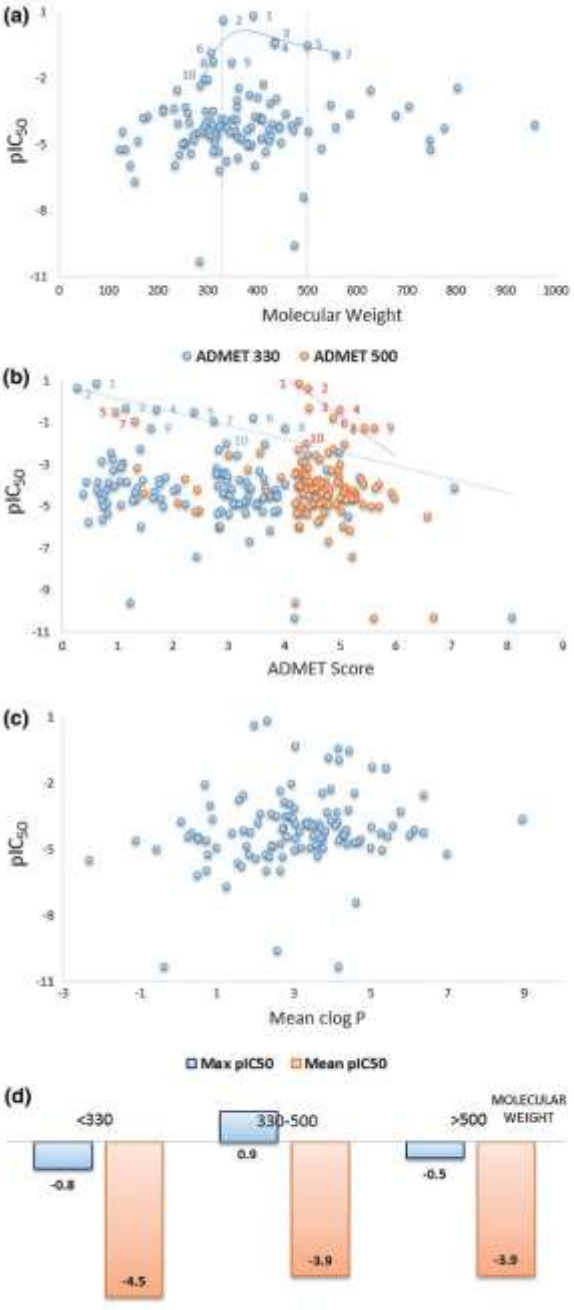


Fig. 3 Maximum and mean values of pIC_{50} in the top 100 list plotted vs. MW (**a**), ADMET 330, ADMET 500 scores (**b**) and clog *P* (**c**). The comparison of the maximum and mean values of pIC_{50} for the data binned according to the MW range of below 330 between 330 and 500 and higher than 500 (**d**). The data in (**a**) could be best described by a polynomial model with the highest values somewhere in the 300–400 range (compare supplementary materials)

Table 1 Drugs located along the highest pIC_{50} line in Fig. 3a, b

No	Trade name	Active principle	IC_{50}	pIC_{50}	MW	ADMET 330	ADMET 500
1	Spiriva	Tiotropium bromide	0.13	0.87	392.51	0.62	4.26
2	Cipro	Ciprofloxacin	0.22	0.66	331.34	0.27	4.43
3	Xarelto	Rivaroxaban	1.90	−0.28	435.88	1.15	4.44
4	Diovan	Valsartan	2.39	−0.38	435.52	1.71	5.00
5	Flovent HFA	Fluticasone propionate	3.00	−0.48	500.57	2.39	0.97
6	Gilenya	Fingolimod	6.10	−0.79	307.47	3.45	4.87
7	Benicar	Olmesartan medoxomil	7.97	−0.90	558.59	2.74	1.32
8	Detrol LA	Tolterodine	16.98	−1.23	311.46	4.02	5.43
9	Zytiga	Abiraterone acetate	17.50	−1.24	349.51	1.61	5.62
10	Arimidex	Anastrozole	94.52	−1.98	293.37	2.96	4.38

reliable repository of the mixed IC_{50} data [15]. Accordingly, we explored this database extracting the needed IC_{50} for the drugs investigated. This was available for 171 drugs, i.e., 82 % of the population investigated. Moreover, because we are comparing *polypharmacological* population with different targets, we averaged all available values for each drug molecule for all assays reported. We believe this should provide much reliable results than individual assays. The results are given in Fig. 3. The relationship between pIC_{50} versus MW (Fig. 3a) is especially illustrative here. We can see that if we divide the MW into 3 main ranges (defined by the 330 and 500 Da, which are used as Gleeson or Lipinski thresholds) the pIC_{50} decreases with the increase of MW. This can be clearly seen if we follow the maximum value line. The individual drugs located along the max value line are specified in Table 1. In particular, the maximum values of activity are the lowest within the lowest MW range, significantly higher in the medium MW range and slightly higher (vs. the lowest MW range) in the highest MW range (Fig. 3d). If we perform similar analysis for the clog *P* values, then maximal pIC_{50} are generated by the medium range clog *P* between ca. 2 and 4 (Fig. 3c).

It should be quite obvious that we should use maximal values of activity to characterize the potential of a drug for the interactions with different targets, if we realize that not all assays are optimal and not always drug can find an optimal bind into the target. Thus, our analysis seems to indicate that although it is easier to bind the target by the low molecular drug it is higher complexity and MW that provides higher activity (lower IC_{50} value or higher pIC_{50} values). Similarly, if we analyze the pIC_{50} versus ADMET scores we can see that the lower ADMET scores give higher pIC_{50} values for the maximal pIC_{50} values' line. Generally, with some minor exclusions this is obeyed for both indexes using thresholds of 330 and 500 Da. Since a definition of the ADMET score decides that this indicator increases with the increase of the difference from some optimal MW and clog *P* values, we can see that, more or less, this result follows this in Fig. 3a.

The above discussed effect can be better understand if we follow here the explanation of the probability of the fit between molecular fragments and targets [16], where a general disagreement between the lead structure activity versus its complexity is a substantial dichotomy realized just recently. Since the development of chemical, physical and biological technologies allowed the identification of compounds with increasingly lower activities, the population of potential lead structures significantly increased. However, the activity level drastically decreases for low molecular compounds. Similar rule, but of some more complicated origins seems to be obeyed for commercial drugs. The lower the MW the more probable it is to find the active compound that can interact with the target because a lower number of molecular features should fit the receptor. However, with the increase of MW the drug molecule (in optimal conditions, i.e., for maximal pIC_{50} values) can better fit the target. Eventually, after exceeding some optimal MW value we observe the decrease in pIC_{50} because the larger molecule makes the probability of drug-target fit much lower.

It is worth mentioning here that the origins for this effect should be seen more in the categories of the probability of finding the optimal molecule interacting with target than a real physico-chemical rationale. Thus, it is easier to find a molecule fitting the target for lower MW. In this region, independent of what effects we will take into account, even if a value of pIC_{50} is not high we can still hit a successful commercial drug. Instead, for higher MWs where the probability for the optimal fit with the receptor is decreasing also the probability of hitting a successful drug is much lower. However, if optimal fit is found, then a molecule would better fit the target, triggering also higher activity, i.e., generating the drugs of the higher pIC_{50}. This is however true only for the highest activity level.

It would be interesting to analyze the top 100 winners not only by the number of representatives but also by the incomes. In Fig. 2c, d we plotted the mean sales value per drug versus drug age for this population. This was

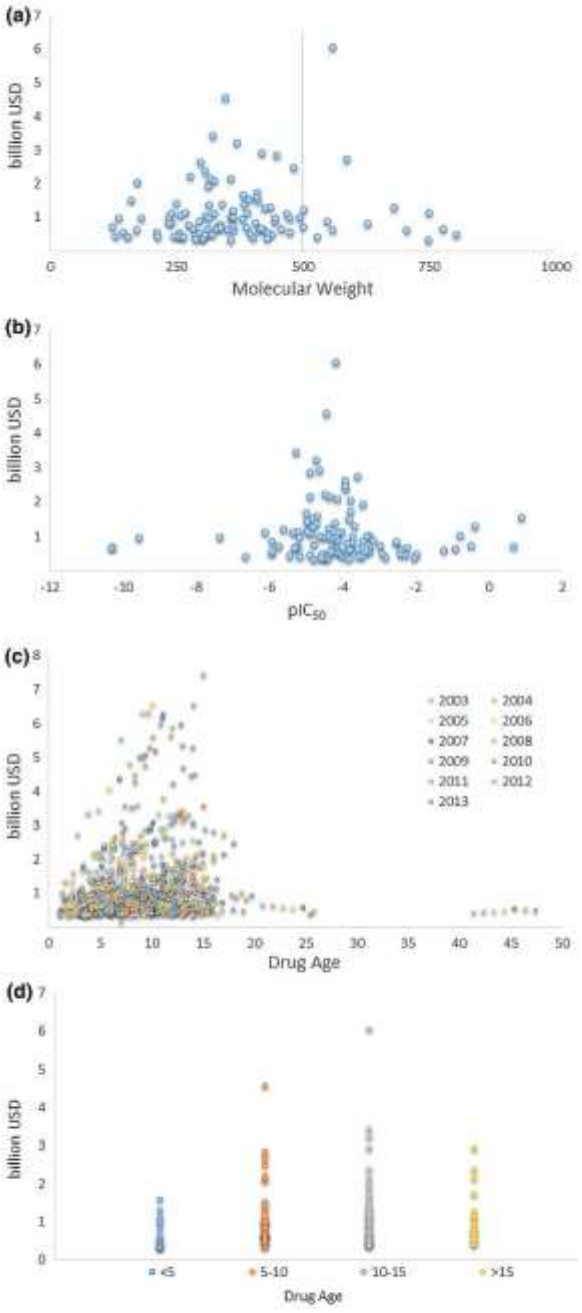
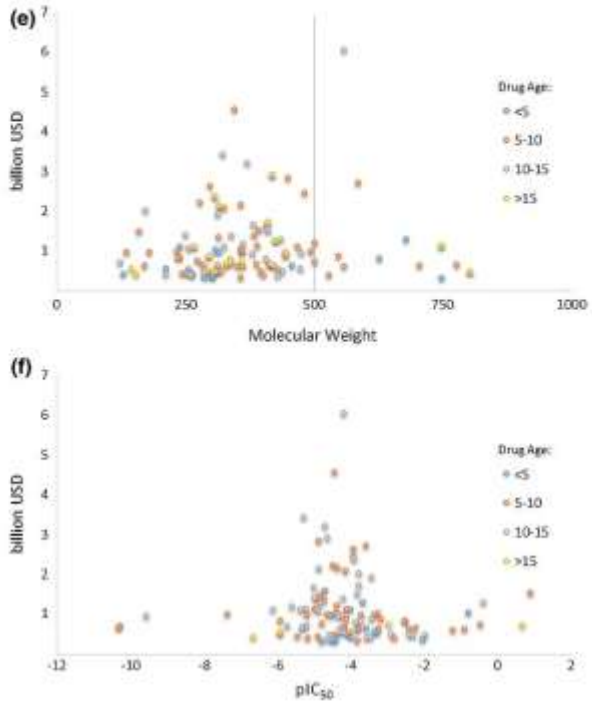


Fig. 4 Mean sales value within the top 100 list plotted versus the molecular weight MW (**a**), pIC_{50} values (**b**), and drug age (**c**), compared to the similar relationships versus drug age (**d**), MW (**e**) and pIC50 (**f**), where drug age data were split into 4 color coded classes (<5, 5–10, 10–15, >15), respectively

Fig. 4 continued



illustrated for two classes of drugs of the higher and lower complexity, indicated by the MWs of 330 (c) and 500 (d), respectively. Interestingly, market success measured by mean sales changes according to the drug age. First, the medium age drugs (4–16 years) profited better than younger or older populations. Second, for the MW threshold of 330 we observe that higher MW generally gives higher incomes. In particular, this is true for all drug age classes excluding the oldest drugs. Remarkable, for the Lipinski MW threshold (500) this keeps to be true for the drugs between 8 and 20 years old, while within the youngest and oldest drugs (below 8 and over 20 years old) we observe the reverse relationship, i.e., lower MW (below 500) drugs provide higher mean incomes than higher MW drugs.

Actually, if we analyze the success ratio for the binned data, the lower complexity drugs are the winners, while on average summing the incomes we got slightly more income from the higher complexity (Fig. 2). However, if we plot individual incomes of all drugs vs. MW (Fig. 4a), pIC_{50} (Fig. 4b) and drug age (Fig. 4c) we can see that there is always some optimal value where incomes are taking the highest value. Moreover, it is a single drug Lipitor (Atorvastatin), that skews the binning statistics (Fig. 4a) shifting this in the advantage of the highest MWs. In Fig. 4b we analyze directly the influence of pIC_{50} on the market success. We can see that similarly to the MW and drug age the highest incomes are drugs of the intermediate pIC_{50} values in the range of -3.5 to -5.

In Fig. 4d–f we presented the relationship between sales and MW, and sales and pIC50 after adjustment for drug age (compare supplementary results for the respective analyses of covariance). Formally, this means we are binning drug age into four ranges examining the relationship between sales and MW, and sales and pIC50 after adjustment for drug age using analysis of covariance. However, the analyses of the plots of all data (Fig. 4d–f) indicted that in the general sense there is no correlation or clear patterns here.

Summing up, the influence of the molecular complexity on market success is not fully clear. With some exclusion a higher molecular complexity can pay off by providing higher incomes for the compounds of the highest pIC_{50} values. The reverse of this trend for the youngest drugs for the Lipinski threshold is remarkable (Fig. 2c). Since this threshold currently determines drug design decisions we

can hypothesize that presently target oriented strategies are less efficient that the traditional drug-likeness criterion. More or less, the drug age in this area (>8 years) corresponds to the strategy switch from the phenotypic to the target oriented scheme. The profits for the medium age drugs are generally higher, which may be connected to marketing strategies, while the traditional low MW drug success for the youngest drugs may indicate the engagement in the high risk strategies where new drug targets are sought after which significantly increases the drug development risk. Since this innovative strategy is not an easy path, apparently we need more time to fully have a handle of target oriented new technologies which will provide the molecules capable of wining the top 100 competition.

Concluding remarks

The in vitro potency of drugs, when probed as a function of the ADMET parameters, indicated that many drugs have a considerable off-target activity, and that in vitro potency does not correlate strongly with therapeutic dose. Thus, "the benefit of high in vitro potency" (target based strategy) "may be negated by the ADMET properties" [11]. When analyzing the list of top 100 drug bestsellers versus all of the FDA approvals, we discovered that on average lower-complexity (MW, ADMET score) drugs are winners of the top 100 list in terms of numbers but that, especially, up to some optimal MW value, a higher molecular complexity pays off with higher incomes. This indicates that slim drugs are doing better but that fat drugs are bigger fishes to catch. Furthermore, drug complexity is increasing, i.e., the mean MW of the top 100 list indicates an increasing trend which means pharma R&D continues to go the target-oriented strategy by increasing drug complexity despite the fact that this is still a bumpy road to market success.

Acknowledgments Anonymous reviewers were kindly acknowledged for a valuable remarks allowing for a substantial improvement of the original text. JP kindly thank the financial support of NCBR Grants ORGANOMET No: PBS2/A5/40/2014 and TANGO1/266384/NCBR/2015.

References

1. Scannell JW, Blanckley A, Boldon H, Warrington B (2012) Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 11:191–200. doi:10.1038/nrd3681
2. Lindborg SR, Persinger CC, Sashegyi A, Mallinckrodt C, Ruberg SJ (2014) Statistical refocusing in the design of Phase II trials offers promise of increased R&D productivity. Nat Rev Drug Discov 13:638–640. doi:10.1038/nrd3681-c1
3. Pammolli F, Magazzini L, Riccaboni M (2011) The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov 10:428–438. doi:10.1038/nrd3405
4. Lendrem D, Senn SJ, Lendrem BC, Isaacs JD (2015) R&D productivity rides again? Pharm Stat 14:1–3. doi:10.1002/pst.1653
5. Polanski J, Bogocz J, Tkocz A (2015) Top 100 bestselling drugs represent an arena struggling for new FDA approvals: drug age as an efficiency indicator. Drug Discov Today 20:1300–1304. doi:10.1016/j.drudis.2015.06.015
6. Swinney DC, Anthony J (2011) How were new medicines discovered? Nat Rev Drug Discov 10:507–519. doi:10.1038/nrd3480
7. Walters WP, Green J, Weiss JR, Murcko MA (2011) What do medicinal chemists actually make? a 50-year retrospective. J Med Chem 54:6405–6416. doi:10.1021/jm200504p
8. Hann MM, Keseru GM (2012) Finding the sweet spot: the role of nature and nurture in medicinal chemistry. Nat Rev Drug Discov 11:355–365. doi:10.1038/nrd3701
9. Hann MM (2011) Molecular obesity, potency and other addictions in drug discovery. Medchemcomm 2:349–355. doi:10.1039/c1md00017a
10. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432:855–861. doi:10.1038/nature03193
11. Gleeson MP, Hersey A, Montanari D, Overington J (2011) Probing the links between in vitro potency, ADMET and physicochemical parameters. Nat Rev Drug Discov 10:197–208. doi:10.1038/nrd3367
12. Ferenczy GG, Keserű GM (2013) How are fragments optimized? a retrospective analysis of 145 fragment optimizations. J Med Chem 56:2478–2486. doi:10.1021/jm301851v
13. Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 6:881–890. doi:10.1038/nrd2445
14. Kenny PW, Montanari CA (2013) Inflation of correlation in the pursuit of drug-likeness. J Comput Aided Mol Des 27:1–13. doi:10.1007/s10822-012-9631-5
15. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC_{50} data: a statistical analysis. PLoS One 8(4):e61007. doi:10.1371/journal.pone.0061007
16. Zartler ER, Shapiro MJ (2005) Fragonomics: fragment-based drug discovery. Curr Opin Chem Biol 9:366–370. doi:10.1016/j.cbpa.2005.05.002