КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри супрамолекулярної хімії
проф. Сергій Вікторович Рябухін
Протокол №____ засідання кафедри
від “____” _____20___ р.

Критична оцінка баз даних синтетично доступних молекул

Випускна кваліфікаційна робота магістра
студентки спеціальності
102 Хімія
ОП «Хемоінформатика»
Перебийніс Мар’яни Юріївни

Науковий керівник від кафедри
професор кафедри
супрамолекулярної хімії
д.х.н. **Комаров Ігор Володимирович**

Робота виконана у Laboratory for Therapeutic Innovation
Faculty of Pharmacy, University of Strasbourg
під керівництвом Dr. **Didier Rognan**

Оцінка захисту роботи

_____

Київ − 2022 р.

Анотація

Ключові слова: хімічний простір, база даних, скринінгові сполуки, віртуальний скринінг

Хімічний простір – це неймовірно велика бібліотека людського хімічного знання – віртуального чи фізичного. Але справжні його розміри уявити надзвичайно важко. Лише кількість сполук, які відповідають правилам Ліпінського (drug-like compounds) становить 10^{33}, а тих що потенційно можуть проявляти біологічну активність навіть більше -близько 10^{60}. Нині основними та найбільшими хімічними просторами є Enamine REAL Space, Otava CHEMriya та WuXi GalaXi. Але скільки молекул із цих хімічних просторів є реальними? Скільки існує фізично та є уже доступними для скринінгу? Реально доступні сполуки були зібрані у in-house базу даних (Bioinfo DB 22.1) наявних скринінгових молекул від 25-ти емпірично відібраних постачальників. У даній роботі ми досліджували як саме розподілені уже існуючі скринінгові сполуки між цими трьома просторами, який із них містить більше реальних сполук, як перекриваються ті частини хімічного простору між собою. Було також досліджено в якому із просторів знаходиться більше унікальних молекул, а також який із них є кращим вибором для проведення віртуального скринінгу.

Annotation

Keywords: chemical space, database, screening compounds, virtual screening

Chemical space is a vast library of human chemical knowledge - virtual or physical. But its actual size is challenging to imagine. For example, only the number of compounds that meet the Lipin rules (drug-like compounds) is 1033, and those that can potentially show biological activity even more - about 1060. Currently, the leading and largest chemical spaces are Enamine REAL Space,

Otava CHEMriya, and WuXi GalaXi. But how many molecules from these chemical spaces are real? How many exist physically and are already available for screening? Actually, available compounds were collected in an in-house database (Bioinfo-DB 22.1) of functional screening molecules from 25 empirically selected suppliers. In this work, we investigated how exactly the existing screening compounds are distributed between these three spaces, which of them contains more real compounds, and how those parts of the chemical space overlap. It was also investigated in which of the spaces there are more unique molecules and which of them is the best choice for virtual screening.

Мета роботи: використовуючи алгоритм пошуку SpaceMACS, здійснити пошук по 3-х основних хімічних просторах по параметру максимальної подібності підструктури по Танімото (MCS Tanimoto similarity). Для результатів пошуку було знайти порогове значення подібності, порівняти між собою результати, проаналізувавши розподіл скафолдів між хімічними просторами.

Було поставлено такі **завдання**:

1. Оновити дані на актуальні у Bioinfo DB станом на лютий 2022 року.
2. Згенерувати на основі цієї бази даних Мурко скафолди.
3. Використовуючи SpaceMACS алгоритм знайти для кожного скафолду максимальну спільну підструктурну подібність в ультра-великих комбінаторних просторах (Enamine REAL Space, Otava CHEMriya та WuXi GalaXi)
4. Встановити порогове значення подібності, базуючись на розподілі результатів (значення подібності-кількість результатів) та емпіричному аналізі результатів. Порівняти між собою на перекривання результати із значенням подібності вище порогового та 1.0.
5. Проаналізувати отримані результати.

Об’єктом дослідження є взаємозв’язок між трьома хімічними просторами та базою даних скринінгових сполук.

Предмет дослідження: порівняння та аналіз результатів на основі пошуку по максимальній спільній підструктурі Мурко скафолдів in-house бази даних.

Методи дослідження: хемоінформатичний аналіз даних скриптовими методами із використанням мови програмування Python та командної

оболонки Bash на персональному комп’ютері і розрахунковому центрі (CC.IN2P3).

Структура роботи. Розділ 1-2. Огляд літератури. Розділ 3. Експериментальна частина. Розділ 4. Результати та їх обговорення. Висновки. Список використаних джерел. Додатки

Зміст

Розділ 1. Хімічний простір та розробка лікарських засобів

1.1 Поняття хімічного простору

Неможливо заперечувати, що медична хімія входить в еру «big data». Data-driven підхід займає вже зараз важливу нішу в новітній фазі розробки лікарських засобів. Кількість інформації, з якою працюють зараз медхіміки та хемоінформатики зростає лавиноподібно, тому цінність і складність у правильному збиранні та аналізі цих даних лише підвищується. Ми зараз створюємо нову інформацію (data) швидше, аніж встигаємо її проаналізувати, інтерпретувати та зробити на її основі висновки[1]. Тому варто прийняти той факт, що з часом роль науковців, що працюють з великими даними (data scientists) буде лише посилюватися, а ніша буде ставати ще більш популярною.[2].

Нині, для опису варіацій хімічних молекул використовується поняття «хімічного простору». Органічні молекули (а саме ними ми оперуємо при розробці лікарських засобів) визначаються кількістю, типом, топологічною зв'язністю та стереохімією атомів, описаними їх структурною формулою. Кількість цих молекул зростає із кожним роком, а їх використання є переважно в контексті медичної хімії[3]. Хемоінформатика надає різноманітні обчислювальні інструменти для обробки величезної кількості інформації, створеної цими мільйонами молекул, зокрема для забезпечення класифікації бази даних. і передбачення біоактивності. Власне розвиток та інструментарій хемоінформатики дозволяє зараз говорити про хімічний простір в контексті big data, оскільки хімічний простір є поєднанням не лише реальних, але й віртуальних сполук.

Наскільки великим є хімічний простір? Зрозуміло, що кількість потенційних молекул, які можуть бути синтезовані, є великим, але як щодо якихось конкретних чисел? Одним із найперших «осягнути» хімічний простір був Weininger, який використав сет із 150 замісників гексану та оцінив розмір простору для малих молекул у 10^{29}. Незалежно від нього, Bohacek та співробітники оцінювали кількість можливих органічних сполук, використовуючи лінійне нарощування хімічного ланцюга атомами C, N, O або S[4]. Враховуючи можливість утворення подвійних та потрійних зв'язків, автори дійшли до 6 варіантів (в середньому) росту ланцюга. Встановивши ліміт у 30 атомів (для збереження визначення простору малих молекул), було досягнено значення у 6^{30} або 2×10^{23} молекул. Додаючи ще можливість утворення циклу, кількість молекул росла до 10^{40} із максимальною кількістю у 4 цикли та 10 розгалужень. Комбінація уже лінійних частин та розгалужень утворювала 10^{63} молекул.

У наступній статті, Ertl використав аналіз органічних замісників, щоб та оцінив кількість можливих органічних молекул десь між 10^{20} і 10^{24} базуючись на аналізі замісників із менш ніж 13 атомів на базі даних із 3 мільйонів комерційно доступних сполук[5].

Напевно однією із найбільш популярних оцінок розміру хімічного простору, є дослідження Polishchuk та співробітників, де вони використали розподіл сполук, згенерованих GDB, яка енумерувала усі можливі молекули із кількістю атомів до 17[6]. Побудувавши графік кількості важких атомів проти логарифмічної кількості перерахованих молекул (**Рис 1.1**), автори вивели рівняння:

$$\log M = 0.584 N log(N) + 0.356$$

де M – кількість молекул, N – кількість атомів в молекулі. Використовуючи це рівняння та екстраполюючи його на молекули із 36 атомами, автори дійшли до оцінки хімічного простору drug-like сполук у 10^{33}.

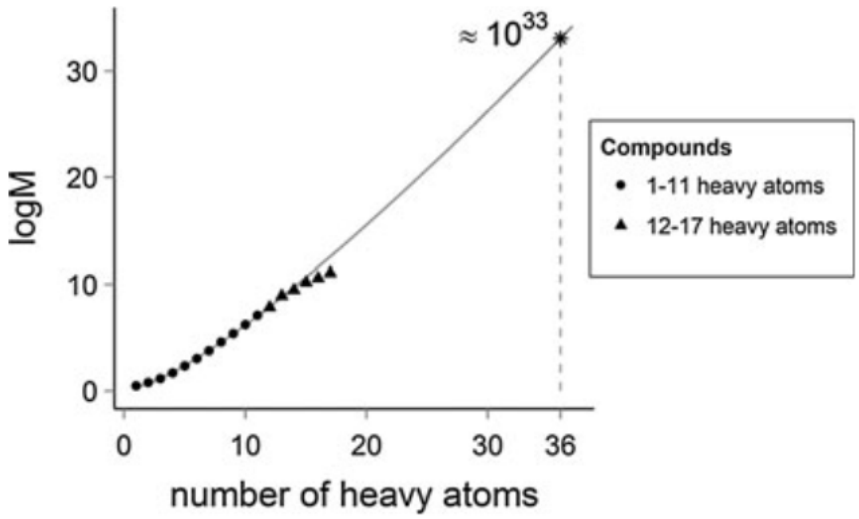

Рис 1.1. Екстраполяція числа сполук (M) як функції кількості важких атомів (N) на основі даних, взятих з GDB-17. Криву підігнано для сполук з N=1–11 атомами, оскільки сполуки з N > 12 були отримані за допомогою іншого правила відбору

Таким чином, кількість молекул у хімічному просторі варіюється між 10^{30} та 10^{60}, але не існує узгодженості щодо їх точного числа. Хоча очевидно, що сполуки, які потенційно можуть стати лікарськими засобами – лише крихта серед цієї міріади можливостей[7].

1.2 Великі дані = великі проблеми

Перш за все, потрібно розграничити деякі терміни, такі як «простір», «бібліотека» та «база даних». У своєму огляді, Warr і колеги пропонують цю диференціацію задля чіткості в їх визначенні[8]. Так, вони пропонують називати «простором» комбінаторно побудовані колекції сполук, які є досить великими, що робить складним задачу енумерації усіх сполук у цій колекції. «Бібліотеки» ж у цей час є енумерованою колекцією повних структур, яка зазвичай має розмір менше 10^9. «База даних» -це спосіб зберігання бібліотеки, наприклад, у вигляді реляційної системи управління базами даних. До важливості цього питання ми повернемось пізніше, але важливим на даний момент є саме розмежування між простором та бібліотекою. Тобто, коли бібліотека може вважатися простором, а простір – бібліотекою. Як ми

бачимо, основними є два параметри – спосіб зберігання та розмір. Так, простір має в собі молекули, які представлені не в прямому вигляді – вони є «закодованими» та являють собою комбінацію будівельних блоків, які можуть взаємодіяти за визначеною реакцією. У бібліотеці сполука уже представлена чітко. У той же час, якщо енумерувати хімічний простір (зробити перетворення із блоками, записані у реакціях) – він стане бібліотекою. Але чи так це? Тут приходить на допомогу другий параметр – розмір. Так, бібліотекою може бути колекція сполук, менша аніж 10^9. Нині існуючі хімічні простори мають орієнтовно такі розміри. Тому концептуально, якщо спробувати енумерувати сучасні хімічні простори, більшість із них не можна вважати бібліотекою через обмеження у розмірі, якщо дотримуватися визначень, наведених в огляді. Та існує одне «але» - існуюча тенденція до росту хімічних просторів (**Рис 1.2**)[8]. Збільшенню розміру хімічних просторів є дуже просте пояснення – зі збільшенням загальної кількості молекул, ми підвищуємо вірогідність знайти хороші хіти, а значить – створити саме ту, єдину та неповторну сполуку, що стане новим лікарським засобом. Ми бачимо з графіку, що з кожним роком відбувається ріст деяких колекцій хімічних сполук, але якщо у випадку колекцій, що мають в собі сполуки, що можуть з середньою чи дещо високою ймовірністю бути синтезовані, ріст відбувається постійно та досить різко, то у випадку сполук, які можуть бути синтезовані із дуже високою ймовірністю – ріст більш поступовий. Абсолютним лідером серед них є GSK XXL із кількістю сполук близько 10^{26}, у той час коли колекції сполук більш «реальних» мають розміри лише в діапазоні 10^8-10^{10}. Реальність цих чисел можна підтвердити наприклад тим, що станом на 2020 рік унікальний CAS номер уже мали 159 мільйонів сполук (не наведено на рисунку).

Звичайно, ми можемо нарощувати розміри хімічних просторів до того моменту, поки не покриємо абсолютно весь хімічний простір. Але чи насправді це так просто і чи є в цьому сенс?

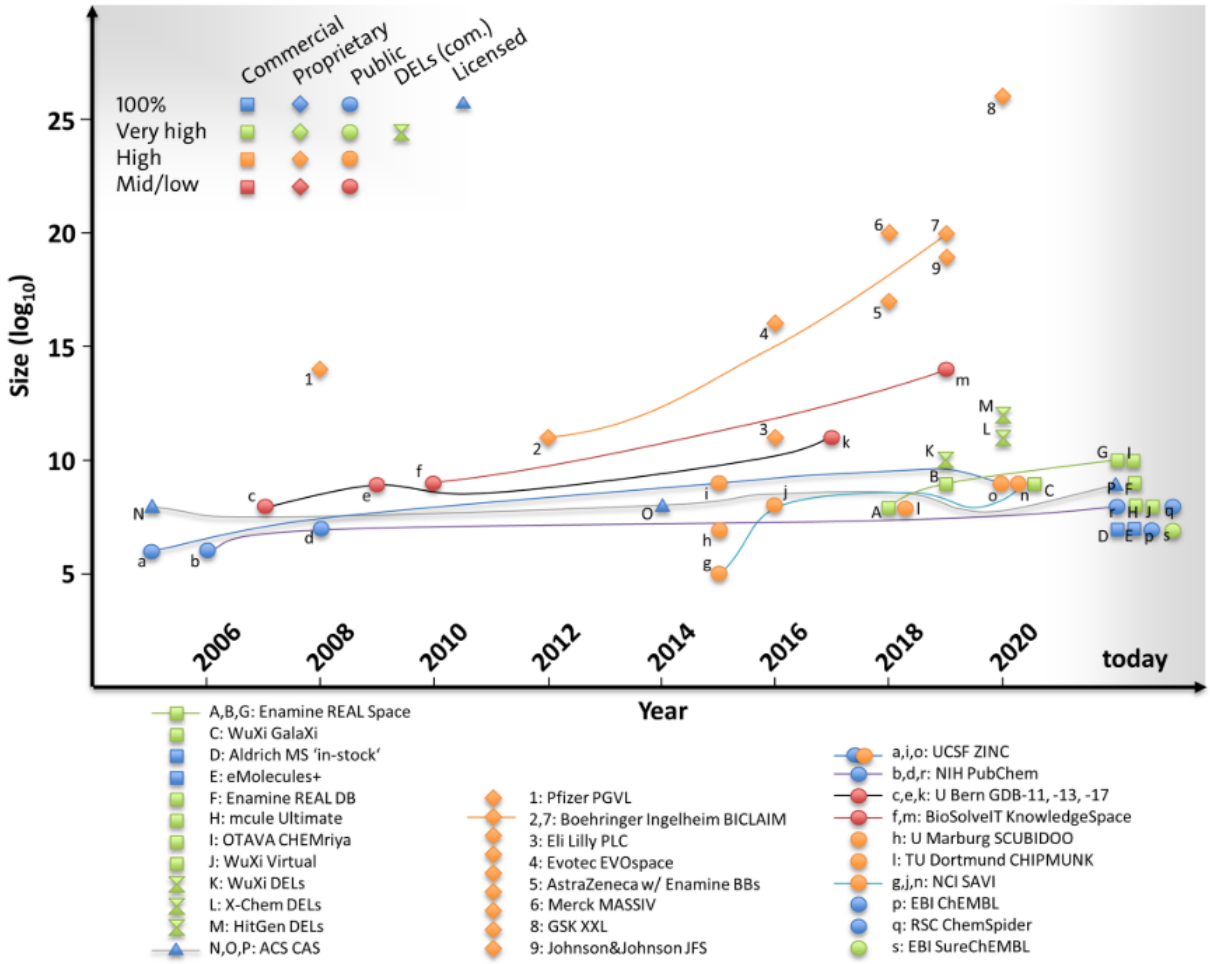

Рис 1.2. Розвиток великих колекцій хімічних сполук з часом: комерційні постачальники (великі літери, квадрати), комерційні DNA encoded бібліотеки (великі літери, подвійні трикутники), власні простори (цифри, ромби) та загальнодоступні колекції (маленькі літери, сфери). Колір показує припущення щодо ймовірності доступності синтетичних сполук у категоріях 100% (синій), дуже високий (зелений), високий (помаранчевий) і середній або низький (червоний). Лінії з'єднують версії однієї колекції. Кілька точок даних з однаковими координатами малюються поруч у довільному порядку.

Насправді ж, використання великих масивів даних – це завжди проблема. У випадку із хімічним простором – це компроміс між вартістю, розміром та часом[9][10]. Таким чином, якщо ми захочемо енумерувати наш хімічний простір, виникне проблема зберігання фінальної бібліотеки, що вилиється у фінансову проблему. Згідно із даними компанії Гугл, зберігання 100 петабайт даних (орієнтовно 10^{11} сполук) буде коштувати 1.1 мільйон доларів щорічно.

Тому науковці дійшли до висновку, що зберігати хімічний простір в енумерованому вигляді – не розумно. Виходом із цієї ситуації стало використання комбінаторного підходу для запису простору.

Отож, замість того, щоб записувати хімічний простір у вигляді «плоскої» (flat) бібліотеки взагалі усіх сполук, тепер було достатньо зберігати лише будівельні блоки (яких значно менше) та реакції, за якими вони можуть взаємодіяти. І при запиті до цього хімічного простору, спочатку відбувається пошук по будівельних блоках, а далі «взаємодія» обраних будівельних блоків за відповідною реакцією із генерацією фінальної сполуки, яка власне і є нашим запитом.

Серед цих комбінаторних хімічних просторів, варто виділити важливу нішу – хімічні простори від окремих постачальників, так звані каталоги сполук «на вимогу» (make-on-demand). Вони представляють собою ультравеликі хімічні простори. Так, основними просторами для досліджень є Enamine REAL Space (19×10^9 сполук, версія 04/2021), Otava CHEMriya (11×10^9 сполук, версія 04/2021) та WuXi GalaXi (2.1×10^9 сполук, версія 11/2020).

В основі Enamine REAL Space було покладено 156 синтетичних протоколів із кількістю будівельних блоків більше 104 тисяч. Створення сполуки відбувається внаслідок взаємодії двох чи трьох білдинг блоків[11]. Для Otava CHEMriya було використано 44 внутрішніх (in-house) реакцій для 30 тисяч будівельних блоків. Реакції були застосовані від 2 до 4 будівельних блоків за раз. WuXi GalaXi було створено за допомогою 30 різних типів реакцій для 155 тисяч блоків і для комбінації 2-3 блоків. Було також досліджено, наскільки перекриваються дані хімічні простори[12], оскільки кількість сполук досить значна, та за значного перекривання цінність цих розмірів зменшується. Встановили, що перекривання між просторами значно менше, аніж очікувалось (**Рис 1.3**):

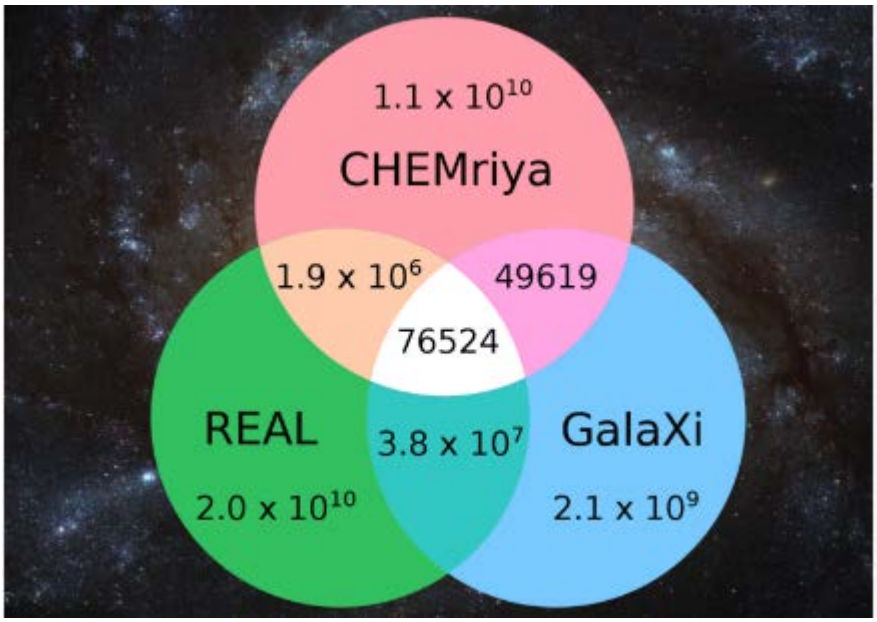

Рис 1.3. Показано перекриття Enamine REAL Space (зелений) і GalaXi Space (блакитний) з CHEMriya Space (кораловий). Цифри в сферах позначають кількість сполук, що містяться в хімічних просторах. Цифри в перекритті сфер вказують на перекриття відповідних двох або трьох хімічних просторів. Наприклад, рожеве перекриття вказує на набір сполук, що містяться в просторі CHEMriya і просторі GalaXi, а не в Enamine REAL Space.

Та варто зазначити, що ці хімічні простори є віртуальними. Серед них Enamine REAL позиціонує себе як хімічний простір сполук, які можуть бути синтезовані з високою долею вірогідністю. Та постає питання, а скільки із цього різноманіття сполук існує в реальності? Якщо ми знайдемо хіт при віртуальному скринінгу в якомусь із цих просторів, чи є він вже синтезованим? Чи питання можливості його синтезу постане після замовлення?

1.3 Бібліотеки скринінгових сполук

Скринінг хімічних сполук – основоположна частина розробки нових лікарських засобів, оскільки цей етап встановлює, чи має сполука біологічну активність чи ні. Однак, варто зазначити, що нині існують два взаємодоповнюючі підходи[13] (**Рис 1.4**) – високопропускний скринінг (high-throughput screening, HTS) із використанням біологічних есеїв та реальних хімічних сполук, що дають можливість оцінити активність сполуки *in vitro* та

віртуальний скринінг, який оперує бібліотеками віртуальних сполук шукаючи активність *in silico*.

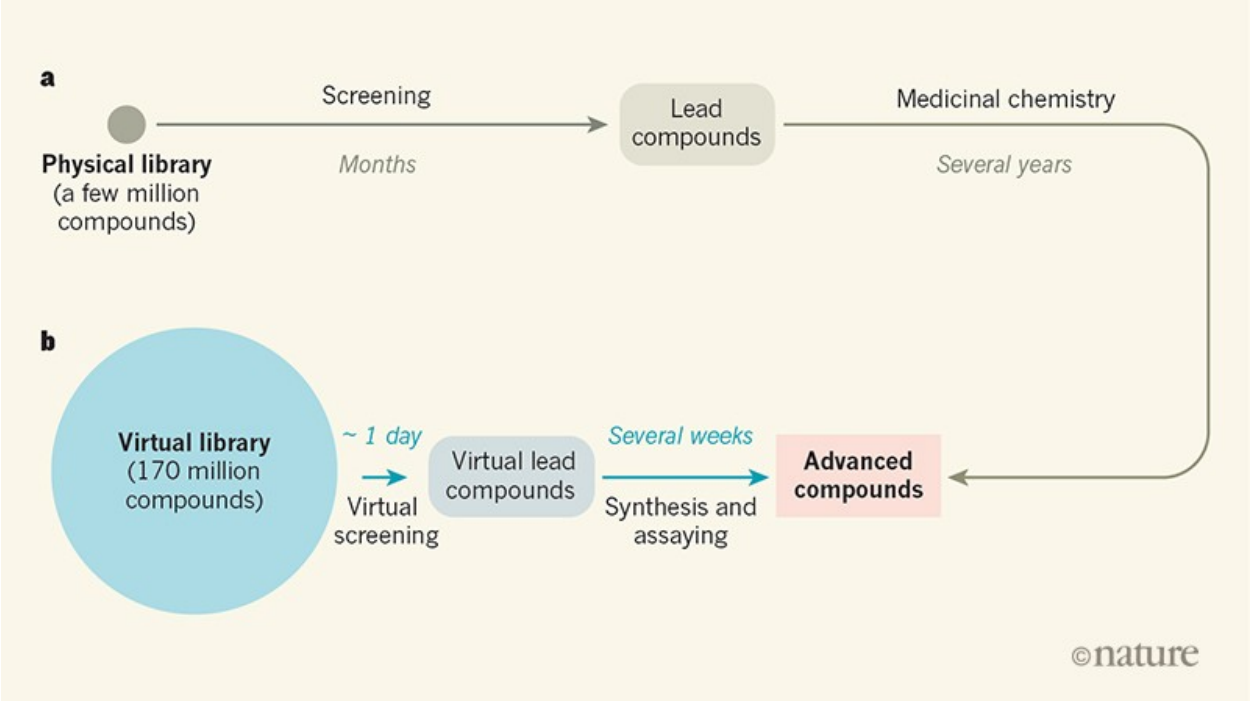

Рис 1.4. Віртуальний скринінг надвеликих бібліотек може підвищити ефективність виявлення ліків. Скринінг фізичної бібліотеки сполук (a) зазвичай дає більш точні результати, оскільки взаємодіє фізично із біологічною мішенню, але є декілька проблем, як обмеженість цієї скринінгової бібліотеки, дороговизна синтезу цієї бібліотеки та підготовка великої кількості біологічних мішеней, а також – довготривалість. Тоді на допомогу приходить віртуальний скринінг (b), який оперує зі значно більшими масивами сполук, хоча це займає значно менше часу і значну кількість коштів.

Пошук та розробка лікарських засобів завжди була справою довгою, дорогою та ризиковою щодо інвестицій. З року в рік ця галузь стає все більш інноваційною та технологічною, щоб зменшити вартість пошуку нових лікарських засобів та підвищити кількість успішних кейсів. Однак, дослідження показують, що починаючи з 1950-х років, кількість ліків затверджених на один мільярд доларів США, які були вкладені в розробку, зменшувались вдвічі, з урахуванням рівню інфляції[14]. Досить показовим був період спаду HTS, оскільки технологія була досить хороша, на неї покладали багато надій, але через не зовсім коректне її використання – це вилилось в досить кризовий період для фармацевтичних компаній[15]. Проблема була в

тому, що технологія була настільки багатообіцяюча, що на біологічних есеях тестували абсолютно всі сполуки, які тільки знаходили. Таким чином, був дуже значний відсоток псевдо позитивних результатів (це зараз ми уже знаємо про PAINS), часто бібліотеки були не різноманітні, тому результати давали невеликий вклад, або ж це були великі молекули, модифікувати які вже не мало сенсу. На той час, внутрішні колекції фармацевтичних компаній сягали мільйонів сполук. Тому можна лиш уявляти, які величезні суми були втрачені компаніями через недостатнє розуміння застосування технології. З іншого боку, це дало значний розвиток для досліджень вибору бібліотек для скринінгу.

Бібліотеки скринінгових сполук із тих часів пройшли досить багато модифікацій через раціоналізацію підходів. Був час переосмислити: кількість чи якість? І було вирішено спочатку зійтися на другому, оскільки саме низька якість сполук та біологічних есеїв була основною помилкою минулого. Зараз же, постачальники знову можуть повертатися до збільшення розмірів своїх колекцій, не випускаючи з виду якість своїх бібліотек у значенні їх різноманітності та фізико-хімічних параметрів, які грають важливу роль при пошуку нових ліків. Так, у дослідженні [16] було показано аналіз 33 постачальників скринінгових сполук, з 8 з яких мають 80% унікальних сполук. Ними є Abamachem, AnalytiCon Discovery, BCH Research, Enamine, FCH Group, Intermed, Selenachem, UOrSy. Усі бібліотеки, окрім AnalytiCon Discovery, містять більше мільйону сполук. В загальному, після проведення процедур стандартизації та видалення дублікатів, було отримано 16 902 208 унікальних сполук. Також було побудовано графіки розподілу молекулярних властивостей, що дали змогу візуалізувати, що дійсно нині бібліотеки скринінгових сполук є різноманітними, але в той же час можуть бути піддані кластеризації, оскільки є тенденція до перекривання областей розподілу деяких із властивостей між різними постачальниками.

Таким чином, у даній статті було проведено аналіз наявних на той час постачальників скринінгових сполук, проаналізовано різноманітність сполук та скафолдів, зміни, які зазнали ці бібліотеки протягом 2010-2017 років у кількісному та якісному значенні, а також показано перекривання сполук між постачальниками із базою даних ChEMBL.

Але на даний момент не існує дослідження, яке показує перекривання віртуального простору хімічних сполук із бібліотеками скринінгових сполук. Це і стало ціллю нашого дослідження – знайти серед віртуальних сполук реальні лікарськоподібні сполуки, у нашому випадку - які можуть бути замовлені у вигляді скринінгових сполук («on-the-shelf»).

1.4 Роль віртуального скринінгу в сучасному пошуку лікарських засобів

Віртуальний скринінг (VS) є широко використовуваним підходом у відкритті сучасних ліків, який відіграє важливу роль у ідентифікації нових хітів. Загалом, підхід складається з швидкої оцінки in silico ймовірності того, що молекула має бажану біологічну активність. Метод, як правило, застосовується до великої бази даних лігандів, а результатом є рейтинговий список на основі ймовірності відповідної активності[17].

Віртуальна комбінаторна хімія та різні інструменти віртуального скринінгу представлені як ключовий інструмент у розробці нових ліків, оскільки є найбільш ефективними методами дослідження у комбінації час-кошти.[18] Ці методи *in silico,* у комплексі чи самі по собі, прискорюють процес виявлення ліків. діючи як фільтри та дозволяючи зосередити експериментальну оцінку лише на сполуках, які є drug-like.

Першим етапом у віртуальному скринінгу є вибір або створення бібліотеки, де буде відбуватися скринінг (**Рис 1.5.**). Які ж існують проблеми на цьому етапі? Перш за все, це компроміс між повнотою бібліотеки та можливістю скринінгу[19]. Все просто – чим більша бібліотека, тим важча навігація в ній. Тому для більш ефективного скринінгу віртуальні бібліотеки

мають бути більш сфокусованими на таргеті, підібраними під запит користувача, оскільки це підвищує ймовірність знаходження істинно позитивних хітів.

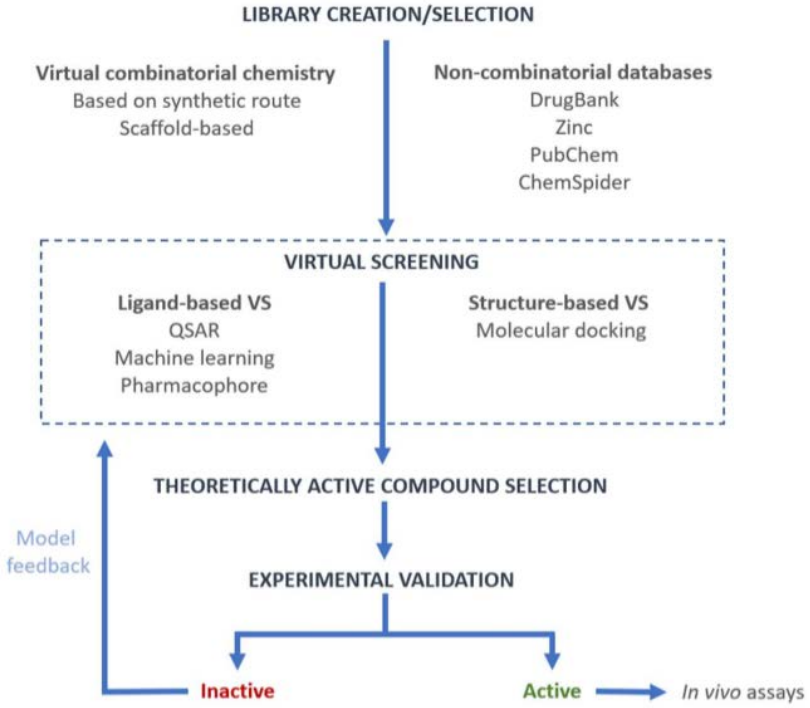

Рис 1.5. Загальна схема процедури віртуального скринінгу.

Саме тому настільки важливо є створювати та обирати правильні об’єкти для віртуального скринінгу, оскільки його результати, успіх та репрезентативність на біологічних моделях залежать напряму від якості етапу підготовки.

Особливо це питання стоїть гостро із розвитком методів машинного навчання та штучного інтелекту для пошуку лікарських засобів, оскільки фактор підбору набору молекул для навчання та тестування є визначальним. Причому використання цих технологій настільки широке, що може зустрічатися від генерування абсолютно нових SMILES базуючись на обробці існуючих SMILES як природньої мови (NLP, natural language processing) до передбачення активності молекули *in silico*, а також передбачення результатів скринінгу *in vitro* (розрахункова токсикологія або

переплетення хемоінформатичних та біоінформатичних результатів із навчанням нейромережі).

Розділ 2. Інструменти дослідження хімічного простору

2.1 Підходи та проблеми із пошуком молекул в хімічних просторах

У віртуальному пошуку лікарських засобів існують два основні підходи – ліганд-базований (серед них, QSAR/QSPR (Quantitative Structure–Activity/Property Relationship) та фармакофорбазоване моделювання) та структуро-базований (як от молекулярний докінг)[18]. Використання другого є більш переважаючою стратегією, оскільки відразу відбувається підбір ліків під конкретний таргет. Але бувають випадки, коли структура цілі є невідомою (зазвичай це пов'язано із проблемами із фізико-хімічними дослідженням, як наприклад кристалізація протеїну для отримання кристалографічного зображення[20]). Тоді на допомогу приходить ліганд-базований підхід.

Ранні підходи із ліганд-базованим *de novo* дизайном включають в себе фрагментацію компонентів бази даних на унікальні білдинг блоки (фрагмент-базований спосіб), які потім можуть бути скомбіновані як нові молекули[21][22]. У фрагмент-базованій стратегії пошуку нових лікарських засобів (**Рис 2.1**) виділять декілька основних підходів: нарощування фрагментів, з'єднання фрагментів або ж їх зливання[23].

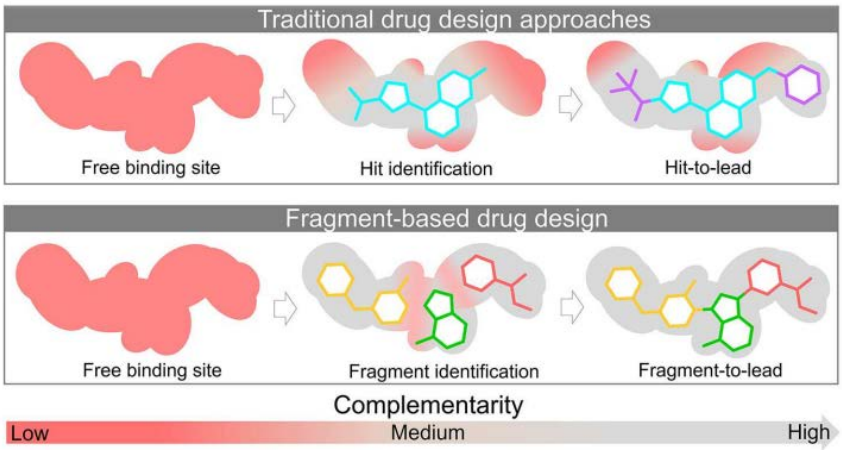

Рис 2.1 Різниця між традиційним та фрагмент базованим підходом у дизайні лідів

Та проблема цього підходу у тому, що неможливо передбачити наскільки будуть синтетично доступними результуючі молекули.

Для вирішення такої критичної проблеми, було вирішено поглянути на проблему під дещо іншим кутом – чому б не почати зберігати фрагменти і комбінувати їх у систематизований спосіб, так щоб фрагменти могли бути з'єднані по конкретній хімічній реакції? Ця ідея була втілена у життя більш ніж десять років тому компанією Файзер (Pfizer) у вигляді PGVL простору (див. **Рис. 1.2**).

Хемоінформатичний двигун в основі PGVL став ранньою версією CoLibri – інструмент для трансформації «синтетичного» знання у так званий хімічний фрагментний простір, який містив велику кількість віртуальних молекул у неенумерованому форматі[24]. Хімічний фрагментний простір містить молекулярні фрагменти і відповідні правила для їх з'єднання. Перевагою використання саме комбінаторного хімічного простору, незважаючи на більшу кількість віртуальних сполук при більшому покритті простору, є значне зменшення кількості пошукових запитів, порівнюючи із класичним простором через унікальність результуючого фрагментного простору (**Рис 2.2**).

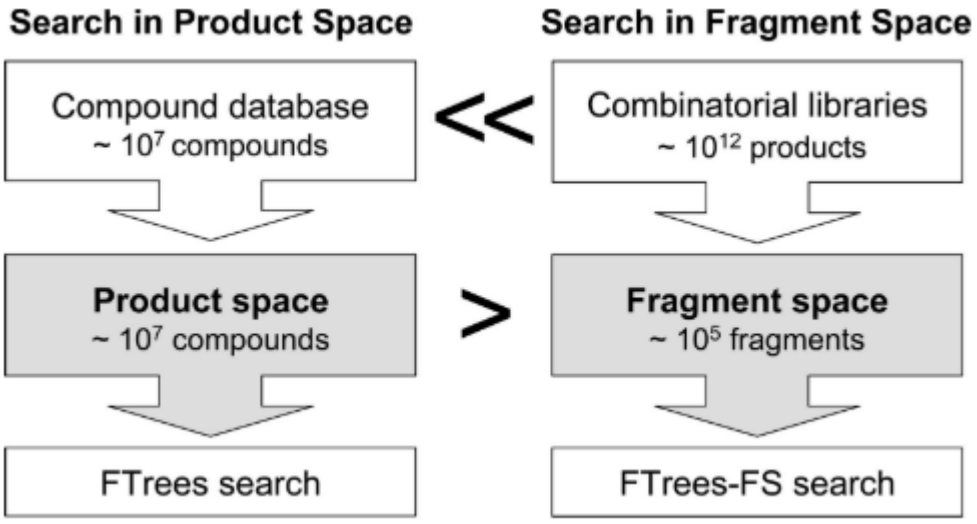

Рис 2.2. Порівняння пошуків подібності, виконаних у просторі продуктів та у просторі фрагментів. Кількість перерахованих комбінаторних бібліотечних продуктів значно перевищує кількість сполук у корпоративних базах даних. Однак кількість унікальних фрагментів менше, ніж кількість молекул у базі даних сполук. Шукаючи у фрагментному просторі, FTrees-FS може ефективно охопити величезний простір комбінаторних сполук.

Нещодавно був презентований open-source інструмент для створення хімічного простору[25]. Він може створювати хімічний простір із декількох мільярдів сполук із комерційно доступних будівельних блоків і списку поширених органічних реакцій.

Історично першим та досі одним із найбільш популярних методів пошуку у фрагментних просторах був FTrees-FS (Feature Trees – Fragment Space)[26][27]. Він базується на Feature Trees – фармакофороподібний пошук по подібності, який використовує репрезентацію у вигляді зменшеного графу (reduced graph). Замість того, щоб енумерувати весь простір і робити скринінг поміж мільярдів сполук, FTrees-FS конструює молекули напряму із простору, що підходить запиту (**Рис 2.3**). Результатом пошуку є список компонентів, що є схожими до запиту, але також ці сполуки мають посилання до білдинг блоків та хімічних реакцій, за якими можна ці сполуки синтезувати.

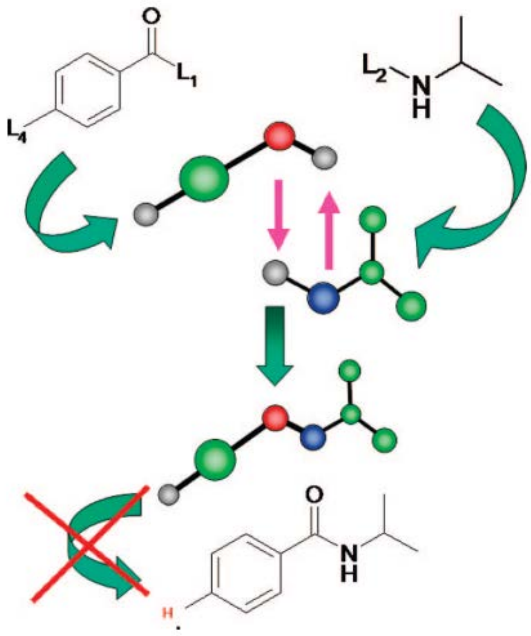

Рис.2.3. Генерування віртуальних структур за допомогою FTrees-FS. Заміщення незавершених лінкерів атомом Гідрогену є забороненим.

Також подібний підхід має додаткову перевагу – якщо просто синтезувати одну молекулу, значить просто синтезувати і її аналоги, оскільки пошук ведеться на рівні функціональних груп.

Нещодавно розроблений пошуковий алгоритм, SpaceLight, виконує пошук по подібності топологічних фінгерпринтів у масивному комбінаторному фрагментному просторі[28][29]. Порівнюючи із традиційними фінгерпринтовинми методами, SpaceLight може використовувати комбінаторний характер цих просторів для ефективності, зберігаючи високу кореляцію опису молекулярної схожості з молекулярними фінгерпринтами як от з ECFPs (extended-connectivity fingerprints)[30]. Результуюче програмне забезпечення здатне виконувати пошуки як от REAL просторі із 20-ма мільярдами сполук за секунди на персональному ком'ютері. Схоже до FTrees-FS, SpaceLight може бути масштабований і виконувати пошуки у просторах більш ніж 10^{20} сполук, оскільки виконує пошук саме серед фрагментів, а не продуктів[8].

Незважаючи на те, що і FTrees-FS, і SpaceLight мають широке практичне значення, їх недолік у недостатку можливостей пошуку точних структурних фрагментів на атомному рівні. Нова технологія, SpaceMACS, робить можливим ефективний і точний пошук по максимальній спільній підструктурі (maximum common substructure, MCS) із підходом пошуку по подібності або підструктурі[31]. SpaceMACS енумерує ту кількість результуючих сполук (не інтермедіатів), яка була обрана користувачем. Це відбувається за багатокроковою процедурою і дає можливість проводити раціональний дизайн лікарських засобів базуючись на комбінаторних make-on-demand каталогах. Після того, як простір було завантажено у пам'ять, результат навіть для Enamine REAL Space може бути отриманий у секунди. Оскільки цей алгоритм має в собі досить ефективний огляд хімічної

структури, він потенційно може стати мостиком між будь-якою технікою молекулярного дизайну та хімічного фрагментного простору.

Варто зазначити, що наведені алгоритми пошуку стосуються саме двовимірного пошуку, тобто топологічного. І зараз це нова реальність та рутина для віртуального скринінгу. Пошук по тривимірній структурі у хімічному просторі – задача досі нетривіальна.

2.2 Порівняння алгоритмів пошуку

Для того, щоб виконати наше дослідження, потрібно було визначити, який із алгоритмів пошуку варто використати, оскільки це буде безпосередньо впливати на результат. Вибір був між трьома найбільш популярними методами, описаними вище: FTrees-FS, SpaceLight та SpaceMACS. Група Rarey, безпосередньо творці останніх двох алгоритмів, у оригінальній статті[31], де було вперше описано принцип SpaceMACS, було також порівняно результати пошуку із використанням цих трьох інструментів, щоб дослідити перекривання. Із використанням кожного із методів було зроблено запит із отриманням найкращих 150 000 результуючих молекул із трьох хімічних просторів (GalaXi, REAL Space, CHEMriya) та було порівняно ці результати (**Рис 2.4**).

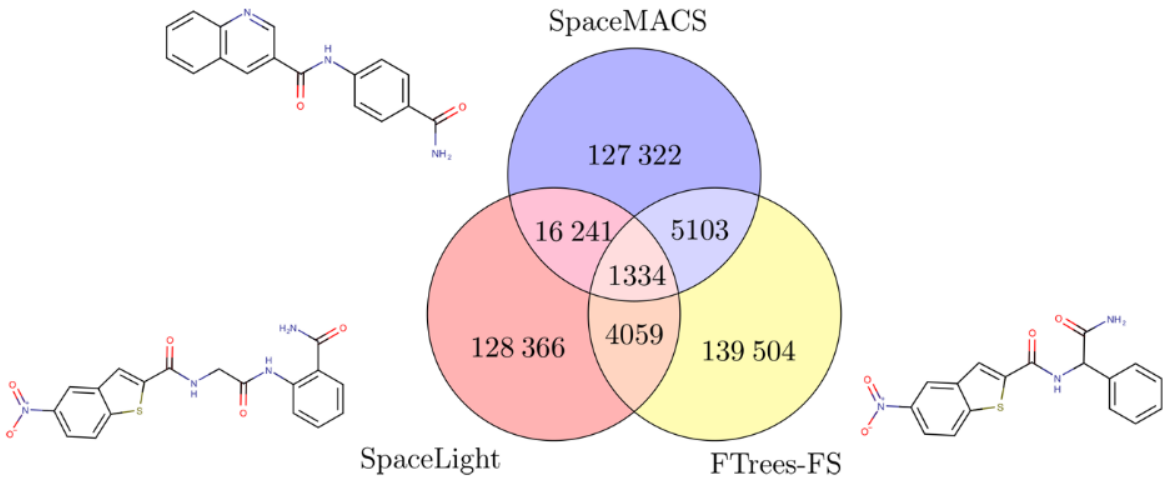

Рис 2.4. Накладання результатів пошуку між SpaceMACS (синій; пошук по подібності), SpaceLight (червоний; fcsfp2.5) і FTrees-FS (жовтий). Подібність на основі MCS

(SpaceMACS) забезпечує точне відповідність підструктури основної області та ігнорує решту молекули. Подібність на основі фінгерпринтів (SpaceLight) також залежить від структури, але допускає помилки, як-от повторювані патерни. Нарешті, подібність зменшеного графа (FTrees-FS) має найменшу залежність від структури і знаходить молекули з подібним розташуванням функціональних груп.

Найбільше перекривання було знайдено між SpaceMACS та SpaceLight (16 241 молекула, 11%), найменше – між FTrees-FS та SpaceLight (4 509 молекул, 2.7%). Перекривання між методами становить 0.9% (1 334 молекули). В загальному, методи показали себе як досить добре для доповнення результатів одне одного, оскільки результати пошуку варіюються по точному хімічному співпадінню від хімічної до фармакофороподібної схожості. Відштовхуючись від походження молекул, було показано також, що перекривання результатів між всіма методами для всіх хімічних просторів є всього 5 молекул (0.01%). Найбільше перекривання є між REAL Space та CHEMriya у 1 425 молекул (2.83%), але також є можливим те, що одні й ті ж молекули були отримані різними синтетичними шляхами.

Розділ 3. Експериментальна частина

Отже, суть нашого дослідження була в тому, щоб визначити перекривання колекції скринінгових сполук із віртуальнимим хімічними просторами, а також перекривання результатів цього пошуку між самими просторами. Це дослідження поділялось на 4 основних частини:

1. Оновити дані та згенерувати оновлену версію Bioinfo DB.
2. На основі цієї бази даних згенерувати Мурко скафолди.
3. Використовуючи алгоритм пошуку SpaceMACS виконати пошук у хімічних просторах.
4. Порівняти та проаналізувати результати пошуку.

3.1 Оновлення Bioinfo DB

Що ж собою являє Bioinfo DB? Це емпірично згенерована база даних скринігових сполук, яка використовується всередині лабораторії («in-house») для кампаній віртуального скринінгу. Ця база даних має майже двадцятирічну історію і постійно піддавалась змінам. Емпіричною її можна назвати, оскільки вибір постачальників базується на досвіді співпраці та взаємодії із компаніями. Тобто, у нинішній версії бази даних фігурують лише ті постачальники, з якими колись співпрацювали і вони показали себе як ті, які варті довіри (trusted). Тому цей список із 25 постачальників є емпіричним списком базуючись на співвідношенні ціна/якість/швидкість.

Цінність цієї бази даних проявляється під час віртуального скринінгу, оскільки якщо сполука із цієї колекції проявить себе як хіт, можна одразу знайти її ідентифікаційний номер та постачальника і моментально замовити для *in vitro* тестування. Тобто використання цієї бази даних є своєрідною

гарантією того, що отриманий хіт є реальною сполукою, яка може бути протестована на біологічних тестах.

Також варто зазначити, що серед постачальників фігурують тільки ті, хто може постачати сполуки у вигляді порошкоподібних субстанцій, а не тільки в платах. Саме тому такі великі постачальники як наприклад FCH group (2 030 974 сполук), Selena chemicals (1 509 367 сполук) чи UOrSy (1 758 216 сполук) не представлені серед постачальників у цій базі даних.

Оновлення даних відбувалось на базі опублікованих на сайті постачальників .zip архівів із .sdf файлами скринінгових сполук. До цього, останнє оновлення бази даних відбувалось у жовтні 2021 року (приблизно 6 місяців до теперішнього оновлення). Представлені дані є актуальними станом на лютий 2022 року.

Нові дані були запущені у вже існуючий протокол у програмі Pipeline Pilot, що є інструментом для workflow менеджменту, тобто процесів, в яких фігурують великі об’єми даних, щодо яких здійснюються повторювані дії.

Першим етапом цього протоколу було послідовне зчитування кожного файлу від окремого постачальника, збереження незмінними таких даних як ідентифікаційний номер та назва самого постачальника. Далі була проведена попередня очистка даних із видаленням сполук, що містять в собі п’ятивалентний карбон, солей чи сполук із більш ніж одним не визначеним стереоцентром. Було підраховано кількість сполук, яка була початково та після цього етапу, а дані про сполуки, які пройшли фільтр були записані у новий файл .sdf. Отож, неорганізовані файли від окремих постачальників були об’єднані в один файл.

Другим етапом даного протоколу стало перевірка на канонічність представлених SMILES і генерування таких, якщо вимога не є дотриманою. Також було видалено дублікати поміж різних постачальників. Кількість сполук опісля була підрахована, кінцевий результат записано у новий файл.

Третім етапом було пропускання файлу через внутрішній фільтр для лікарськоподібних сполук від програми OpenEye, а також іонізація сполук при pH =7.4 (рівень у крові). Наступна перевірка показала, що на цьому етапі не було утворено дублікатів (кількість унікальних сполук збігається із загальною кількістю сполук). Також було підраховано зміни у цьому етапі, а сполуки записані у новий файл.

Останнім, четвертим етапом, стало стандартизація для деяких проблемних фрагментів та груп (як от для пірідіонів, піридинів та сульфонамідів), розрахунок дескрипторів (MW, PSA, HBA, HBD, NRot, Nrings, Solubility) та підрахунок кількості порушень класичних правил Ліпінського для кожної сполуки. Було підраховано сумарну кількість фінальних сполук та записано у вигляді .sdf файлу.

3.2 Особливості фільтрів перевірки на drug likeness

У 1997 році Ліпінським було представлено перший емпіричний збір параметрів, якими мала б володіти молекула, щоб вона була лікарськоподібною, тобто не просто виявляла біологічну дію, а й мала хорошу розчинність та доступність[32]. Цей набір правил було названо правилами Ліпінського, або ж правилами п’яти (Rule of 5, RO5). Ці правила були протягом дуже довгого часу must have при розробці лікарських засобів, їх застосовували для попереднього фільтрування молекул. Однак, досить часто забувалось, що ці правила є суто емпіричними, а також мають бути застосованими лише для сполук, що мають оральну біодоступність. Тому сполуки по типу пептидів, антибіотиків, чи просто сполук природнього походження були поза цими правилами. Це означало, що вони були відсіяні ще на самих ранніх етапах віртуальних досліджень.

Однак нині тенденція йде до переосмислення та вихід за правила п’яти (beyond rule of 5, bRO5)[33][34]. Саме тому у програмі OpenEye, фільтри якої були застосовані для перевірки на drug likeness є набагато більш гнучкими та

варіативними, аніж класичні правила Ліпінського. Список правил, які були застосовані, наведені у **Додатку 1**. Тут порівняю лиш основні (**Таблиця 3.1**)

Таблиця 3.1 Порівняння RO5 і правил, які були застосовані у фільтрі

	MW	**logP**	**HBD**	**HBA**
RO5	≤500	≤5	≤5	≤10
bRO5	150≤MW≤750	-2.0≤logP≤6.0	≤6	≤10

Таким чином, відібрані молекули мають значно ширший діапазон значень. Також було застосовано PAINS фільтр і відкидання сполук із реактивними групами. Максимально допустиме значення порушень правил Ліпінського було встановлено на рівні ≤ 2.

3.3 Генерування скафолдів Беміса-Мурко

Далі було вирішено працювати не із окремими молекулами, а із скафолдами, оскільки це не тільки зменшить кількість запитів для пошуку, але й матиме більший хімічний зміст. Враховуючи те, що для пошуку в хімічних просторах було вирішено використовувати алгоритм SpaceMACS, генерування скафолдів було найбільш логічним кроком, бо пошук відбувається по найбільшій спільній підструктурі, а замісники будуть лише заважати, бо основна наша цікавість на саме наявності конкретних скафолдів, а окремих молекул (**Рис 3.1**).

Базуючись на оригінальній статті, скафолдом Беміса-Мурко є комбінація кільцевої системи та лінкерів між ними[35]. Усі побічні ланцюги прибираються.

При цьому, лінкерними атомами називають атоми, що лежать на прямому шляху між кільцями. А атомами бічних ланцюгів є атоми, які не входять ні до кільцевої системи, ні до лінкеру.

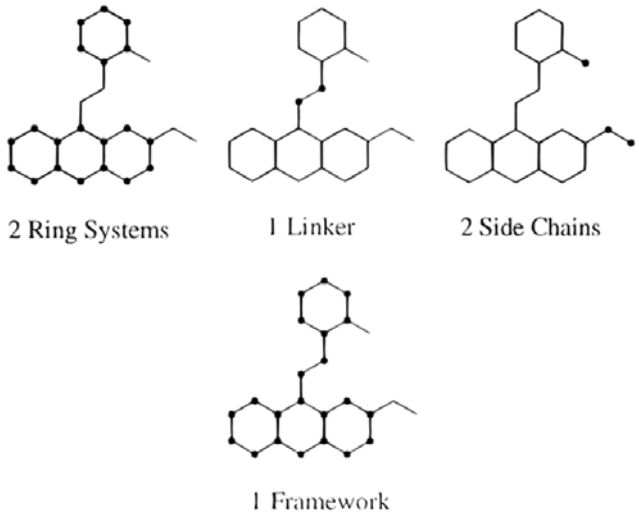

Рис 3.1 Розрізнення між кільцевою системою, лінкерами та бічними ланцюгами.

Однак, під час тестування, було виявлено, що використання двох різних підходів (програма Pipeline Pilot та скрипт мовою програмування Python із використанням бібліотеки RDkit) у деяких випадках приводили до дещо різних результатів.

Було виявлено, що різниця між методами зумовлена різними інтерпретацією правил утворення скафолдів (**Рис 3.2**). Таким чином, була помічена різниця у генеруванні скафолдів для тестового набору сполук, які мають термінальний подвійний зв'язок, що зв'язаний із циклом (хоча у випадку сульфонільної групи у циклі проблеми відсутні), сульфонільні/сульфонамідні групи чи просто амідний зв'язок у лінкері.

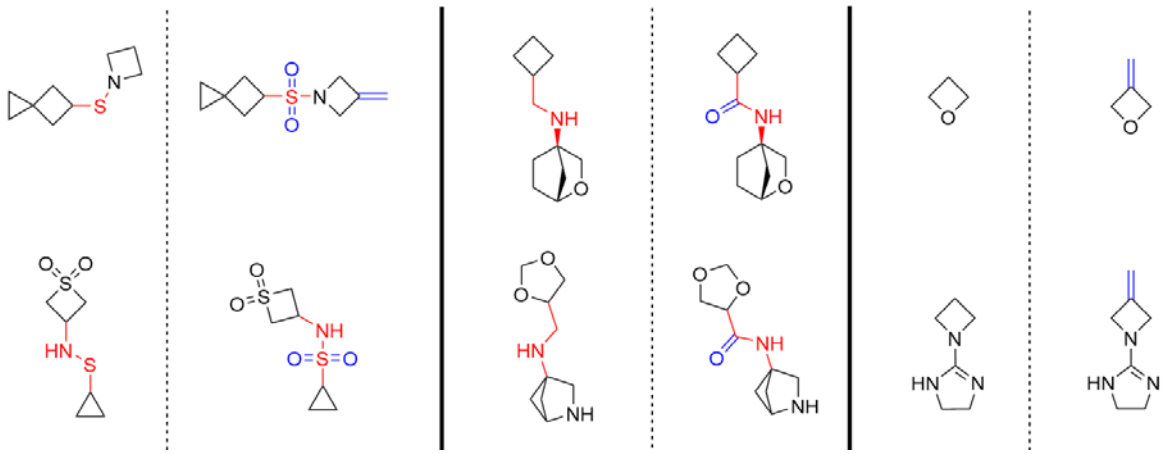

Рис 3.2 Різниця у генерування скафолдів Беміса-Мурко із використанням двох підходів – Pipeline Pilot (ліворуч) та RDkit (праворуч). Атоми-лінкери виділено червоним.

Постає вибір – який із методів обрати? І хоча Pipeline Pilot більш чітко використовує трактування із оригінальної статті, однак RDkit все ж зберігає хімічну функцію лінкерних атомів. Тому було вирішено використати останній.

3.4 Пошук у хімічних просторах

Через значну кількість скафолдів для пошуку, а значить значні розрахункові ресурси. Тому пошук у хімічних просторах за допомогою алгоритму SpaceMACS було виконано із використанням ресурсів обчислюваного центру (CC-IN2P3).

Навіть після генерування скафолдів Беміса-Мурко, кількість SMILES як для одного запиту до хімічного простору була досить значною, тому було вирішено поділити файл на 246 пакетів по 10 000 молекул у кожному. Був написаний скрипт на мові Python, який для кожного з пакетів генерував скрипт на Bash для передачі завдання на ноду розрахункового центру (обчислювальна одиниця). Також був написаний Bash скрипт для одночасного запуску усіх завдань для паралельного обчислення.

Кількісні параметри для пошукового запиту для SpaceMACS були задані базуючись на результатах масштабування для цього алгоритму групою Rarey[31]. Так, групування SMILES у пакети по 10 000 було більш ефективним, аніж по 100 чи 1000, оскільки час на пошук у фрагментах у всіх трьох випадках є однаковим, різниця лиш у швидкості отримання та енумерації результатів. Але оскільки збільшення кількості часу на ці операції зі збільшенням кількості запитів не є лінійним, тому результати для 10 000 запитів будуть найбільш ефективним використанням ресурсів (**Рис 3.3**).

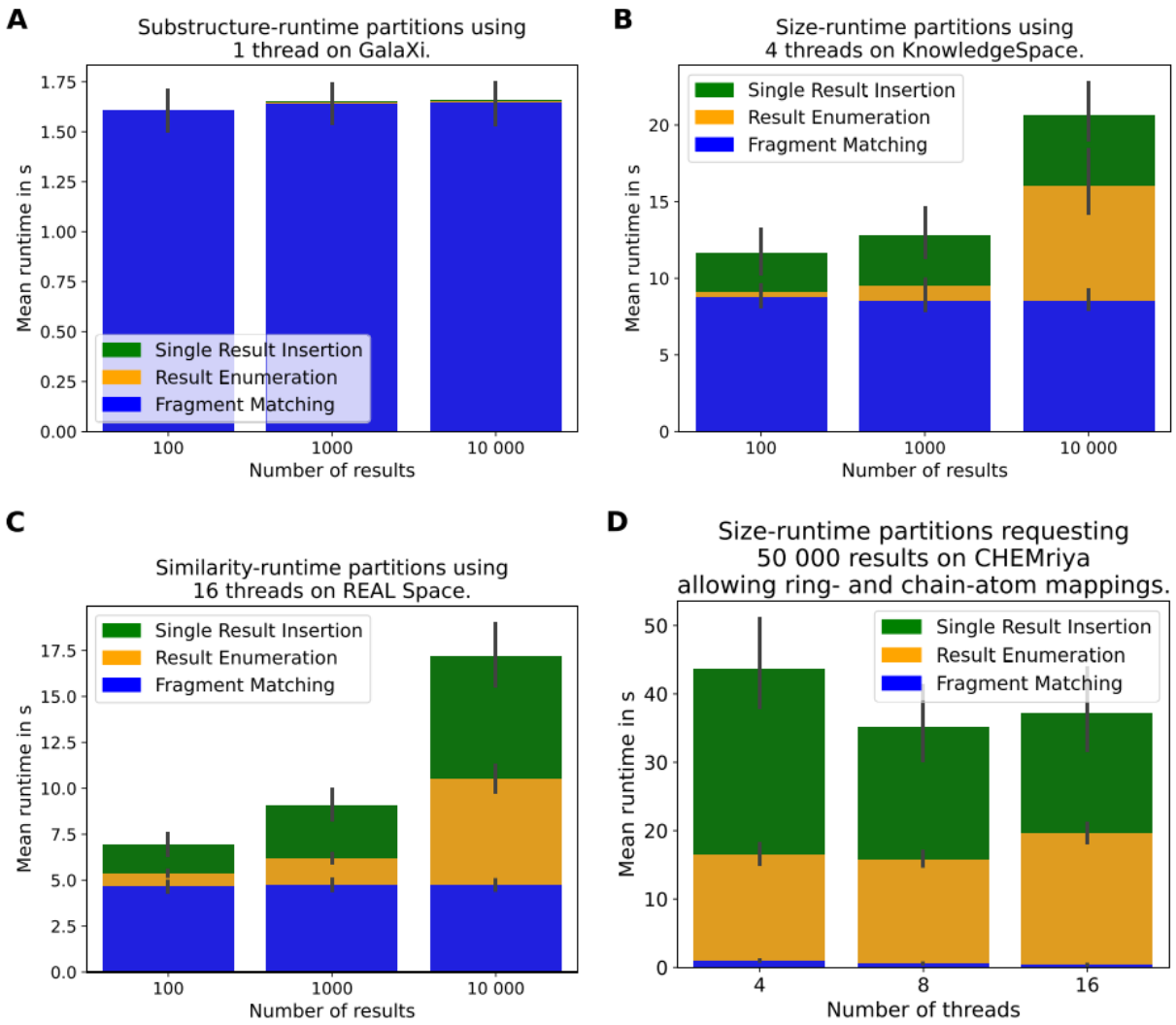

Рис 3.3 Розподіл часу виконання трьох кроків алгоритму SpaceMACS із запитом від 100 до 10 000 результатів (**A-C**) і 50 000 результатів (**D**). У **A** 100, 1000 і 10 000 результати отримані із GalaXi, що виконані в один потік, використовуючи режим підструктури. Розрізнення кроків алгоритму недоступне для запитів без результату. У **B** 100, 1000 і 10 000 результатів отримані з KnowledgeSpace, що виконані у чотири паралельні потоки за допомогою MCS-Size. У **C** 100, 1000 і 10 000 результати отримані з REAL Space, що виконані у 16 потоків паралельно за допомогою MCS-Similarity. У **D** 50 000 результатів отримуються з CHEMriya, що виконані у 4, 8 і 16 потоків паралельно за допомогою MCS-Similarity.

Однак потрібно також вирішити, у скільки потоків будуть йти розрахунки, оскільки використання одного потоку не завжди виправдане по часу і ресурсах. В рамках цієї статті[31] було проведено також дослідження із масштабування алгоритму базуючись на сталій кількості запитів (10 000), але змінюючи кількість паралельних потоків для різних типів пошуку(**Рис 3.4**).

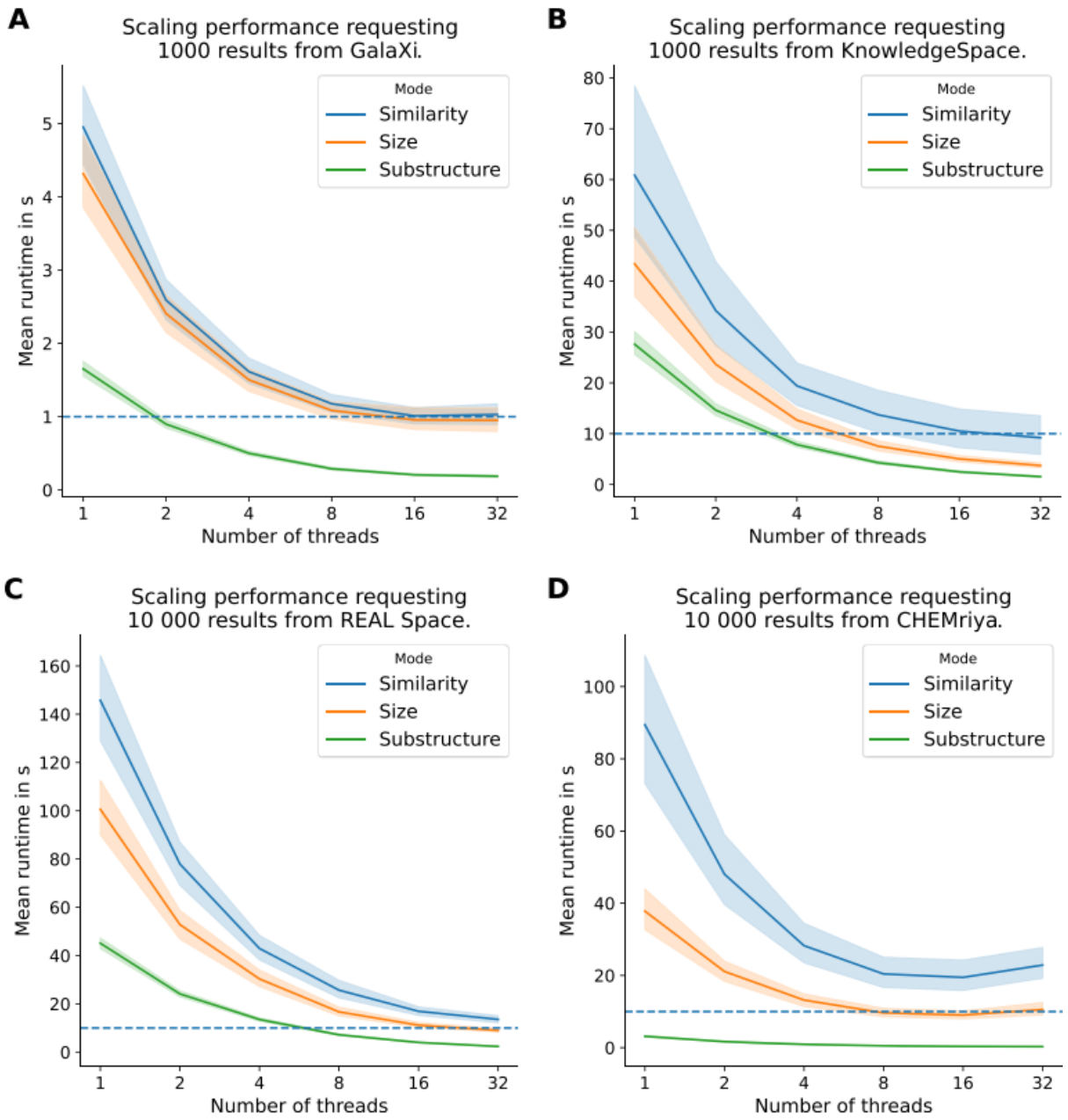

Рис 3.4 Огляд масштабування SpaceMACS, що із 1000 запитами до комерційного GalaXi (**A**) і загальнодоступного KnowledgeSpace (**B**) і 10 000 запитами до комерційних REAL Space (**C**) і CHEMriya (**D**), які виконуються із використанням від 1 до 32 потоків паралельно. Референсом для GalaXi (**A**) показано час роботи 1 с. Для KnowledgeSpace (**B**), REAL Space (**C**) і CHEMriya (**D**) показано 10 с як референс.

Базуючи на результатах цього дослідження, було вирішено використовувати пакети запитів по 10 000 SMILES із використанням 1 потоку для GalaXi та 16 потоків для CHEMriya та REAL Space для отримання одного найкращого результату на запит із пошуком по MCS-similarity.

3.5 Обробка результатів

Отримані результати були опрацьовані скриптом на Python із екстракцією лише потрібних значень (вхідний SMILES, значення MCS-similarity та SMILES знайдений у хімічному просторі) та об'єднанням їх у єдиний файл для кожного простору.

Для кожного із хімічних просторів було побудовано графік розподілу за значеннями MCS-similarity та їх часткою серед результатів (**Рис 3.5**). Було також встановлено порогове значення у 0.85 і далі аналізувалися результати, які мають значення вище порогового. Окремо було проаналізовано результати із значенням подібності 1.00.

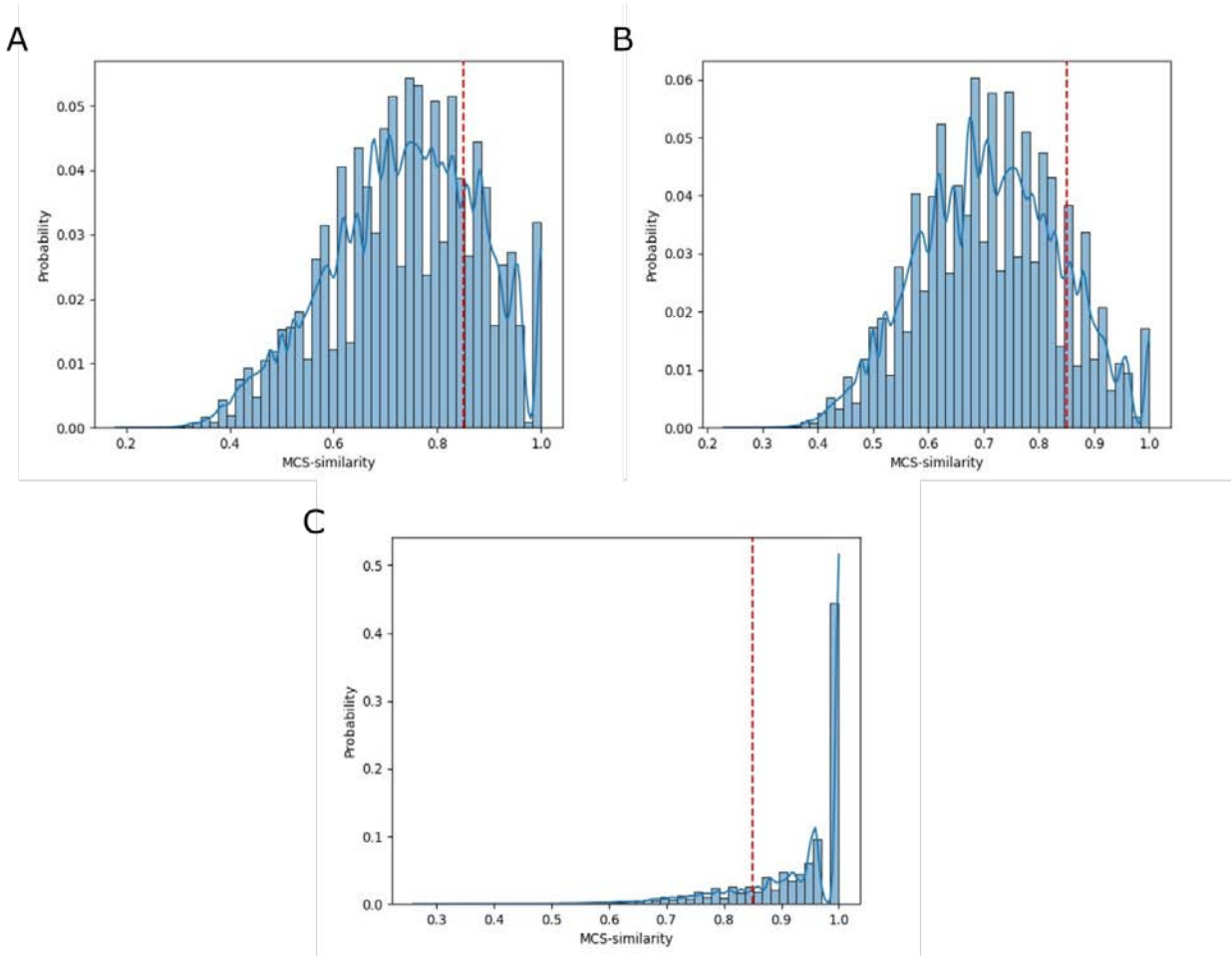

Рис 3.5 Розподіл результатів пошуку по MCS-similarity для (**A**) GalaXi, (**B**) CHEMriya та (**C**) REAL Space. Червона пунктирна лінія – порогове значення подібності у 0.85. Ширина стовпчика – 0.02.

Порогове значення було відібрано емпірично, на основі мануального перегляду результатів. Так, саме для значення 0.85 різниця між вхідним та результуючим SMIELS була більш хімічно обґрунтованою, як от заміна замісників, ізостерія циклів чи незначна модифікація лінкера.

Далі було побудовано діаграми Венна, щоб побачити як перекриваються результати для різних хімічних просторів між собою. Створено дві діаграми – для всіх значень вище порогового та окремо для 1.00. Детальний аналіз результатів буде наведено в наступному розділі.

Розділ 4. Результати та обговорення

4.1 Підготовка даних та пошук зі SpaceMACS

Оновлені дані бібліотек скринінгових сполук були взяті із офіційних сайтів постачальників. Тому інформація щодо бібліотек актуальна станом на 02/02/2022. У підсумковій таблиці (**Рис. 4.1**) представлено як змінювалась кількість молекул на кожному з етапів для різних бібліотек. Так, у колонці **#cpds** представлені «сирі» дані, тобто кількість молекул, які були в початкових файлах. **#clean** – це кількість молекул після видалення молекул із п’ятивалентним карбоном, солями та із невизначеними стереоцентрами. **#unique 2D** – кількість молекул після видалення дублікатів всередині бібліотеки та між постачальниками. **#filtered 2D** – кількість сполук після фільтрування на drug-likeness у OpenEye. Ну і **#final 2D** – після стандартизації замісників, видалення дублікатів, підрахунку кількості порушення правил Ліпінського та видалення тих сполук, що мають значення більше 2-х. Як можна побачити, майже для кожного постачальника за півроку кількість молекул змінилась (або було дещо реорганізовано бібліотеку). Після усіх фільтрів та чисток, кількість молекул зменшилась більш ніж вдвічі. Якщо порівнювати із попередньою версією Bioinfo DB, кількість фінальних молекул у новій версії збільшилась на 1.7% (~150 тисяч молекул).

Bioinfo 2022.1

	Suppliers	# cpds	# clean	#unique 2D	#filtered 2D	#final 2D
1	AbamaChem	1,453,532	1,354,932	1,321,858	1,067,958	1,067,958
2	Alinda	892,907	883,601	99,576	58,947	58,855
3	AnalytiCon	45,536	43,181	41,905	23,803	23,803
4	Aronis	26,848	26,757	65	31	31
5	Asinex	522,390	516,653	516,646	334,935	334,925
6	AsisChem	2,109,738	2,089,223	1,767,506	612,933	612,666
7	BCH Research	1,465,065	1,423,662	1,337,874	1,085,393	1,085,390
8	Bionet	276,701	274,204	258,133	110,367	110,233
9	Cayman	18,069	17,858	10,597	3,106	3,088
10	Chembridge	1,312,796	1,304,946	1,193,737	923,164	923,119
11	ChemDiv	2,196,216	2,172,979	1,532,060	879,955	879,451
12	CNRS	82,569	78,514	69,316	33,296	33,162
13	Enamine	2,912,316	2,866,860	2,757,229	1,935,754	1,935,633
14	ExiMed	61,009	60,845	4,759	3,384	3,382
15	InterBioScreen	555,304	545,146	354,931	176,303	175,838
16	Intermed	1,966,445	1,905,365	1,748,239	1,382,194	1,382,184
17	LifeChemicals	519,016	516,567	373,378	252,222	252,153
18	Maybridge	53,352	52,777	41,621	17,081	17,067
19	Otava	277,932	275,617	77,693	39,506	39,349
20	PBMR_Labs	1,532,541	1,505,095	60,211	39,166	38,706
21	Pharmeks	407,318	395,884	119,403	68,797	63,135
22	Specs	207,684	204,257	173,883	92,856	92,835
23	Synthon_Lab	32,275	32,063	9,398	3,571	3,409
24	TimTec	994,852	972,738	252,841	97,423	96,502
25	Vitas-M	1,414,553	1,384,071	158,679	74,453	73,501
	Total	**21,336,964**	**20,903,795**	**14,281,538**	**9,316,598**	**9,306,375**

Рис 4.1 Підсумкова таблиця поетапної зміни кількості молекул після drug likeness фільтрування та стандартизації по кожному постачальнику. Кольором виділено значення, які змінилися із жовтня 2021 (зелений – кількість молекул зросла, червоний – зменшилась, жовтий – різка зміна кількості молекул, майже двічі)

Далі для роботи із цією базою даних необхідно було екстрагувати лише потрібні для наступних кроків дані – у даному випадку, це лише SMILES. Тому за допомогою скрипта на Python було зчитано .sdf файл із Bioinfo DB, екстраговано лише значення SMILES та записано у текстовий файл з розширенням .smi. Це дозволило, по-перше, оперувати лише з необхідною інформацією, по-друге – зменшити навантаження на оперативну пам’ять при роботі із базою даних, оскільки початковий файл мав розмір у 17 гігабайт, а після обробки – лише 404 мегабайт, що значно полегшує подальшу роботу.

Наступним етапом було генерування скафолдів Беміса-Мурко. Враховуючи інформацію з попереднього розділу, дана частина роботи виконувалась за допомогою RDkit. Таким чином, із 9 306 375 молекул було згенеровано 2 453 944 скафолдів. Окрім концептуального значення,

використання скафолдів Беміса-Мурко дозволяє значно зменшити кількість запитів до хімічного простору (на 73.63%).

Пошук у хімічному просторі із використанням SpaceMACS задається у терміналі у такому вигляді:

директорія, де знаходиться SpaceMACS/**SpaceMACS -t** **кількість потоків** **-f** **директорія, де знаходиться хімічний простір**/**name.space -n** **кількість результатів, які ми хочемо отримати для кожного SMILES** **-m** **тип пошуку** **директорія, де зберігається вхідний файл із запитами** **-o** **назва файлу, в якому будуть зберігатися результати**

Для прикладу, якщо ми хочемо задати пошук у REAL Space, який знаходиться у нинішній директорії, як і SpaceMACS, із використанням 16 потоків із пошуком по MCS-similarity із отриманням одного найкращого результату, команда буде виглядати так:

./SpaceMACS -t 16 -f ./19bn-REALSpace_2021-04.space – n 1 -m Similarity ./bioinfo221_10k_part0001.smi -o bioinfo221_10k_REAL_part_0001.out

Варто зазначити, що якщо не зазначити кількість потоків, то за визначенням, буде використовуватися мультипотоковий режим, тобто будуть використовуватися всі доступні потоки.

Таким чином, як було зазначено в попередньому розділі, пошук відбувався із використанням ресурсів обчислювального центру. Файл із SMILES скафолдів був поділений на 246 частин (10 000 скафолдів для кожної частини), для кожної з яких було написано файл із запуском пошуку. Таким чином, за допомогою скрипту Bash було запущено 246 робіт одночасно для паралельного обчислення. Для GalaXi було запущено роботи на 1 потоці, що зайняло ~5 годин сумарно. Для CHEMriya та REAL Space було використано 16 потоків, із сумарним часом пошуку 3 та 18 годин відповідно.

Отримані результати далі були об’єднані в окремі для кожного простору об’єднані файли. Було зроблено перевірку та виявлено, що дублікатів серед результатів всередині одного простору немає, тобто всі результати є унікальними та не відбулось збою із дублюванням рядків.

Оскільки результуючий файл містить досить багато інформації (**Рис 4.2**), було екстраговано лише потрібну для подальшого аналізу.

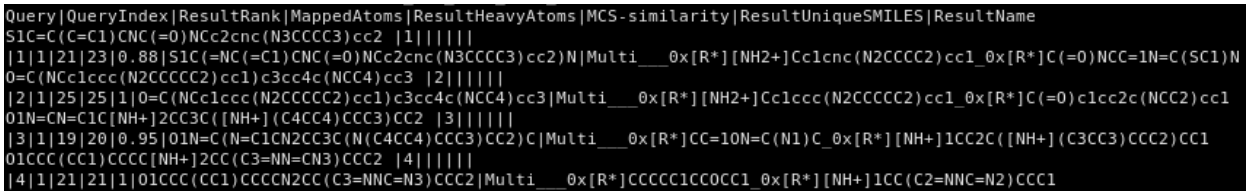

Рис 4.2 Приклад файлу із першими 4 результатами. Для подальшої роботи потрібно екстрагувати лише значення Query, MCS-Similarity та ResultUniqueSMILES.

4.2 Аналіз результатів

Для аналізу отриманої інформації першочергово було побудовано графіки розподілу результатів для кожного із хімічних просторів (див. **Рис 3.5**). Для просторів GalaXi та CHEMriya загальний вигляд розподілу був передбачуваний (схожий до нормального розподілу). Хоча результати для GalaXi зміщені у сторону більших значень подібності. Таким чином, обравши порогове значення у 0.85, для GalaXi, при меншому загальному розмірі простору, кількість результатів вищих за порогове більша (606 323 скафолди), аніж для CHEMriya (395 370 скафолдів). Із цих чисел, значення подібності у 1.00 мають ~11% для CHEMriya і ~13% для GalaXi. Таким чином, із вигляду розподілу можна зробити висновок, що ці хімічні простори, через широкий діапазон значень MCS-similarity, є хорошим варіантом для початкових етапах віртуального скринінгу (пошук хітів) та розробки лікарських засобів в цілому, оскільки це свідчить про різноманітність наявних там скафолдів.

Однак досить неочікуваними стали результати для REAL Space, оскільки ймовірність отримати результат ідентичний до нашого запиту – майже 50%. Ми очікували отримати дещо зміщені результати до вищих значень для цього простору, оскільки він позиціонує себе як той, молекули з якого можна синтезувати із високою ймовірністю. Але результат перевершив очікування. Для REAL Space із 2 453 944 запитів було отримано 2 006 061

результат із значенням подібності більше 0.85, тобто 81.7% від усіх результатів перетинають порогове значення. Із значень вище порогового, майже половину складають ті, що мають 1.00 (1 089 906 результатів, або 44.4% від усіх результатів).

Оскільки значення для REAL Space є дещо завищеними, є можливість, що початково наше дослідження було упередженим щодо цього простору, оскільки Enamine REAL DB є одним із постачальників, які складають Bioinfo DB (майже 20% від фінальної кількості сполук). Саме тому, подальші дослідження будуть направлені на визначенні та оцінці упередження.

Однак, варто зазначити, що не включати Enamine REAL DB у список постачальників є не правильним із декількох точок зору. По-перше, це один із найбільших постачальників будівельних блоків та скринінгових сполук у світі, оскільки має найбільш оптимальне співвідношення ціна/швидкість/якість, а також надає надзвичайно широкий вибір сполук різної складності. По-друге, нам початково невідомо, наскільки Enamine REAL DB перекривається із Enamine REAL Space. Зрозуміло, що оскільки вони мають походження із однієї компанії, значна частка сполук із Enamine REAL DB будуть знаходитися у їх просторі, але все ж, це не енумерований простір, а комбінаторний, тому інколи молекули можуть бути не включені у хімічний простір через складнощі або невигідністю її включення у простір. По-третє, ми також не знаємо, наскільки перекриваються скафолди між постачальниками всередині Bioinfo DB. Тобто не можна сказати, абсолютно всі сполуки із Enamine REAL DB мають унікальні скафолди. Саме тому, в подальшому, ми хочемо оцінити, який вклад має Enamine REAL DB у створення скафолдів (скільки із них мають походження Enamine), що допоможе нам дати оцінку упередженості.

Але якщо виходити із наявних результатів і вважати упередженість не критичною, тоді REAL Space є кращим варіантом для використання на стадії

hit-to-lead оптимізації, оскільки він містить значну частку доступних унікальних скафолдів із варіаціями замісників у ньому.

Останнім етапом було дослідження перекривання результатів між просторами (**Рис. 4.3**).

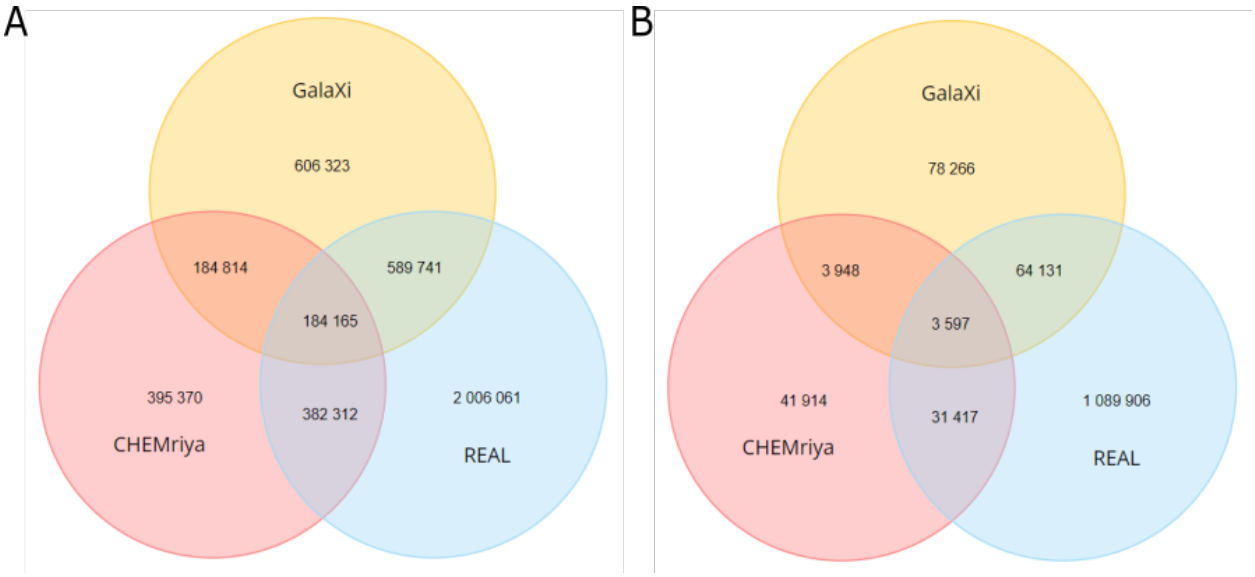

Рис. 3.4 Діаграми Венна розподілу результатів із значеннями MCS-similarity ≥ 0.85 (**A**) та 1.00 (**B**) між ультра-великими хімічними просторами.

Ми бачимо, у обох випадках, і GalaXi, і CHEMriya значно перекриваються із REAL Space, причому перекривання між GalaXi і CHEMriya є лімітуючим для перекривання всіх трьох просторів. Цікавим також є те, що якщо перекривання між трьома просторами становить 31% та 47% від усіх результатів ≥ 0.85 для GalaXi та CHEMriya відповідно, то для REAL Space цей показник уже всього 9.2%. Однак, для MCS-similarity = 1.00 цей показник значно падає до рівня 4.6% (GalaXi), 8.6% (CHEMriya) і всього 0.3% для REAL Space. Виходячи із цього, перекривання між унікальними скафолдами мізерне.

Висновки

У даній роботі було виконано пошук реальних скринінгових сполук на основі скафолдів Беміса-Мурко в ультра-великих віртуальних комбінаторних хімічних просторах (WuXi-GalaXi, Otava-CHEMriya, Enamine REAL Space) за допомогою алгоритму пошуку, який базується на понятті максимальної спільної підструктури (SpaceMACS). Для кожного із хімічних просторів було отримано кількість скафолдів синтетично доступних сполук у них, а також було показано, як між цими просторами перекриваються результати.

Початково було оновлено дані для об’єднаної бази даних скринінгових сполук (Bioinfo-DB) із подальшими процедурами фільтрування (у тому числі, із розширеними правилами drug likeness), стандартизації та розрахунками основних фізико-хімічних параметрів. У такому вигляді база даних використовується для in-house потреб, у тому числі – для процедур віртуального скринінгу.

Далі на основі цієї бази даних (9.3M молекул) було згенеровано скафолди Беміса-Мурко (2.5M скафолдів). Використання саме скафолдів є концептуально необхідним, оскільки воно дає можливість оцінити наявність в хімічному просторі не просто конкретної молекули, але ряду молекул із даним скафолдом. Також зменшення кількості запитів до хімічного простору робить виконання завдання більш ефективним за рахунок зменшення часу та обчислювальних ресурсів.

Пошук у хімічних просторах відбувався у якості 246 паралельних завдань, по 10 000 запитів у кожному, із використанням потужностей обчислювального центру. Результати для кожного простору були об’єднані

та екстрагована лише необхідна інформація для подальшого аналізу результатів.

Для кожного хімічного простору був побудований графік розподілу результатів за значеннями MCS-similarity та емпірично обране порогове значення у 0.85 як найбільш доцільне з точки зору органічної та медичної хімії.

Якщо для хімічних просторів GalaXi та CHEMriya значення розподілу нагадують нормальний (хоча результати GalaXi зсунуті до більших значень, тому при меншому розмірі простір має більше результатів із значенням подібності більше порогового), то розподіл для REAL Space дещо аномальний, що буде ціллю подальшого дослідження та аналізу для виключення упередженості дослідження щодо цього простору.

Було також досліджено перекривання результатів між просторами при значенні подібності ≥0.85 та окремо для 1.00. Виявлено, що у обох випадках окреме перекривання GalaXi та CHEMriya із REAL Space є значним. Однак, якщо у першому випадку перекривання усіх трьох просторів відносно результатів ≥0.85 становить у рамках 30-50% для GalaXi та CHEMriya, і 9.2% для REAL, то перекривання унікальних скафолдів є на рівні 0.03-8.6%.

На основі цих досліджень можна сказати, що GalaXi та CHEMriya краще підходять для початкових етапів віртуального скринінгу (пошуку хітів), а REAL – як для пошуку хітів, так і оптимізації до ліду. Результати перекривання між просторів свідчить про те, що кожен простір містить в собі значну долю унікальних скафолдів, які не перекриваються між просторами, а також простори є взаємодоповнюваними для подібних скафолдів.

Список використаних джерел

[1] S. J. Lusher, R. McGuire, R. C. Van Schaik, C. D. Nicholson, J. De Vlieg, *Drug Discov. Today* **2014**, *19*, 859–868.

[2] J. G. Lombardino, J. A. Lowe, *Nat. Rev. Drug Discov.* **2004**, *3*, 853–862.

[3] S. Renner, M. Popov, A. Schuffenhauer, H. J. Roth, W. Breitenstein, A. Marzinzik, I. Lewis, P. Krastel, F. Nigsch, J. Jenkins, E. Jacoby, *Future Med. Chem.* **2011**, *3*, 751–766.

[4] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50.

[5] P. Ertl, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

[6] P. G. Polishchuk, T. I. Madzhidov, A. Varnek, *J. Comput. Aided. Mol. Des.* **2013**, *27*, 675–679.

[7] W. P. Walters, *J. Med. Chem.* **2019**, *62*, 1116–1124.

[8] W. A. Warr, M. C. Nicklaus, C. A. Nicolaou, M. Rarey, *J. Chem. Inf. Model.* **2022**, DOI 10.1021/acs.jcim.2c00224.

[9] D. E. Clark, *J. Chem. Inf. Model.* **2020**, *60*, 4120–4123.

[10] C. Grebner, E. Malmerberg, A. Shewmaker, J. Batista, A. Nicholls, J. Sadowski, *J. Chem. Inf. Model.* **2020**, *60*, 4274–4282.

[11] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina, Y. S. Moroz, *iScience* **2020**, *23*, 101681.

[12] L. Bellmann, P. Penner, M. Gastreich, M. Rarey, *J. Chem. Inf. Model.* **2022**, DOI 10.1021/acs.jcim.1c01378.

[13] D. E. Gloriam, *Nature* **2019**, *566*, 193–194.

[14] J. W. Scannell, A. Blanckley, H. Boldon, B. Warrington, *Nat. Rev. Drug Discov.* **2012**, *11*, 191–200.

[15] L. M. Mayr, P. Fuerst, *J. Biomol. Screen.* **2008**, *13*, 443–448.

[16] D. M. Volochnyuk, S. V. Ryabukhin, Y. S. Moroz, O. Savych, A. Chuprina, D. Horvath, Y. Zabolotna, A. Varnek, D. B. Judd, *Drug Discov. Today* **2019**, *24*, 390–402.

[17] Q. Li, *Virtual Screening of Small-Molecule Libraries*, Elsevier Inc., **2019**.

[18] B. Suay-García, J. I. Bueso-Bordils, A. Falcó, G. M. Antón-Fos, P. A. Alemán-López, *Int. J. Mol. Sci.* **2022**, *23*, DOI 10.3390/ijms23031620.

[19] N. Van Hilten, F. Chevillard, P. Kolb, *J. Chem. Inf. Model.* **2019**, *59*, 644–651.

[20] J. L. Grey, D. H. Thompson, *Expert Opin. Drug Discov.* **2010**, *5*, 1039–1045.

[21] G. Schneider, M. L. Lee, M. Stahl, P. Schneider, *J. Comput. Aided. Mol. Des.* **2000**, *14*, 487–494.

[22] G. Schneider, O. Clément-Chomienne, L. Hilfiger, P. Schneider, S. Kirsch, H.-J. Böhm, W. Neidhart, *Angew. Chemie* **2000**, *39*, 4130–4133.

[23] L. R. de Souza Neto, J. T. Moreira-Filho, B. J. Neves, R. L. B. R. Maidana, A. C. R. Guimarães, N. Furnham, C. H. Andrade, F. P. Silva, *Front. Chem.* **2020**, *8*, 1–18.

[24] M. Boehm, T. Y. Wu, H. Haussen, C. Lemmen, *J. Med. Chem.* **2008**, *51*, 2468–2480.

[25] J. Wahl, T. Sander, *J. Chem. Inf. Model.* **2021**, DOI 10.1021/acs.jcim.1c01041.

[26] M. Rarey, M. Stahl, *J. Comput. Aided. Mol. Des.* **2001**, *15*, 497–520.

[27] U. Lessel, B. Wellenzohn, M. Lilienthal, H. Claussen, *J. Chem. Inf. Model.* **2009**, *49*, 270–279.

[28] L. Bellmann, P. Penner, M. Rarey, *J. Chem. Inf. Model.* **2021**, *61*, 238–251.

[29] L. Bellmann, P. Penner, M. Rarey, *J. Chem. Inf. Model.* **2019**, *59*, 4625–4635.

[30] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[31] R. Schmidt, R. Klein, M. Rarey, *J. Chem. Inf. Model.* **2021**, DOI 10.1021/acs.jcim.1c00640.

[32] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.

[33] B. C. Doak, B. Over, F. Giordanetto, J. Kihlberg, *Chem. Biol.* **2014**, *21*, 1115–1142.

[34] B. C. Doak, J. Kihlberg, *Expert Opin. Drug Discov.* **2017**, *12*, 115–119.

[35] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

Додатки

Додаток 1

Список правил, які застосовувались фільтром OpenEye

```
#/*****************************************************************************
#Copyright (C) 2000-2005 by OpenEye Scientific Software, Inc.
#*****************************************************************************/
#This file defines the rules for filtering multi-structure files based on
#properties and substructure patterns.
MIN_MOLWT      150        "Minimum molecular weight"
MAX_MOLWT      750        "Maximum molecular weight"

MIN_NUM_HVY    10        "Minimum number of heavy atoms"
MAX_NUM_HVY    55        "Maximum number of heavy atoms"

MIN_RING_SYS    0      "Minumum number of ring systems"
MAX_RING_SYS    7      "Maximum number of ring systems"

MIN_RING_SIZE    0      "Minimum atoms in any ring system"
MAX_RING_SIZE    20      "Maximum atoms in any ring system"

MIN_CON_NON_RING    0      "Minimum number of connected non-ring atoms"
MAX_CON_NON_RING    20     "Maximum number of connected non-ring atoms"

MIN_FCNGRP       0      "Minimum number of functional groups"
MAX_FCNGRP       20      "Maximum number of functional groups"

MIN_UNBRANCHED   0      "Minimum number of connected unbranched non-ring atoms"
```

MAX_UNBRANCHED 8 "Maximum number of connected unbranched non-ring atoms"

MIN_CARBONS 5 "Minimum number of carbons"

MAX_CARBONS 40 "Maximum number of carbons"

MIN_HETEROATOMS 2 "Minimum number of heteroatoms"

MAX_HETEROATOMS 20 "Maximum number of heteroatoms"

MIN_Het_C_Ratio 0.10 "Minimum heteroatom to carbon ratio"

MAX_Het_C_Ratio 1.0 "Maximum heteroatom to carbon ratio"

MIN_HALIDE_FRACTION 0.0 "Minimum Halide Fraction"

MAX_HALIDE_FRACTION 0.5 "Maximum Halide Fraction"

#count ring degrees of freedom = (#BondsInRing) - 4 - (RigidBondsInRing) - (BondsSharedWithOtherRings)

#must be >= 0, from JCAMD 14:251-265,2000.

ADJUST_ROT_FOR_RING true "BOOLEAN for whether to estimate degrees of freedom in rings"

MIN_ROT_BONDS 0 "Minimum number of rotatable bonds"

MAX_ROT_BONDS 20 "Maximum number of rotatable bonds"

MIN_RIGID_BONDS 0 "Minimum number of rigid bonds"

MAX_RIGID_BONDS 50 "Maximum number of rigid bonds"

MIN_HBOND_DONORS 0 "Minimum number of hydrogen-bond donors"

MAX_HBOND_DONORS 6 "Maximum number of hydrogen-bond donors"

MIN_HBOND_ACCEPTORS 0 "Minimum number of hydrogen-bond acceptors"

MAX_HBOND_ACCEPTORS 10 "Maximum number of hydrogen-bond acceptors"

MIN_LIPINSKI_DONORS 0 "Minimum number of hydrogens on O & N atoms"

MAX_LIPINSKI_DONORS 5 "Maximum number of hydrogens on O & N atoms"

```
MIN_LIPINSKI_ACCEPTORS  0      "Minimum number of oxygen & nitrogen atoms"
MAX_LIPINSKI_ACCEPTORS  10     "Maximum number of oxygen & nitrogen atoms"

MIN_COUNT_FORMAL_CRG    0      "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG    3      "Maximum number of formal charges"

MIN_SUM_FORMAL_CRG   -2      "Minimum sum of formal charges"
MAX_SUM_FORMAL_CRG    2      "Maximum sum of formal charges"

MIN_CHIRAL_CENTERS                0                    "Minimum chiral centers"
MAX_CHIRAL_CENTERS                4                    "Maximum chiral centers"

MIN_XLOGP      -2.0      "Minimum XLogP"
MAX_XLOGP       6.0      "Maximum XLogP"

#choices are insoluble<poorly<moderately<soluble<very<highly
MIN_SOLUBILITY      moderately      "Minimum solubility"

PSA_USE_SandP   true    "Count S and P as polar atoms"
MIN_2D_PSA      0.0     "Minimum 2-Dimensional (SMILES) Polar Surface Area"
MAX_2D_PSA      150.0   "Maximum 2-Dimensional (SMILES) Polar Surface Area"

AGGREGATORS     true      "Eliminate known aggregators"
PRED_AGG        false     "Eliminate predicted aggregators"

#secondary filters (based on multiple primary filters)
GSK_VEBER       true      "PSA>140 or >10 rot bonds"
MAX_LIPINSKI   2         "Maximum number of Lipinski violations"
MIN_ABS 0.5             "Minimum probability F>10% in rats"
PHARMACOPIA     true      "LogP > 5.88 or PSA > 131.6"

ALLOWED_ELEMENTS  H,C,N,O,F,S,Cl,Br,I,P
```

ELIMINATE_METALS Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd

#acceptable molecules must have <= instances of each of the patterns below

#specific, undesirable functional groups

RULE 0 quinone
RULE 0 pentafluorophenyl_esters
RULE 0 paranitrophenyl_esters
RULE 0 HOBT_esters
RULE 0 triflates
RULE 0 lawesson_s_reagent
RULE 0 phosphoramides
RULE 0 beta_carbonyl_quat_nitrogen
RULE 0 acylhydrazide
RULE 0 cation_C_Cl_I_P_or_S
RULE 0 phosphoryl
RULE 0 alkyl_phosphate
RULE 0 phosphinic_acid
RULE 0 phosphanes
RULE 0 phosphoranes
RULE 0 imidoyl_chlorides
RULE 0 nitroso
RULE 0 N_P_S_Halides
RULE 0 carbodiimide
RULE 0 isonitrile
RULE 0 triacyloxime
RULE 0 cyanohydrins
RULE 0 acyl_cyanides
RULE 0 sulfonylnitrile
RULE 0 phosphonylnitrile
RULE 0 azocyanamides
RULE 0 beta_azo_carbonyl

RULE 0 polyenes

RULE 0 saponin_derivatives

RULE 0 cytochalasin_derivatives

RULE 0 cycloheximide_derivatives

RULE 0 monensin_derivatives

RULE 0 squalestatin_derivatives

#functional groups which often eliminate compounds from consideration

RULE 0 acid_halide

RULE 0 aldehyde

RULE 0 alkyl_halide

RULE 0 anhydride

RULE 0 azide

RULE 0 azo

RULE 0 di_peptide

RULE 0 michael_acceptor

RULE 0 beta_halo_carbonyl

RULE 0 nitro

RULE 0 oxygen_cation

RULE 0 peroxide

RULE 0 phosphonic_acid

RULE 0 phosphonic_ester

RULE 0 phosphoric_acid

RULE 0 phosphoric_ester

RULE 0 sulfonic_acid

RULE 0 sulfonic_ester

RULE 0 tricarbo_phosphene

RULE 0 epoxide

RULE 0 sulfonyl_halide

RULE 0 halopyrimidine

RULE 0 perhalo_ketone

RULE 0 aziridine

RULE 1 oxalyl

RULE 0 alphahalo_amine

RULE 0 halo_amine

RULE 0 halo_alkene

RULE 0 acyclic_NCN

RULE 0 acyclic_NS

RULE 0 SCN2

RULE 0 terminal_vinyl

RULE 0 hetero_hetero

RULE 0 hydrazine

RULE 0 N_methoyl

RULE 0 NS_beta_halothyl

RULE 0 propiolactones

RULE 0 iodoso

RULE 0 iodoxy

RULE 0 noxide

#groups of molecules

RULE 0 dye

#functional groups which are allowed, but may not be wanted in high quantities

#common functional groups

RULE 6 alcohol

RULE 8 alkene

RULE 4 amide

RULE 4 amino_acid

RULE 4 amine

RULE 4 primary_amine

RULE 4 secondary_amine

RULE 4 tertiary_amine

RULE 4 carboxylic_acid

RULE 6 halide

RULE 1 iodine

RULE 4 ketone

RULE 4 phenol

RULE 2 imine

RULE 1 methyl_ketone

RULE 1 alkylaniline

RULE 4 sulfonamide

RULE 1 sulfonylurea

RULE 0 phosphonamide

RULE 0 alphahalo_ketone

RULE 0 oxaziridine

RULE 1 cyclopropyl

RULE 2 guanidine

RULE 0 sulfonimine

RULE 0 sulfinimine

RULE 1 hydroxamic_acid

RULE 0 sulfinylthio

RULE 0 disulfide

RULE 0 enol_ether

RULE 0 enamine

RULE 0 organometallic

RULE 0 dithioacetal

RULE 1 oxime

RULE 0 isothiocyanate

RULE 0 isocyanate

RULE 3 lactone

RULE 3 lactam

RULE 1 thioester

RULE 1 carbonate

RULE 0 carbamic_acid

RULE 1 thiocarbamate

RULE 0 triazine

RULE 1 malonic

#other functional groups

RULE 2 alkyne

RULE 4 aniline

RULE 4 aryl_halide

RULE 4 carbamate

RULE 4 ester

RULE 4 ether

RULE 1 hydrazone

RULE 0 nonacylhydrazone

RULE 1 hydroxylamine

RULE 2 nitrile

RULE 2 sulfide

RULE 2 sulfone

RULE 2 sulfoxide

RULE 0 thiourea

RULE 1 thioamide

RULE 1 thiol

RULE 2 urea

RULE 0 hemiketal

RULE 0 hemiacetal

RULE 0 ketal

RULE 1 acetal

RULE 0 aminal

RULE 0 hemiaminal

#protecting groups

RULE 0 benzyloxycarbonyl_CBZ

RULE 0 t_butoxycarbonyl_tBOC

RULE 0 fluorenylmethoxycarbonyl_Fmoc

RULE 1 dioxolane_5MR

RULE 1 dioxane_6MR

RULE 1 tetrahydropyran_THP

RULE 1 methoxyethoxymethyl_MEM

RULE 2 benzyl_ether

RULE 2 t_butyl_ether

RULE 0 trimethylsilyl_TMS

RULE 0 t_butyldimethylsilyl_TBDMS

RULE 0 triisopropylsilyl_TIPS

RULE 0 t_butyldiphenylsilyl_TBDPS

RULE 1 phthalimides_PHT

RULE 2 arenesulfonyl