КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри супрамолекулярної хімії
проф. Рябухін Сергій Вікторович
Протокол №____ засідання кафедри
від “____” _____2022 р.

СТВОРЕННЯ МОДЕЛІ ХІМІЧНОГО ПРОСТОРУ ДЛЯ ПОШУКУ НОВИХ ЕФЕКТОРІВ ТУБУЛІНУ

Випускна кваліфікаційна робота магістра
студента спеціальності
102 Хімія
ОП «Хемоінформатика»
Миронова Богдана Сергійовича

Науковий керівник від кафедри
доцент кафедри молекулярної
біотехнології та біоінформатики
к.б.н. **Самофалова Дарія Олексіївна**

Оцінка захисту роботи
_____

Робота виконана під керівництвом наукового співробітника відділу геноміки та молекулярної біотехнології Державної Установи «Інститут харчової біотехнології та геноміки НАН України», к.б.н Самофалової Д.О.

Київ − 2022 р.

АНОТАЦІЯ

Миронов Б.С. Створення моделі хімічного простору для пошуку нових ефекторів тубуліну. – Випускна кваліфікаційна робота магістра за спеціальністю 102 Хімія ОП «Хемоінформатика».

В роботі проведено аналіз трьох датасетів бази даних Lifechemicals з метою розробки моделі для пошуку нових ефекторів тубуліну. Для побудови моделі аналізу були використані декілька фільтрів та *tanimoto similarity fingerprint.* В результаті аналізу було знайдено 16 кандидатів ефекторів тубуліну. Отримані дані були використані для побудови хімічного простору та дослідження моделей побудови хімічного простору за різними параметрами.

Ключові слова: ефектори тубуліну, хімічний простір, tanimoto similarity fingerprint, аналіз датасетів.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

AI – Artificial intelligence.

ML – Machine learning.

SOM – Self-organizing map.

PCA – Principal component analysis.

QSAR – Quantitative structure–activity relationship.

MFTA – Molecular field topology analysis.

ЗМІСТ

ВСТУП

Концепція хімічного простору широко використовується у відкритті ліків через багато можливостей застосування. Зокрема у розробці бібліотеки, класифікації сполук або наборів даних, виборі з'єднань, дослідженні зв'язків структура-діяльність (SAR) та навігації через зв'язки структура-властивість (SPR) загалом. Тому створення моделі хімічного простору для пошуку нових ефекторів тубуліну може дати новий поштовх в методах розробки ліків у багатьох напрямках. Візуалізація хімічного простору має зменшити розмірність проблеми порівняння молекулярних структур, що можна зробити за допомогою таких алгоритмів, як аналіз головних компонентів і стохастичний t-розподілений сусідній вклад, тощо.

Об'єкт дослідженнях: відомі та нові ефектори скелетних структур клітини про- та еукаріот.

Основна мета роботи полягає в пошуку відомих ефекторів тубуліну та на основі цього побудови моделі аналізу датасетів для пошуку потенційних ефекторів тубуліну.

Відповідно до мети були поставлені наступні ***завдання*** до виконання магістерської кваліфікаційної дипломної роботи:

1. Проаналізувати методи первинного відбору потенційно біологічно-активних сполук.
2. Здійснити пошук відомих ефекторів тубуліну встановлених експериментально та проаналізувати їх властивості.
3. Створити модель аналізу датасетів та пошуку нових потенційних ефекторів тубуліну.
4. Порівняти методи створення моделей віртуального простору задля пошуку нових сполук.

РОЗДІЛ 1. ОГЛЯД ЛІТЕРАТУРИ

1.1. Концепція хімічного простору. Маловимірні методи

Хімічний простір зазвичай визначають як сукупність усіх можливих органічних сполук [1]. Загальновідомо, що віртуальний хімічний простір більш ніж астрономічно великий, оскільки навіть не всіх атомів у Всесвіті було б достатньо, щоб синтезувати одну молекулу з усіх 10^{63} можливих органічних сполук розміром до 30 атомів. Припущення використовується для оцінки кількості потенційних фармакологічно активних молекул, однак, використовують Правила Ліпінського, зокрема обмеження молекулярної маси 500. Оцінка також обмежує хімічні елементи, якими раніше були вуглець, водень, кисень, азот та сірка. Це також робить припущення, що максимум 30 атомів залишатиметься нижче 500 Дальтон, дозволяє розгалуження і максимум 4 кільця і прибуває приблизно за 10^{63}. У наступних публікаціях це число часто неправильно цитується, щоб оцінювати розмір всього простору органічної хімії, яка була б набагато більшою, якщо включити галогени та інші елементи. Тим не менш, були зроблені величезні зусилля, щоб перерахувати мільярди гіпотетичних органічних сполук, що дозволило проводити великі віртуальні кампанії скринінгу [2].

Разом зі збільшенням розміру хімічного простору зріс інтерес до застосування картографічних методів для візуалізації простору [3]. В результаті з’явилися численні підходи до візуалізації та концептуалізації хімічного простору [4-6]. Наріжним каменем і ключовим аспектом усіх запропонованих методів є молекулярне представлення та параметри, які використовуються визначити простір, де будуть знаходитися сполуки.

Добре відомо, що хімічний простір дійсно великий. Обсяг доступної хімічної інформації постійно зростає і досягає настільки великих обсягів, що люди не можуть проаналізувати, таким чином потрапляючи у світ «big data» [7, 8].

Пошук і проектування нових молекул із бажаними властивостями зазвичай називають метафоричними термінами дослідження та навігація в хімічному просторі [9], і ефективна навігація вимагає певних карт та інструментів для позиціонування в цьому величезному хімічному всесвіті. Термін «chemography» був запропонований для цієї галузі за аналогією з географією [3], але він не став широковживаним.

Поняття хімічного простору не дуже формалізоване і включає кілька аспектів. По-перше, він може інтуїтивно представляти набір усіх можливих молекул або всіх молекул, що мають відношення до певної галузі чи проекту [10, 11]. Відчутна близькість точок у цьому хімічному просторі (або зірок у «хімічному всесвіті») може відображати ідею хіміка про структурну подібність, легкість синтетичного перетворення та/або схожість властивостей відповідних молекул. Одна з спроб формалізації цієї концепції визначила хімічний простір більш вузько як «комбінаторний набір усіх сполук, які можна виділити і сконструювати з можливих комбінацій та конфігурацій атомів N_I та електронів N_e в реальному просторі» [8].

По-друге, відповідно до загальної парадигми математичної хімії та аналізу структурно-активності, ми можемо представити кожну молекулу набором числових дескрипторів, що відповідають точкам у реальному просторі на основі координат. За даними Web of Knowledge, найперші публікації, в яких використовувався термін «хімічний простір» у значенні, подібному до нинішнього розуміння, з’явилися від геохіміків у 1980-х роках, які вільно визначили його як набір усіх хімічних даних, що відповідають різним мінералам. зразки. У контексті відкриття ліків ідеологічно подібне визначення запропонував Добсон [12], який прирівняв хімічний простір до «загального простору дескрипторів, що охоплює всі малі молекули на основі вуглецю, які в принципі можуть бути створені», або «багато- просторовий дескрипторний простір», де практично будь-яку цифрову характеристику молекули можна розглядати як дескриптор.

Незважаючи на те, що ці описи легко зрозуміти, ці описи важко використовувати в контексті комп’ютерного проектування ліків через відсутність математичної строгості. Аналіз, включаючи візуальний аналіз, зазвичай виконується для певних реалізацій вищезгаданих принципів.

Вчені добре визнають силу візуального представлення даних [13, 14]. Хороші зображення сприяють творчому мисленню та дозволяють легше розуміти великі обсяги складних даних. У візуалізації хімічного простору використовуються обидва аспекти концепції: вони неминуче вимагають набору молекул для зображення, і вони використовують дескриптори для генерування точкових координат у площині візуалізації. У найпростіших випадках візуалізація базується на координатах молекул у чистих (під)просторах дескрипторів, а в більш просунутих дослідженнях інші простори генеруються на основі значень дескриптора. У контексті аналізу хімічного простору дескриптор слід розуміти не лише як суто математичну функцію, а й у більш загальному вигляді як будь-яку характеристику молекули чи речовини, наприклад, будь-який експериментально отриманий параметр (температура плавлення, температура кипіння, щільність, кислотність, ліпофільність тощо) або навіть біологічну активність, представлену в будь-якій прийнятній формі.

1.2. Діаграми Ван Кревелена

Методи візуалізації, корисні для аналізу хімічного простору, з’явилися за роки до того, як термін «хімічний простір» був введений у широке наукове використання. Проблема аналізу складу природних сумішей була важливою в області аналізу викопного палива, і в 1950 році ван Кревелен запропонував рішення, яке пізніше стало добре відомим під його ім'ям.

Діаграма ван Кревелена спочатку була розроблена для характеристики складних зразків за співвідношенням елементів, тобто за кількістю вуглецю (C) до водню (H). Завдяки прогресу в мас-спектрометрії, що забезпечує точні вимірювання маси, аналіз окремих компонентів у складній суміші став можливим. Необхідною умовою є перетворення точних виміряних мас в елементарні формули. Діаграма ван Кревелена — це дво- або тривимірний графічний аналіз, у якому елементний склад сполук зображено відповідно до їхніх атомних співвідношень, наприклад, графік співвідношення H/C та O/C. Співвідношення H/C розділяє сполуки за ступенем їх насичення, тоді як співвідношення O/C та N/C відображає класи O та N відповідно.

Діаграми Ван Кревелена були запропоновані як метод візуалізації результатів елементного аналізу для різних зразків вугілля. Оскільки елементний аналіз дає лише дані про вміст вуглецю, водню та кисню у зразках, атомні співвідношення H/C та O/C можна використовувати як координати, а послідовний аналіз різних зразків дозволив дослідникам вивчати розподіл вмісту елемента або аналізувати процеси окислення або відновлення. Нове життя в цю техніку влилося під час розробки методів мас-спектрометричного аналізу таких складних сумішей, як гумінові речовини [15] та інші природні органічні речовини. Склад сумішей являє собою характерний малюнок в координатах O/C-H/C, який також можна покращити шляхом додавання третього виміру N/C [16] і фарбування точок на основі наявність інших елементів, таких як сірка, фосфор тощо. Ідентифікація окремих сполук у таких сумішах є нетривіальним завданням, і зазвичай воно не потрібно для обґрунтованої характеристики хімічного простору, представленого в конкретних зразках.

1.3. Координати дескриптора

Ймовірно, найпростіший метод візуалізації розподілу властивостей для наборів даних використовує самі значення дескриптора як координати. Цей метод широко застосовується для класифікації та характеристики бібліотек з малими молекулами: якщо розподіл однієї властивості є важливим, то його залежність від простих параметрів, таких як молекулярна маса (MW), є більш інформативною. Це спостереження було елегантно реалізовано в інструменті візуалізації «Золотий трикутник» [17], що дозволяє користувачеві знаходити сполуки з оптимальним кліренсом та оральним всмоктуванням на основі їх положення в координатах MW та оціненого коефіцієнта розподілу октанол-буфер (pH 7,4) (eLogD).

Візуалізація наборів даних з координатами дескриптора легко інтерпретується та дуже інтуїтивно зрозуміла; однак інформаційний вміст на картах, отриманих таким чином, дуже низький. Були розроблені більш досконалі методи для збільшення обсягу інформації, упакованої у візуальні зображення хімічного простору та підпростору. Дві основні групи цих методів – це методи, засновані на карті, і методи на основі графіків.

1.4. Розробка дата-сету на основі штучного інтелекту (AI)

Штучний інтелект (AI) пропонує потенціал для трансформації відкриття ліків. За останні кілька років відкриття ліків за допомогою штучного інтелекту значно зросло завдяки технологічному прогресу, такому як використання нейронних мереж для розробки молекул і застосування графіків знань для розуміння цільової біології.

Декілька компаній, що займаються розробкою ліків, створених на основі штучного інтелекту, розпочали клінічні випробування молекул, у деяких випадках повідомляючи про значно прискорені терміни та знижені витрати, що викликає високі очікування в спільноті досліджень і розробок. Крім того, багато відомих фармацевтичних компаній створили партнерські відносини з компаніями з

штучного інтелекту, щоб досліджувати технологію. Незважаючи на цей прогрес, AI ще тільки починає відкривати ліки, і багато відкритих питань щодо його впливу та майбутнього потенціалу.

1.5. Хімічні структури та їх властивості, побудовані за допомогою AI

Загальнодоступні дані про хімічну структуру активів, отриманих від штучного інтелекту, наразі обмежені. Як наслідок, систематичний статистичний аналіз на даний момент неможливий. Однак аналіз прикладів з деякими розкритими даними може дати проблиск майбутнього.

Одним із таких прикладів є інгібітори TYK2. TYK2 є членом сімейства кіназ Янус (JAK), які мають кілька існуючих інгібіторів, включаючи 10 продуктів, що продаються. Однією з поширених проблем цих молекул є їх обмежена селективність щодо однієї ізоформи JAK, що впливає на їх профіль безпеки. Зусилля щодо виявлення за допомогою штучного інтелекту нещодавно виявили актив з новим алостеричним способом дії, який, як видається, принаймні в 20 разів селективний щодо TYK2 порівняно з іншими членами сімейства JAK.

Цікаво, що порівнюючи структуру отриманих AI, TYK2-селективних інгібіторів з класично відкритими, менш селективними інгібіторами JAK в хімічному просторі, ми не спостерігаємо вражаючих відмінностей. Навпаки, отримані від штучного інтелекту селективні інгібітори TYK2, здається, поширюються на недостатньо представлені області хімічного простору.

1.6. Хронологічні рамки відкриття молекул за допомогою AI

Однією з найбільших надій на відкриття ліків за допомогою штучного інтелекту є прискорення термінів відкриття — наприклад, швидка ідентифікація та перевірка цілі або менша кількість циклів розробки та оптимізації молекул.

Хоча, як відомо, важко виміряти терміни відкриття, використовуючи загальнодоступні дані, ми змогли відновити приблизні терміни для вибраних партнерств фармакології та штучного інтелекту та програм відкриття. Виходячи з часу отримання патентів, публікацій та публічних повідомлень, ми знаходимо численні програми з підтримкою штучного інтелекту, які завершують все відкриття та доклінічний шлях менш ніж за чотири роки. Такі початкові дані вигідно порівнюються з історичними часовими рамками в індустрії за п’ять-шість років і здаються особливо вражаючими, враховуючи, що AI все ще на стадії відкриття і, ймовірно, буде прискорюватися далі, коли компанії AI дозрівають.

1.7. Map-based уявлення хімічного простору

У map-based представленнях хімічного простору багатовимірні дескрипторні простори сполук згортаються (відображаються) у двовимірний простір, подібно до побудови плоских карт Землі з тривимірних даних. Отже, необхідно застосовувати методи зменшення розмірності, серед яких найбільш помітними є аналіз головних компонентів (PCA), самоорганізуючі карти (SOM), стохастичне вбудовування близькості (SPE), стохастичне вбудовування сусідів (SNE) та генеративне топографічне відображення (GTM [18]).

PCA є потужним методом лінійного зменшення розмірності для великих наборів даних. Основна ідея цього методу полягає в тому, щоб витягти та пояснити якомога більше варіацій з багатовимірної матриці (зазвичай взаємокорельованих) дескрипторів шляхом отримання оптимальних лінійних комбінацій цих векторів дескрипторів. Основною перевагою цього методу при застосуванні до проблем хімічного аналізу простору є його нестохастичність,

тобто один і той же набір дескрипторів для однієї складеної множини забезпечуватиме ті самі основні компоненти та те саме відображення. Основні головні компоненти, що накопичують найбільше варіацій, використовуються як осі і зазвичай мають фізично значущі інтерпретації.

1.8. Карти подібності

Метрики і функції подібності, особливо ті, що базуються на молекулярних фрагментах, також можна використовувати для візуалізації та аналізу хімічних бібліотек. Нижча обчислювальна ефективність порівняно з 2D-дескрипторами обмежує застосовність карт подібності до менших баз даних, зазвичай анотованих даними про біоактивність, що дозволяє моделювати ландшафти біоактивності на основі порівняння профілів активності. Фрагменти (fingerprints) і значення схожості можуть використовуватися як вхідні дані для PCA [19], але також використовуються іншими способами. У картах подібності мультизлиття [19] сполуки представлені у вигляді точок у координатах максимального злиття та середнього злиття, визначених для кожної молекули тестового набору проти молекул еталонного набору. Цей підхід показав свою корисність у кількох тестових випадках, будучи більш дискримінаційним, ніж PCA.

Карти подібності структури і активності зображують подібність активності проти молекулярної подібності, таким чином роблячи представлення всієї хімічної інформації одновимірною. Наступним очевидним кроком є відкидання хімічних координат та фарбування точок за схожістю, що дозволяє побудувати подвійні та потрійні карти активності-різниці, що є дуже ефективним для аналізу поліфармакології проти обмежених наборів цілей [20]. Інша назва того ж підходу до аналізу — хіміографія [21].

1.9. Стохастичні карти

Інша велика група методів зменшення розмірності для візуалізації заснована на стохастичних процедурах. Незважаючи на їх нижчу візуальну інтерпретацію порівняно з вищезгаданими методами, які зазвичай забезпечують фізично значущі координати, стохастичні карти дозволяють краще кластеризувати сполуки та мають більш високу передбачувану здатність. Вміст інформації в картах, отриманих стохастичними методами, зазвичай також вищий порівняно з картами PCA, в яких кілька перших компонентів можуть пояснити до 60 - 80% варіації.

Мабуть, найбільш поширеним стохастичним методом візуалізації хімічного простору та нелінійного зменшення розмірності є метод SOM [22]. Ключовим інструментом цього підходу є двовимірний масив або мережа процесорних блоків («нейронів»), що надають один одному локальний зворотний зв’язок. Ці одиниці ініціалізуються випадковим чином у просторі дескрипторів, що використовується для навчання, і коли вибірка представлена в мережу, визначається «переможна» одиниця, яка найбільш схожа на представлену вибірку. Сусідні одиниці потім оптимізуються для збільшення схожості. Коли всі зразки представлені в мережу, вони розподіляються між підрозділами. Щоб уникнути граничних проблем, мережу можна зробити топологічно еквівалентною тору з одиницями на протилежних сторонах карти, які розглядаються як сусіди (подібно до періодичних граничних умов при моделюванні молекулярної динаміки) або сфери [23]. Оскільки під час навчання немає цільової функції, яку потрібно оптимізувати, класичний SOM розглядається як метод навчання без нагляду.

Залежно від співвідношення між кількістю сполук, а також розподілу властивостей у навчальному наборі, результуюча карта може містити незайняті або зайняті нейрони різними сполуками [24]. Під час навчання кожна сполука проектується точно в один нейрон, і лише один нейрон забезпечує відповідь на етапі передбачення. Варіанти візуалізації для SOM включають забарвлення за кількістю нейронів або молекулярними особливостями. Прогноз властивостей за

допомогою SOM заснований на подібності: коли нова сполука «вистрілює» нейрон, зайнятий сполуками з необхідною активністю, очікується, що ця сполука матиме таку саму активність. Віртуальний третій вимір може бути введений шляхом попарного порівняння подібності нейронів для різних SOM в ході підходу з використанням мульти SOM, що дозволяє користувачеві ідентифікувати нейрони, зайняті сполуками, що володіють кількома необхідними властивостями, відображеними на різних SOM.

Аналіз глобального хімічного простору («Small Molecule Universe», SMU) проведено за допомогою SOM [25]. Масштабну стохастичну генерацію сполук 9M проводили з урахуванням правил Ліпінського та синтетичної доступності; отримана бібліотека була візуалізована та порівняна з PubChem за допомогою PCA та SOM, що продемонструвало набагато ефективніше використання простору сюжету в останньому випадку. Найбільш вражаючою особливістю SMU. SMU була заповненість його вузлів сполуками PubChem, 98% з яких займають лише 2% вузлів, що підкреслює, що більша частина хімічного простору SMU (а саме 84%) є абсолютно невивченою. Коли простори натуральних продуктів і ліків порівнювали за допомогою SOM, лише кілька вузлів містили лише ліки, тоді як значна частина вузлів містила лише натуральні продукти, що підкреслює взаємодоповнюваність цих підпросторів [26].

1.10. Побудова хімічного простору на основі графів

У вищезгаданих підходах до візуалізації хімічного простору сполуки представлені у вигляді точок на графіку або проектуються у вузли сітки. Кожна молекула розглядається як сама по собі, і лише показники подібності використовуються для аналізу та порівняння її з подібними. Тим не менш, під час аналізу зв'язків структура-діяльність (SAR) у серії близькоспоріднених аналогів

важливо порівнювати їх не лише за загальними показниками, а й за парними відношеннями. Мережевий і, загалом, графовий аналіз тривалий час широко використовується в обчислювальній та математичній хімії і застосовувався для численних завдань, таких як класифікація хімічних реакцій, перерахування малих молекул, організація даних та генерація моделі QSAR [27]. Мережі, пов’язані з відкриттям ліків, у парадигмі мережевої фармакології розглядають пари з’єднання-мішень. Значно менше уваги приділяється складено-складеним відношенням [28]. Дві основні причини роблять мережеві представлення хімічного простору важливими: (1) вони позбавлені координат для окремих сполук і (2) вони представляють дискретні хімічні набори даних природним, дискретним чином [28]. Розроблено багато ефективних підходів для аналізу та візуалізації графів у дуже різних областях, а зі зростанням обчислювальної потужності можна аналізувати все більш і більш досконалі графіки. Мережні методи аналізу хімічного простору зазвичай забезпечують значний рівень інтерактивності в реалізації.

Інший клас мережевих уявлень хімічного простору вводить напрямок до ребер, перетворюючи таким чином плоскі графи в дерева. Ідея риштування була використана у двох підходах: Scaffold Hunter [29] та Scaffold Explorer [30]. Обидва підходи дозволяють аналізувати послідовні декорації риштувань і зведення різних сполук до батьківських риштувань. Scaffold Hunter пропонує розташовувати сполуки на концентричних колах, кожне з яких представляє наступний рівень функціоналізації центрального фрагмента, тоді як Scaffold Explorer групує сполуки відповідно до каркасів і створює дерева шляхом послідовної функціоналізації. Подібний підхід реалізовано в модулі CardView пакету програм StarDrop, що дозволяє користувачеві виконувати всі необхідні кластеризації та групування сполук на основі власних експертних знань, розширюючи таким чином можливості формальної кремнієвої логіки до нечітких. людська логіка. У цьому підході спрямовані ребра, забарвлені подібністю або іншою попарною

мірою, можуть представляти послідовність оптимізації властивостей у ході проекту або принципу організації спеціального дерева.

Підходи, засновані на графіках, можуть бути використані не тільки на молекулярному, а й на атомному рівні для роботи з обмеженими хімічними підпросторами. Цікавий підхід до хімічно значущої візуалізації та навігації в хімічному просторі заснований на аналізі топології молекулярного поля (MFTA) [27]. Цей надструктурний метод має на меті звести локальні структурні особливості сполук у загальну систему відліку, визначену молекулярним суперграфом, тобто таку топологічну мережу, що всі структури, що належать до досліджуваного ряду, можуть бути накладені на суперграф для того, щоб охарактеризувати їх уніфіковано. Як молекулярні дескриптори MFTA використовує локальні фізико-хімічні параметри (властивості атома та зв'язку), які можна швидко оцінити за структурною формулою та відображати основні типи міжмолекулярних взаємодій, що беруть участь у зв'язуванні лігандів невеликих молекул з мішенями.

РОЗДІЛ 2. МАТЕРІАЛИ ТА МЕТОДИ

Об’єктами даного дослідження були вже відомі ефектори з підтвердженою експериментально активністю та нові ефектори скелетних структур клітини про- та еукаріот, отримані з відкритих комерційних бібліотек.

Для пошуку сполук з підтвердженою активністю були використані відкриті хімічні бази даних та бази даних наукової латератури, зокрема база даних ChEMBL, PubMed, тощо. Так, задля проведення даної роботи потрібно проаналізувати методи первинного відбору потенційно біологічно-активних сполук, відібрано датасет для його подальшого тестування та аналізу.

Для створення головних параметрів моделі хімічного простору було проведено аналіз фізико-хімічних властивостей та структурних особливостей відповідних датасетів - відомих ефекторів тубуліну встановлених експериментально. Отримання референсних даних відкриває можливість написання моделі аналізу датасетів та пошуку нових потенційних ефекторів тубуліну із використанням різноманітних фільтрів та фінгерпринтів.

Пошук нових сполук та побудова моделі хімічного простору була за проведена за допомогою програми з відкритим кодом типу CIME із використанням необхідних параметрів скринінгових сполук.

2.1. Вибір датасетів

Для аналізу датасетів було використано базу даних LifeChemicals. Були вибрані такі датасети:

- Stock HTS Compounds (550k)
- General Fragment Library (50k)
- Pre Plated Diversity Set PS6 (50k)

Lifechemicals Stock HTS Compounds — це колекції малих органічних молекул для високопродуктивного скринінгу на даний момент містить понад 490 000 оригінальних лікарських сполук, доступних на складі.

Середні фізико-хімічні значення та їх відносний розподіл наведені на **рис. 2.1**. Усі продукти проходять суворий контроль якості, щоб гарантувати їх чистоту > 90 %, за допомогою ЯМР 400 МГц та/або LCMS.

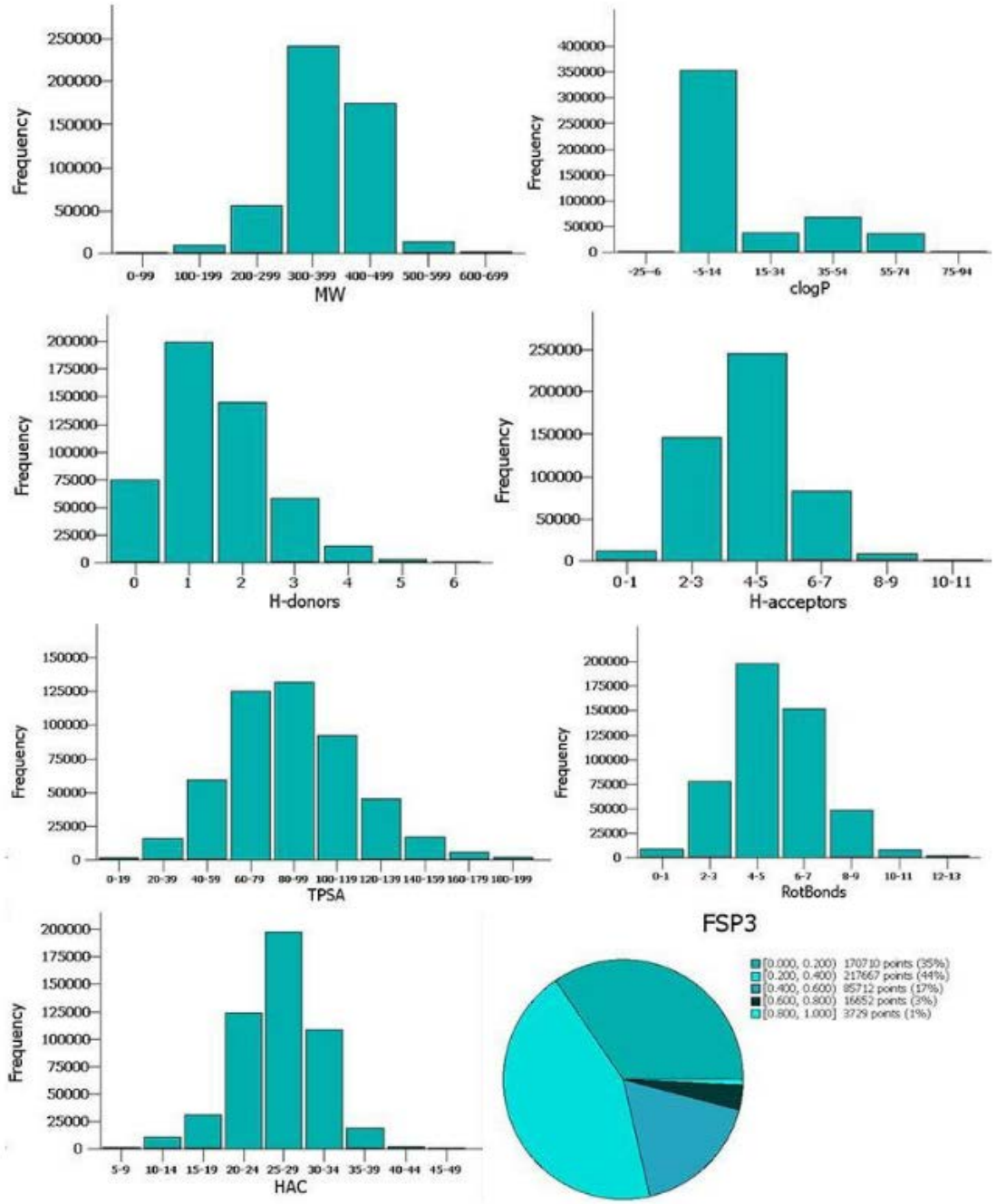


Рис. 2.1. Розподіл фізико-хімічних значень колекції HTS Compound.

Бібліотека Life Chemicals General Fragment Library містить близько 54 900 фрагментів з MW ≤ 300 і ClogP ≤ 3,0, які легко доступні на складі для проектів пошуку ліків на основі фрагментів. Усі реакційноздатні та нестабільні молекули відфільтровані.

Fragment-based lead discovery (FBLD) є ефективним сучасним підходом до виявлення ліків. Він заснований на скринінгу відносно невеликих бібліотек фрагментів (як правило, кілька тисяч сполук) і ідентифікації потенційних влучень, які можуть лише слабко зв'язуватися з біологічною мішенню (мілімолярну спорідненість можна вважати достатньо значною). Зменшений розмір і складність цих молекул дозволяють ефективніше відбирати хімічні проби простору. Подальше створення та/або об'єднання фрагментів залишає більше можливостей для отримання нових сполук з більш високою спорідненістю та покращеними фізико-хімічними властивостями.

Life Chemicals Pre-plated Diversity Set (PS6) – це ексклюзивний набір, що складається загалом із 50 000 нових скринінгових сполук з оптимальними фізико-хімічними властивостями, відібраних шляхом пошуку відмінності з колекції новосинтезованих молекул Life Chemicals HTS Compound. Він був розроблений для полегшення фенотипового (на основі клітин) і цільового високопродуктивного скринінгу (HTS) шляхом надання доступу до колекції структурно різноманітних малих молекул у зручний спосіб.

The Pre-plated Diversity Set є ідеальними інструментом для проектів виявлення ліків, які вимагають широкого діапазону хімічної структури, схожості та якісних лікарських і lead-like сполук для скринінгу на нові або підтверджені біологічні цілі з різних класів, або де мало інформації про цільову структуру білка.

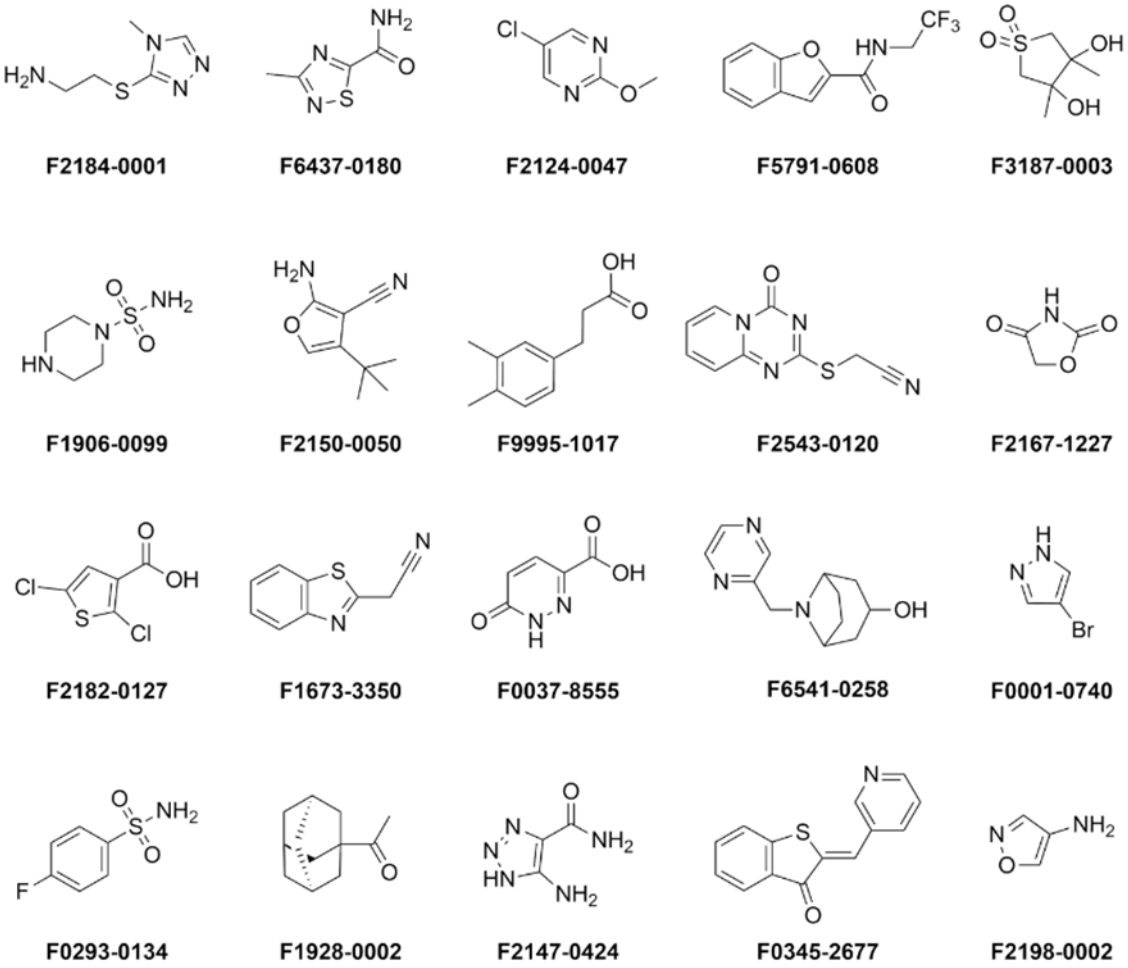


Рис. 2.2. Фрагменти молекул із General Fragment Library

2.2. Вихідні дані

З молекулярно-біологічної точки зору, найбільш підходящою стратегією конструювання лікарських засобів на основі структури (SBDD) є виявлення та розробка нових лікарських засобів, націлених на унікальні білки патогенного процесу, зберігаючи максимум консерватизму і не маючи прямих ортологів. Цей підхід покращує шанси подолання побічних ефектів, пов'язаних з інгібуванням подібних білкових мішеней. У відповідності з метою проекту було проведено пошук референтних активних речовин та визначено фізико-хімічні параметри, що

забезпечують інгібуючий вплив досліджуваних речовин на цитоскелет **таблиця 2.1.**

За результатами проведеного пошуку в базах даних Google Patents, PubChem, ZINC, eMolecules і ChEMBL було відібрано первинний сет (IC50, Ki та ін., менше ніж 10 мкМ, інгібування> 25%) депонованих біологічно-активних речовин з селективною дією по відношенню до α-, β- і γ-тубулінів. Зведену групу потенційних інгібіторів було перевірено згідно з правилами конструювання лікарських засобів (drug design rules), а також на наявність певних фрагментів – характерних фармакофорів. Застосовуючи метод пошуку за гомологією (2Д-фінгерпринти, алгоритм Танімото і Тверські з порогом подібності до 85%) до отриманого набору сполук з доступних комерційних баз даних нами було розроблено бібліотеку низькомолекулярних сполук з прогнозованою активністю до відповідних мішеней. Обробка бази, що після аналізу можливих конформацій налічувала близько 15 мільйонів речовин зводилась до фільтрації за медико-хімічними та PAINS фільтрами, що видаляють небажані з біологічної точки зору структури. Вся ця процедура була проведена за допомогою програми DataWarrior та набору скриптів RDKit **рис 2.3.**

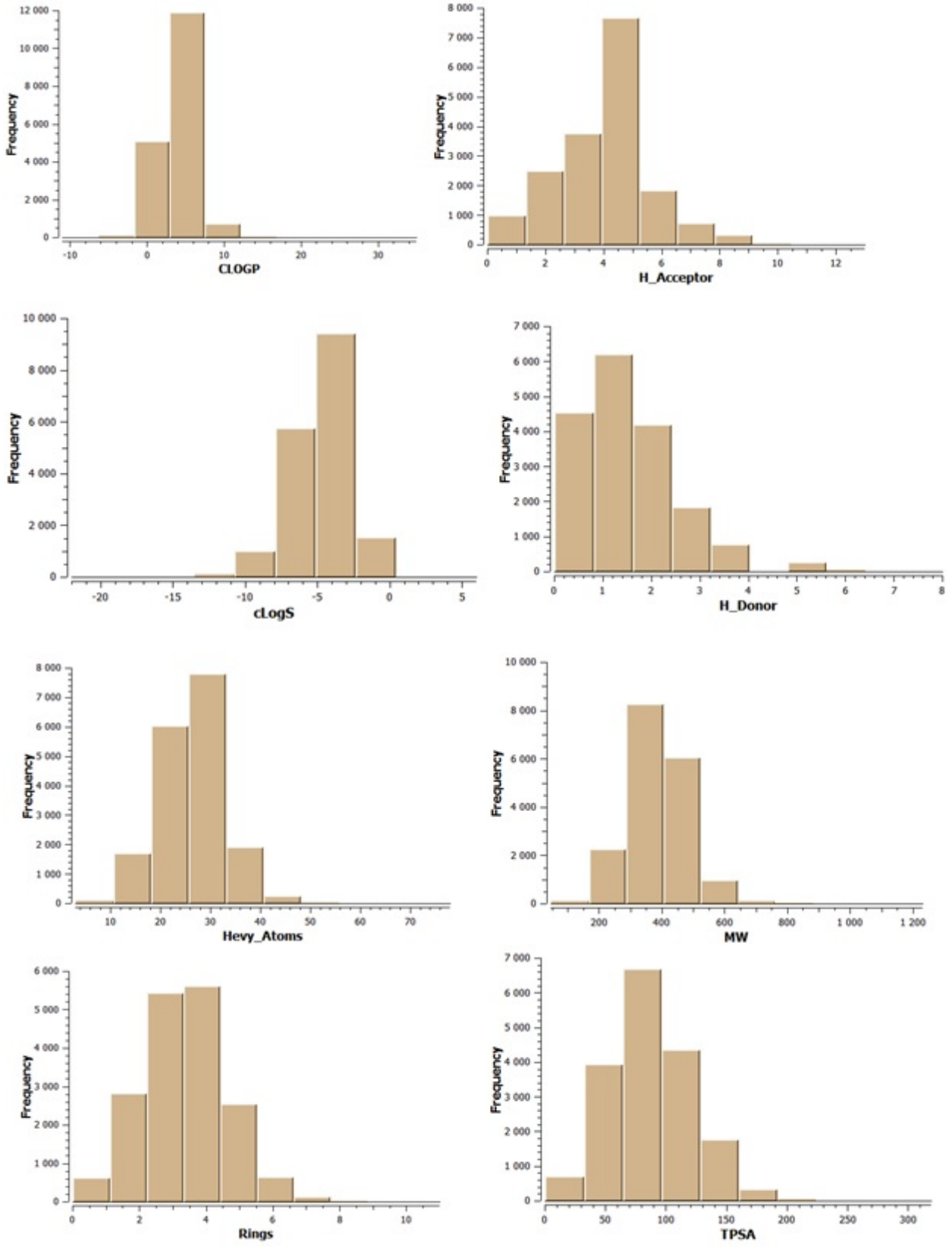
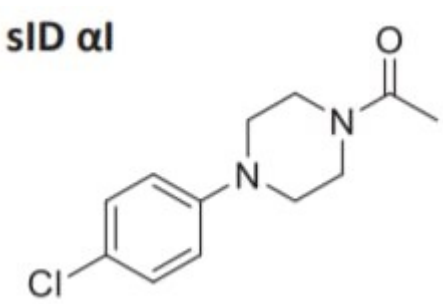
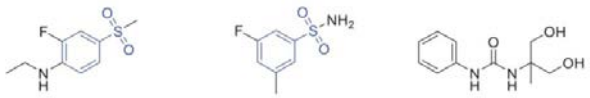
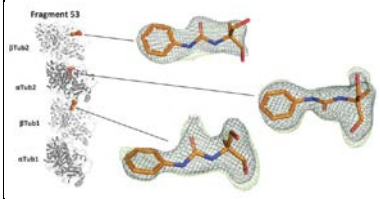
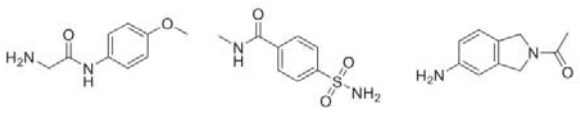
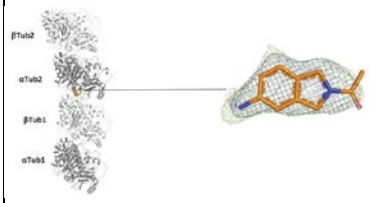
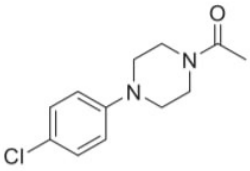
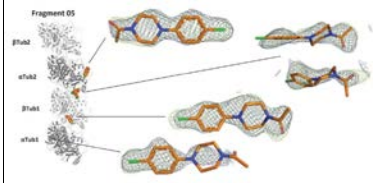
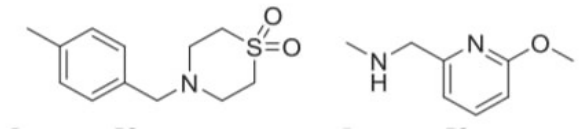
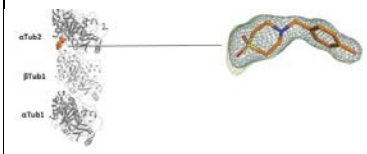
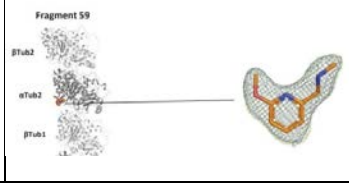


Рис. 2.3. Розподіл сполук за фізико-хімічними властивостями, що будуть використані під час віртуального скринінгу та молекулярного докінгу

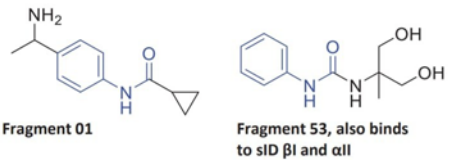
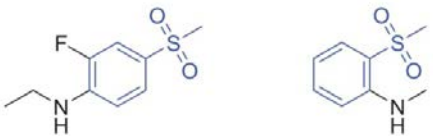
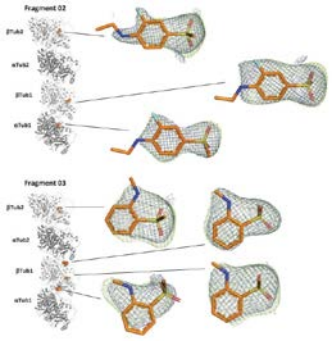
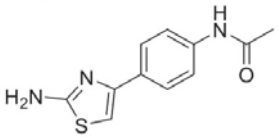
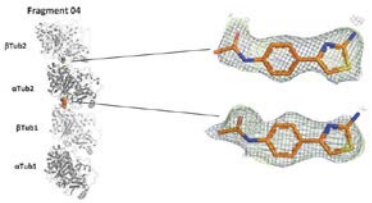
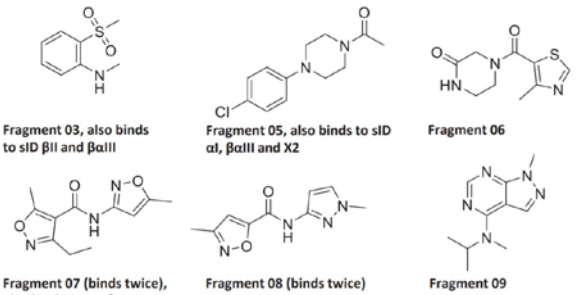
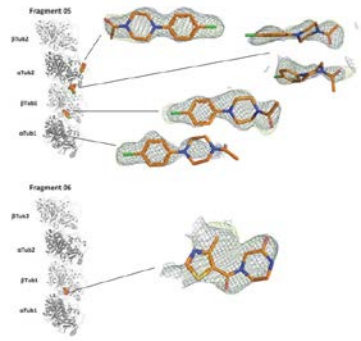
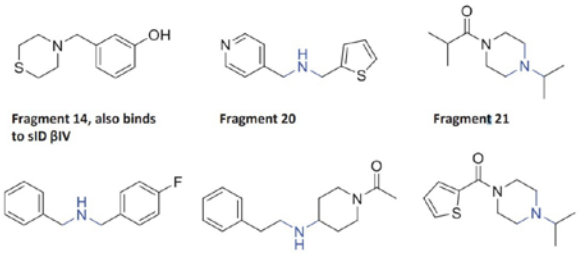
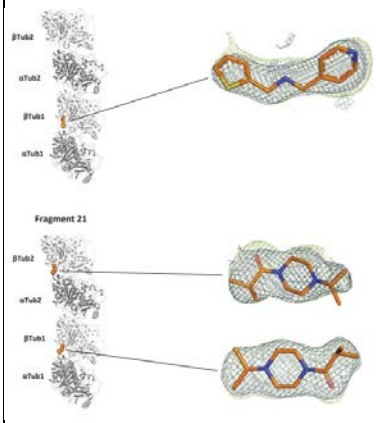
Таблиця 2.1.

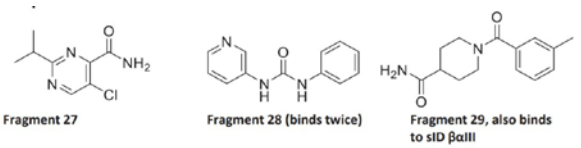
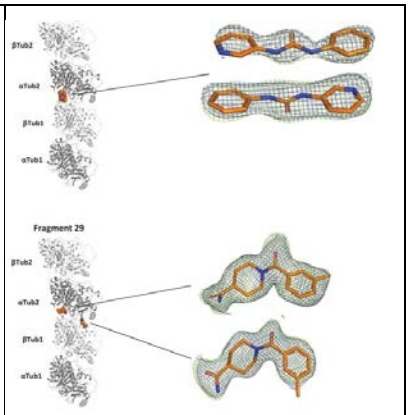
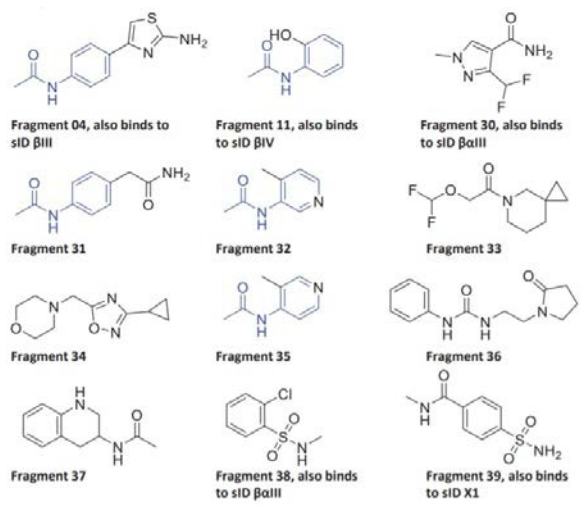
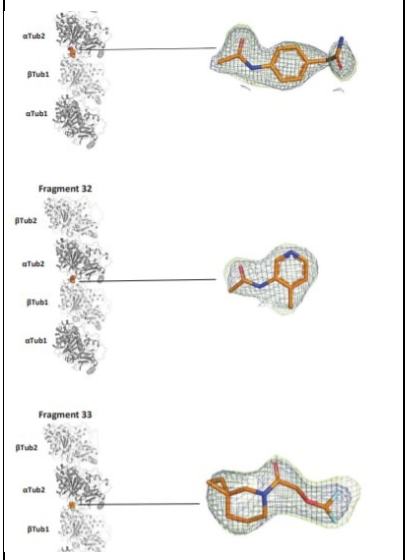
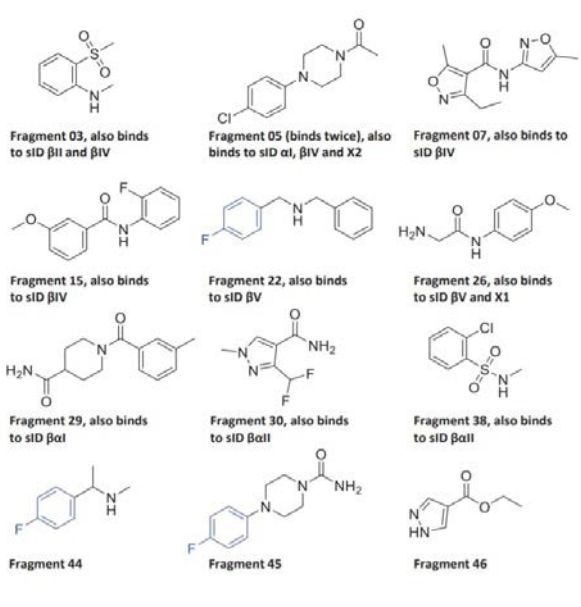
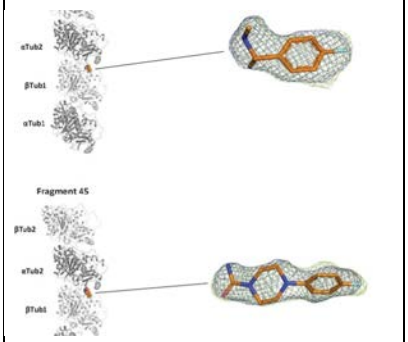
Класифікація інгібіторів тубуліну за сайтами

Субодиниця	Сайт	Фрагмент (згенерований ліганд)	Візуалізація
α	sID αI		
	sID αII		
Crystal contact/binding partner in T2R-TTL	sID X1		
	sID X2		
	sID X3		 

Таблиця 2.2.

Класифікація інгібіторів тубуліну за сайтами

Субодиниця	Сайт	Фрагмент (згенерований ліганд)	Візуалізація
β	sID βI	 Fragment 01 Fragment 53, also binds to sID βI and αII	
	sID βII		
	sID βIII		
	sID βIV	 Fragment 03, also binds to sID βII and βαIII Fragment 05, also binds to sID αI, βαIII and X2 Fragment 06 Fragment 07 (binds twice), also binds to sID βαIII Fragment 08 (binds twice) Fragment 09	
	sID βV	 Fragment 14, also binds to sID βIV Fragment 20 Fragment 21 Fragment 22, also binds to sID βαIII Fragment 23 Fragment 24	

β1**α**2-тубулін	sID βαI	 Fragment 27 Fragment 28 (binds twice) Fragment 29, also binds to sID βαIII	
	sID βαII	 Fragment 04, also binds to sID βIII Fragment 11, also binds to sID βIV Fragment 30, also binds to sID βαIII Fragment 31 Fragment 32 Fragment 33 Fragment 34 Fragment 35 Fragment 36 Fragment 37 Fragment 38, also binds to sID βαIII Fragment 39, also binds to sID X1	
	sID βαIII	 Fragment 03, also binds to sID βII and βIV Fragment 05 (binds twice), also binds to sID αI, βIV and X2 Fragment 07, also binds to sID βIV Fragment 15, also binds to sID βIV Fragment 22, also binds to sID βV Fragment 26, also binds to sID βV and X1 Fragment 29, also binds to sID βαI Fragment 30, also binds to sID βαII Fragment 38, also binds to sID βαII Fragment 44 Fragment 45 Fragment 46	

2.3. Аналіз датасетів

Головною ідеєю роботи є аналіз датасетів та подільшої розробки моделі AI для побудови хімічного простору для подальшого пошуку кандидатів ефекторів тубуліну. Для цього потрібно було взяти якісь вихідні дані (властивості) від яких можна відштовхуватись та проводити подальший аналіз. Після цього йде етап написання скриптів для пошуку ідеального кандидату. В якості еталону були взяті уже відсортовані та згенеровані ліганди. Після цього було взято декілька фільтрів для аналізу молекул в датасетах та використаний метод *Tanimoto similarity* для пошуку кандидатів зв'язування з сайтами субидиниць тубуліну.

Заглиблюючись у світ хемоінформатики, я нещодавно наштовхнувся на інші фільтри, які розширюють правило Ліпінського. Список можна продовжевати довго (в залежності від методу дослідження: ML та моделі на основі якої ML побудований, fingerprints які використовуються і тд), але основні, це:

- Lipinski Rule of 5 [31]
- Ghose Filter [32]
- Veber Filter [33]
- Rule of 3 Filter [34]
- REOS Filter [35]
- Drug-like Filter (QED) [36]

Для аналізу датасетів було використано бібліотеку RDKit, а також для написання скриптів Python 3.10. Спочатку було завантажено датасети у форматі SDF. Після цього було вибрано усі необхідні параметри та фінгерпринти.

Для аналізу вибраних датасетів було використано лише 4 з 6 фільтрів: Lipinski Rule of 5, Ghose Filter, Veber Filter та REOS Filter. Ці фільтри є одні із найвживаніших. Було вирішино викинути Rule of 3 Filter та Drug-like Filter, так як вони б обмежували прогін датасету. Ці фільтри зменшили б нам кількість сполук за молекулярною масою (менше або дорівнює 300) та за донор-акцепторними

зв’язками (до 3). Для Drug-like Filter ці зачення трішки більші, але все одно, менші ніж у інших (молекулярна маса до 400 та донорні зв’язки до 5, а акцепторні зв’язки до 10). Для сучасних лікарських засобів такі значення молекулярної маси є замалі.

Lipinski Rule of 5:

- Молекулярна маса <= 500
- LogP <= 5
- Кількість донорів водневого зв'язку <= 5
- Кількість акцепторів водневого зв'язку <= 10

Ghose Filter:

- Молекулярна маса від 160 до 480
- LogP від -0,4 до +5,6
- Кількість атомів від 20 до 70
- Молярна рефракційна здатність від 40 до 130

Veber Filter:

- Поворотні зв’язки <= 10
- Топологічна площа полярної поверхні (TPSA) <= 140

REOS Filter:

- Молекулярна маса від 200 до 500
- LogP від -5,0 до +5,0
- Кількість донорів водневого зв'язку від 0 до 5
- Кількість акцепторів водневого зв'язку від 0 до 10
- Заряд від -2 до +2
- Поворотні зв’язки від 0 до 8
- Кількість важких атомів від 15 до 50

Для пошуку кандидатів ефекторів тубуліну було вирішено використовувати *Tanimoto similarity fingerprint* та порівнювати молекули в датасеті з референсними даними (сполуки які вже були отримані в результаті дата майнингу та перевірені докингом). В результаті усі отримані дані були записані в новий SDF датасет. Були вписані всі необхідні параметри для подальшого аналізу за допомогою програми CIME.

CIME (ChemInformatics Model Explorer), інтерактивна веб-система, яка дозволяє користувачам будувати хімічний простір з набору **рис 2.4**, завантаженого в нього, хімічних даних, візуалізувати пояснення моделей, порівнювати методи інтерпретації та досліджувати підгрупи сполук **рис 2.5**. Інструмент не залежить від моделі і може запускатися на сервері або робочій станції.

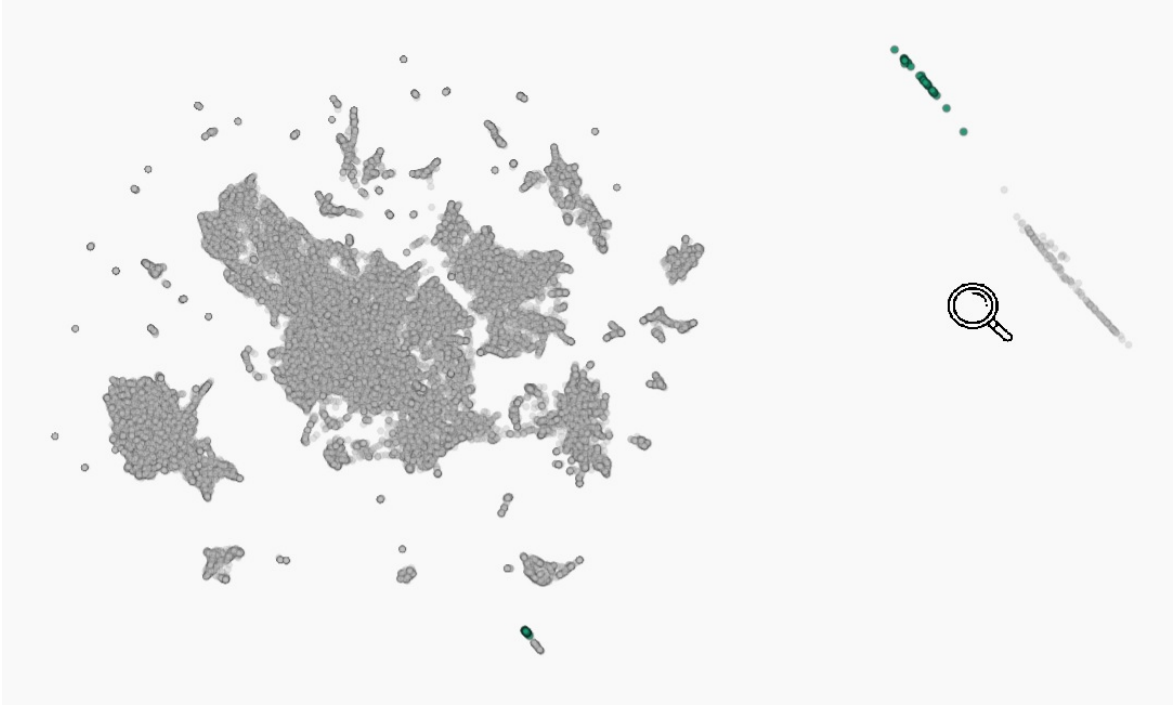


Рис 2.4. Приклад побудови хімічного простору з використанням програми CIME (зеленим кольором виділений набір сполук згрупований за певними fingerprints. В правій верхній частині збільшений маштаб цієї ділянки простору)

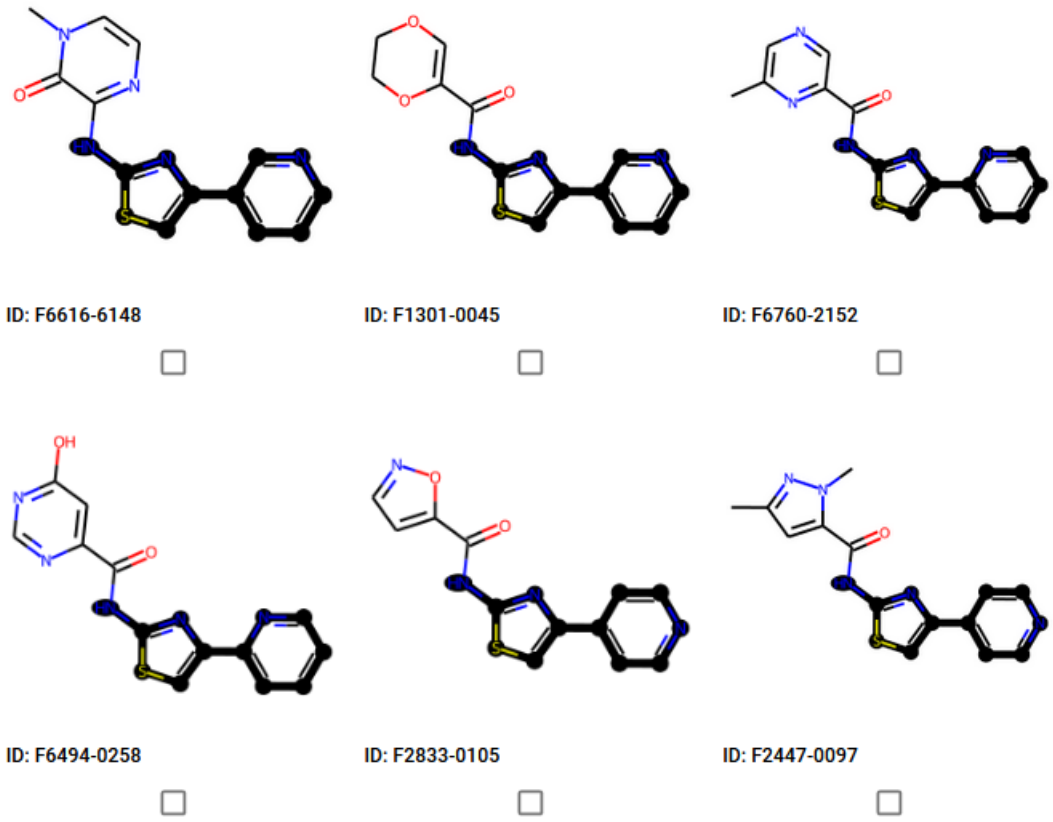


Рис 2.5. Приклад групування сполук за фрагментами в хімічному просторі в програмі CIME

2.4. Підготовка до початку роботи

Для використання та запуску CIME потрібен Docker. Для комфортного користування Docker потрібна *nix система. В даній роботі використовував wsl2 для віртуалізації лінукса на віндовс. Звичайно, можна використовувати і лінукс, але в даному випадку я використав wsl2. Зазвичай я надаю перевагу використунню нативного лінукса, особливо для контеризації. Але в доному випадку, wsl2 більш ніж достатньо.

Я використовував Ubuntu 22.04 LTS. Встановити можна через Microsoft Store. Також для зручності встановив Windows Terminal для комфортного

користування wsl2. Офіційна документація по встановленню wsl2 https://docs.microsoft.com/en-us/windows/wsl/install.

З самого початку потрібно оновити систему. Для цього в терміналі протрібно вписати наступну команду:

$ sudo apt update && sudo apt upgrade –y

Після цього можна встановлювати все саме необхідне. Встановлюємо docker. Офіційна документація по встановленню https://docs.docker.com/engine/install/ubuntu/

Після цього потрібно добавити docker в автозапуск. Для wsl це робиться так, в /etc/wsl.conf потрібно добавити ось такі рядки:

[boot]

command="service docker start"

а також потрібно виконати цю команду для виправлення помилки *nftables used instead of iptables* запуску dockerd:

$ sudo update-alternatives --set iptables /usr/sbin/iptables-legacy

$ sudo dockerd

$ ^c

На linux це робиться ще простіше, командою:

$ sudo systemctl enable docker

$ sudo systemctl start docker

Встановлюємо останню версію CIME:

$ docker pull jkuvdslab/cime

Запускаємо CIME контейнер з додатковими параметрами для покращення роботи (виділяємо контейнеру 4 ядра та 6 Гб ОЗУ)

```
$ sudo docker run -d -p 8080:8080 \
    --cpus=4 \
    --memory=6g \
    --memory-swap=-1 \
```

--name cime \

--detach jkuvdslab/cime

Після цього для того, щоб використовувати CIME потрібно перейти за посиланням http://127.0.0.1:8080/ або http://localhost:8080/

Також встановлюємо Jupyter Notebook для комфортної роботи з RDKit:

$ sudo apt update

$ sudo apt install python3-pip python3-dev

$ sudo -H pip3 install --upgrade pip

$ sudo apt install python3-venv

$ python3 -m venv venv

$ source venv/bin/activate

$ pip install jupyter

$ jupyter notebook

Після цього переходимо за посиланням: http://127.0.0.1:8888/tree

РОЗДІЛ 3. РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ ТА ЇХНЄ ОБГОВОРЕННЯ

3.1. Аналіз датасетів за фільтрами

Після прогону усіх 3-х датасетів із використанням фільтрів та *Tanimoto similarity fingerprint* отримані дані були записані в **таблицю 3.1**. Для більш точних результатів, під час прогону датасетів, солі амінів були видалені та використані чисті аміни. В таблиці можемо побачити, що було проаналізовано 3 датасети: Stock HTS Compounds (550k), General Fragment Library (50k) та Pre Plated Diversity Set PS6 (50k) із використанням фільтрів: Lipinski Rule of 5, Ghose Filter, Veber Filter, REOS Filter. Також я добавив в таблицю дані про те, скільки сполук пройшли всі ці фільтри. Остання колонка вказує на те, скільки сполук із даного датасету пройшли відбір та можуть бути потенційними інгібіторами тубуліну (T – коефіцієнт *Tanimoto*).

Таблиця 3.1.

Результати аналізу датасетів із використанням фільтрів та *Tanimoto similarity fingerprint.*

Датасет	Lipinski Rule of 5	Ghose Filter	Veber Filter	REOS Filter	Passes All Filters	T >= 0.85 < 1
Stock HTS Compounds (550k)	200839	415720	515199	448914	156981	15
General Fragment Library (50k)	40887	18388	50524	42404	13227	15
Pre Plated	24309	44984	50191	48567	21038	0

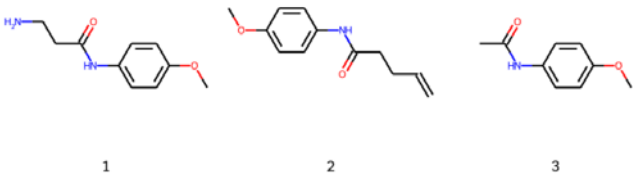
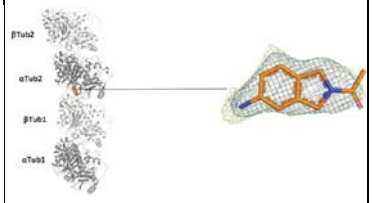
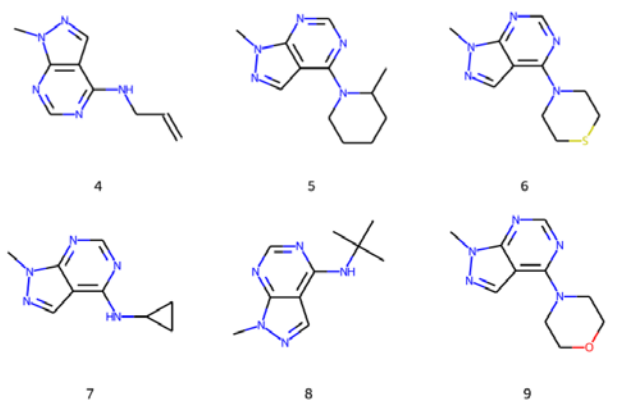
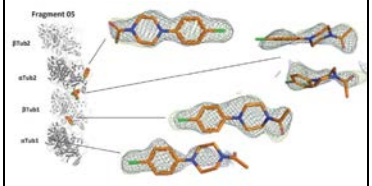
Diversity Set PS6 (50k)						

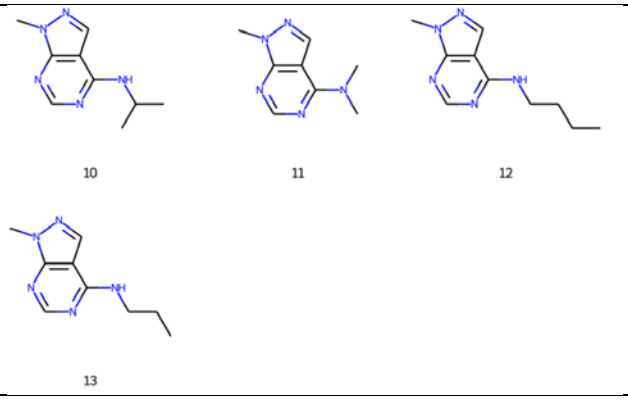
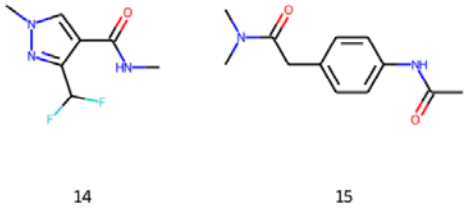
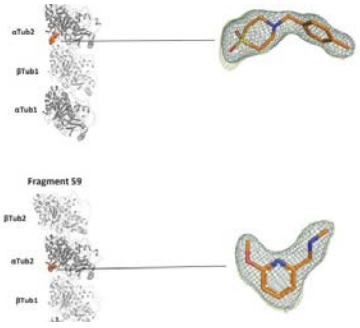
3.2. Сполуки, які пройшли відбір

В **таблиці 3.2** та **тиблиці 3.3** наведена інформація про фрагменти, які пройшли відбір. В обох цих таблицях сполуки відсортовані по сайтах зв'язування та субодициях тубуліну. Колонка з візуалізацією відображає теоретично можливий докінг даного фрагменту з сайтом зв'язування.

Таблиця 3.2.

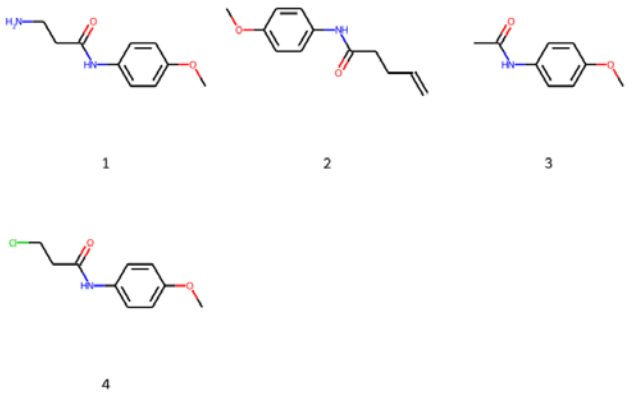
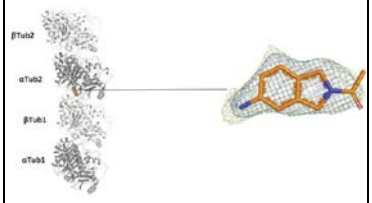
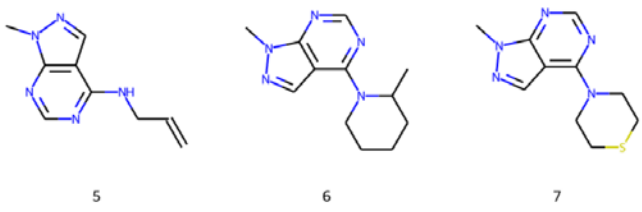
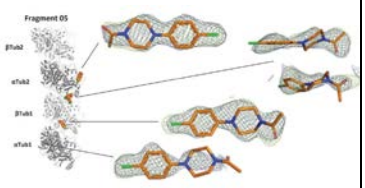
Фрагменти з датасету General Fragment Library (50k), які пройшли відбір.

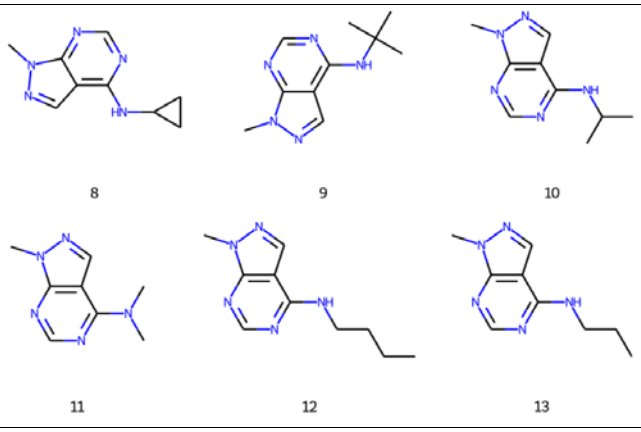
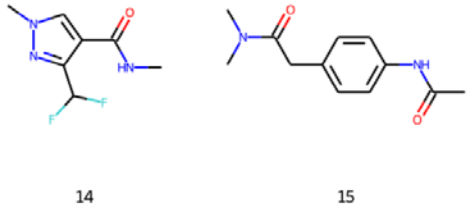
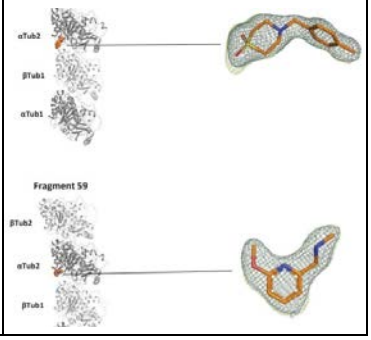
Субодиниця	Сайт	Фрагмент	Візуалізація
α	sID X1		
β	sID βIV		

			
β	sID βαII		

Таблиця 3.3.

Фрагменти з датасету Stock HTS Compounds (550k), які пройшли відбір.

Субодиниця	Сайт	Фрагмент	Візуалізація
α	sID X1		
β	sID βIV		

			
β	sID βαII		

Як ми бачимо, сполуки з датасету Stock HTS Compounds (550k) та General Fragment Library (50k), що пройшли відбір, схожі та відрізняються лише на 1 сполуку для sID X1 сайту зв'язування α- субодиниці тубуліну та 1 сполуку sID βIV сайту зв'язування β- субодиниці тубуліну.

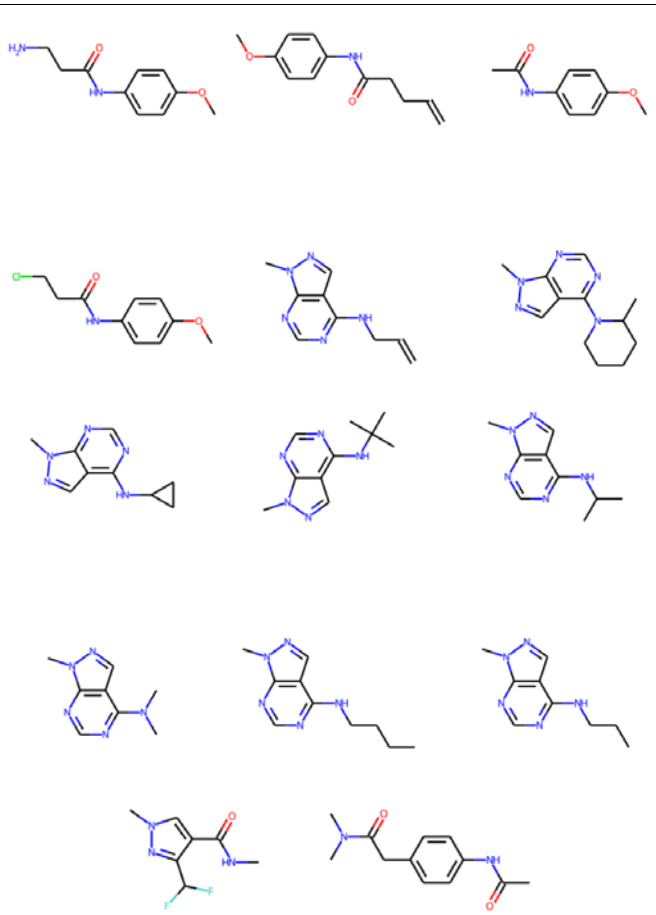
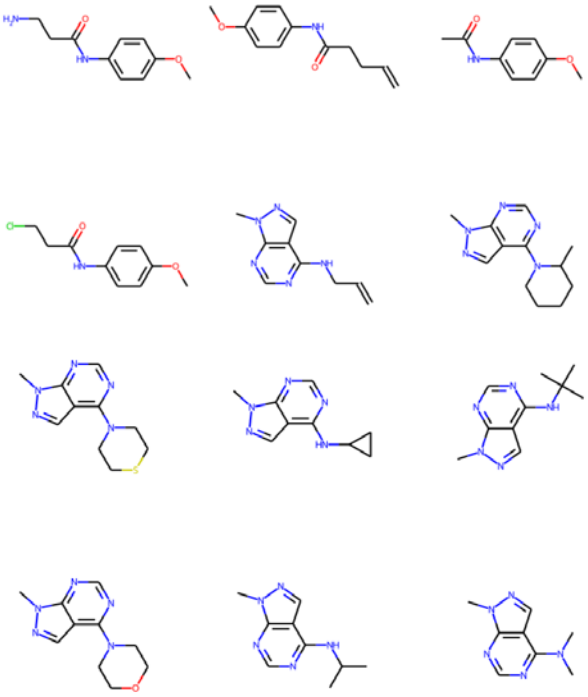
3.3. Аналіз відібраних сполук

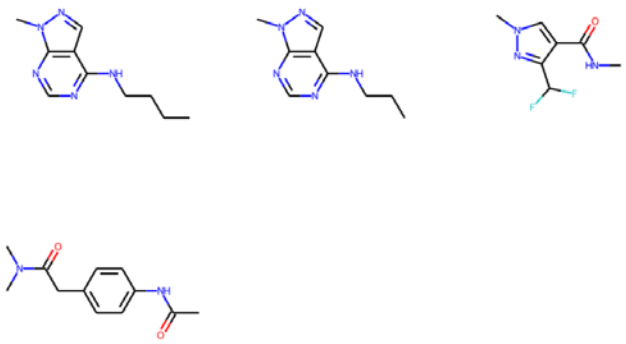
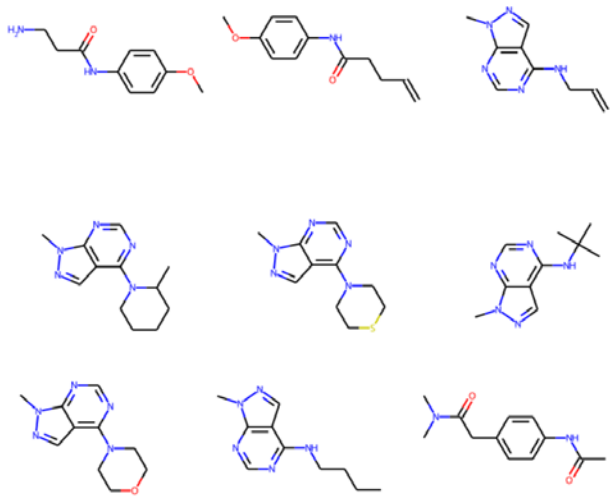
Отримавши перші результати та сполуки, які пройшли відбір, я вирішив проаналізувати і їх також. Подивитися, які з цих фрегментів змогли пройти фільтри. В **таблиці 3.4** я описав результати після прогону 16 сполук отриманих врезультаті прогону датасетів та використання *Tanimoto similarity fingerprint.*

Таблиця 3.4.

Результати фільтрації отриманих сполук вибранеми фільтрами.

Фільтр	Фрагменти	К-ть

Lipinski Rule of 5		14
Veber Filter		16

		
REOS Filter		9

Як ми бачимо, жодна з цих сполук не пройшла фільтрацію Ghose Filter, REOS Filter змогли пройти лише 9 сполук, Lipinski Filter of 5 пройшли 14 сполук, та Veber Filter пройшли усі відібрані сполуки.

3.4. Побудова хімічного простору та аналіз сполук за допомогою програми CIME

Для побудови хімічного був створений новий датасет на основі вже прогнаного датасету General Fragment Library (50k). Новий датасет був модифікований, в нього були добавлені додаткові параметри *Tanimoto coefficient* по кожному із можливих сайтів зв'язування α- та β- тубуліну, а також, результати проходження фільтрів. Після імпортування новоствореного датасету в програму

CIME були вибрані необхідні нам параметри та запущено UMAP оптимізацію **рис. 3.1.**

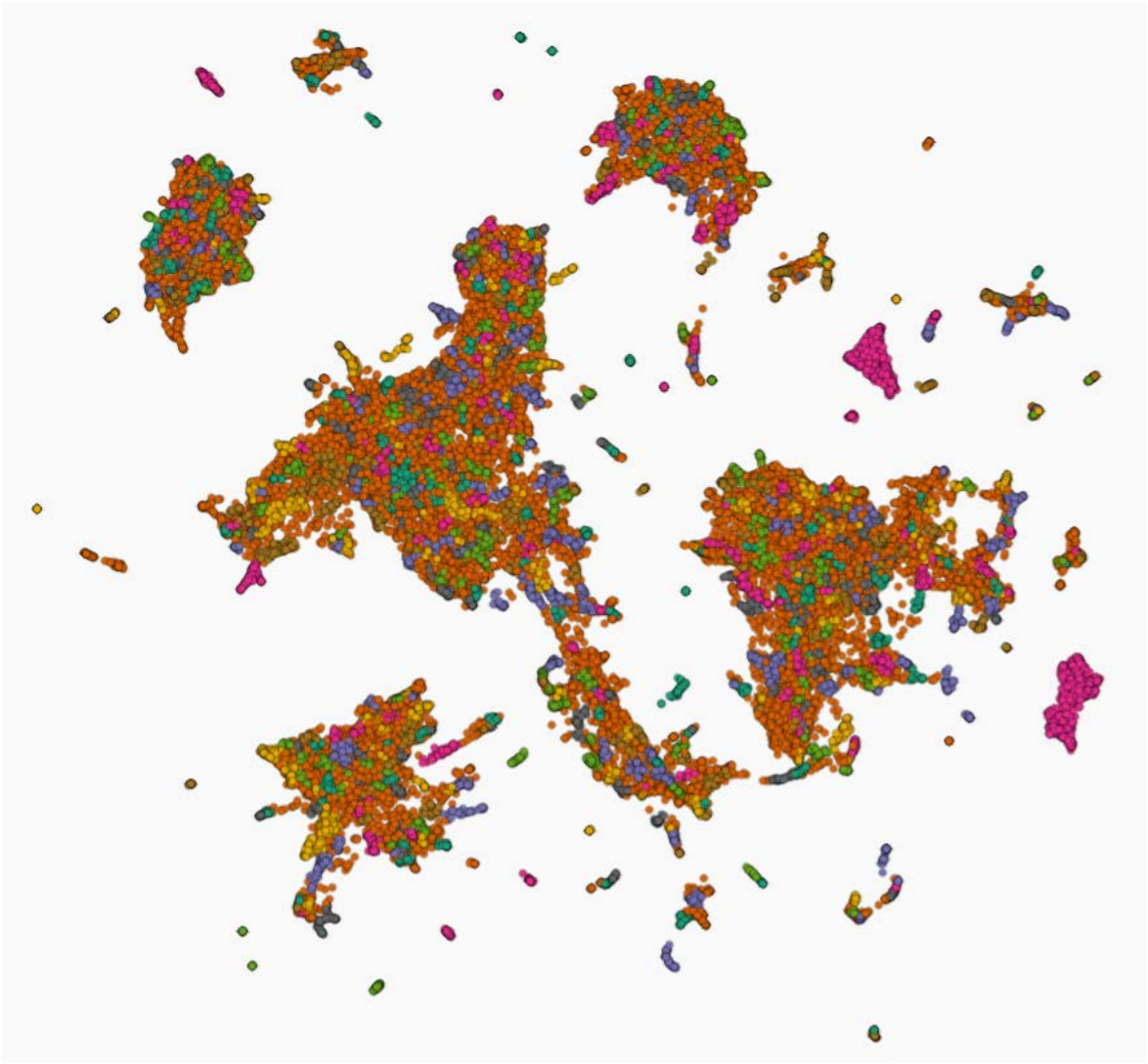


Рис. 3.1. Результат UMAP оптимізації новоствореного датасету за параметри *Tanimoto coefficient* по кожному із можливих сайтів зв'язування α- та β- тубуліну.

Після візуалізації хімічного простору програма надає нам можливість вибрати ділянку із цікавими для нас сполуками **рис. 3.2.** Ми можемо побачити, що ми вибрали самі цікаві для нас сполуки (які пройшли відбір) та бачимо, як саме CIME будує хімічний простір і групує елементи. А саме, вона групує їх за фрагментами (фінгерпринтами).

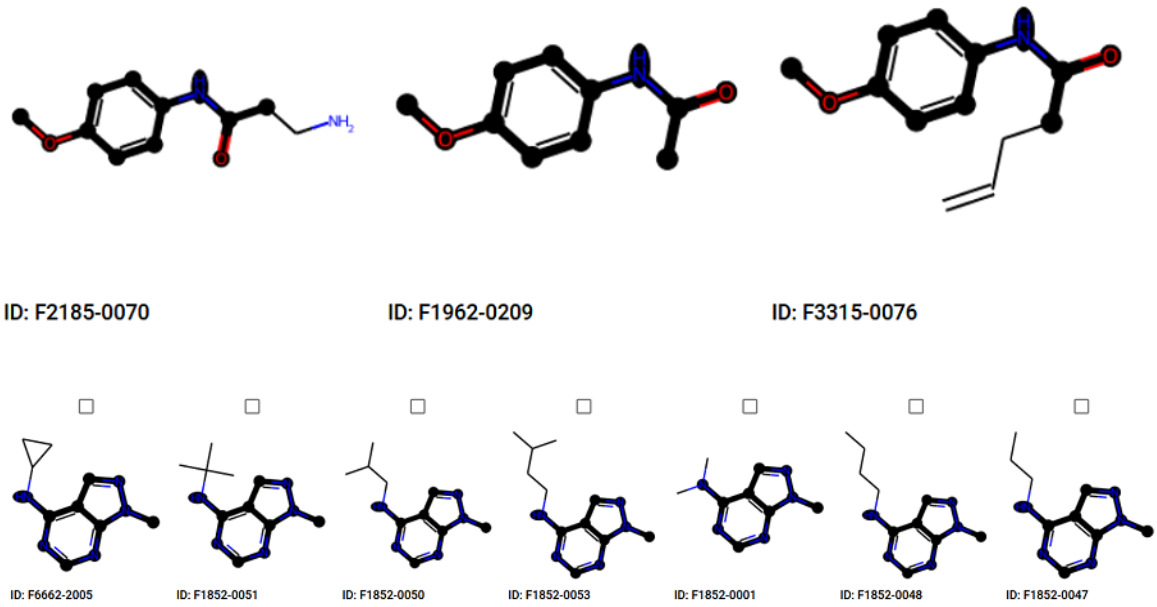


Рис. 3.2. Приклад групування за фрагментами двох різних груп сполук, які пройшли відбір

3.5. Порівняння методів візуалізації хімічного простору в програмі CIME

CIME оптимізує хімічний простір використовуючи оптимізацію UMAP. Змінюючи параметри оптимізації можна отримати різні результати візуалізації хімічного простору, що відкриває нові можливості аналізу та пошуку цікавих нам біологічно-активних сполук. Було побудовано хімічний простір використовуючи стандартні властивості молекул (молекулярна маса, logP, кількість донорів / акцепторів водневого зв'язку, поворотні зв’язки, кількість атомів, tpsa, важкі атоми, кількість кілець) **рис. 3.3**, хімічний простір на основі tanimoto coefficient за сайтам зв’язування **рис 3.4** та хімічний простір на основі вибраних фільтрів **рис 3.5**.

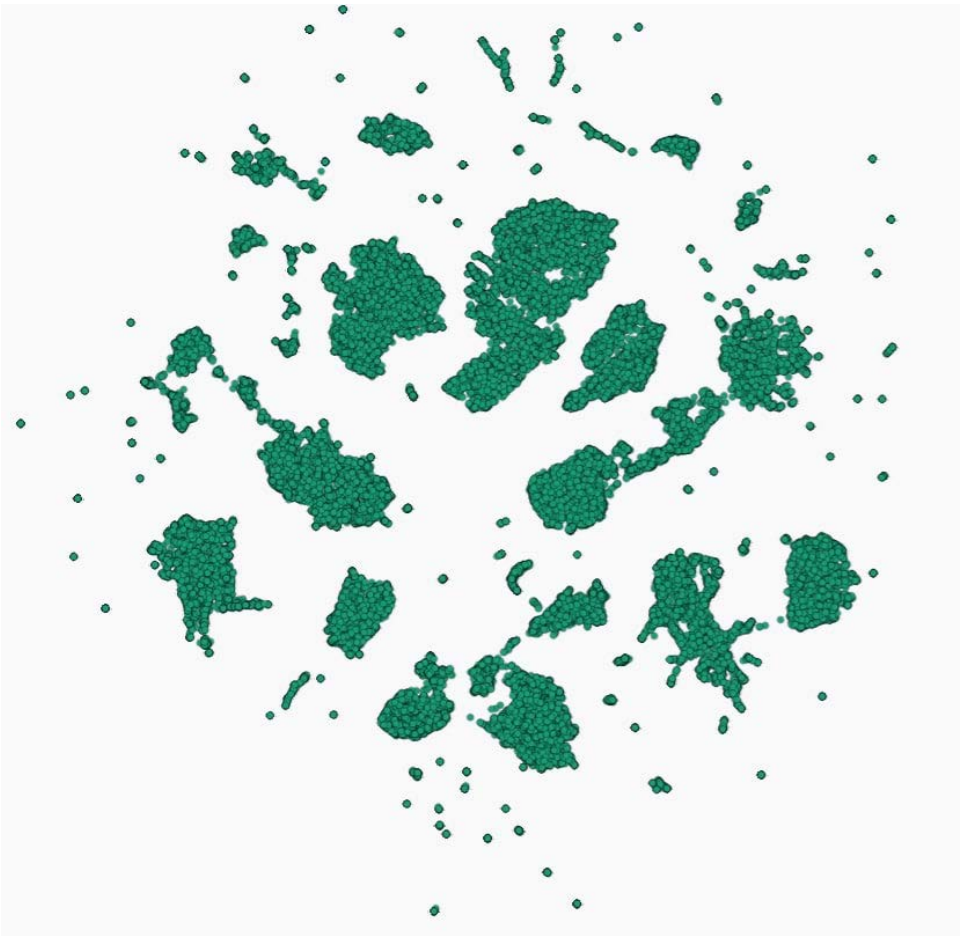


Рис 3.3. Візуалізація хімічного простору за стандартними властивостями молекул

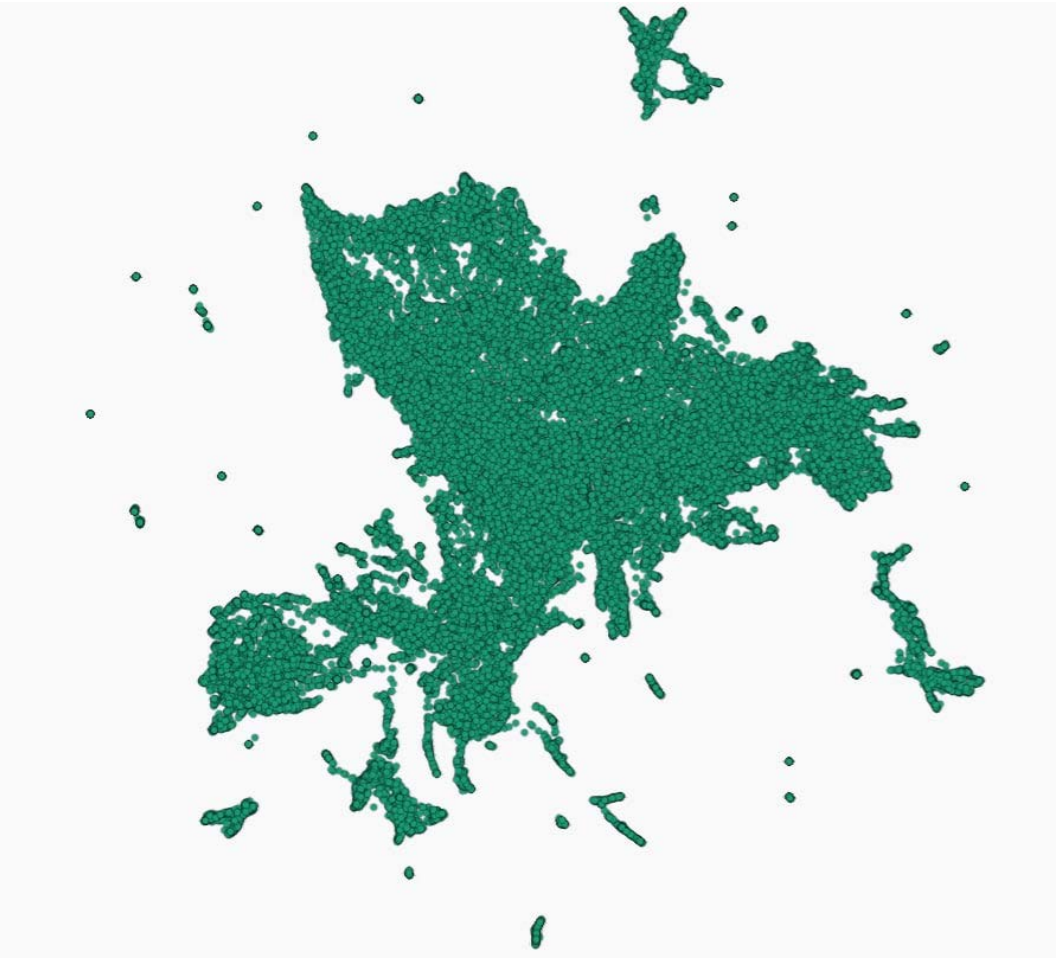


Рис 3.5. Візуалізація хімічного простору за сайтами зв’язування

Тож для пошуку потенційних ефекторів тубуліну була розроблена модель аналізу датасету на основі просторових фінгерпринтів (tanimoto similarity fingerprint). Цей метод побудований на пошуку подібних сполук за схожістю їх структур. В основі методу закладений пошук уже існуючих інгібіторів тубуліну, використовуючи їх як референсні дані для подальшого аналізу та відбору. Для аналізу датасету ми дослідили, що найкращими фільтрами для вібору сполук є Lipinski rule of 5, Veber filter та REOS filter. Ці фільтри дають нам можливість для кращого групування сполук за їх властивостями і більш точного пошуку

потенційних біологічного-активних сполук. Використання даного методу дало нам можливість пошуку відповідних сполук схожих до тих, що уже перевірені, а також додатково кластеризації їх за групами. Це спрощує етап попереднього генерування «кандидатів» та дає змогу побудувати модель відбору потенційних ефекторів тубуліну за потрібними нам параметрами, наприклад, за сайтами зв'язування. Це дає змогу зібрати основний пул параметрів та фільтрів для побудови моделі ML/AI для автоматичної генерації сполук використовуючи morgan fingerprint та створення хімічного простору сполук для пошуку потенційних ефекторів тубуліну.

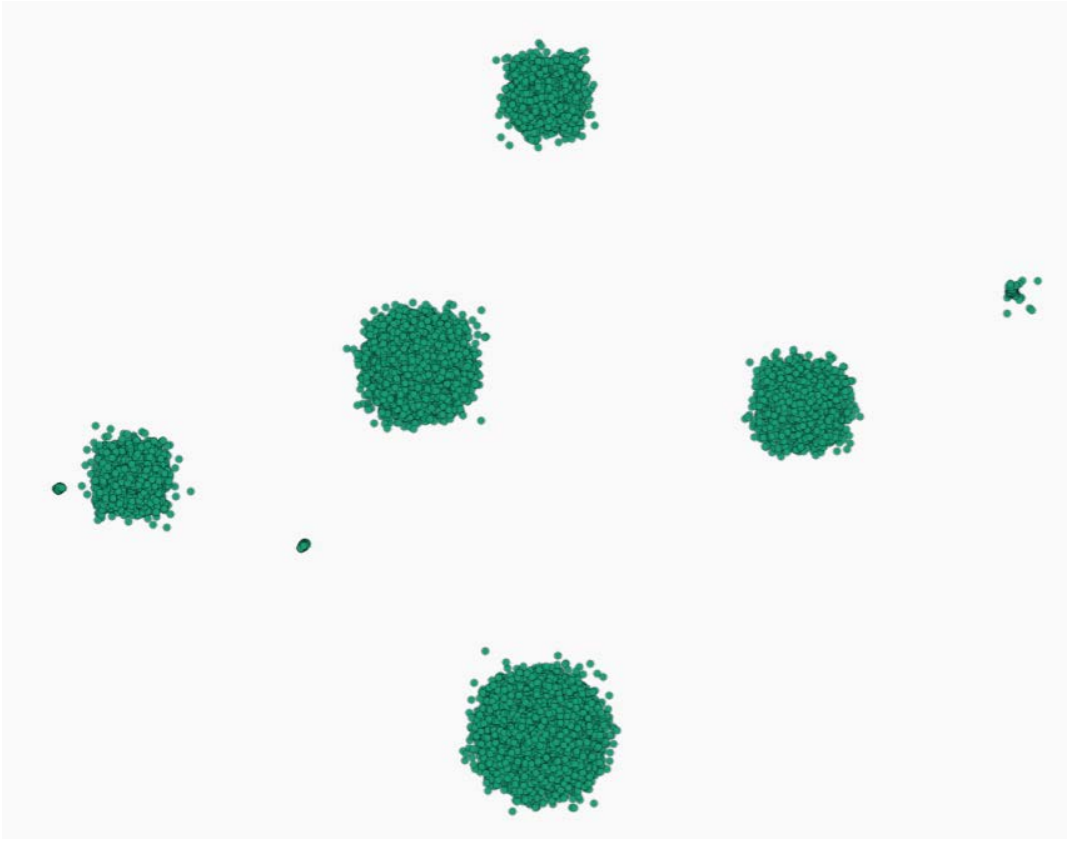


Рис 3.5. Візуалізація хімічного простору за результатами проходження вибраних фільтрів

Використаний метод має перевагу над іншими, коли беремо готову базу даних або хімічний простір, так як ми вже маємо референсні дані. Це зменшує пул сполук доволі сильно. Інша перевага це використання зібраних даних та параметрі для «навчання» моделі ML/AI для більш точної побудови нових сполук з важливим для нас властивостями. А перевага використання саме програми CIME над іншими аналогами (CarcinoPred-EL, ChemGPS-NP та ін) полягає в простоті роботи та у виборі параментрі візуалізації хімічного простору імпортованого датасету. Крім цього, CIME має вбудований функціонал для аналізу вибраної ділянки простору, а також групування кластеризаціїю, що спрощує візуальний пошук наєбхідного нам пулу сполук.

Створення такого датасету та використання його в програмі CIME для побудови хімічного простору на основі стохастичних методів та картах подібностей дало можливість відібрати нові потенційні ефектори тубуліну. Модель та відібрана бібілотека запропонована для подальшого тестування в лабораторію біоінформатики та структурної біології Інституту.

ВИСНОВКИ

1. На основі аналізу існуючих методів хемоінформатичного пошуку сполук з певною біологічною активністю та бібліотек хімічних сполук компанії LifeChemicals було показано, що створення моделі віртуального простору має переваги над існуючими.

2. В результаті аналізу уже відомих ефекторів тубуліну встановлених експериментально було поділено ефектори на основі сайтів їх зв’язування. Це дало нам можливість реалізувати пошук потенційних кандидатів ефекторів тубуліну за конкретними сайтами зв’язування без використання мішеней.

3. Базуючись на тестування бібліотек хімічних сполук за відомими фільтрами було обрано декілька: Veber Filter, Lipinski rule of 5 та REOS, як найбільш кращі параметри моделі для пошуку ефекторів тубуліну. За допомогою створеної моделі було отримано 16 нових сполук для майбутнього докінгу.

4. За результатами побудови хімічного простору в програмі CIME ми дослідили залежність візуалізації хімічного простору від параметрами UMAP групування та різними типами фінгерпринтів. Було встановлено, що набір цих даних відкриває можливість краще навчати AI для більш точної побудови нових ефекторів тубуліну в майбутньому.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. Nature, 432(7019), 855–861. doi: 10.1038/nature03193

2. Learning, H. O. W. M., Data, B. I. G., Helping, A. R. E., & Search, C. (2016). How Machine Learning and Big Data Are Helping Chemists Search the Vast Chemical Universe for Better Medicines. doi: 10.1021/ar500432k

3. Oprea, T. I., & Gottfries, J. (2001). Chemography: The art of navigating in chemical space. Journal of Combinatorial Chemistry, 3(2), 157–166. doi: 10.1021/cc0000388

4. Larsson, J., Gottfries, J., Muresan, S., & Backlund, A. (2007). ChemGPS-NP: Tuned for navigation in biologically relevant chemical space. Journal of Natural Products, 70(5), 789–794. doi: 10.1021/np070002y

5. Osolodkin, D. I., Radchenko, E. V., Orlov, A. A., Voronkov, A. E., Palyulin, V. A., & Zefirov, N. S. (2015). Progress in visual representations of chemical space. Expert Opinion on Drug Discovery, 10(9), 959–973. doi: 10.1517/17460441.2015.1060216

6. Medina-Franco, J. L., & Naveja, J. J. (2017). ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. F1000Research, 6. doi: 10.12688/f1000research.12095.2

7. Lusher, S. J., McGuire, R., Van Schaik, R. C., Nicholson, C. D., & De Vlieg, J. (2014). Data-driven medicinal chemistry in the era of big data. Drug Discovery Today, 19(7), 859–868. doi: 10.1016/j.drudis.2013.12.004

8. Szlezák, N., Evers, M., Wang, J., & Pérez, L. (2014). The role of big data and advanced analytics in drug discovery, development, and commercialization. Clinical Pharmacology and Therapeutics, 95(5), 492–495. doi: 10.1038/clpt.2014.29

9. Medina-Franco, J. L. (2012). Interrogating novel areas of chemical space for drug discovery using chemoinformatics. Drug Development Research, 73(7), 430–438. doi: 10.1002/ddr.21034

10. Polishchuk, P. G., Madzhidov, T. I., & Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. Journal of Computer-Aided Molecular Design, 27(8), 675–679. doi: 10.1007/s10822-013-9672-4

11. Rosén, J., Gottfries, J., Muresan, S., Backlund, A., & Oprea, T. I. (2009). Novel Chemical Space Exploration via Natural Products. Journal of Medicinal Chemistry, 52(7), 1953–1962. doi: 10.1021/jm801514w

12. Dobson, C. M. (2004). Chemical space and biology. Nature, 432(7019), 824–828. doi: doi.org/10.1038/nature03192

13. Ritchie, T. J., Ertl, P., & Lewis, R. (2011). The graphical representation of ADME-related molecule properties for medicinal chemists. *Drug Discovery Today*, *16*(1–2), 65–72. doi: doi.org/10.1016/j.drudis.2010.11.002

14. Oellien, F., Ihlenfeldt, W. D., & Gasteiger, J. (2005). InfVis - Platform-independent visual data mining of multidimensional chemical data sets. *Journal of Chemical Information and Modeling*, *45*(5), 1456–1467. doi: doi.org/10.1021/ci050202k

15. Vlsser, S. A. (1983). Application of Van Krevelen's Graphical-Statistical Method for the Study of Aquatic Humic Material. *Environmental Science and Technology*, *17*(7), 412–417. doi: doi.org/10.1021/es00113a010

16. Wu, Z., Rodgers, R. P., Marshall, A. G., & Fourier, U. (2004). Two- and three-dimensional van krevelen diagrams: a graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance mass measurements. *Wu et al 2004-2 and 3D Van Krevlen Diagrams*. *76*(9), 2511–2516. doi: 10.1021/ac0355449

17. Johnson, T. W., Dress, K. R., & Edwards, M. (2009). Using the Golden Triangle to optimize clearance and oral absorption. *Bioorganic and Medicinal Chemistry Letters*, *19*(19), 5560–5564. doi: doi.org/10.1016/j.bmcl.2009.08.045

18. Reutlinger, M., & Schneider, G. (2012). Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *Journal of Molecular Graphics and Modelling*, *34*, 108–117. doi: 10.1016/j.jmgm.2011.12.006

19. Medina-Franco, J. L., Maggiora, G. M., Giulianotti, M. A., Pinilla, C., & Houghten, R. A. (2007). A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chemical Biology and Drug Design*, *70*(5), 393–412. doi: 10.1111/j.1747-0285.2007.00579.x

20. Medina-Franco, J. L., Yongye, A. B., Pérez-Villanueva, J., Houghten, R. A., & Martínez-Mayorga, K. (2011). Multitarget structure-activity relationships characterized by activity-difference maps and consensus similarity measure. *Journal of Chemical Information and Modeling*, *51*(9), 2427–2439. doi: 10.1021/ci200281v

21. Lounkine, E., Kutchukian, P., Petrone, P., Davies, J. W., & Glick, M. (2012). Chemotography for multi-target SAR analysis in the context of biological pathways. *Bioorganic and Medicinal Chemistry*, *20*(18), 5416–5427. doi: 10.1016/j.bmc.2012.02.034

22. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. doi: 10.1007/BF00337288

23. Schmuker, M., Schwarte, F., Brück, A., Proschak, E., Tanrikulu, Y., Givehchi, A., Scheiffele, K., & Schneider, G. (2007). SOMMER: Self-organising maps for education and research. *Journal of Molecular Modeling*, *13*(1), 225–228. doi: 10.1007/s00894-006-0140-0

24. Bonachera, F., Marcou, G., Kireeva, N., Varnek, A., & Horvath, D. (2012). Using self-organizing maps to accelerate similarity search. *Bioorganic and Medicinal Chemistry*, *20*(18), 5396–5409. doi: 10.1016/j.bmc.2012.04.024

25. Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W., & Beratan, D. N. (2013). Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *Journal of the American Chemical Society*, *135*(19), 7296–7303. doi: 10.1021/ja401184g

26. Reker, D., Perna, A. M., Rodrigues, T., Schneider, P., Reutlinger, M., Mönch, B., Koeberle, A., Lamers, C., Gabler, M., Steinmetz, H., Müller, R., Schubert-Zsilavecz, M., Werz, O., & Schneider, G. (2014). Revealing the macromolecular targets of complex natural products. *Nature Chemistry*, *6*(12), 1072–1078. doi: 10.1038/nchem.2095

27. Palyulin, V. A., Radchenko, E. V., & Zefirov, N. S. (2000). Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds. *Journal of Chemical Information and Computer Sciences*, *40*(3), 659–667. doi: 10.1021/ci980114i

28. Maggiora, G. M., & Bajorath, J. (2014). Chemical space networks: A powerful new paradigm for the description of chemical space. *Journal of Computer-Aided Molecular Design*, *28*(8), 795–802. doi: 10.1007/s10822-014-9760-0

29. Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T. I., Mutzel, P., & Waldmann, H. (2009). Interactive exploration of chemical space with Scaffold Hunter. *Nature Chemical Biology*, *5*(8), 581–583. doi: 10.1038/nchembio.187

30. Agrafiotis, D. K., & Wiener, J. J. M. (2010). Scaffold Explorer: An interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. *Journal of Medicinal Chemistry*, *53*(13), 5002–5011. doi: 10.1021/jm1004495

31. Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *64*(SUPPL.), 4–17. doi: 10.1016/j.addr.2012.09.019

32. Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *Journal of Combinatorial Chemistry*, *1*(1), 55–68. doi: 10.1021/cc9800071

33. Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug

candidates. *Journal of Medicinal Chemistry, 45(12), 2615–2623.* doi: 10.1021/jm020017n

34. Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A “Rule of Three” for fragment-based lead discovery? Drug Discovery Today, 8(19), 876–877. doi:10.1016/S1359-6446(03)02831-9. doi: 10.1016/S1359-6446(03)02831-9

35. Walters, W. P., & Namchuk, M. (2003). Designing screens: How to make your hits a hit. *Nature Reviews Drug Discovery*, *2*(4), 259–266. doi: 10.1038/nrd1063

36. Muegge, I., Heald, S. L., & Brittelli, D. (2001). Simple selection criteria for drug-like chemical matter. *Journal of Medicinal Chemistry*, *44*(12), 1841–1846. doi: 10.1021/jm015507e