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

Завідувач кафедри супрамолекулярної хімії
проф. Рябухін Сергій Вікторович
Протокол №___ засідання кафедри
від «___» _____2023 р.

АНАЛІЗ СУЧАСНИХ БАЗ ДАНИХ ЯК ЗРУЧНОГО ІНСТРУМЕНТУ ДЛЯ ПОШУКУ ІНФОРМАЦІЇ ПРО БІОЛОГІЧНУ АКТИВНІСТЬ ХІМІЧНИХ СПОЛУК

Випускна кваліфікаційна робота магістра
студента спеціальності 102 Хімія
ОП «Хемоінформатика»
Кирсанова Андрія Володимировича

Науковий керівник
Професор кафедри супрамолекулярної хімії
ННЦ Інституту Високих Технологій КНУ імені Тараса Шевченка
д.х.н. **Рябухін Сергій Вікторович**

Оцінка захисту роботи
_____

Київ - 2023р.

АНОТАЦІЯ

Кирсанов А. В. Аналіз сучасних баз даних як зручного інструменту для пошуку інформації про біологічну активність хімічних сполук *Кваліфікаційна робота магістра за спеціальністю хемоінформатика. – Київський національний університет імені Тараса Шевченка, ННЦ Інститут високих технологій, кафедра супрамолекулярної хімії. – Київ, 2023.*

Науковий керівник: доктор хімічних наук Рябухін С. В., Професор кафедри супрамолекулярної хімії ННЦ Інституту Високих Технологій КНУ імені Тараса Шевченка.

У ході роботи було проаналізовано доступні академічні, комерційні та ресурси провідних постачальників на якість і доступність даних про основні характеристики, біологічну активність і терапевтичну дію хімічних сполук. Для аналізу було використано метод опитування за допомогою анкетування 28 людей, серед яких були студенти бакалаври, магістри та аспіранти хімічних спеціальностей.

Серед розроблених критеріїв для анкетування було приділено особливу увагу зручності пошуку, а також повноцінності опису біологічної активності. Отримані дані було проаналізовано статистичними методами, підраховано частоти критеріїв і виявлено основні ключові патерни анотацій біоданих сполук і мішеней у досліджуваних базах даних.

Знайдено, що для ефективного пошуку інформації необхідно наразі користуватись декількома ресурсами одночасно, що не є ефективним і економічно доцільним. Виявлено необхідність створення сучасної бази даних яка відповідала б основним вимогам провідних фахівців у галузях хімії, біології та фармакології.

Ключові слова: бази даних, опитування, біохімічні дані, медико-хімічні дані.

ABSTRACT

Kyrsanov A. V. Analysis of modern databases as a convenient tool for searching for information on the biological activity of chemical compounds.

Qualifying work of the master on a speciality chemoinformatics. – Taras Shevchenko National University of Kyiv, Institute of High Technologies, Department of Supramolecular Chemistry. – Kyiv, 2023.

Research supervisor: Doctor of Chemical Sciences, S. V. Ryabukhin, Professor of the Department of Supramolecular Chemistry of the Institute of High Technologies of Taras Shevchenko KNU.

During the work, the available academic, commercial and resources of leading suppliers were analyzed for the quality and availability of data on the main characteristics, biological activity and therapeutic effect of chemical compounds. The survey method was used for the analysis, with the help of a questionnaire of 28 people, among whom there were bachelor's, master's and postgraduate students of chemical specialties.

Among the developed criteria for the questionnaire, special attention was paid to ease of search, as well as to the completeness of the description of biological activity. The obtained data were analyzed by statistical methods, the frequencies of criteria were calculated and the main key patterns of annotations of biodata of compounds and targets in the studied databases were identified.

It was found that for the effective search of information it is necessary to use several resources at the same time, which is not efficient and economically feasible. The need to create a modern database that would meet the basic requirements of leading specialists in the fields of chemistry, biology and pharmacology was revealed.

Keywords: databases, crowd review, biochemical data, medicinal chemistry data.

Зміст

ВСТУП

При розробці лікарських засобів дослідники покладаються на хімічні бази даних, які є швидким джерелом інформації про велику кількість біохімічних сполук. Однак із наявністю численних ресурсів може бути складно вибрати найбільш підходящий для конкретного питання дослідження, більше того вони можуть бути незручними та складними у використанні. У такому разі постає проблема пошуку та аналізу якості біоданих і, як результат, вибору найкращого ресурсу для пошуку біологічних даних.

Для більш детального аналізу існуючих ресурсів, було вирішено провести порівняння 13 популярних баз даних, серед яких були ресурси постачальників, комерційні та академічні бази. Для отримання інформативної і неупередженої оцінки баз даних був обраний метод опитування із створенням анкети. Для цього була створена вибірка мішеней (84 шт.) та сполук (140 шт.) та розроблені критерії оцінки пошукової системи і якості біоданих. У дослідженні прийняло участь 28 студентів.

РОЗДІЛ 1. ОГЛЯД ЛІТЕРАТУРИ

1.1 Великі дані

В останні роки концепція Великих Даних (Big Data) привертає все більшу увагу в різних сферах, включаючи фармацевтичну промисловість [1]. Оскільки розробка лікарських засобів стає все складнішим процесом, великі дані відіграють вирішальну роль у виявленні потенційних мішеней, виконанні високопродуктивного скринінгу, проведенні віртуального скрінінгу. Це дозволяє вченим хімікам і біологам приймати обґрунтовані рішення.

Біг Дата або Великі Дані – це відносно нове поняття, яке означає велику кількість різних видів даних, та включає аналіз і роботу з ними. Кількість інформації може сягати такого розміру, що ручний перебір даних буде неможливим або скоріше нераціональним. По-перше, з очевидних причин зростає час обробки, а по-друге, якість результату знижується через зниження уваги [2].

Одним з факторів завдяки якому це зростання стало можливим – це збільшення обчислювальних здібностей комп'ютерів разом із їх здешевленням. За допомогою звичайного ПК стало можливо проводити такі колись ресурсомісткі операції як докінг, навчання нейромереж, віртуальний скринінг та інше.

Як наслідок нові методи роботи з Біг Дата виникли, саме IT підходи, що ґрунтуються на знаннях математичної статистики, мов програмування та знанні алгоритмів і структур даних. Сучасний підхід полягає в автоматичному аналізі наборів даних за допомогою програмного коду, який є зрозумілим для людини, а результат виконання відтворюваним, на відміну від ручного підходу (помилки, пов'язані з людським фактором). Як наслідок виникає потреба у ефективній взаємодії людини з машиною. Це досягається за допомогою мов програмування, наприклад, Python або R.

Крім взаємодії з даними, їх ще необхідно якимось чином зберігати, звідси виникають бази даних. З феноменологічного погляду це набір таблиць із набором інформації, що пов'язані між собою якимось чином. Зазвичай, таблиці мають посилання на інші із використанням ідентифікаційних номерів, Id. Таке уявлення дозволяє зберігати інформацію у більш сприйнятливому та зручному для читання форматі. Такий тип БД називається **реляційний**, від англійского слова *relation*, тобто відношення.

1.1.1 Характеристики Big Data

Для більш точного визначення Big Data можна звернутись до характеристик цього поняття. У 2001 році американська IT компанії Gartner визначила концепт Біг Дата за допомогою 3V [3]: Volume, Velocity, Variety. А у 2013 році аналітиком *van Rijmenam* було запропоновано розширити це до 7V [4]: Volume, Velocity, Variety, Value, Veracity, Variability, Visualisation (**Рисунок 1.1**).

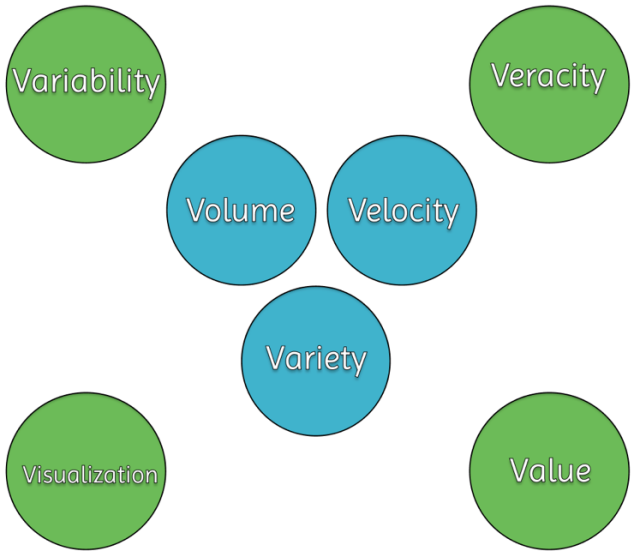

Рисунок 1.1. 7Vs Великих даних

Volume (укр. Об’єм). Цей пункт є ключовим у понятті Біг Дата, тому що відповідає на питання чому саме дані є великими (Big). На даний

момент, як було зазначено вище, обсяги даних обчислюються зеттабітами, де зетта означає 10^{21} порядок (**Рисунок 1.2**), і це число продовжує зростати.

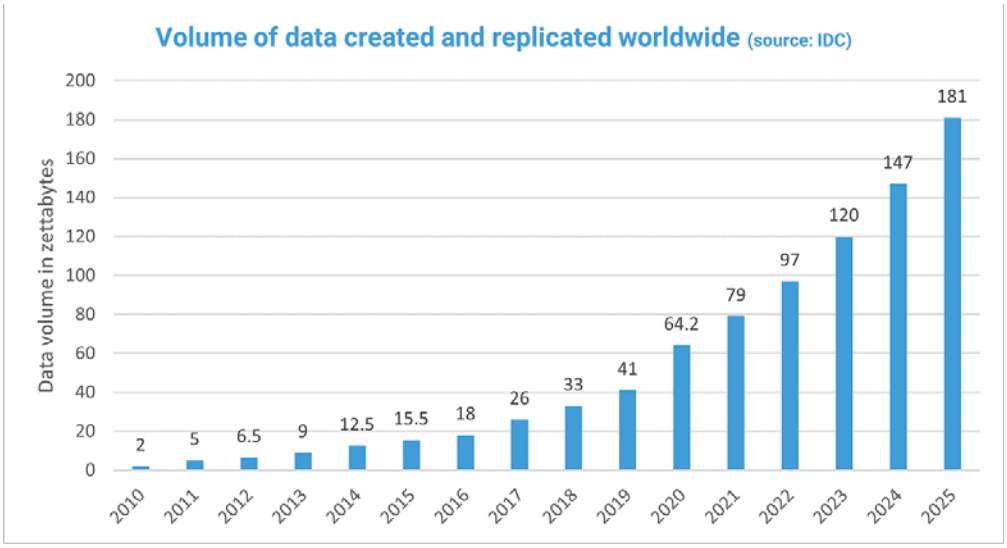

Рисунок 1.2. Обсяг створених даних по рокам

Velocity (укр. Швидкість). Паралельно з обсягом даних продовжує зростати і швидкість інтернету. Грунтуючись на даних Futuretimeline [5], глобальна швидкість Інтернету в 2020 становила 100 Мбіт/с (**Рисунок 1.3**), що було на порядок вище, ніж було в 2010. Наприклад, ґрунтуючись на аналітиці Youtube, наведеній на сайті semrush [6], на момент 2019 року щогодини завантажується близько 30 000 тис. годин відео.

Таким чином це непрямим чином свідчить про те, що дані створюються, обробляються та зберігаються в реальному часі, у той час коли раніше це було неможливим.

Variety (укр. різноманітність). При великій кількості даних вони можуть бути різних видів. По-перше, дані можуть відрізнятися за структурованістю. У структурованих даних будуть присутні зв'язки між собою. Прикладом може бути організація інформації в реляційних базах даних, де записи в таблицях мають взаємозв'язок між таблицями через

ідентифікаційні номери. По-друге, дані можуть бути представлені як у текстовому, так і в медіа форматі (аудіо, відео та зображення).

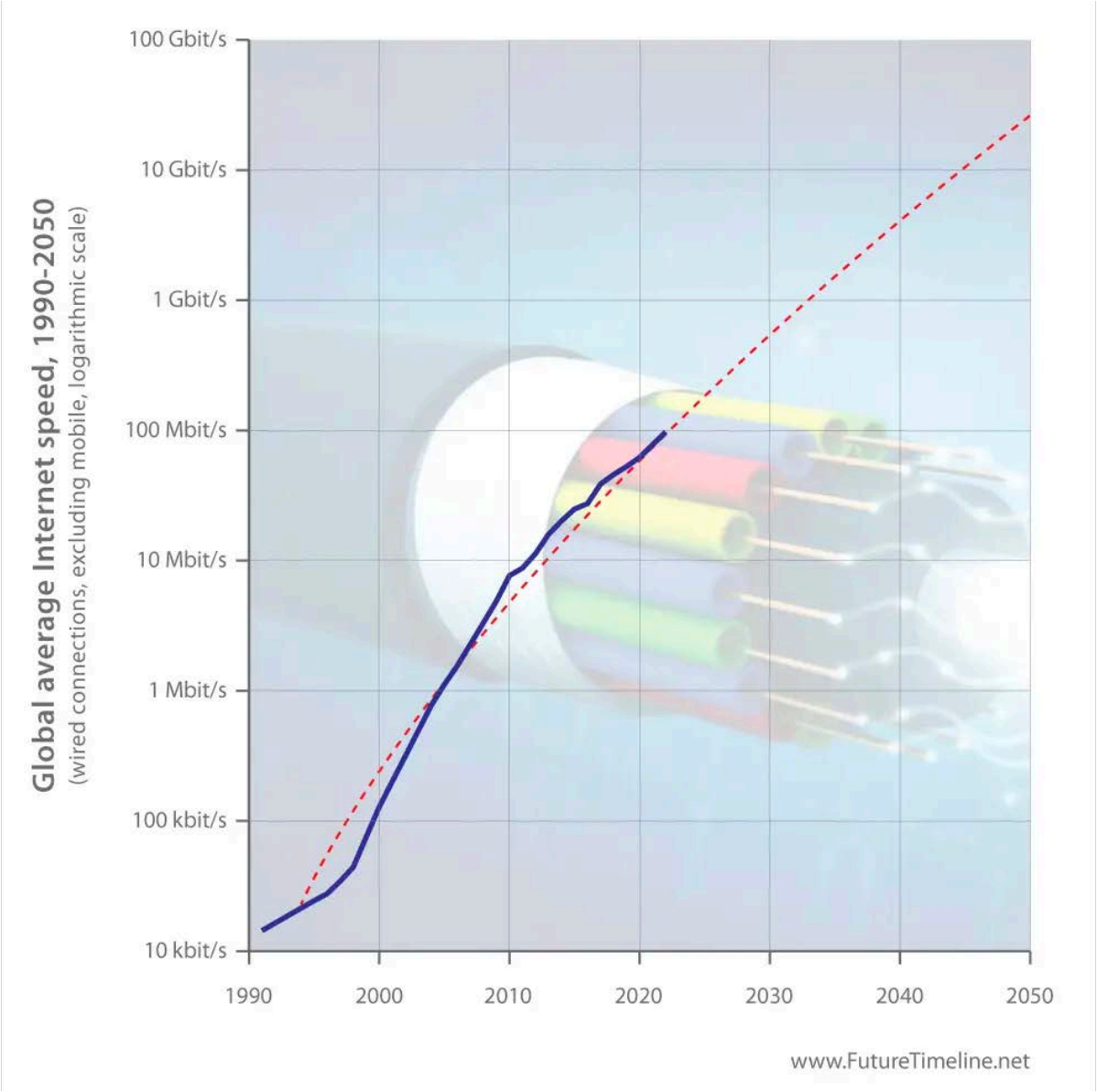

Рисунок 1.3. Швидкість Інтернету по рокам за даними futeretimeline.net

Залежно від типу даних варіюються і методи їх використання та аналізу. Наприклад, текстова інформація може бути оброблена за допомогою методів Nature Language Processing (NLP), а зображення проаналізовані за допомогою згорткових нейронних мереж [7].

Variability (укр. Мінливість). Ця характеристика вказує на те, що одні й самі дані можуть мати зовсім різні значення залежно від контексту. З останнього випливає, що аналіз необхідно проводити "в контексті", а саме із залученням додаткової та уточнюючої інформації. Вочевидь, збільшення обсягу аналізованих даних має у свою чергу негативний вплив на швидкість їх аналізу.

Прикладом аналізу мінливих даних може бути аналіз текстових послідовностей з NLP — word2vec, заснований на skip-gram [8]. Word2vec – це метод векторизації слів, що базується на принципі того, що семантичне значення слово набуває лише в контексті, тобто в оточенні інших слів, що досягається за допомогою алгоритму skip-gram. В результаті після тренування відповідної нейронної мережі виходять точні вектори, що відображають зміст слова і далі можуть бути використані у завданні класифікації текстів.

Варто також відзначити, що мінливість та різноманітність можуть бути легко сплутані один з одним. Для їх розрізнення можна навести приклад, що таке поняття як "червоний" може бути представлене у вигляді написаного слова, зображення або набору звуків - це все відноситься до різноманітності. Мінливість полягає в тому, що слово може бути написане різними шрифтами, зображені можуть бути різні відтінки (наприклад, рубіновий і червоний), а слово вимовлено різною інтонацією.

Veracity (укр. Правдивість). Очевидно, що дані повинні відображати суть явища, яке вони представляють, а також бути якомога точнішими, щоб можна було робити на підставі їх аналізу певні судження. Насправді ситуація така, що будь-яка інформація містить у собі шум, який є її невід'ємною частиною. Як результат, мають існувати методи та інструменти, які дозволяють знайти серед нього цінну інформацію. Одним із таких методів є статистичний аналіз.

Далі необхідно представити отримані результати у зручному форматі, наприклад, як діаграми, що приводить до наступного пункту — візуалізації.

Visualisation (укр. візуалізація). Величезні таблиці даних можуть містити важливу інформацію, проте вона важко сприймається людиною. Тому є сенс у пошуці закономірностей, які можуть бути представлені у вигляді графіків, таблиць, рівнянь регресії тощо.

Value (укр. цінність чи значимість). Це кінцевий результат аналізу Великих Даних, а саме, що ми отримуємо на виході. У контексті бізнесу дані перетворюються на прибуток. Досліджуючи настрої покупців, можна постаратися точніше передбачити їхні інтереси та попит. З погляду наукових досліджень ми отримуємо на виході закономірність або певний результат, який у свою чергу перетворюється на Value. Наприклад, у ході розробки лікарського препарату за підсумком світ отримує не тільки «пігулку», а й знання про біохімію, пов’язану з цими ліками, можливо нові підходи до синтезу.

1.2 Сучасний погляд на розробку лікарських засобів

Велики Дані грають важливу роль у розробці лікарських препаратів оскільки дають можливість аналізувати вже існуючу інформацію про сполуки і протеїни, а також знаходити потенційні зв’язки між ними. Сучасний цикл пошуку лікарських засобів можна коротко описати наступними етапами [9]: вибір хвороби, дослідження патогенезу на молекулярному рівні (пошук терапевтичної мішені), *розробка біоесею на знайдену мішень, скрінінгова компанія*, валідація та оптимізація хітів у ліди, доклінічні та клінічні випробовування. Увесь цикл розробки одного

препарату займає близько 10 років і коштує фармкомпанії близько 5 млрд. доларів за даними журналу Forbes [10].

Не дивлячись на те, що майже кожна стадія розробки вимагає швидкого, а головне ефективного аналізу великої кількості інформації, варто розглянути детальніше ранні етапи, такі як пошук мішені та скрінінг. На цих стадіях можуть бути застосовані методи *in silico,* через те, що вони є дешевими можуть звузити коло пошуку хімічних речовин.

Стадія пошуку мішені зазвичай включає в себе біоінформатичний [9] аналіз, а зокрема обробку геномних послідовностей і аналіз мутацій. Після валідації мішені розробляється біоесей на неї. Біоесей являє собою тест систему, що використовується для вивчення зв'язування протеїну із лігандом. Така система складається із досліджуваного протеїну та допоміжних речовин, завдяки яким можна зареєструвати їх зв'язування. Як правило аналітичним сигналом є або активність у ході радіоактивного розпаду, або флуоресценція. Далі необхідним етапом є калібровка, вона у свою чергу передбачає наявність сполуки, що селективно зв'язується з мішенню та сполуки, що не зв'язується.

Готовий біоесей є основою для проведення високоефективного скрінігу. Як і на стадії його розробки, так і під час скрінігової кампанії необхідно проаналізувати велику кількість інформації про мішень. Це можуть бути власне результати скрінінгу, аналіз відношення структура-активність, а також передбачення активності, розчинності, токсичності сполук методами машинного навчання. Увесь цей процес є трудозатратний і займає багато часу та в деяких випадках може бути невдачним, більше того, дослідник не застрахований від помилок, які можуть бути наявними у базі [11].

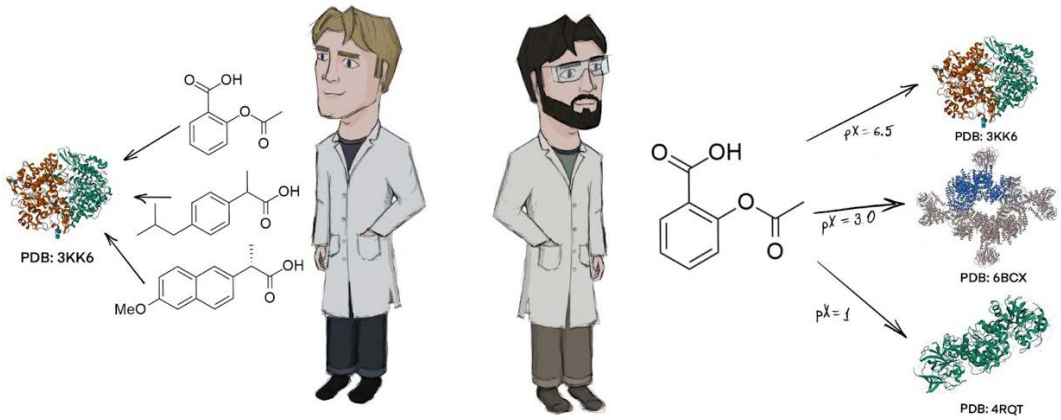

Рисунок 1.4. Погляд науковців до їх досліджуваної матерії. Біолог зліва, хімік справа.

Хіміки грають важливу у процесі розробки лікарських препаратів, оскільки вони відповідальні за дизайн, розробку та вдосконалення методу синтезу і отримання необхідних речовин, що будуть взаємодіяти із специфічною мішенню. Для цього хіміка-вченого в першу чергу хвилює питання: «А із якими протеїнами взаємодія синтезована мною речовина і наскільки ефективно?». Це включає у себе вивчення таких зв’язків структури активності, дослідження стабільності в живому організмі, розчинність і так далі.

З іншого боку, протеїни, приймаючи участь у різних біохімічних шляхах, можуть виступати у якості терапевтичних мішеней, на які треба діяти хімічними сполуками. Через це, вчених біологів цікавить питання: «А які хімічні сполуки взаємодіють із протеїном, з яким я зараз працюю і як ефективно?». Для цього досліджуються структурні особливості білкової молекули, її взаємодія із відомими лігандами, взаємодія з іншими протеїнами тощо.

Разом вони працюють над розробкою біологічно активних молекул, що вимагає їх тісної співпраці, а також знань у хімії, біології та медицини. Щоб ефективно використовувати великі дані для розробки ліків, дослідники покладаються на хімічні бази даних, які надають біохімічну інформацію для досліджуваних сполук. Однак з наявністю великого

різномаїття баз даних може бути складно вибрати найбільш підходящий ресурс для конкретного питання дослідження.

1.3 Огляд порівнянь баз даних

Незважаючи на те, що в літературі існують різні дослідження з приводу порівняння баз даних, на даний момент не було відомостей про порівняння методом опитування. Для початку варто розглянути існуючі публікації, щоб оцінити сучасний стан справ у цій галузі.

У публікації 2001 року [12] порівнювалсь різні хімічні та геномні бази даних шляхом їх вигрузки та перевірки на наявність спільних елементів (дослідження перекриття даних). Варто відмітити, що порівняння включало базу даних від постачальника, а саме каталог Sigma Aldrich. В іншій публікації [13] порівнювались як джерела хімічної інформації, так і ресурси, що містили інформацію про статті. Були також досліджені ресурси PubChem, Crystallography Open Database, PubMed, ZINC, ChemSpider і Google Scholar. Стаття несла описовий характер та не порівнювала їх іншими методами. У 2013 році був проведений аналіз [14] курованих баз даних, які мали відношення до хемогеномних досліджень. Вченими були порівняні ресурси такі як ChEMBL, DrugBank та інші.

У публікації 2016 Bharti [15] були розглянуті комерційні і відкриті хімічні бази даних, зокрема, SciFinder, Reaxys та Web of Science. Вони були досліджені на зручність пошуку інформації, їх інтерфейсу, а також була приведена аналітика. Варто зазначити, що інтерфейс з моменту публікації був змінений.

В області біологічних досліджень був проведений аналіз геномних «баз знань» (knowledge bases) [16]. У цій публікації були розглянуті популярні джерела в області досліджень ракового геному, наприклад, Cancer, OncoKB, Cancer Gene Census та інші. Їх порівняння проводилось біоінформатичними методами, а також було розглянуто перекриття

спільних даних. Як результат було показано, що бази містять багато спільної інформації та для того, щоб зібрати повну картину необхідно використання усіх ресурсів.

1.4 Задача дослідження

Розробка лікарських засобів це складний і різнобічний процес, який включає в себе обробку великої кількості даних з різних ресурсів, наприклад, хімічних і біологічних баз даних, результатів клінічних випробовувань та аналізу наукових публікацій. Ефективний пошук потенційних хітів сильно залежить від якості інформації, а якість скрінінгу від чистоти хімічної речовини. Розмір інформації беззупинно зростає і як наслідок із цим зростає важливість перевірки її якості в базах даних, а також швидкість її використання дослідником. У такому випадку постає питання, чи існує такий ресурс, що містить хімічну та біологічну інформації і дозволяє науковцю придбати ці речовини? Більше того, чи є такий ресурс юзер-френдлі?

Як вже було зазначено вище, у літературі немає згадок про проведення дослідження біохімічних баз даних методом опитування. Тому попередньо необхідно розробити критерії оцінки якості інформації, створити вибірку сполук і мішеней.

РОЗДІЛ 2. МЕТОДИ

У якості методу порівння баз даних був використаний метод опитування. Його розробка шлобально була здійсненна у 2 етапи: створення форми для опитування і створення вибірки сполук та мішеней. Для дослідження були обрані наступні ресурси Reaxys [17], SciFinder [18], DrugBank [19] and ChEMBL [20] як ті, що стосувались хімічної інформації; Guide to Pharmacology [21] and Probes and Drugs [22], як біологічної. А також топ постачальники біологічно активних речовин, такі як MedChemExpress [23], TargetMol [24], SelleckChem [25], Toronto Research Chemicals [26], Tocris [27], Santa Cruz [28] та Cayman [29].

2.1 Структура форми для опитування

Збір даних було вирішено проводити у бінарному форматі, тобто реєструвалась лише наявність певного критерію за допомогою базі GoogleForms [30]. Відсутність відміченої відповіді означала відсутність цього критерію. Обрана відповідь «Не знайдено» (див. нижче) означала повну відсутність інформації для сполуки або мішені.

Перша сторінка була ознайомчою і слугувала вхідною точкою опитування, де учасник вводив номер своєї вибірки. Далі слідували секції пошуку та перевірки якості біоданих, а на завершення була проведена оцінка усіх баз разом. Розділи пошуку та оцінки даних були поділені окремо на підрозділи по сполукам і мішеням.

Підсекції сполук і мішеней у свою чергу складалися із 13 таблиць, кожна із яких відповідала досліджуваному ресурсу. У рядках таблиці були розташовані пронумеровані сполуки (5 шт.) і мішені (3 шт.) відповідно до їх секції. У стовпчиках розташовувались питання (критерії), що перечислені нижче, і разом з цим контрольне питання «Не знайдено» для відслідковування відсутності даних про сполуку/мішень. Кожне завдання і

критерій мали детальний опис та містили приклади для кращого розуміння. Назви таблиць відповідали назвам ресурсів, були клікабельними та містили посилання на базу даних. Усі ресурси були обов'язковими для заповнення, окрім Reaxys та SciFinder, через те, що не всі учасники мали до них доступ.

У розділі пошуку учасник мав галочками обрати ті питання за якими йому вдалось знайти сполуку або мішень. Якщо ні один із перелічених пошукових запитів не давав позитивного результату, то учасник мав поставити «Не знайдено». Сполука або мішень вважалися знайденими лише у тому випадку, коли пошук тривав не більше 10 хвилин і бажана сторінка була у топ 10 результатів. На початку учасник отримував таблицю в pdf форматі із номером своєї вибірки, список молекул та мішеней. Кожна сполука та мішень містили вже готові пошукові запити, які необхідно було скопіювати і провести за ними пошук. Структурний пошук здійснювався копіюванням наведеного в таблиці SMILES у відповідно форму для структурного пошуку. У разі якщо її не було або учасник її не знаходив, він мав скопіювати SMILES у поле для звичайного пошуку.

Розділ біоданих включав питання про наявність певної інформації на сторінці. У разі її знаходження на сторінці учасник мав її помітити галочкою. Заключна секція складалась із таблиць де рядки були базами даних, а критерії оцінки по стовпчикам. Розділ містив таблиці із оцінкою позитивного і негативного. Разом із цим була форма зворотнього зв'язку, де учасник міг поділитися досвідом користування ресурсом.

Після заповнення учасником відповідей, робота була перевірена на наявність фальсифікацій. Для цього випадковим чином обиралась мішень та сполука та перевірялись досвідченою людиною.

2.2 Критерії оцінки

На початку створення опитування були розроблені критерії, які стосувались пошуку, якості біологічних даних та суб'єктивної оцінки учасником його роботи з ресурсом.

2.2.1 Пошукова система

Для пошуку сполук були використані наступні критерії (пошукові запити): синонім 1, синонім 2, синонім 3, CAS номер і SMILES для структурного пошуку. Синонім 1 – це звичайна назва сполуки, тобто канонічна або інша популярна назва, у випадку лікарського засобу його назва або, у випадку погано анотованої сполуки, одна з назв. Синонім 2 – назва IUPAC, згенерована за допомогою ChemAxon. Синонім 3 – синонімічна назва сполуки, яка може бути іншим варіантом написання назви (з або без «е» на кінці); для лікарських засобів це може бути назва дженеріку; кодова назва сполуки під час її (до-)клінічних випробовувань. У випадку погано анотованих сполук Синонім 3 міг бути відсутнім.

Для терапевтичних мішеней були використані наступні пошукові запити: синонім 1, синонім 2, синонім 3 та ідентифікатор UniprotID з популярної бази даних протеїнів Uniprot [31]. Синонім 1 – це конвенційна назва протеїну у довгій формі. Синонім 2 – «approved symbol» із геномної бази даних HGNC [32], яка відповідала назві гену, що кодує даний протеїн. Синонім 3 – інша назва протеїну.

2.2.2 Анотація біоданих

Щоб оцінити якість анотації сполук, в анкеті були використані питання, що стосувались фармакодинаміки (терапевтична мішень, дія на мішень, ефект сполуки, коефіцієнт активності, захворювання),

фармакокінетики (метаболізм, токсичність), скринінгу (*in vivo*, біоесей) та інші (посилання, інша форма).

Терапевтична мішень (Target) – назва мішені проти якої досліджувана сполука проявляє активність. Дія на мішень (Action on target) – вказаний тип активності сполуки на мішень, наприклад, агоніст, інгібітор, блокатор каналу тощо. Ефект сполуки (Substance effect) – вказівка на терапевтичний ефект сполуки, наприклад, протипухлинний препарат, протизапальний тощо. Коефіцієнт активності (Activity coefficient) – будь-який виміряний кількісний параметр, наприклад, константа дисоціації або IC_{50}. Захворювання (Disease) – назва хвороби або стану проти якого дана речовина використовується. Метаболізм (Metabolism) – будь-які дані про метаболізм сполуки, це могли бути реакції метаболізму в організмі, інформація про цитохроми, що оксинюють сполуку тощо. Токсичність (Toxicology) – будь-яка інформація про токсичність, наприклад, LD_{50} або побічні ефекти. In vivo – дані про те, що дана сполука проходила випробовування у живих організмах. Біоесей (Assay details) – будь-який опис процедури біоесею.

Оцінка якості даних по терапевтичним мішеням була здійснена за допомогою наступних параметрів: опис мішені, список лігандів, класифікаційне дерево. Опис мішені (Target description) – будь-яка загальна текстова інформація про протеїн. Список лігандів (Ligand list) – сторінка мала містити список сполук які проявляють активність відносно мішені. Класифікаційне дерево (Classification tree) – інформація про класифікацію даного протеїну, наприклад, до якого сімейства він належить.

2.2.3 Оцінка баз даних

Суб’єктивна оцінка проводилась за рахунок трьох критеріїв: якість представленої інформації, якість пошукової системи, привабливість

інтерфейсу та складний результат пошуку. Якість представленої інформації (Interface representation) – пункт оцінював власне якість та зручність представлення інформації. Пошукова система (Search system) – чи стикався учасник із проблемою довгого пошуку. Привабливість інтерфейсу (Visually attractive interface) – чи подобався інтерфейс сайту. Складний результат пошуку (Complicated search outcome) – чи стикався учасник із проблемою того, що в результаті пошуку виникало багато погано анотованих форм сполук (солей, гідратів і т. д.). В якості негативного контролю був присутній пункт «нічого особливого».

Перші два питання були представлені у першій та другій таблицях. Складний результат пошуку – у таблиці із негативною оцінкою, а Привабливість інтерфейсу – позитивною оцінкою

Учасники також мали змогу залишити текстовий коментар з приводу того, що не було включено у вишезгаданих питаннях.

2.3 Створення вибірки

2.3.1 Розрахунок виборок

Приблизна кількість сполук була визначена за формулою (1), що використовується для розрахунку величини довірчого інтервалу генеральної сукупності нескінченного розміру. Формула (2) була застосована для обрахунку кількості мішеней і є коригованим варіантом формули (1) із поправкою на скінчений розмір.

$$\Delta = Z\sqrt{\frac{p(1-p)}{n}} \quad (1)$$

$$\Delta_f = \Delta\sqrt{\frac{N-n}{N-1}} \quad (2)$$

Δ – довірчий інтервал для нескінченої генеральної сукупності

Δ_f – довірчий інтервал для скінченої генеральної сукупності

Z – z-критерій (z-score)

p – вірогідність зустріти бажаний критерій

n – розрахований розмір вибірки

2.3.2 Створення вибірки

У вибірку сполук, яка складалась із 140 молекул, увійшли такі категорії молекул як

- препарати затверджені ФДА за останні 5 років (за даними DrugBank)
- сполуки, що проходять клінічні випробовування на стадіях 1-3, включаючи й ФДА препарати, які зараз знаходяться на Drug repurposing (перепрофілювання ліків)
- Вилучені з ринку препарати (withdrawn) через їх токсичність або за інших причин (за даними DrugBank)
- Хімічні зонди [33]
- Сполуки-ліди на пізніх стадіях виробництва

Вибірка мішеней була зібрана із основних терапевтичних мішеней для затверджених ФДА ліків у 2019-2021 роках. Протеїни були позначені наступним чином: рецептор (receptor), фермент (enzyme), рецептор, зв’язаний із g-білком (GPCR), йонний канал (ionic channel), транспортер (transporter), інший (other). Усього було обрано 84 мішені

2.3.3 Учасники

У дослідженні приймали участь респонденти із наступних категорій: 14 студентів Бакалаврів, що прослухали курс «Інформаційні технології в хімії», 7 студентів магістратури, які відвідали курс «Бібліотеки хімічних

сполук для біологічного скринінгу», 7 аспірантів спеціальності «Органічна хімія» КНУ імені Тараса Шевченка, ІОХ НАНУ. Перевірка здійснювалась групою досвічених науковців КНУ, ІОХ та ТОВ «Єнамін». Загальна кількість респондентів 28 осіб. Опитування проводилось онлайн з жовтня по грудень 2022 року

2.4 Обробка результатів

Первинні результати опитування були отримані у вигляді таблиці Excel у Google Sheets [34] і далі були переведені у більш зручні форму для роботи як результат була створена база даних із відповідями.

У результуючій базі даних відповідей були наступні колонки: Id – ідентифікаційний номер сполуки або мішені; section – назва секції у формі для опитування; instance_type – тип даного рядку (сполука, мішень або оцінювання БД); instance_category – категорія до якої належить тип даного рядку; participant – код вибірки учасника; database – назва бази даних; answer – назва питання (див. вище), яке було відмічено; isgood – критерій, що вказує чи були помилки при заповненні таблиць учасником, у разі наявності помилок стоїть «N» і така сполука виключалась із аналізу; database_type – тип бази даних (комерційна, академічна або постачальник).

Перед цим була здійснена ручна перевірка на фальсифікацію робіт, а також було проведено відсіювання помилкових позицій. Перевірені були наступні помилки: помилка відповідності, помилка, що стосувалась синоніму 3 та помилка пов’язана із «Не знайдено». Відповідність полягала у тому, що сполуки або мішені, знайдені у секції пошуку, мали містити хоч якусь інформацію у розділі присвяченому біоданим. Як зазначалось вище, деякі сполуки не містили синоніму 3, через погану анотованість в літературі. Тому була зроблена додаткова перевірка того, щоб поля пошуку сполук лише із синонімами 1 та 2 не містили заповнені поля

синонім 3. Разом із цим було перевірено. щоб при виборі «Не знайдено» не було заповнено інших позицій у цьому рядку.

Аналіз популярних патернів анотації біоданих у базах даних здійснювався наступним чином. З відповідей під кожну сполуку у певній базі даних були сформовані усі можливі комбінації розміром від 4 до 7 позицій та від 3 до 5 для мішеней. Для речовин Reference був виключений. Далі були підраховані частоти сформованих комбінацій та відсортовані по спаданню.

РОЗДІЛ 3. РЕЗУЛЬТАТИ ТА ЇХ ОБГОВОРЕННЯ

3.1 Форма для опитування

Для порівняння баз даних було обрано 13 ресурсів, а саме Reaxys, SciFinder, DrugBank і ChEMBL як джерела, пов’язані із хімією; Guide to Pharmacology і Probes and Drugs, як пов’язані із біологією. Крім того, ми обрали провідних комерційних гравців у галузі біоданих, зокрема MedChemExpress, TargetMol, SelleckChem, Toronto Research Chemicals, Tocris, Santa Cruz і Cayman як джерела біохімічної інформації, отриманої від постачальників.

Питання для кожного розділу обирались із власного досвіду роботи з біологічною інформацією, а також орієнтуючись на ту анотацію, що представлена на популярих сервісах, наприклад, Reaxy або ChEMBL. На базі цих питань була розроблена форма для опитування, яке було створене у Google Forms і проходило онлайн з жовтня по грудень 2022 року. Його структура зображена на **Рисунок 3.1**. На початку учасник знайомився із форматом опитування і отримував короткий інструктаж як заповнювати форму та обирав номер своєї вибірки. На схемі це показано розгалудженям.

Далі учасники переходили на секцію пошуку де їм пропонувалось провести пошук кожної сполуки із індивідуальної вибірки за трьома синонімами, CAS номером і структурою. Пошук мішеней здійснювався також за трьома синонімами та UniprotID. Вхідні дані були розіслані кожному учаснику для відтворюваності експерименту. Наступним етапом був пошук наявності певних біоданих для сполуки або мішені. І в завершення після ознайомлення із кожним ресурсом учасники мали змогу дати суб’єктивну оцінки роботи з кожним ресурсом в останній секції. Також у кінці була форма зворотнього зв’язку для того, щоб залишити

власну думку про свій досвід, яка була поза межами представлених критеріїв.

Рисунок 3.1. Структура форми для опитування

На **Рисунок 3.2** наведений зразок таблиці для заповнення. Назва таблиці відповідала назві досліджуваного ресурсу. Вона була клікабельною і містила посилання на головну сторінку бази даних для запобігання непорозумінь. Рядки були пронумеровані і відповідали сполукам 1-5 з розданої кожному вибірки, а стовпчики були пошуковими запитами. У разі знаходження сполуки за певним запитом учасник відмічав це за допомогою галочки.

DrugBank *

	Synonym 1	Synonym 2	Synonym 3	CAS number	Structure search	Not found
Compound 1	☑	☑	☑	☑	☑	☐
Compound 2	☑	☑	☑	☑	☑	☐
Compound 3	☐	☐	☐	☐	☐	☑
Compound 4	☑	☑	☐	☑	☑	☐
Compound 5	☑	☑	☑	☑	☑	☐

Рисунок 3.2. Зразок таблиці на прикладі розділу пошуку сполук.

3.2 Вибірка

Паралельно із цим була створена вибірка сполук і мішеней для статистично вірного порівняння баз даних (див. розділ 2.3). Розмір вибірки сполук був розрахований за формулою 1 (див. розділ 2.3.1) кількість мішеней була встановлена за формулою 2, приймаючи до уваги, що кількість відомих на даний момент мішеней є приблизно 3000 [35]. При розрахунках вважалося, що величина інтервалу мала бути не більше 10%, а Z = 1.96 (95% рівень надійності). Істинне значення p було невідоме, тому прийнявши в гіршому випадку p = 0.5 (рівноймовірно зустріти бажаний критерій або не зустріти його) були розраховані розміри виборок сполук. Кількість молекул мала бути не менше 97, а мішеней не менше 95.

Вибірку молекул було вирішено сформувати за медхімічним спрямуванням, тому що досліджувані ресурси містили переважно інформацію про біологічно активні сполуки. До неї входили такі категорії речовин: затверджені FDA препарати (останні 5 років); сполуки, що

проходять клінічні випробовування на 1-3 стадії [36]; вилучені з ринку препарати (статус withdrawn); хімічні зонди (Probes) та сполуки ліди на пізніх стадіях виробництва. У якості мішеней були обрані протеїни, на які основним чином діяли затверджені FDA препарати у 2019–2021 роках. Вони були представлені в наступних категоріях: ферменти, транспортери, рецептори, шо зв'язані з g-білком (GPCR), йонні канали та інші. Розподіли зображені на **Рисунок 3.3**.

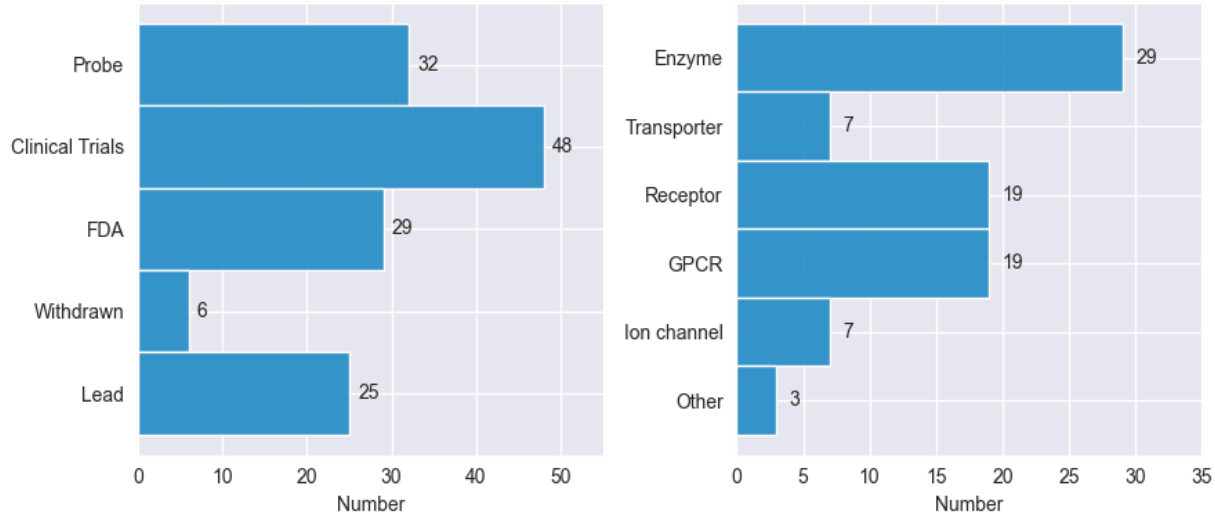

Рисунок 3.3. Розподіли сполук (зліва) і мішеней (справа) по категоріям. По осі у вказані назви категорій, по осі х – їх кількість

Далі створена вибірка сполук була анотована такими пошуковими запитами як синоніми 1-3, CAS номер і SMILES для структурного пошуку. Для мішеней це були синонім 1-3 та UniprotID. Усі дані були взяті із відкритих джерел (SciFinder, DrugBank, Uniprot, HGNC). Готова вибірка була випадковим чином розподілена між учасниками, таким чином, щоб кожний респондент отримав набір п'яти сполук (не менше ніж з трьох різних категорій) та трьох випадкових мішеней. Таким мінімізувалися упередження пов'язані із роботою з одним типом сполук. Наприклад,

ФДА затверджені ліки, як правило, досліджені краще, а тому їх легше знайти і вони краще анотовані.

На наступному етапі інформація була об'єднана у таблиці (**Рисунок 3.4**), що імітували таблицю пошуку у гугл формі (**Рисунок 3.2**) і розіслані кожному учаснику. Зверху був зазначений (Sample barcode) номер індивідуальної вибірки сполук та мішеней. На місці червоного прямокутника було вказано ім'я учасника. На перетині знаходились пошукові запити для кожної сполуки, які респонденти мали скопіювати і вставити у пошукову строку досліджуваного ресурсу. Аналогічна таблиця була створена для мішеней.

SAMPLE BARCODE: 402					
#	Synonym 1	Synonym 2	Synonym 3	CAS	SMILES
Compound 1	Tacrolimus	(1R,9S,12S,13R,14S,17R,21S,23S,24R,25S,27R)-1,14-dihydroxy-12-{1-[(1R,3R,4R)-4-hydroxy-3-methoxycyclohexyl]prop-1-en-2-yl}-23,25-dimethoxy-13,19,21,27-tetramethyl-17-(prop-2-en-1-yl)-11,28-dioxa-4-azatricyclo[22.3.1.0⁴,⁹]octacos-18-ene-2,3,10,16-tetrone	FK-506	104987-11-3	CO[C@@H]1C[C@@H](CC[C@H]1O)C=C(C)[C@H]2OC(=O)[C@@H]3CCCCN3C(=O)C(=O)[C@]4(O)O[C@@H]([C@H](C[C@H]4C)OC)[C@H](C[C@@H](C)CC(=C[C@@H](CC=C)C(=O)C[C@H](O)[C@H]2C)C)OC
Compound 2	Lanicemine	(1S)-1-phenyl-2-(pyridin-2-yl)ethan-1-amine	AZD-6765	153322-05-5	N[C@@H](CC=1C=CC=CN1)C=2C=CC=CC2
Compound 3	Benactyzine	2-(diethylamino)ethyl 2-hydroxy-2,2-diphenylacetate	Benactyzin	302-40-9	CCN(CC)CCOC(=O)C(O)(C=1C=CC=CC1)C=2C=CC=CC2
Compound 4	Epothilone B	(1S,3S,7S,10R,11S,12S,16R)-7,11-dihydroxy-8,8,10,12,16-pentamethyl-3-[(1E)-1-(2-methyl-1,3-thiazol-4-yl)prop-1-en-2-yl]-4,17-dioxabicyclo[14.1.0]heptadecane-5,9-dione	EPO 906	152044-54-7	C[C@H]1CCC[C@@]2(C)O[C@H]2C[C@H](OC(=O)C[C@H](O)C(C)(C)C(=O)[C@H](C)[C@H]1O)/C(=C/C3=CSC(C)=N3)/C
Compound 5	Prednisone	(1R,3aS,3bS,9aR,9bS,11aS)-1-hydroxy-1-(2-hydroxyacetyl)-9a,11a-dimethyl-1H,2H,3H,3aH,3bH,4H,5H,7H,9aH,9bH,10H,11H,11aH-cyclopenta[a]phenanthrene-7,10-dione	Dehydrocortisone	53-03-2	C[C@@]12CC(=O)[C@H]3[C@@H](CCC4=CC(=O)C=C[C@]34C)[C@@H]2CC[C@]1(O)C(=O)CO

Рисунок 3.4. Зразок таблиці із завданням на прикладі сполук. Для учасника із вибіркою номер 402 і пошуковими запитами. На місці червоної стрічки було ім'я учасника

3.3 Статистика

В опитуванні прийняло участь 28 учасників-студентів, що прослуховували курси, які були пов'язані із хімічними базами даних і роботою з ними. Із них 12 та 6 людей заповнило секцію біоданих у Reaxys і SciFinder відповідно.

В ***Таблиця 1*** наведені загальні дані з пошуку сполук і мішеней. Рядки, які містили помилки (див. розділ 2.4) були виключені із результуючої статистики і як результат загальна кількість протестованих сполук у різних базах відрізнялась. Як наслідок кількість сполук усього і мішеней усього була різною, не зважаючи на однаковий розмір вибірки.

Таблиця 1. Загальна статистика пошуку сполук і мішеней

DATABASE	FOUND/ALL CPDS	FOUND CPDS, %	ERR CPD, %	FOUND/ALL TAR	FOUND TAR, %	ERR TAR, %
CHEMBL	131/139	94.2	8.3	77/84	91.7	10.6
GTF[1]	88/135	65.2	8.4	84/84	100.0	10.6
PROBES AND DRUGS	67/140	47.9	8.3	33/84	39.3	10.6
DRUGBANK	98/136	72.1	8.4	69/84	82.1	10.6
TARGETMOL	130/138	94.2	8.3	43/80	53.8	10.8
MCE[2]	129/139	92.8	8.3	63/80	78.8	10.8
SELLECKCHEM	107/139	77.0	8.3	35/83	42.2	10.6
TRC[3]	111/132	84.1	8.5	1/79	1.3	10.9
TOCRIS	54/138	39.1	8.3	48/79	60.8	10.9
CAYMAN	114/140	81.4	8.3	28/84	33.3	10.6
SANTA CRUZ	93/134	69.4	8.5	28/84	33.3	10.6
REAXYS	69/70	98.6	**11.7**	24/39	61.5	**15.6**
SCIFINDER	35/35	100.0	**16.6**	7/15	46.7	**25.3**

1 – Guide to Pharmacology, 2 – MedChemExpress, 3 – Toronto Research Chemicals

За даними можна відмітити, що в базах Tocris, Probes and Drugs, Guide to Pharmacology і Santa Cruz спостерігаються мінімуми, які можна пояснити або поганою анотацією сполук, або малим розміром бази даних сполук. Варто відмітити, що далі буде показано, що учасники стикалися з проблемами пошуку у Probes and Drugs і про цей ресурс вимагає подальшого вивчення.

За даними пошуку мішеней найменше усього було знайдено у Toronto Research Chemicals, яке вкладається в похибку і можна стверджувати, що в ресурсі і справді немає жодних даних про

терапевтичні мішені. У Cayman і Santa Cruz учасники помилково знаходили мішені, що детальніше буде розглянуто нижче.

Бажаний інтервал похибки дослідження був досягнутий для академічних баз і ресурсів постачальників. І становив менше 10% для сполук, а у випадку мішеней не перевищував 10.9%. Комерційні бази даних Reaxys і SciFinder були заповнені лише частиною учасників і похибки становили більше 10%. Вони мають бути досліджені детальніше на більшій виборці учасників.

3.3.1 Пошук

Гістограми із частотами відповідей зображені на **Рисунок 3.5**. Для зручності порівняння бази були об’єднані за своїм типом по кольору і «Не знайдено» виділено більш яскравим відтінком. Також наведена кумулятивна гістограма із сумарними результатами пошуку в усіх базах.

При досліджені учасники стикались з проблемою пошуку у базі Probes and Drugs, яка полягала у тривалій обробці запиту (більше 10 хвилин), що вважалося негативним результатом. Таким чином, ці дані можуть не зовсім коректно відображати реальну ситуацію, пов’язану із ресурсом.

Складніше усього було знайти речовину за допомогою синоніму 2 (назви ІЮПАК). Це може бути пояснено тим, що при пошуці алгоритми канонізації рядків видаляють важливі токени, які несуть інформацію. Для прикладу, це можуть бути, такі символи як знак «-», різні дужки та інше. Також ІЮПАК рядки для великих молекул можуть мати декілька назв, при цьому може існувати проблема того, що сполука має не канонізовану ІЮПАК назву.

Найкращим способами знайти сполуку були пошук за синонімом 1 (конвенційна або популярна назва) або за її CAS номером. Останній звичайному користувачеві, як правило, дістати складно, тому що в статтях

він зазвичай не приводиться. Однак CAS номер можна було б розглядати як універсальний хімічний ідентифікатор сполуки. Пошук по структурі також дає гарні результати і концептуально є одним із найкращих варіантів пошуку речовини. Варто відмітити, що у базах даних постачальників такі низькі результати структурного пошуку обумовлені відсутністю відповідної форми. Вона була присутня лише у Toronto Research Chemicals, але вона була досить недосконалою, що можна побачити на відповідному графіку. Серед комерційних і академічних ресурсів форма структурного пошуку була відсутня у Guide to Pharmacology.

Критерій «Не знайдено» може говорити про дві важливі речі. По-перше, таким чином можливо оцінити наповненість бази даних, оскільки вибірка для дослідження була створена репрезентативною. По-друге, якість анотації синонімів речовини та її ідентифікаторів. Великі значення цього параметру можна було відмітити у базах Tocris і Probes and Drugs.

Мішень вважалась знайденою лише у тому випадку, коли була знайдена відповідна їй веб-сторінка з даними про цей протеїн. Наявність сторінки може слугувати критерієм того, що у розробників бази даних є реляційний зв’язок між мішенню та сполукою і ступінь організації даних є вищим ніж у тих, в кого вона відсутня. Сторінками не вважалися, наприклад, список молекул, який видавав при пошук за назвою мішені, тому що містив назву протеїну у своїй анотації.

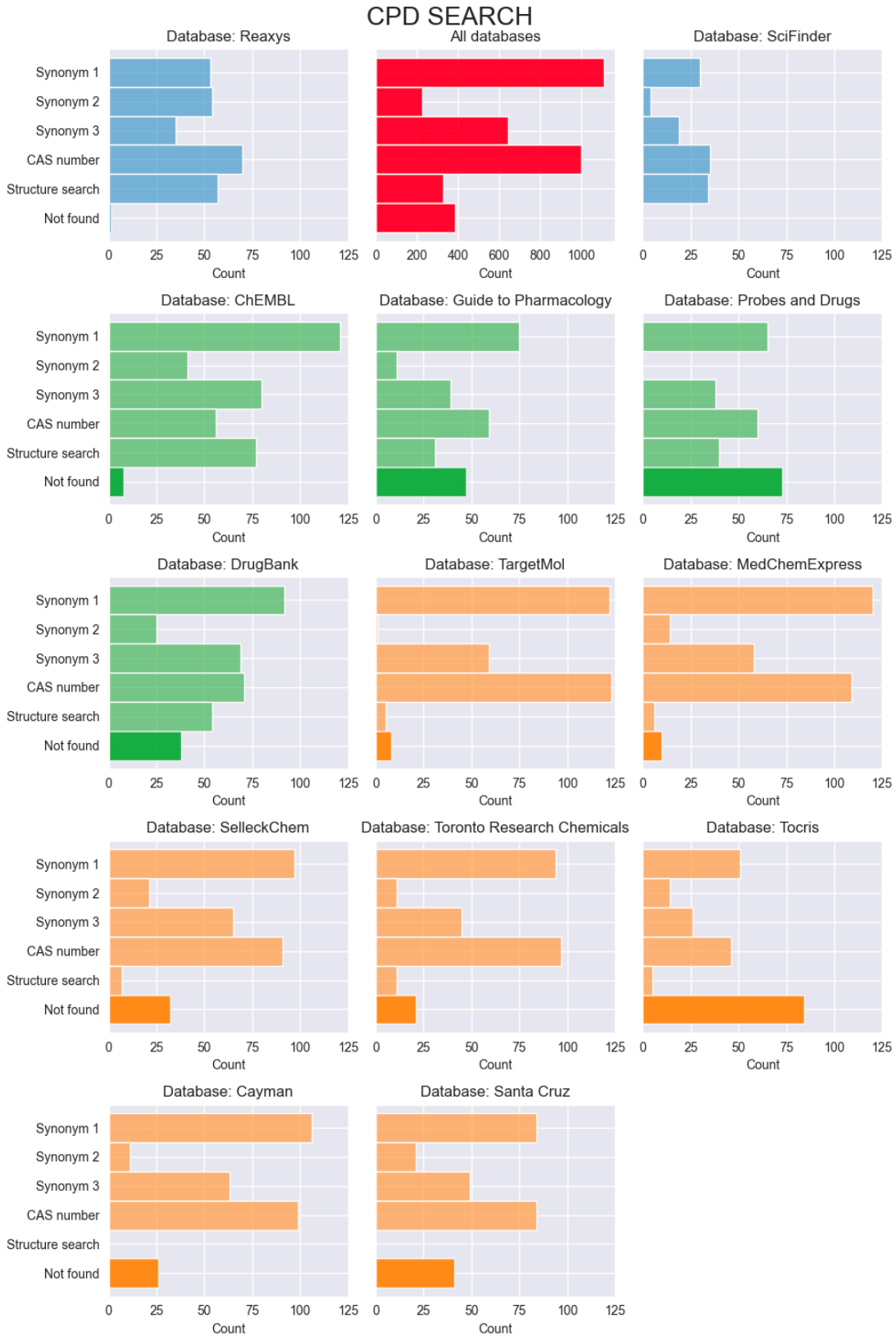

Рисунок 3.5. Гістограми результатів пошуку сполук. Червоним показані кумулятивні результати по усім 13 базам даних. Синім відмічені комерційні ресурси, зеленим – академічні бази даних, жовтим – бази постачальників

Іншою типовою помилкою були сторінки антитіл до мішені, рекомбінанті протеїни або тест-системи, які деякі учасники вважали сторінками мішені. Усі перелічені варіанти містили певну інформацію про протеїн, але з боку даного дослідження не вважалися коректними. Як наслідок частика учасників помічали такі сторінки знайденими, не зважаючи на те, що дана умова у описі завдань була оголошена. В реальності бази Cayman, Santa Cruz та TRC не містили веб-сторінки із анотованими мішенями.

Результуючі гістограми приведені на **Рисунок 3.6**. Майже усюди можна відмітити рівномірний розподіл пошуку за синонімами 1-3. У базах даних постачальників та Drugbank пошук за альтернативною назвою протеїну (синонім 3) рідше призводив до позитивних результатів пошуку у порівнянні з синонімами 1-2. UniprotId є популярним ідентифікатором протеїнів, який можна співставити із CAS-номером для хімічних сполук. Пошук за ним був присутній у всіх академічних базах та у Reaxys. Із графіків видно, що також пошук по UniprotId є в MCE та у Tocris, але у першому випадку за ідентифікатором шукались не сторінки протеїнів, а рекомбінантні та інші.

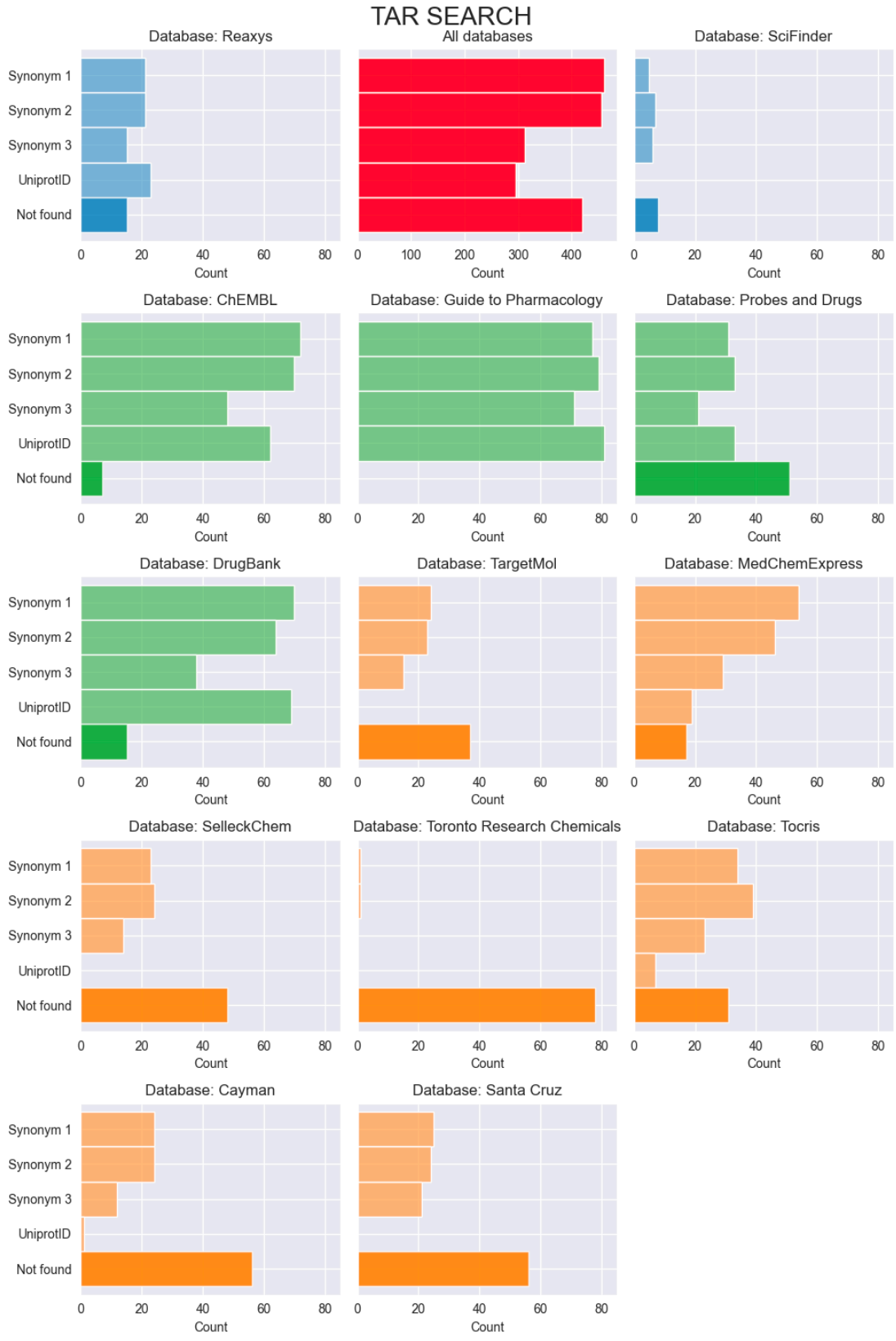

Рисунок 3.6. Гістограми результатів пошуку терапевтичних мішеней. Червоним показані кумулятивні результати по усім 13 базам даних. Синім відмічені комерційні ресурси, зеленим – академічні бази даних, жовтим – бази постачальників

3.3.2 Біодані

Відомо, що у скрінінгових компаніях можуть використовуватись як вільні основи, так і різні сольові форми хімічних сполук. Це робиться для варіації їх розчинності. В переважній більшості, протийон підбирається таким чином, щоб він не сильно змінював профіль біологічної активності (наприклад, це солі натрію або хлориди). Як наслідок у базах даних можуть зберігатись різні форми молекул, і по деяким причинам, певні із цих форм можуть бути краще анотованими ніж інші. Критерій other form був створений, щоб оцінити те, яку з форм обирали учасники. Якщо форма відрізнялась від тої, що наведена у таблицях із завданнями, то учасники його відмічали. І у коментарях під таблицею і вільній формі зазначали обрану форму

Топовими категоріями, які з'являлись частіше усього були References, Target and action on target, на наступному місці переважно було activity coefficient. Не зважаючи на те, що із заповненням Probes and Drugs у учасників виникали проблеми, за результатами видно те, що більшість респондентів не помітила коефіцієнти активності та інші дані пов'язані із мішенями.

Цікаво, що у DrugBank, немає даних по активності мішені, не зважаючи на те, що є інформація про мішень та її дію. При більш детальному досліджені виявляється, що ця інформація може лише *іноді* міститися в описі сполук, у розділі фармакодинаміки, але відсутня у розділі із мішенню.

Метаболізм був тим критерієм, що зустрічався менше усього. Дані з нього були присутні у Reaxys і DrugBank та в деякій мірі у ChEMBL,.

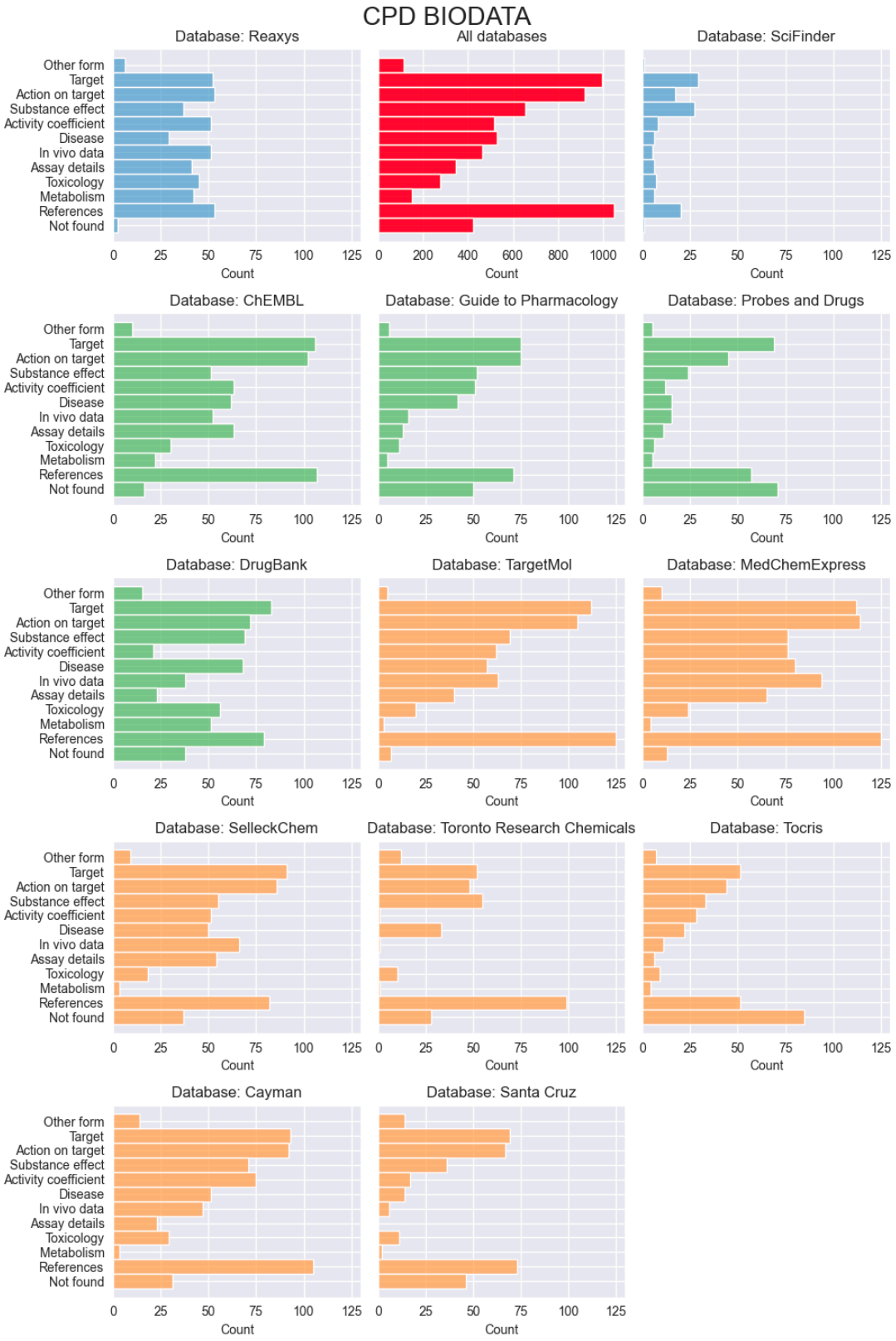

Рисунок 3.7. Гістограми результатів анотації біоданих сполук. Червоним показані кумулятивні результати по усім 13 базам даних. Синім відмічені комерційні ресурси, зеленим – академічні бази даних, жовтим – бази постачальників

Анотація мішеней в академічних базах даних від баз постачальників відрізнялась тим, що останні об'єднуть багато протеїнів в одну групу, так зване суперсімейство, для того, щоб спростити пошук і простіше організувати дані. Наприклад, замість чіткого розділення групи 5-НТ рецепторів на 5-НТ 1a, 5-НТ 2b і т.д., постачальники їх об'єднують на одній веб-сторінці присвяченій 5-НТ рецепторам. Для оцінки цього явища, був розроблений критерій «Target family page», який необхідно було обирати у випадку, коли ресурс не видавав сторінку із точною назвою, а учасник при описі заповнював анотацію для сторінки суперсімейства. Як видно із графіків зображених на **Рисунок 3.8**, відносна кількість сполук із відміченим «Target family page» є більшою для постачальників у порівнянні із академічними ресурсами.

Деякі учасники «знаходили» мішені там, де їх насправді не було, а саме у базах Cayman та Santa Cruz, які не містили веб-сторінок із мішенями. Те що було сприйнято за сторінку мішені було скоріше за все або список сполук (пошук за ключовими словами в описі), або рекомбінантні протеїні чи антитіла. Бази даних SciFinder і Toronto Research Chemicals не містили інформації про терапевтичні мішені.

Топовими позиціями були target description та ligand list. Таким чином ці два критерії були обов'язковими для анотації мішені. Другою важливою позицією було наявність дерева класифікації протеїну. Цей критерій показує високорівневу та складнішу організацію даних і дає змогу досліднику набагато простіше знайти ліганди на близькі мішені. Класифікаційне дерево було присутнє у наступних ресурсах MCE, Reaxys, ChEMBL, GTF, Tocris. В деяких із ресурсів цей параметр був заниженим, через те, що він міг бути важко помітним.

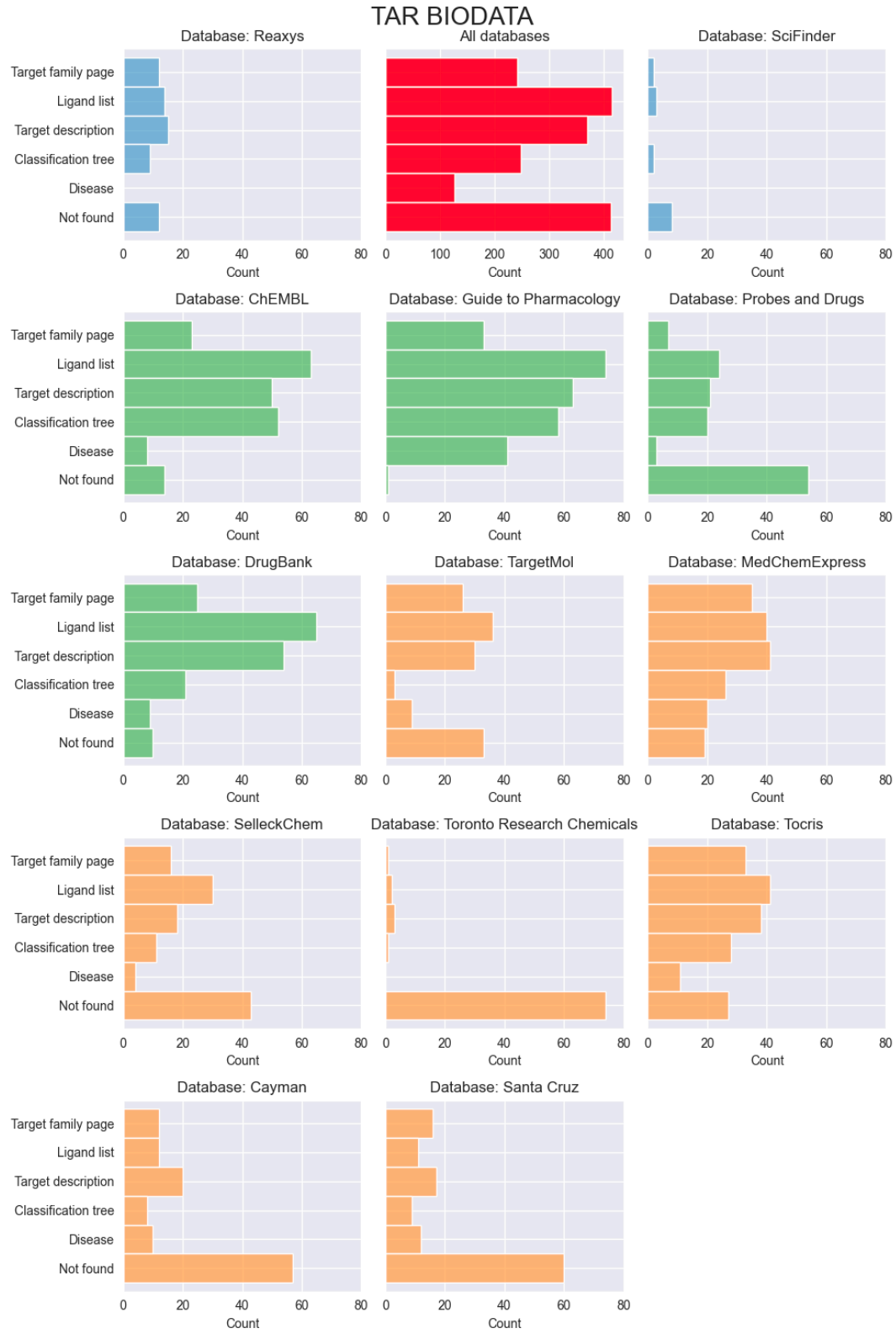

Рисунок 3.8. Гістограми результатів анотації біоданих мішеней. Червоним показані кумулятивні результати по усім 13 базам даних. Синім відмічені комерційні ресурси, зеленим – академічні бази даних, жовтим – бази постачальників

3.3.3 Аналіз зв'язок (патернів) біоданих

Гістограми із розподілом відповідей можуть бути складними для аналізу, тому було вирішено провести аналіз, який дозволить виявити певні шаблони (патерни) за якими найчастіше усього представлені дані у досліджуваних ресурсах.

Для цього були вибрані комбінації із відповідей розміром із 4 позицій. Для речовин Reference був виключений через те, що був самим популярним критерієм і майже завжди супроводжував будь-яку анотацію, але мається на увазі, що базово він присутній усюди. Для побудови графіку були обрані топ 3 популярні патерни для кожної із баз даних і були представлені на **Рисунок 3.9.**

В першу чергу можна помітити, що в більшості випадків спостерігається перші 2 позиції це Target + Action on target, а далі в залежності від ресурсу це можуть бути Activity coefficient, In vivo data, Assay details, Disease. У єдиному випадку самий рідкий критерій, Metabolism, входив до патерну в DrugBank. У топ патернів ввійшла токсикологія у базах Reaxys та DrugBank.

Далі видно, що академічні бази та ресурси постачальників є представленими в більшості категорій. У перших спільних патернів між базами майже немає і таким чином вони є гетерогенними. Якщо порівняти популярні патерни постачальників, то зв'язка **Target + Action on target + Substance effect + Activity coefficient** зустрічається у 5 базах одночасно та у 7, дивлячись на усі ресурси. Спільні пари були популярними у MCE та SelleckChem; TargetMol та Cayman; Toronto Research Chemicals і Santa Cruz.

На кумулятивній гістограмі по всім базам даних видно, що найпопулярнішим патерном був *Target + Action on target + Activity coefficient + In vivo data*, далі за ним слідують *Target + Action on target + In*

vivo data + *Assay details* та *Target* + *Action on target* + *Substance effect* + *Activity coefficient*.

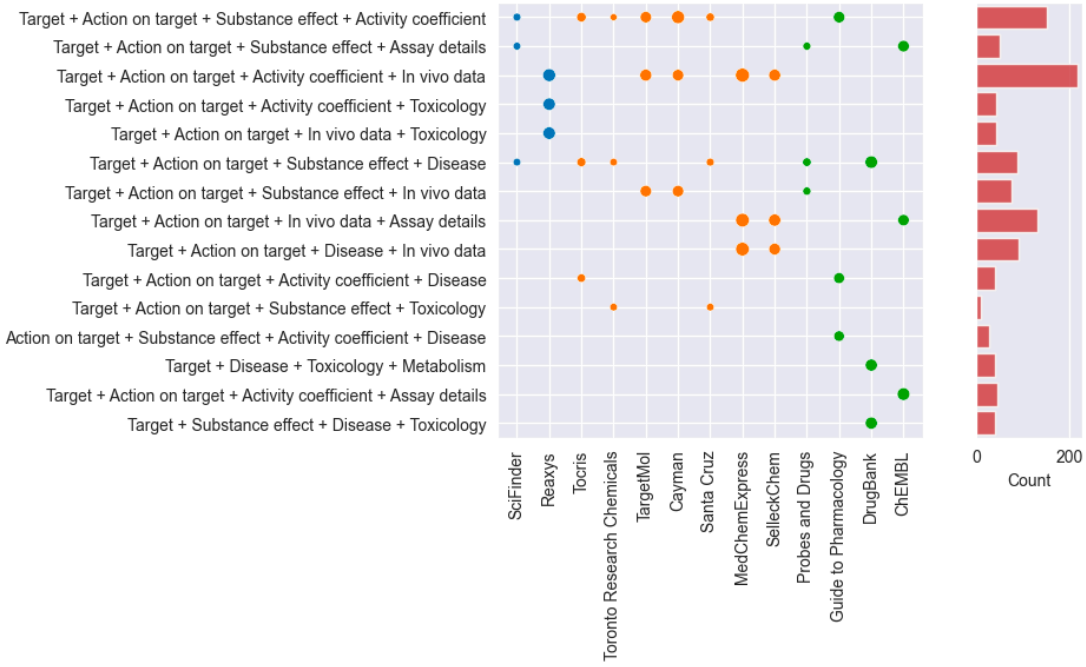

Рисунок 3.9. Топ 3 популярних зв'язок біоданих речовин. На осі Y показані патерни заповнення біоданих, на вісі X – бази даних. Розмір кружечка пропорційний частоті цього патерну. Жовтим вказані бази постачальників, синім – комерційні, зеленим – академічні. Справа червоним наведена гістограма із даними по всім базам даних.

У порівнянні із речовинами для побудови графіків було обрано топ 3 результати, які зображені на **Рисунок 3.10**. Розмір зв'язок мішеней був зменшений через меншу кількість критеріїв, що категоризують їх біодані.

В першу чергу варто відмітити, що у ресурсах SciFinder та Toronto Research Chemicals не було виявлено жодних патернів при заданих умовах. Це не є дивним оскільки дані ресурси виявились виключно хімічними і якщо і мастили біологічну анотацію, то вона була лише на сторінках речовин. Бази даних Cayman і Santa Cruz також не варто розглядати в цьому розділі із причин перелічених вище.

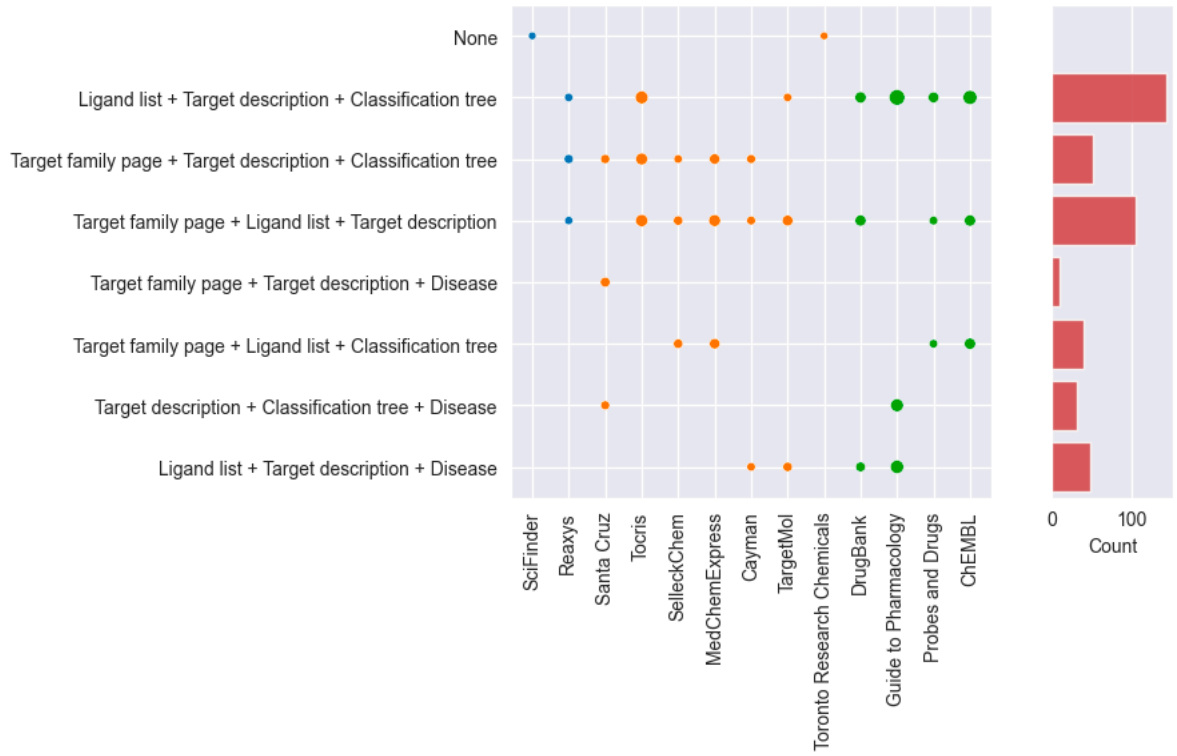

Рисунок 3.10. Топ 3 популярних зв'язок біоданих мішеней. На осі Y показані патерни заповнення біоданих, на вісі X – бази даних. Розмір кружечка пропорційний частоті цього патерну. Жовтим вказані бази постачальників, синім – комерційні, зеленим – академічні. Справа червоним наведена гістограма із даними по всім базам даних.

Найпопулярнішим шаблоном анотації мішеней був **Ligand list + Target description + Classification tree**. Далі йшли *Target family page + Ligand list + Target description* і *Target family page + Target description + Classification tree*. Що є цікавим так це те, що на 2му та 3му місцях фігурував *Target family page*. Це свідчить, що в більшості випадків мішені представлені у неточному з біологічної точки зору вигляді, тобто у вигляді певного суперсімейства. Коли вченого цікавить в більшій мірі цікавить точна назва протеїні. Варто зауважити, що детальну інформацію можна зустріти в описі сполуки, але при аналізі великих масивів даних такий формат не є зручним і вимагає «ручної» обробки результатів. У противагу цьому назва протеїну може бути винесена в окрему колонку.

Наведені патерни в більшій мірі представлені у більшості баз даних. Причиною цього може бути невелике різноманіття категоризації даних по мішеням.

3.3.4 Оцінка баз

Рисунок 3.11 показує результати оцінки баз. На ньому можна побачити, що учасники стикались із проблемами пов'язаними із пошуком. У більшості учасників в базі Probes and Drugs пошук тривав довше визначеного часу і не приводив до жодних результатів (за результатами форми зворотного зв'язку), тому об'єктивне оцінювання цього ресурсу було утрудненим. Видно, що думки розділились при оцінці бази ChEMBL, де з невеликим розривом розподілені позитивні і негативні оцінки. Це було пов'язано із повільним пошуковим процесом.

Невелика частина опитаних відмічала, що їм було складно знайти бажану сторінку із гарною анотацією через те, що пошук видавав багато різних сольових форм (Complicated search outcome). Таке спостерігалось у ChEMBL, Probes and Drugs, Toronto Research Chemicals.

Проводячи межу на рівні 15, яка є більшою за половину кількості опитуваних (усього 28), були зроблені наступні спостереження. Учасники оцінили інтерфейси ресурсів MedChemExpress, ChEMBL, DrugBank, Reaxys привабливими. Інформація була добре структурована, що дозволяла легко заповнити секцію біоданих була у Reaxys, ChEMBL, Guide to Pharmacology, DrugBank, MedChemExpress. Як наслідок, дані, що знаходились у чітко диференційованих розділах або в таблицях надавали можливість більш ефективного заповнення таблиць. У протилежність, у таких ресурсах як Cayman, Toronto Research Chemicals, Santa Cruz інформація була представлена у вигляді простого тексту. Як наслідок, можна, припустити, що це є більш прийнятним варіантом для сприйняття інформації.

.

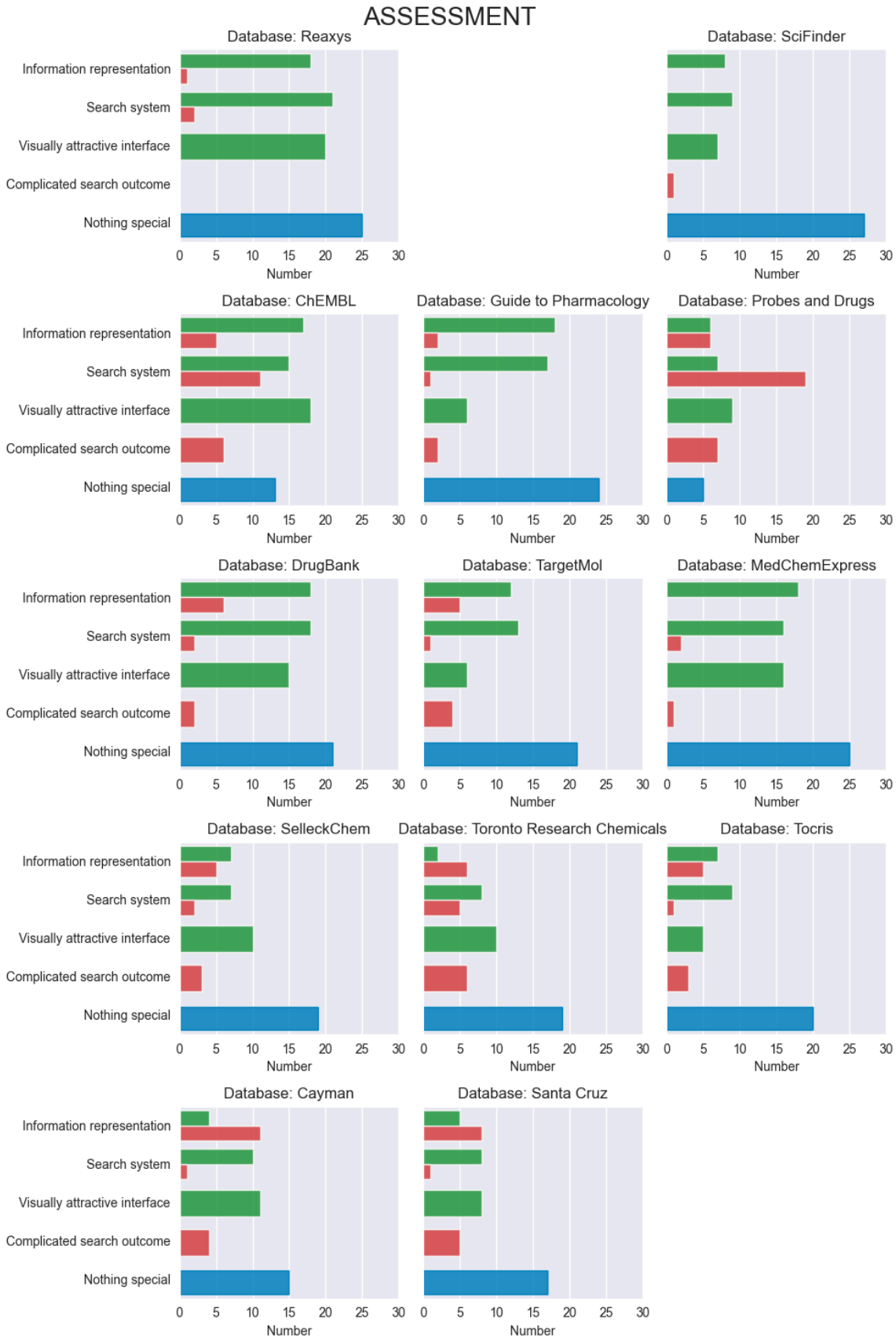

Рисунок 3.11. Результати оцінювання баз даних. Зеленим показана позитивна оцінка, червоним – негативна, синім – нейтральна.

Сукупно за трьома критеріями Information representation, search system visually attractive interface більше половини учасників виділили бази даних Reaxys, ChEMBL, DrugBank, MedChemExpress

РОЗДІЛ 4. ВИСНОВКИ

1. Було проаналізовано ієрархію зберігання даних в академічних, комерційних і ресурсах постачальників за допомогою методу опитування анкетуванням. Розроблено критерії для оцінки якості пошуку і анотації біоданих, залучено до опитування 28 студентів.
2. Знайдено, що серед академічних баз найкращим ресурсом для пошуку біоактивних сполук та терапевтичних мішеней був ChEMBL, серед комерційних – Reaxys, серед постачальників – MedChemExpress.
3. Встановлено, що найкращим джерелом інформації з біоактивності сполук серед академічних є ChEMBL та DrugBank, серед комерційних – Reaxys, серед постачальників – MedChemExpress.
4. Знайдено, що повноцінно анотованим ресурсом по терапевтичним мішеням серед академічних є Guide to Pharmacology, серед комерційних – Reaxys, серед постачальників – MedChemExpress.
5. Показано, що патерном анотації біоданих для сполук, який найчастіше зустрічався, є *Target + Action on target + Activity coefficient + In vivo data*, а для мішеней – *Ligand list + Target description + Classification tree.*

РОЗДІЛ 5. ДЖЕРЕЛА

[1] «Hu, Y., & Bajorath, J. (2017). Entering the “big data” era in medicinal chemistry: molecular promiscuity analysis revisited. Future Science OA, 3(2), FSO179. doi:10.4155/fsoa-2017-0001».

[2] «Fortenbaugh FC, DeGutis J, Germine L, Wilmer JB, Grosso M, Russo K, Esterman M. Sustained Attention Across the Life Span in a Sample of 10,000: Dissociating Ability and Strategy. Psychol Sci. 2015 Sep;26(9):1497-510. doi: 10.1177/0956797615594896. Epub 20».

[3] https://www.gartner.com/en/information-technology/glossary/big-data. [Онлайновий].

[4] https://datafloq.com/read/3vs-sufficient-describe-big-data/. [Онлайновий].

[5] https://www.futuretimeline.net/data-trends/2050-future-internet-speed-predictions.htm. [Онлайновий].

[6] https://www.semrush.com/blog/youtube-stats/. [Онлайновий].

[7] Sebastian Raschka. 2015. Python Machine Learning. Packt Publishing..

[8] «Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781».

[9] Patrick, G. (2017). An introduction to medicinal chemistry (6th ed.). Oxford University Press..

[10] https://www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine/?sh=70a50dbc13c3. [Онлайновий].

[11] «Tiikkainen, P., Bellis, L., Light, Y., & Franke, L. (2013). Estimating Error Rates in Bioactivity Databases. Journal of Chemical Information and Modeling, 53(10), 2499–2505. doi:10.1021/ci400099q».

[12] «Voigt, J. H., Bienfait, B., Wang, S., & Nicklaus, M. C. (2001). Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. Journal of Chemical Information and Computer Sciences, 41(3), 702–712. doi:10.1021/ci000150t».

[13] «Ghani, Syed. (2020). A comprehensive review of database resources in chemistry. Eclética Química Journal. 45. 57-68. 10.26850/1678-4618eqj.v45.3.2020.p57-68.».

[14] «Southan, C., Sitzmann, M., & Muresan, S. (2013). Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. Molecular Informatics, 32(11-12), 881–897. doi:10.1002/minf.201300103».

[15] «Bharti, N., Leonard, M., & Singh, S. (2016). Review and Comparison of the Search Effectiveness and User Interface of Three Major Online Chemical Databases. Journal of Chemical Education, 93(5), 852–863. doi:10.1021/acs.jchemed.5b00601».

[16] «Pallarz, S., Benary, M., Lamping, M., Rieke, D., Starlinger, J., Sers, C., … Leser, U. (2019). Comparative Analysis of Public Knowledge Bases for Precision Oncology. JCO Precision Oncology, (3), 1–8. doi:10.1200/po.18.00371».

[17] https://www.reaxys.com/. [Онлайновий].

[18] https://scifinder-n.cas.org/. [Онлайновий].

[19] https://go.drugbank.com/. [Онлайновий].

[20] https://www.ebi.ac.uk/chembl/. [Онлайновий].

[21] https://www.guidetopharmacology.org/. [Онлайновий].

[22] https://www.probes-drugs.org/home/. [Онлайновий].

[23] https://www.medchemexpress.com/. [Онлайновий].

[24] https://www.targetmol.com/. [Онлайновий].

[25] https://www.selleckchem.com/. [Онлайновий].

[26] https://www.trc-canada.com/. [Онлайновий].

[27] http://www.tocris.com/. [Онлайновий].

[28] https://www.scbt.com/. [Онлайновий].

[29] https://www.caymanchem.com/. [Онлайновий].

[30] https://forms.gle/nHKfKNQvHg1DfTH59. [Онлайновий].

[31] https://www.uniprot.org. [Онлайновий].

[32] https://www.genenames.org/. [Онлайновий].

[33] •. https://www.probes-drugs.org/compounds/standardized#compoundset=408@AND. [Онлайновий].

[34] https://docs.google.com/spreadsheets/d/1h5UfLLdezkl2eZ7RqCZN-UoA3c27RB_CWXssCC2lFi4/edit?usp=sharing. [Онлайновий].

[35] «https://www.guidetopharmacology.org/about.jsp,» [Онлайновий].

[36] «https://clinicaltrials.gov/,» [Онлайновий].

[37] https://www.statista.com/statistics/871513/worldwide-data-created/. [Онлайновий].