

Міністерство освіти і науки України  
Харківський національний університет радіоелектроніки

Факультет \_\_\_\_\_ комп'ютерних наук (або центр післядипломної освіти, або навчально-науковий центр заочної форми навчання)  
(повна назва)

Кафедра \_\_\_\_\_ програмної інженерії  
(повна назва)

**КВАЛІФІКАЦІЙНА РОБОТА**  
**Пояснювальна записка**

рівень вищої освіти \_\_\_\_\_ другий (магістерський)

Дослідження методів програмної реалізації  
оцінки властивостей складних структур  
(тема)

Виконав:

студентка 2 курсу, групи ІПЗм-21-1

Сергієнко О.С.

(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення

(код і повна назва спеціальності)

Тип програми освітньо-наукова

Керівник доц. Турута О.П.

(посада, прізвище, ініціали)

Допускається до захисту  
Зав. кафедри

\_\_\_\_\_  
(підпис)

З.В.Дудар

(прізвище, ініціали)

2024 р.

## Харківський національний університет радіоелектроніки

Факультет \_\_\_\_\_ комп'ютерних наук (або центр післядипломної освіти, або навчально-науковий центр заочної форми навчання)  
 Кафедра \_\_\_\_\_ програмної інженерії  
 Рівень вищої освіти \_\_\_\_\_ другий (магістерський)  
 Спеціальність \_\_\_\_\_ 121 – Інженерія програмного забезпечення  
 Тип програми \_\_\_\_\_ освітньо-наукова програма  
 Освітня програма \_\_\_\_\_ Інженерія програмного забезпечення  
 (шифр і назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри \_\_\_\_\_

(підпис)

«\_\_\_\_» \_\_\_\_\_ 2024 р.

### ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові \_\_\_\_\_ Сергієнко Олександрі Сергіївні \_\_\_\_\_

(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів програмної реалізації оцінки властивостей складних структур»

Затверджена наказом по університету від 29.03.2024р. № 250 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 20.06.2024

3. Вихідні дані до роботи хімічні структури, їх властивості, методи та моделі прогнозування, прогнозування ADME/Т властивостей

4. Перелік питань, що потрібно опрацювати в роботі

аналіз стану сучасних методів визначення властивостей хімічних структур, проблеми даних у хімічних дослідження, огляд методів машинного навчання, що використовуються для прогнозування ADME/Т, проведення експериментальних досліджень, розробка API

## КАЛЕНДАРНИЙ ПЛАН

| №  | Назва етапів роботи                                                      | Термін виконання етапів роботи | Примітка |
|----|--------------------------------------------------------------------------|--------------------------------|----------|
| 1  | Аналіз предметної галузі та постановка задачі                            | 20.01 – 11.02.24               | виконано |
| 2  | Аналіз існуючих методів                                                  | 12.02 – 25.02.24               | виконано |
| 3  | Аналіз та моделювання предметної області                                 | 26.02 – 10.03.24               | виконано |
| 4  | Планування експериментів та постановка задачі                            | 11.03 – 15.03.24               | виконано |
| 5  | Експериментальні дослідження                                             | 16.03 – 11.04.24               | виконано |
| 6  | Аналіз результатів експериментальних досліджень та розробка рекомендацій | 12.04 – 26.04.23               | виконано |
| 7  | Розробка API                                                             | 27.04 – 13.05.24               | виконано |
| 8  | Написання та оформлення статті та тез доповіді                           | 13.05 – 17.05.24               | виконано |
| 9  | Підготовка пояснювальної записки                                         | 17.05 – 02.06.24               | виконано |
| 10 | Підготовка презентації та доповіді                                       | 03.06 – 13.06.24               | виконано |
| 11 | Нормоконтроль                                                            | 14.06 – 17.06.24               | виконано |
| 12 | Рецензування                                                             | 18.06 – 19.06.24               | виконано |
| 13 | Занесення диплома в електронний архів                                    | 19.06.2024                     | виконано |
| 14 | Попередній захист                                                        | 19.06.2024                     | виконано |
| 15 | Допуск до захисту у зав. кафедри                                         | 19.06.2024                     | виконано |

Дата видачі завдання 29 березня 2024р.

Студентка

\_\_\_\_\_  
(підпис)

\_\_\_\_\_  
Сергієнко О.С.

Керівник роботи

\_\_\_\_\_  
(підпис)

\_\_\_\_\_  
доц. Турута О.П.,  
(посада, прізвище, ініціали)

## РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 61 с., 11 рис., 10 табл., 10 джерел.

ХІМІЧНІ СТРУКТУРИ, ADME/T ВЛАСТИВОСТІ, ПРОГНОЗУВАННЯ  
ВЛАСТИВОСТЕЙ, CHEMPROP, MGA, RDKit.

Об'єктом дослідження є методи та моделі прогнозування властивостей хімічних сполук.

Метою роботи є підвищення ефективності розробки лікарських препаратів, зокрема для лікування COVID-19.

У результаті кваліфікаційної роботи було розглянуто методи визначення ADME/T властивостей, зокрема квантово-хімічні розрахунки, молекулярне моделювання, машинне навчання, використання експертних оцінок та їх інтеграція. Також було досліджено моделі та інструменти, що спеціалізуються на прогнозуванні різних хімічних властивостей молекул, таких як біологічна активність і токсичність. Було розроблено пайплайн для обробки хімічних структур, який буде доповнено в подальшому та API для дослідження роботи Chemprop, MGA та інших моделей та фреймворків.

CHEMICAL STRUCTURES, ADME/T PROPERTIES, PROPERTIES  
PREDICTION, CHEMPROP, MGA, RDKit.

The object of research is methods and models for predicting the properties of chemical compounds.

The purpose of the work is to enhance the efficiency of drug development, particularly for the treatment of COVID-19.

As a result of the qualification work, methods for determining ADME/T properties were considered, including quantum chemical calculations, molecular modeling, machine learning, the use of expert assessments, and their integration.

Models and tools specializing in predicting various chemical properties of molecules, such as biological activity and toxicity, were also investigated. A pipeline for processing chemical structures has been developed, which will be further expanded, along with an API for investigating the operation of Chemprop, MGA, and other models and frameworks.

Заява щодо самостійного виконання кваліфікаційної роботи та можливості її публікації в електронному архіві відкритого доступу ElArKhNURE.

Я, Сергієнко Олександра Сергіївна, студентка гр. ПЗм-21-1, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів програмної реалізації оцінки властивостей складних структур», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлена з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

## ЗМІСТ

|                                                                             |    |
|-----------------------------------------------------------------------------|----|
| Перелік скорочень                                                           | 10 |
| Вступ                                                                       | 11 |
| 1 Аналіз предметної галузі                                                  | 12 |
| 1.1 Аналіз стану сучасних методів визначення властивостей хімічних структур | 12 |
| 1.1.1 Квантово-хімічні розрахунки                                           | 12 |
| 1.1.2 Молекулярне моделювання                                               | 13 |
| 1.1.3 Методи машинного навчання                                             | 13 |
| 1.1.4 Використання експертних оцінок                                        | 13 |
| 1.1.5 Інтеграція різних методів                                             | 14 |
| 1.1.6 Ефективність методів визначення властивостей                          | 14 |
| 1.2 Проблеми даних у хімічних дослідженнях                                  | 15 |
| 1.2.1 Недостатність даних                                                   | 15 |
| 1.2.2 Варіабельність форматів представлення даних                           | 16 |
| 1.2.3 Неточності в експериментальних даних                                  | 17 |
| 1.2.4 Обмеження обчислювальних ресурсів                                     | 17 |
| 1.3 Аналіз існуючих аналогів                                                | 17 |
| 2 Аналіз існуючих методів                                                   | 21 |
| 2.1 Перелік критеріїв та їх опис                                            | 21 |
| 2.2 Векторний опис альтернатив за обраними критеріями                       | 22 |
| 2.3 Перетворення якісних шкал в кількісні                                   | 23 |
| 2.4 Нормування оцінок за шкалами та оцінка корисності                       | 24 |
| 3 Постановка задачі                                                         | 28 |
| 3.1 Збір хімічних даних                                                     | 28 |
| 3.2 Пайплайн для обробки складних структур                                  | 29 |
| 3.3 Аналіз зібраних даних                                                   | 34 |
| 3.4 Огляд поточних рішень                                                   | 36 |
| 3.4.1 Chemprop-RDKit6                                                       | 36 |
| 3.4.2 Фреймворк MGA                                                         | 37 |

|                                                                                                                                  |    |
|----------------------------------------------------------------------------------------------------------------------------------|----|
|                                                                                                                                  | 9  |
| 3.4.3 Моделі на основі дескрипторів                                                                                              | 39 |
| 3.5 Експериментальне дослідження обраних рішень                                                                                  | 40 |
| 3.5.1 Метрики для валідації моделей                                                                                              | 40 |
| 3.5.2 Результати експериментів з Chemprop-RDKit6                                                                                 | 41 |
| 3.5.3 Результати експериментів з моделями на основі дескрипторів                                                                 | 42 |
| 3.6 Розробка API для інтеграції                                                                                                  | 44 |
| Висновки                                                                                                                         | 46 |
| Перелік джерел посилання                                                                                                         | 47 |
| Перелік джерел посилання за науковими напрямами керівника та науковців<br>кафедри програмної інженерії                           | 49 |
| Додаток А Звіт результатів перевірки на унікальність тексту в мережі інтернет та<br>базі ХНУРЕ                                   | 50 |
| Додаток Б Слайди презентації                                                                                                     | 51 |
| ДОДАТОК В Апробація результатів                                                                                                  | 58 |
| ДОДАТОК Г Експертний висновок результатів перевірки кваліфікаційної роботи<br>на відповідність оформлення вимогам ДСТУ 3008:2015 | 61 |

## **ПЕРЕЛІК СКОРОЧЕНЬ**

ADMET – Absorption, Distribution, Metabolism, Excretion, Toxicity.

MOLFILE – Molecular Structure Data File

QSAR – Quantitative Structure-Activity Relationship

SAR – Structure-Activity Relationship

SDF – Spatial Data File

SMILES – Simplified Molecular Input Line Entry System

InChi – International Chemical Identifier

GNN – Graph Neural Networks

API – Application Programming Interface

## ВСТУП

У сучасному світі зростає потреба в нових лікарських засобах, особливо в умовах пандемій, таких як COVID-19. Сучасні комп'ютерні методи аналізу даних дозволяють ефективно оцінювати властивості молекул, що дозволяє оптимізувати процес досліджень і знижує ризики невдалих експериментів. Такі дослідження відіграють ключову роль у розвитку фармацевтичної науки, сприяючи створенню більш ефективних і безпечних лікарських препаратів.

Метою дослідження є валідація моделей QSAR [1], які є важливим інструментом для передбачення молекулярної активності на основі структурних властивостей. Дослідження включає аналіз ефективності наступних технологій: Chemprop-RDKit, фреймворку MGA та моделей, розроблених на основі дескрипторів. Проводиться порівняння їх ефективності, аналіз прогностичної здатності та виокремлення значущих особливостей.

Об'єктом дослідження є хімічні сполуки, їхні властивості та взаємозв'язки.

Наукова новизна дослідження полягає у застосуванні інноваційних методів аналізу даних для покращення точності моделей QSAR.

Практичне значення дослідження полягає у можливості значного вдосконалення процесу розробки нових лікарських препаратів, що сприятиме підвищенню їхньої ефективності та безпеки. Розроблені підходи і моделі можуть бути використані фармацевтичною індустрією та науковими установами для прискорення виведення нових лікарських засобів на ринок, зменшення витрат на дослідження та підвищення їхньої надійності.

В рамках роботи було спроектовано пайплайн для обробки хімічних структур, проаналізовано низку поточних рішень та розроблено API для інтеграції з іншими системами.

## 1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

### 1.1 Аналіз стану сучасних методів визначення властивостей хімічних структур

Сучасний етап розвитку хімії та фармацевтики характеризується високою кількістю наукових та технологічних інновацій. Однією з ключових проблем є потреба у точних та ефективних методах визначення властивостей хімічних сполук [6]. Це важливо для розробки нових лікарських препаратів, забезпечення їхньої безпеки та ефективності. У цьому розділі ми детально розглянемо еволюцію розвитку методів, сучасні методи визначення властивостей та їхню ефективність.

Еволюція методів визначення властивостей хімічних сполук включає значний прогрес від класичних хімічних аналізів до сучасних обчислювальних та експериментальних підходів. На початкових етапах розвиток методів базувався на емпіричних даних та ручних обчисленнях. Зі зростанням складності хімічних структур та вимог до точності аналізу, з'явилися більш складні методи, такі як спектроскопія, хроматографія та рентгеноструктурний аналіз.

У другій половині 20-го століття почали активно розвиватися обчислювальні методи, зокрема квантово-хімічні розрахунки та молекулярне моделювання. З появою високопродуктивних обчислювальних систем та алгоритмів машинного навчання ці методи стали більш доступними та точними. Сьогодні інтеграція різних методів дозволяє отримувати комплексну інформацію про властивості хімічних сполук.

#### 1.1.1 Квантово-хімічні розрахунки

Квантово-хімічні розрахунки є одним із важливих підходів до визначення властивостей хімічних сполук. Ці розрахунки базуються на принципах квантової механіки та дозволяють моделювати електронну структуру молекул, взаємодії між атомами та енергетичні рівні. Використання квантово-хімічних розрахунків

дозволяє отримувати точні результати, але вимагає значних обчислювальних ресурсів. Наприклад, моделювання невеликої молекули може займати від кількох годин до кількох днів, залежно від точності обраного методу.

### 1.1.2 Молекулярне моделювання

Молекулярне моделювання дозволяє аналізувати взаємодії молекул на рівні атомів та молекул. Методи, такі як динамічне моделювання та кількісно-структурні відносини, дозволяють досліджувати рух частинок та їхні структурні особливості. Молекулярне моделювання може бути дуже витратним з точки зору обчислювальних ресурсів, але сучасні досягнення в області обчислювальної хімії дозволяють значно підвищити його ефективність.

### 1.1.3 Методи машинного навчання

Методи машинного навчання набувають все більшої популярності у визначенні властивостей хімічних сполук. З використанням великої кількості даних та алгоритмів, таких як глибоке навчання, ці методи дозволяють прогнозувати властивості хімічних сполук з високою точністю. Графові нейронні мережі, наприклад, можуть враховувати складні графові структури молекул [8], роблячи їхнє використання ефективним для аналізу великих та складних молекул. Ефективність методів машинного навчання зростає зі збільшенням обсягу даних, що використовується для їхнього навчання.

### 1.1.4 Використання експертних оцінок

Експертні оцінки використовуються для уточнення та підтвердження результатів, отриманих іншими методами. Експерти можуть надати важливі інсайти в ті питання, які можуть бути недостатньо або неправильно враховані автоматизованими методами. Залучення досвідчених фахівців дозволяє підвищити

надійність та точність отриманих результатів, особливо в складних чи нестандартних випадках.

### 1.1.5 Інтеграція різних методів

Інтеграція різних методів є одним із ефективних способів підвищення точності та ефективності визначення властивостей хімічних сполук. Наприклад, комбінування квантово-хімічних розрахунків та молекулярного моделювання дозволяє отримати детальні та збалансовані результати. Квантово-хімічні розрахунки можуть надати точні дані про електронну структуру молекули, в той час як молекулярне моделювання дозволяє аналізувати взаємодії та динаміку молекул на великих просторах. Такий інтегрований підхід дозволяє враховувати якісні та кількісні аспекти вивчення молекулярних систем.

### 1.1.6 Ефективність методів визначення властивостей

Оцінка ефективності методів визначення властивостей хімічних сполук є важливим аспектом їхнього вибору та застосування [8]. Наприклад:

- квантово-хімічні розрахунки показують високу точність при моделюванні малих молекул. Згідно з деякими дослідженнями, похибка в розрахунках енергій молекул може складати менше 1% від експериментальних значень;
- молекулярне моделювання дозволяє ефективно досліджувати динаміку великих молекул. Точність результатів може залежати від якості вхідних даних та параметрів моделі. У деяких випадках точність прогнозів може досягати 90-95%;
- методи машинного навчання можуть швидко обробляти великі обсяги даних і давати точні прогнози. Для великих наборів даних точність прогнозів може перевищувати 95%, що робить їх дуже ефективними для аналізу великих та складних молекул.

Загальний аналіз сучасних методів визначення властивостей підкреслює важливість інтеграції різних підходів. Квантово-хімічні розрахунки, молекулярне моделювання та методи машинного навчання можуть доповнювати один одного, створюючи комплексний підхід до вивчення властивостей хімічних сполук. Такий інтегрований підхід дозволяє враховувати якісні та кількісні аспекти вивчення молекулярних систем, роблячи його більш придатним для потреб сучасної хімічної науки та промисловості.

## 1.2 Проблеми даних у хімічних дослідженнях

Дані відіграють вирішальну роль у хімічних та фармацевтичних дослідженнях [8], оскільки вони є основою для аналізу, моделювання та передбачення властивостей хімічних сполук. Однак, існує кілька проблем, пов'язаних з даними, які можуть впливати на точність і надійність досліджень. У цьому розділі розглянемо основні проблеми, такі як недостатність даних, варіабельність форматів представлення даних та неточності в експериментальних даних, а також обмеження обчислювальних ресурсів.

### 1.2.1 Недостатність даних

Одна з найбільших проблем у хімічних дослідженнях – це недостатність експериментальних даних. У багатьох випадках дані про властивості нових або рідкісних хімічних сполук є обмеженими або відсутніми. Це може бути пов'язано з високими витратами на проведення експериментів, складністю синтезу сполук або обмеженим доступом до необхідного обладнання. Недостатність даних може призводити до низької точності моделей передбачення властивостей сполук та уповільнювати процес розробки нових ліків.

Для вирішення цієї проблеми використовуються методи створення синтетичних даних, які можуть бути згенеровані за допомогою моделювання або

машинного навчання. Синтетичні дані допомагають заповнити прогалини в експериментальних даних та покращити результати передбачення.

### 1.2.2 Варіабельність форматів представлення даних

Іншою важливою проблемою є варіабельність форматів представлення хімічних структур. Існують різні формати, такі як SMILES, MOLFile, InChI та інші [7], кожен з яких має свої переваги та недоліки. Використання різних форматів може ускладнювати обмін даними між дослідниками та інструментами, що впливає на ефективність обробки та аналізу даних (див. табл. 1.1).

Таблиця 1.1 – Опис форматів представлення даних (таблиця виконана самостійно)

|                         | SMILES                                    | MOLFile                                  | InChi                         |
|-------------------------|-------------------------------------------|------------------------------------------|-------------------------------|
| Тип даних               | Текстовий                                 | Бінарний                                 | Текстовий                     |
| Зберігання координат 3D | Ні                                        | Так                                      | Ні                            |
| Додаткові параметри     | Обмежені                                  | Розширені                                | Стандартні                    |
| Зручність обробки       | Висока                                    | Середня                                  | Висока                        |
| Компактність            | Висока                                    | Низька                                   | Середня                       |
| Застосування            | Хімічна інформатика, віртуальний скринінг | Тривимірне моделювання, детальний аналіз | Обмін даними та ідентифікація |

Вибір формату залежить від конкретних потреб досліджень. Якщо основною задачею є швидкий обмін та обробка даних, то SMILES є більш зручним через свою компактність та простоту. Для детального аналізу молекулярних структур, включаючи тривимірні аспекти, MOLFile надає необхідну гнучкість та інформаційне насичення. InChI є кращим для стандартизації та обміну даними між різними базами та системами.

### 1.2.3 Неточності в експериментальних даних

Неточності в експериментальних даних також є серйозною проблемою. Ці неточності можуть виникати через помилки в вимірюваннях, неоднорідність зразків, людський фактор або несправності обладнання. Невідповідні дані можуть призводити до помилкових висновків та знижувати надійність моделей передбачення.

### 1.2.4 Обмеження обчислювальних ресурсів

Обчислювальні ресурси є ще однією критичною проблемою. Для складних молекулярних розрахунків і моделювання потрібні значні обчислювальні потужності, що може бути витратним і вимагати спеціалізованого обладнання. Це обмежує можливості проведення великомасштабних досліджень та аналізу великих наборів даних.

## 1.3 Аналіз існуючих аналогів

На сьогоднішній день існує широкий вибір програмних засобів та веб-сервісів, які використовують передові методи машинного навчання та квантово-хімічні підходи для прогнозування хімічних властивостей, зокрема ADME (поглинання, розподіл, метаболізм та виведення) хімічних сполук. Один із таких сервісів - Admetlab [1].

Admetlab - це онлайн-платформа, призначена для прогнозування хімічних властивостей та ADME хімічних сполук. Він відкритий для використання та має API, що дозволяє легко інтегрувати його у власні дослідження. Інтерфейс Admetlab приємний та інтуїтивно зрозумілий, що полегшує користування. Для отримання результатів достатньо ввести структуру у форматі SMILES або скористатися альтернативним варіантом - намалювати її за допомогою редактора JMSE (див. рис. 1.1).

Home / Services / ADMET Evaluation

Paste a single SMILES string [Draw a molecule using JMSE editor](#)

SMILES  [Example](#)

[Reset](#) [Submit](#)

**Tip**

- Before submitting, users should prepare a valid SMILES in case of errors.
- Molecules composed of more than 128 atoms are not recommended for evaluation.

Рисунок 1.1 - Інтерфейс AdmetLab (рисунок створено самостійно)

Серед недоліків Admetlab [1] можна відзначити відсутність можливості прогнозування для цілого датасету, що не завжди зручно для користувачів. Також відсутня валідація структур, і неможливо евалювати молекули, що складаються з більш ніж 128 атомів, що може стати обмеженням. Однак кількість моделей на платформі є задовільною (див. рис. 1.2), а швидкість розрахунків висока, що є позитивними моментами.

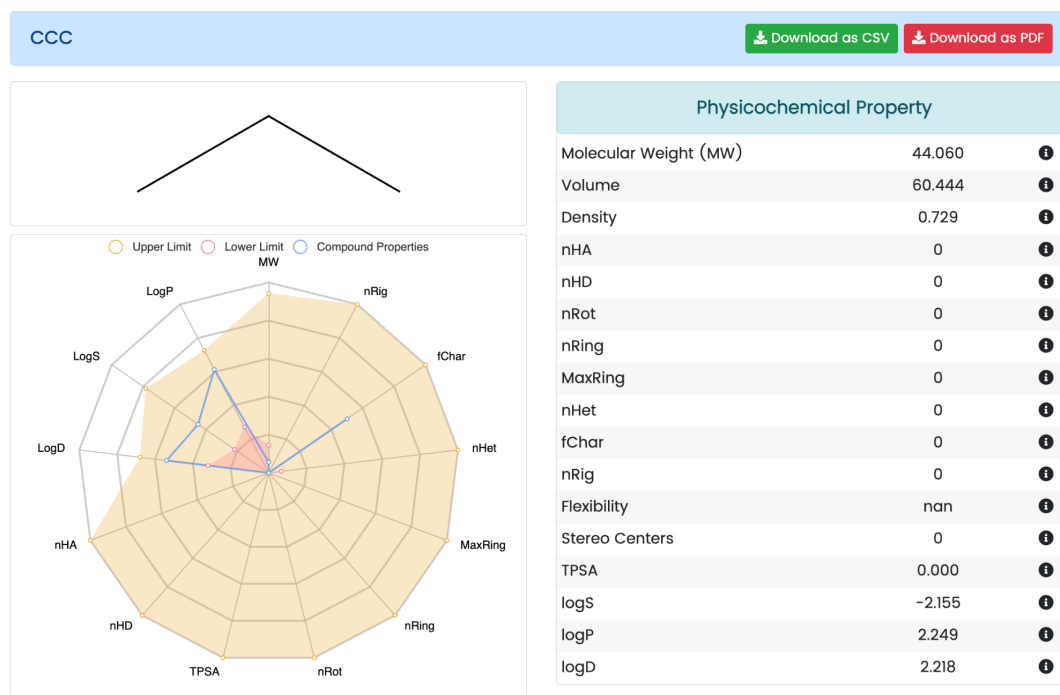


Рисунок 1.2 - Розраховані властивості за допомогою AdmetLab (рисунок створено самостійно)

SwissAdme - це онлайн-інструмент для прогнозування ADME властивостей хімічних сполук. Він надає доступ до різних параметрів та моделей для аналізу та прогнозування характеристик поглинання, розподілу, метаболізму та виведення хімічних сполук. SwissAdme [2] має інтуїтивно зрозумілий інтерфейс, який дозволяє користувачам намалювати структуру за допомогою скетчера MarvinJS або ввести список структур у форматі SMILES (див. рис. 1.3).

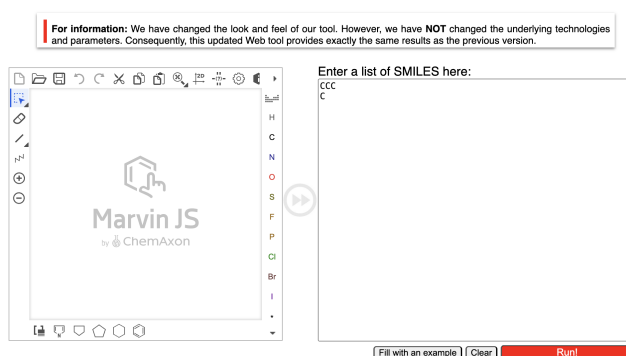


Рисунок 1.3 - Інтерфейс SwissAdme (рисунок створено самостійно)

Результати представлені у формі "властивість-значення", а також у вигляді радара для більш детального аналізу (див. рис. 1.4).

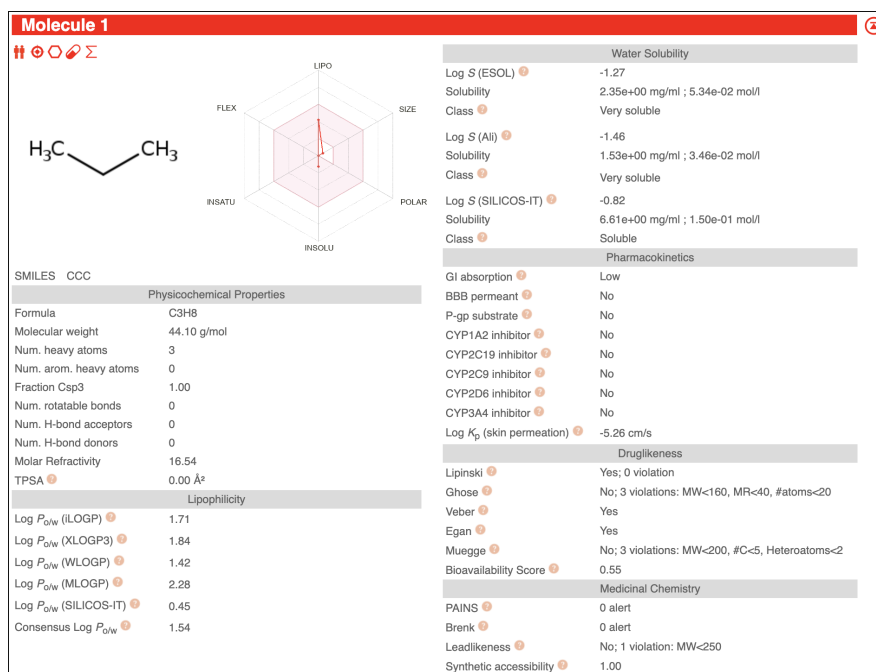


Рисунок 1.4 - Форма результатів SwissAdme (рисунок створено самостійно)

Останнім часом з'явився ще один аналог - ADMET-AI [2], який надає можливість робити прогнози аналогічним чином. Користувачам пропонується ввести список, завантажити файл або намалювати структуру для отримання результатів. Проте, серед його недоліків - обмежена функціональність, оскільки платформа дозволяє робити прогнози лише для 1000 структур, що є надто малою кількістю. Крім того, вона має проблеми з обробкою текстового вводу (див. рис. 1.5) та довгим часом обчислень, що робить її використання менш зручним.

**Predict**

Input up to **1,000** molecules to make ADMET predictions.

Text Input Upload File Draw Molecule Example

SMILES (one per line)

CCC

Predict

**DrugBank Reference**

Create a DrugBank reference set to provide context for ADMET predictions.

Choose an [Anatomical Therapeutic Chemical \(ATC\)](#) code to select a subset of DrugBank approved drugs or choose "all" for all approved drugs.

Selected ATC code "all" includes **2,579** molecules.

Select ATC Code

Error: No valid SMILES strings given.

Рисунок 1.5 - Інтерфейс ADMET-AI (рисунок створено самостійно)

## 2 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ

У попередніх розділах було наведено перелік методів, які можуть використовуватися для вирішення поставленої задачі. Але важливим є обрати той метод, який забезпечить високу точність та продуктивність.

### 2.1 Перелік критеріїв та їх опис

Наведемо перелік критеріїв, за якими буде проведено аналіз:

- точність результатів;

Оцінка точності є важливою для визначення здатності надавати достовірні та точні результати прогнозування властивостей складних сполук.

Цей критерій визначено як найважливіший, оскільки точність є ключовим фактором для надійного використання програмного продукту в області хімічних та фармацевтичних досліджень.

- швидкодія та ефективність обчислень;

Ефективність обчислень та швидкодія є важливими критеріями, оскільки вони визначають продуктивність та придатність продукту для використання в реальному часі.

Забезпечення швидкодії та ефективності обчислень важливо для забезпечення продуктивності та розгляду більшої кількості альтернатив у короткий час.

- масштабованість;

Здатність продукту працювати з різними масштабами даних дозволяє йому впевнено функціонувати як у великих, так і в обмежених за об'ємом дослідженнях.

Масштабованість визначено для того, щоб продукт можна було використовувати як у великих, так і в обмежених за об'ємом дослідженнях.

- актуальність та надійність даних;

Оновлення бази даних та використання актуальних та надійних даних важливо для забезпечення достовірності та актуальності прогнозів.

Актуальність даних та їх надійність визначаються як критичні фактори для забезпечення правильності та достовірності результатів.

– гнучкість та адаптованість.

Гнучкість продукту важлива для використання його в різних умовах та завданнях. Адаптованість до різних типів хімічних сполук визначає його універсальність.

Гнучкість та адаптованість продукту до різних типів хімічних сполук важливі для його універсального використання у різних умовах та дослідженнях.

## 2.2 Векторний опис альтернатив за обраними критеріями

У таблиці нижче (див. табл. 2.1) представлено векторний опис п'яти альтернативних підходів до дослідження моделей та методів прогнозування ADME/T властивостей хімічних сполук [2]. Кожна альтернатива оцінюється за визначеними критеріями, що дозволяє провести порівняльний аналіз їхньої ефективності та придатності для дослідження.

Таблиця 2.1 – Векторний опис альтернатив (таблиця виконана самостійно)

| Альтернатива                | Точність результатів | Швидкість та ефективність обчислень | Масштабованість         | Актуальність та надійність даних | Гнучкість та адаптованість |
|-----------------------------|----------------------|-------------------------------------|-------------------------|----------------------------------|----------------------------|
| Квантово-хімічні розрахунки | Висока точність      | Повільна швидкість обчислень        | Середня масштабованість | Актуальні дані                   | Частково гнучкий           |

Кінець таблиці 2.1

|                                |                  |                             |                          |                                   |                  |
|--------------------------------|------------------|-----------------------------|--------------------------|-----------------------------------|------------------|
| Молекулярне моделювання        | Середня точність | Середня швидкість обчислень | Велика масштабованість   | Дані з оновленням старше 2 тижнів | Повністю гнучкий |
| Машинне навчання               | Середня точність | Висока швидкість обчислень  | Велика масштабованість   | Актуальні дані                    | Повністю гнучкий |
| Інтеграція різних методів      | Висока точність  | Середня швидкість обчислень | Середня масштабованість  | Дані з оновленням старше 2 тижнів | Частково гнучкий |
| Використання експертних оцінок | Низька точність  | Висока швидкість обчислень  | Обмежена масштабованість | Застарілі дані                    | Не гнучкий       |

### 2.3 Перетворення якісних шкал в кількісні

Перетворення якісних шкал в кількісні є необхідним для підвищення точності, об'єктивності та зручності аналізу. Це, в свою чергу, сприяє покращенню якості досліджень та прийняття рішень (див. табл. 2.2).

Таблиця 2.2 – Перетворення якісних шкал (таблиця виконана самостійно)

| Альтернатива                | Точність результатів | Швидкодія та ефективність обчислень | Масштабованість | Актуальність та надійність даних | Гнучкість та адаптованість |
|-----------------------------|----------------------|-------------------------------------|-----------------|----------------------------------|----------------------------|
| Квантово-хімічні розрахунки | 5                    | 1                                   | 3               | 5                                | 3                          |

Кінець таблиці 2.2

|                                          |   |   |   |   |   |
|------------------------------------------|---|---|---|---|---|
| Молекуляр<br>не<br>модельован<br>ня      | 3 | 3 | 5 | 3 | 5 |
| Машинне<br>навчання                      | 3 | 5 | 5 | 5 | 5 |
| Інтеграція<br>різних<br>методів          | 5 | 5 | 3 | 3 | 3 |
| Використан<br>ня<br>експертних<br>оцінок | 1 | 5 | 1 | 1 | 1 |

У цьому векторному описі числові значення кожної альтернативи визначаються за кількісними шкалами, де максимальне значення дорівнює 5.

Усі критерії відповідають принципу – «більше значення - краще», тому додаткові перетворення значень критеріїв не потрібні.

Використовуючи принцип Парето, з таблиці альтернатив було прибрано Використання експертних оцінок через її неконкурентноспроможність у порівнянні з іншими розглянутими моделями. Квантово-хімічні розрахунки було вирішено залишити, незважаючи на загально слабкі показники критеріїв порівняння. Вона не гірше за інших за двома показниками: «Точність результатів» та «Актуальність та надійність даних». Саме тому була залишена для подальшого порівняння (відповідно до принципу «альтернатива має бути не гірше за інші хоча б за одним критерієм»).

## 2.4 Нормування оцінок за шкалами та оцінка корисності

Для забезпечення об'єктивності порівняння альтернатив, ми будемо використовувати метод нормування min-max для кожної шкали. Цей метод

перетворює оцінки таким чином, щоб вони знаходилися в межах визначеного інтервалу (зазвичай від 0 до 1).

Вигляд результуючої таблиці наведено нижче (див. табл. 2.3).

Таблиця 2.3 – Нормування оцінок (таблиця виконана самостійно)

| Альтернатива                | Точність результатів | Швидкість та ефективність обчислень | Масштабованість | Актуальність та надійність даних | Гнучкість та адаптованість |
|-----------------------------|----------------------|-------------------------------------|-----------------|----------------------------------|----------------------------|
| Квантово-хімічні розрахунки | 1                    | 0                                   | 0.5             | 1                                | 0.5                        |
| Молекулярне моделювання     | 0.5                  | 0.5                                 | 1               | 0.5                              | 1                          |
| Машинне навчання            | 0.5                  | 1                                   | 1               | 1                                | 1                          |
| Інтеграція різних методів   | 1                    | 1                                   | 0.5             | 0.5                              | 0.5                        |

У цій таблиці числові значення кожної альтернативи відображають їх нормовані оцінки за кількісними шкалами, де максимальне значення дорівнює 1. Така нормалізація дозволяє провести об'єктивне порівняння альтернатив та визначити їхню конкурентоспроможність за кожним критерієм.

Адитивна згорткова модель є відмінним вибором для вирішення задачі вибору альтернатив у нашому дослідженні. Цей підхід передбачає додавання зважених оцінок за різними критеріями, враховуючи їхню важливість. Одна з основних переваг цієї моделі полягає в її простоті та прозорості. Вагові коефіцієнти можуть легко налаштовуватися для відображення зміни важливості різних критеріїв в залежності від конкретних умов чи уподобань. Адитивна

згорткова модель також відзначається гнучкістю, дозволяючи адаптувати вагові коефіцієнти до різних ситуацій. Ця модель стає ідеальним інструментом для узгодження різноманітних критеріїв в контексті нашого дослідження з метою об'єднання та ранжування альтернатив в області прогнозування ADME/Т властивостей хімічних сполук.

Наведемо вагові коефіцієнти:

- точність результатів ( $W1$ ): 0.25;
- швидкодія та ефективність обчислень ( $W2$ ): 0.1875;
- масштабованість ( $W3$ ): 0.125;
- актуальність та надійність даних ( $W4$ ): 0.25;
- гнучкість та адаптованість ( $W5$ ): 0.1875.

Для розрахунку корисності альтернатив за обраною загортковою моделлю, ми будемо використовувати адитивну згорткову модель із ваговими коефіцієнтами. Розрахунок відбувається за формулою 2.1:

$$Z^* = \max \sum_{j=1}^n \alpha_j \beta_j a_j \quad (2.1)$$

де  $\alpha_j$  – нормуючі множники,

$\beta_j$  – вагові коефіцієнти, що відображають відносний внесок окремих критеріїв до загального критерію.

Нижче наведено результуючу таблицю (див. табл. 2.4) з розрахунками корисності для кожної альтернативи за обраною загортковою моделлю.

Таблиця 2.4 – Оцінка корисності (таблиця виконана самостійно)

| Альтернатива | Точність результатів | Швидкодія та ефективність обчислень | Масштабованість | Актуальність та надійність даних | Гнучкість та адаптованість | $Z^*$ |
|--------------|----------------------|-------------------------------------|-----------------|----------------------------------|----------------------------|-------|
|              |                      |                                     |                 |                                  |                            |       |

Кінець таблиці 2.4

|                             |     |     |     |     |     |         |
|-----------------------------|-----|-----|-----|-----|-----|---------|
| Квантово-хімічні розрахунки | 1   | 0   | 0.5 | 1   | 0.5 | 0.65625 |
| Молекулярне моделювання     | 0.5 | 0.5 | 1   | 0.5 | 1   | 0.62625 |
| Машинне навчання            | 0.5 | 1   | 1   | 1   | 1   | 0.875   |
| Інтеграція різних методів   | 1   | 1   | 0.5 | 0.5 | 0.5 | 0.71875 |

На основі оцінки кінцевих альтернатив із використанням згорткової моделі з ваговими коефіцієнтами, було визначено, що машинне навчання та інтеграція різних методів є найбільш конкурентоспроможними, забезпечуючи високі значення корисності. Молекулярне моделювання та квантово-хімічні розрахунки показали помірні результати, тоді як використання експертних оцінок продемонструвало найменшу корисність серед варіантів. Ці результати оцінювання можуть слугувати важливими вказівками для прийняття рішень щодо вибору оптимального методу для прогнозування властивостей хімічних сполук. Таким чином, на основі вагових коефіцієнтів та корисності кожної альтернативи, машинне навчання та інтеграція різних методів є найбільш відповідними для вирішення поставленої задачі [10].

### 3 ПОСТАНОВКА ЗАДАЧІ

Після того як ми провели аналіз проблемної області дослідження та обрали машинне навчання для дослідження, можна сформулювати список задач для проведення дослідження. Отже, для проведення дослідження необхідно вирішити наступні задачі:

- зібрати актуальні хімічні дані для передбачення метаболізму з таких баз даних як PubChem, ChEMBL та інших;
- розробити пайплайн для обробки складних структур (визначитися з правилами для навчальних та прогнозованих даних, вирішити проблему сумісності та стандартизації даних з різних джерел);
- розглянути поточні рішення (з використанням графових нейронних мереж, фреймворку MGA);
- провести експериментальне дослідження обраних рішень та розробити рекомендації щодо використання того чи іншого рішення;
- запропонувати рішення краще та розробити API, який можна буде інтегрувати з іншими платформами.

#### 3.1 Збір хімічних даних

Збір даних є фундаментальним етапом у проведенні досліджень у галузі кількісних зв'язків структура-активність, що має велике значення для розвитку фармацевтичної науки та промисловості. У нашому дослідженні ми використали ряд відкритих баз даних, які містять інформацію про хімічні сполуки та їх властивості, що є критичними для проведення аналізу QSAR [7].

Однією з ключових баз є PubChem, що є величезною колекцією хімічних сполук з даними про їхні властивості, синтез та біологічну активність. PubChem є важливим ресурсом для дослідників у галузі медичної хімії, фармацевтики та біології, оскільки надає доступ до великої кількості даних про хімічні сполуки, їхні властивості та взаємодії з біологічними системами.

ChEMBL, інша важлива база даних, спеціалізується на біологічних даних про хімічні сполуки. Вона включає інформацію про біологічну активність та механізми дії сполук, що є важливим для розуміння взаємодії хімічних речовин з живими системами.

DrugBank - це ще один важливий ресурс, який надає дані про лікарські засоби та їх взаємодію з білками та метаболізм у людському організмі. Ця база даних дозволяє дослідникам отримати доступ до інформації про властивості лікарських засобів та їх вплив на організм.

Додатково, для нашого дослідження ми використали базу Zinc, яка є важливим джерелом інформації про хімічні сполуки, що можуть мати лікарське застосування. Zinc надає доступ до великої кількості хімічних сполук, які можуть бути потенційними лікарськими засобами.

Також ми звернулися до бази CAS, яка містить ідентифікатори хімічних сполук та дозволяє швидко знаходити інформацію про них. Більш того, ми використали інші бази даних, такі як BindingDB та DSSTox, для отримання додаткової інформації про хімічні сполуки та їх властивості.

Збір даних з цих відкритих баз даних надав нам можливість отримати доступ до широкого спектру хімічних сполук та їх властивостей, що є важливим для проведення аналізу QSAR. Використання відкритих баз даних має значний науковий та практичний потенціал, оскільки воно сприяє швидкому обміну знаннями та співпраці між дослідниками у галузі хімії та фармації. Такі джерела даних допомагають покращити якість досліджень та сприяють розвитку нових технологій та лікарських засобів.

### 3.2 Пайплайн для обробки складних структур

Підготовка даних є важливим кроком у дослідженнях. Тож, ми розробили методологію, яка допомагає забезпечити точність прогнозів моделей [8].

Ми використовуємо однакові правила для навчальних та прогнозованих даних, для того щоб забезпечити єдність у процесі. При створенні моделей ми

ретельно вивчаємо експериментальне середовище джерела даних для підвищення надійності моделей.

Під час прогнозування ми рекомендуємо перевіряти область застосування моделі для забезпечення актуальності та надійності прогнозів.

Перший етап обробки даних забезпечує сумісність та стандартизацію даних з різних джерел для моделювання. Ми видаляємо дані, що не мають відношення до структури та властивостей/активності та додаємо додаткові дані, такі як тип джерела, одиниці виміру та молекулярні мішені. Нижче, наведено таблицю (див. табл. 3.1) з переліком додаткових даних.

Таблиця 3.1 – Перелік додаткових даних (таблиця виконана самостійно)

| Параметр даних | Приклад                                                  | Призначення                                                                                                                                                   | Можливі назви       |
|----------------|----------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Тип джерела    | Патент, рукопис, комерційна база даних з аналізами       | Надає цінні відомості, оскільки ці категорії зазвичай досліджують різні класи хімічних сполук, з помітними відмінностями як у якості, так і в кількості даних | Література, джерело |
| Одиниці виміру | моль/кг, моль/л                                          | Перетворює одиниці на стандарт                                                                                                                                | Одиниці             |
| Відношення     | Дані представлені з відношеннями типу $\geq$ , $<$ , $=$ | Залежно від завдання, ці дані можна використовувати або вилучати                                                                                              | Відношення          |

Кінець таблиці 3.1

|                    |                |                                                                                                                                                                   |                                   |
|--------------------|----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| Тип кінцевої точки | Inh%, IC50, Ki | Допомагає вирішити різноманітні виклики в розвитку моделі за допомогою визнання та врахування відмінностей у кінцевих точках, вимірах та експериментальних умовах | PubChem: назва групи, ціль/мішень |
| Молекулярна мішень | CYPs, hERG     | Визначає, чи пов'язані дані зі спільною біологічною мішенню                                                                                                       | Молекулярна мішень                |

Другий етап - це етап обробки структур. До нього входить:

- вилучення некоректних структур. Ми використовуємо бібліотеку RDKit для визначення, чи є структура правильною;
- вилучення структур з рідкісними елементами. Якщо немає конкретних перешкод для збереження структур з рідкісними елементами, ми видаляємо всі структури, що містять:

"R ", "D ", "He ", "Li ", "Be ", "B ", "Ne ", "Al ", "Si ", "Ar ", "Sc ", "Ti ", "V ", "Cr ", "Co ", "Ni ", "Ga ", "Ge ", "As ", "Se ", "Kr ", "Rb ", "Sr ", "Y ", "Zr ", "Nb ", "Mo ", "Tc ", "Ru ", "Rh ", "Pd ", "Ag ", "Cd ", "In ", "Sn ", "Sb ", "Te ", "Xe ", "Cs ", "Ba ", "La ", "Ce ", "Pr ", "Nd ", "Pm ", "Sm ", "Eu ", "Gd ", "Tb ", "Dy ", "Ho ", "Er ", "Tm ", "Yb ", "Lu ", "Hf ", "Ta ", "W ", "Re ", "Os ", "Ir ", "Pt ", "Au ", "Hg ", "Tl ", "Pb ", "Bi ", "Po ", "At ", "Rn ", "Fr ", "Ra ", "Ac ", "Th ", "Pa ", "U ", "Np ", "Pu ", "Am ", "Cm ", "Bk ", "Cf ", "Es ", "Fm ", "Md ", "No ", "Lr ", "Rf ", "Db ", "Sg ", "Bh ", "Hs ", "Mt ", "Ds ", "Rg ", "Cn ", "Nh ", "Fl ", "Mc ", "Lv ", "Ts ", "Og "

- вибір головного компонента. Бажані властивості зазвичай залежать від однієї структури. Записи в наборі даних, що представляють собою суміші

та солі, містять більше одного структурного компонента. Це може значно вплинути на створення дескрипторів/відбитків. Суміші та окремі молекули матимуть різні набори ознак. Залежно від задачі, це не завжди є важливим для солей (2 фрагменти, один з них - іон одного атома). Солі повинні оброблятися шляхом вибору головного фрагмента і наступною нейтралізацією (видаленням заряду) для більшості моделей, пов'язаних з молекулярною мішенню (наприклад, hERG), і не для властивостей;

- нейтралізація. Для більшості задач ми нейтралізуємо молекули, оскільки їх заряд значно залежить від біологічного/хімічного середовища;
- стандартизація структурних формул. Різні хімічні типи можуть бути представлені по-різному в різних хімічних нотаціях. Ми виконуємо стандартизацію, щоб отримати однакові дескриптори/відбитки, отримані з таких груп;
- рівна ароматизація. Ми перетворюємо весь набір даних до рівної ароматичної або кекулівської форми;
- рівна нормалізація. Для всіх активностей на макромолекулярні мішені і для більшості властивостей (крім pKa) ми виконуємо нормалізацію (не плутати з нейтралізацією або реіонізацією). Нормалізація зберігає загальний заряд молекули;
- рівна таутомеризація. У більшості випадків весь набір даних повинен містити рівну таутомерну форму групи, щоб уникнути відмінного створення дескрипторів/відбитків. Це обґрунтовано тим, що обидві форми зазвичай присутні в реальних системах, але призводять до різного створення дескрипторів/відбитків;
- обробка стерео інформації. Ми зберігаємо стерео інформацію лише тоді, коли всі дані її містять, в іншому випадку виключаємо її;
- обробка інформації про водень. Присутність або відсутність водню в структурах призводить до різного створення дескрипторів/відбитків. Ми включаємо або додаємо експліцитний водень до всіх структур у наборі даних залежно від аналізованої проблеми.

На рисунку (див. рис. 3.1) нижче наведений приклад такої обробки.

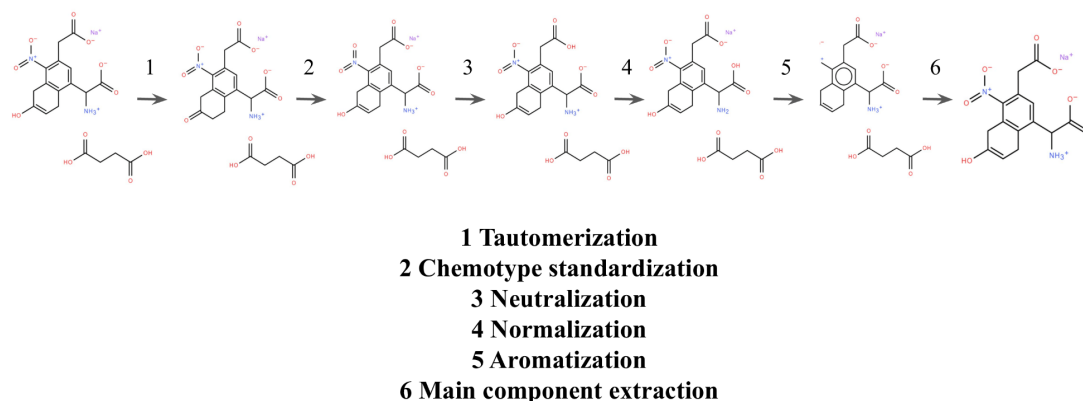


Рисунок 3.1 - Обробка хімічної структури (рисунок створено самостійно)

Обробка експериментальних даних також включає видалення неправильно позначених або нерелевантних даних та обробку дублікатів, що можуть виникнути через кілька значень властивостей/активності для однієї структури.

З огляду на все вищеперелічене, було розроблено скрипт на мові програмування Python з використанням наступних бібліотек: pandas, numpy, rdkit, pubchempy, nltk та інших. Наведемо нижче метод для стандартизації структур:

```
def standardize_molecule(mol):
    # Neutralization
    def neutralize_atoms(mol):
        pattern = Chem.MolFromSmarts("[+1!h0!$([*]~[-1,-2,-3,-4]),-1!$([*]~[+1,+2,+3,+4])]")
        at_matches = mol.GetSubstructMatches(pattern)
        at_matches_list = [y[0] for y in at_matches]
        if len(at_matches_list) > 0:
            for at_idx in at_matches_list:
                atom = mol.GetAtomWithIdx(at_idx)
                chg = atom.GetFormalCharge()
                hcount = atom.GetTotalNumHs()
                atom.SetFormalCharge(0)
                atom.SetNumExplicitHs(hcount - chg)
                atom.UpdatePropertyCache()
        return mol

    mol = neutralize_atoms(mol)

    # Standardization of structural formulas
    mol = rdMolStandardize.Reionize(mol)
    mol = rdMolStandardize.Normalize(mol)
```

```

# Equal tautomerization
enumerator = rdMolStandardize.TautomerEnumerator()
mol = enumerator.Canonicalize(mol)

# Handling stereo information
def handle_stereochemistry(mol):
    if Chem.MolToSmiles(mol, isomericSmiles=True) !=
Chem.MolToSmiles(mol, isomericSmiles=False):
        return mol
    else:
        Chem.RemoveStereochemistry(mol)
        return mol

mol = handle_stereochemistry(mol)

# Handling hydrogen information
def handle_hydrogens(mol, exclude=False):
    if exclude:
        Chem.RemoveHs(mol)
    else:
        Chem.AddHs(mol)
    return mol

mol = handle_hydrogens(mol)

# Removal structures with rare elements
elements_to_remove = ["R", "D", "He", "Li", "Be", "B", "Ne",
"Al", "Si", "Ar", "Sc", "Ti", "V", "Cr", "Co",
                        "Ni", "Ga", "Ge", "As", "Se", "Kr", "Rb",
"Sr", "Y", "Zr", "Nb", "Mo", "Tc", "Ru", "Rh",
                        "Pd", "Ag", "Cd", "In", "Sn", "Sb", "Te",
"Xe", "Cs", "Ba", "La", "Ce", "Pr", "Nd", "Pm",
                        "Sm", "Eu", "Gd", "Tb", "Dy", "Ho", "Er",
"Tm", "Yb", "Lu", "Hf", "Ta", "W", "Re", "Os",
                        "Ir", "Pt", "Au", "Hg", "Tl", "Pb", "Bi",
"Po", "At", "Rn", "Fr", "Ra", "Ac", "Th", "Pa",
                        "U", "Np", "Pu", "Am", "Cm", "Bk", "Cf",
"Es", "Fm", "Md", "No", "Lr", "Rf", "Db", "Sg",
                        "Bh", "Hs", "Mt", "Ds", "Rg", "Cn", "Nh",
"Fl", "Mc", "Lv", "Ts", "Og"]

    if any(atom.GetSymbol() in elements_to_remove for atom in
mol.GetAtoms()):
        return None
    return mol

```

### 3.3 Аналіз зібраних даних

Для проведення експериментів було зібрано дані для метаболізму [6].

Метаболізм - це складна біохімічна система, відповідальна за управління енергетичними та молекулярними процесами в живих організмах. Вона

складається з двох основних компонентів: анаболізму, де будуються складні молекули, вимагаючи енергії, схожої на інвестицію, та катаболізму, який розкладає речовини для вивільнення енергії, подібно до контрольованого хімічного розкладу. Цей високо регульований процес під впливом різних факторів, таких як генетика, вік та вибір способу життя, впливає на використання енергії, зростання, та загальний життєвий тонус у формальному та спрощеному вигляді.

Більшість даних узяті з бази даних PubChem AID 1851, яку ми розглядали раніше. Нижче наведено таблицю (див. табл. 3.2) з описом частини даних.

Таблиця 3.2 – Огляд зібраних даних (таблиця виконана самостійно)

| Назва            | Опис                                                                                                                                                                                           | Обробка                                                                                                                                                                                                                                                               | Кількість | Позитивні | Негативні |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-----------|-----------|
| CYP1A2-Inhibitor | Це ймовірність бути інгібітором цитохрому CYP1A2, який каталізує метаболізм ліків. Інгібування CYP1A2 може бути як бажаною, так і не бажаною властивістю в залежності від цілей розробки ліків | Завантаження<br>Фільтрування за «panel name» у p450-cyp1a2<br>Видалення недостовірних записів у «PUBCHEM_ACTIVITY_OUTCOME»<br>Фільтрування TRUE за «Inhibition Observed»<br>Встановлення Activity = 1, якщо «PUBCHEM_ACTIVITY_OUTCOME» є Active, і 0 у іншому випадку | 9<br>1323 | 5997      | 7242      |

Кінець таблиці 3.2

|                  |                                                                                                                                                                                                |                                                                                                                                                                                                                                                           |   |      |      |      |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|------|------|------|
| CYP3A4-Inhibitor | Це ймовірність бути інгібітором цитохрому CYP3A4, який каталізує метаболізм ліків. Інгібування CYP3A4 може бути як бажаною, так і не бажаною властивістю в залежності від цілей розробки ліків | Завантаження Фільтрування за «panel name» у p450-cyp3a4 Видалення недостовірних записів у «PUBCHEM_ACTIVITY_OUTCOME» Фільтрування TRUE за «Inhibition Observed» Встановлення Activity = 1, якщо «PUBCHEM_ACTIVITY_OUTCOME» є Active, і 0 у іншому випадку | 7 | 1299 | 5265 | 7732 |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|------|------|------|

### 3.4 Огляд поточних рішень

У попередніх розділах ми наголосили на тому, що саме машинне навчання є найбільш відповідним для вирішення задачі оцінки властивостей складних структур. Розглянемо декілька state-of-the-art рішень, якими вже користуються.

#### 3.4.1 Chemprop-RDKit6

Глибоке навчання стало потужним інструментом у сфері прогнозування властивостей хімічних сполук, що відкриває нові можливості у розробці лікарських засобів та хімічних дослідженнях. Один із передових підходів - використання архітектури Chemprop-RDKit6 [8] для прогнозування параметрів ADMET (абсорбція, розподіл, метаболізм, виділення та токсичність).

Chemprop-RDKit6 використовує графову нейронну мережу, яка аналізує хімічні структури молекул. Ця архітектура базується на глибокому навчанні та використовує пакет хемінформатики RDKit для обчислення різноманітних фізико-хімічних ознак для кожного атома та зв'язку у молекулі. Після обчислення ознак графова нейронна мережа виконує декілька кроків передачі повідомлень

між атомами та зв'язками для агрегації цієї інформації та побудови представлення для всієї молекули.

Однією з ключових особливостей Chemprop-RDKit6 є використання 200 фізико-хімічних ознак, які обчислюються за допомогою пакету RDKit. Ці ознаки доповнюють інформацію, отриману з графової нейронної мережі, і розширюють можливості моделі для більш точного прогнозування параметрів ADMET.

Архітектура Chemprop-RDKit6 (див. рис. 3.2) є потужним інструментом для прогнозування параметрів ADMET хімічних сполук. Використання глибокого навчання та фізико-хімічних ознак RDKit дозволяє отримувати точні та надійні прогнози, однак велика кількість параметрів може вимагати значних обчислювальних ресурсів.

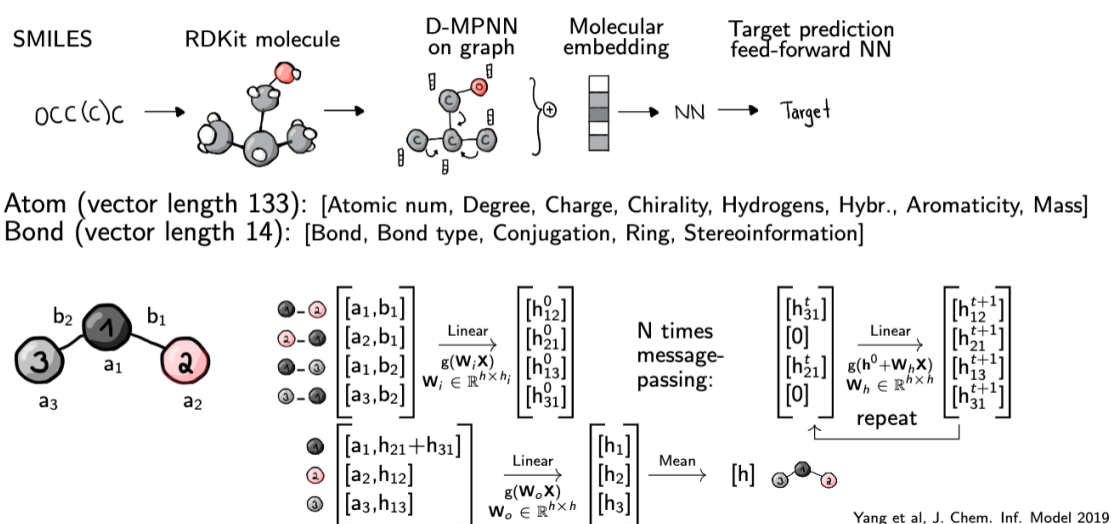


Рисунок 3.2 - Архітектура Chemprop-RDKit6 (за даними [8])

### 3.4.2 Фреймворк MGA

Традиційні методи, такі як графові нейронні мережі (GNN), зазвичай вирішують лише однорідні завдання, такі як регресія або класифікація. Проте, для прогнозування параметрів ADMET потрібно одночасно вирішувати регресійні та класифікаційні завдання.

У даному дослідженні ми розглянемо фреймворк MGA, який дозволяє моделі вирішувати різні типи завдань одночасно та краще моделювати складні залежності у даних ADMET. MGA використовує графову увагу для атомів, що дозволяє враховувати структурні особливості молекул та отримувати більш точні прогнози. Крім того, він використовує комбінований втратний шар для класифікаційних і регресійних завдань, що дозволяє ефективно навчати модель на різних типах даних ADMET.

Проаналізуємо переваги та недоліки фреймворку MGA у контексті прогнозування параметрів ADMET (див. табл. 3.3).

Таблиця 3.3 – Переваги та недоліки фреймворку MGA (таблиця виконана самостійно)

| Переваги                                                                         | Недоліки                                                                          |
|----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Ефективне одночасне вирішення регресійних і кваліфікаційних завдань              | Потреба у великій кількості даних для навчання ефективної моделі                  |
| Здатність враховувати структурні особливості молекул за допомогою графової уваги | Складність налаштування гіперпараметрів для досягнення оптимальної продуктивності |
| Можливість налаштування для різних завдань прогнозування ADMET                   |                                                                                   |

Розглянемо пайплайн роботи MGA на рисунку нижче (див. рис. 3.3).

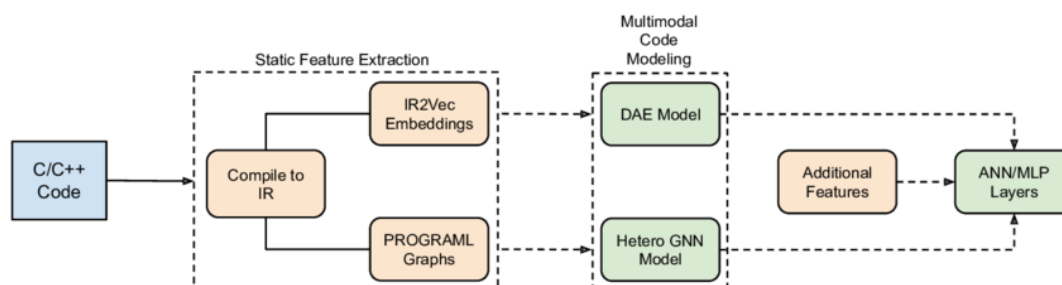


Рисунок 3.3 - Пайплайн роботи MGA (за даними [1])

### 3.4.3 Моделі на основі дескрипторів

Для моделювання властивостей ADME/T [4] застосовуються молекулярні дескриптори та відбитки, що дозволяють побудувати як регресійні, так і класифікаційні моделі.

Зазвичай використовується 11 типів широко використовуваних дескрипторів, таких як топологія, Kappa, basak, burden, autocorrelation, charge, property, дескриптори MOE-типу та інші. Усі ці дескриптори обчислюються за допомогою пакету Chemopy на мові програмування Python. Крім того, використовуються відбитки, такі як FP2, MACCS, ECFP2, ECFP4 та ECFP6, які обчислюються за допомогою пакетів ChemDes і BioTriangle. Ці дескриптори використовуються для побудови регресійних і класифікаційних моделей.

Перед побудовою регресійних моделей проводяться три кроки попереднього відбору дескрипторів, щоб уникнути використання неінформативних дескрипторів. Спочатку видаляються дескриптори, у яких дисперсія нульова або дуже мала. Далі видаляються дескриптори, у яких велика частка однакових значень (понад 95%). Нарешті, якщо кореляція двох дескрипторів більше 0,95, один з них випадково видаляється. Останній набір дескрипторів використовується для подальшого відбору дескрипторів і побудови моделей QSAR.

Для побудови QSAR регресійних або класифікаційних моделей використовуються шість різних алгоритмів: випадковий ліс (RF), машина опорних векторів (SVM), рекурсивне розділення регресії (RP), часткове найменше квадратів (PLS), наївний Баєс (NB) і дерево рішень (DT). Вони використовуються для моделювання регресійних та класифікаційних задач. Перед побудовою моделей параметри кожного з алгоритмів оптимізуються для кращої продуктивності за допомогою кросс-валідації.

Таким чином, використання молекулярних дескрипторів та відбитків у поєднанні з різними моделями також дозволяє ефективно моделювати (див. рис. 3.4) властивості ADME/T.

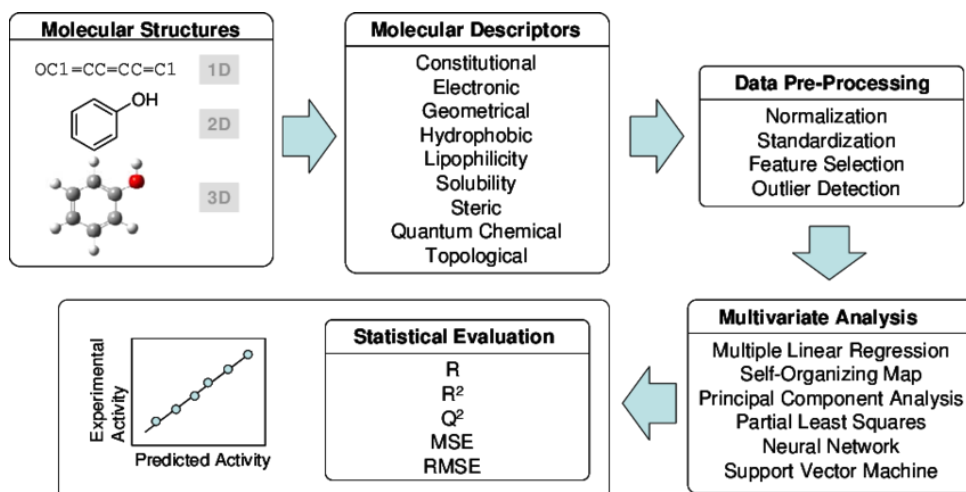


Рисунок 3.4 - Принцип моделювання з використанням дескрипторів (за даними [4])

### 3.5 Експериментальне дослідження обраних рішень

#### 3.5.1 Метрики для валідації моделей

Для оцінки прогностичної здатності моделей на перехресному контролі були використані критерії, такі як збалансована точність та площа під ROC-кривою при кросс-валідації. Ці критерії дозволяють визначити ефективність моделей та порівняти їх точність.

Збалансована точність при кросс-валідації визначається як середнє арифметичне точності для кожного класу за формулою 3.1:

$$BA = \frac{1}{n_{classes}} \sum_{i=1}^{n_{classes}} \frac{TP_i}{TP_i + FN_i}, \quad (3.1)$$

де  $TP_i$  - це кількість вірно передбачених позитивних прикладів для класу

$FN_i$  - це кількість хибно відкинутих прикладів для класу  $i$

Площа під ROC-кривою при кросс-валідації обчислюється як інтеграл від ROC-кривої за формулою 3.2, яка відображає залежність чутливості від специфічності при різних значеннях порогу класифікації. ROC-крива створюється шляхом розрахунку TPR та FPR для різних значень порогу класифікації. Площа під ROC-кривою допомагає визначити загальну прогностичну здатність моделі, незалежно від конкретного порогу класифікації.

$$AUC = \int_0^1 TPR(FPR^{-1}(t)) dt \quad 3.2)$$

де  $TPR$  - чутливість

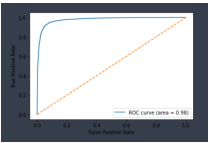
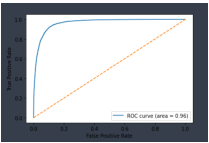
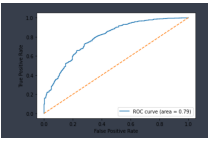
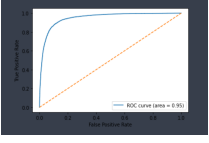
$FPR$  - специфічність

$FPR^{-1}(t)$  - обернена функція до функції, що повертає значення  $FPR$  при заданому порозі  $t$ .

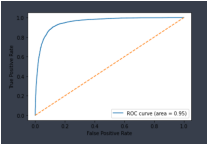
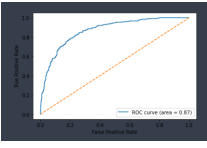
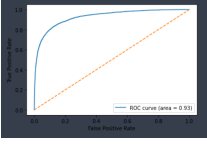
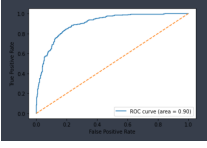
### 3.5.2 Результати експериментів з Chemprop-RDKit6

Було проведено ряд експериментів з використанням Chemprop-RDKit6 для оцінки метаболізму. Таблицю (див. табл. 3.4) з розрахованими метриками наведено нижче.

Таблиця 3.4 – Результати експериментів з Chemprop-RDKit6 (таблиця виконана самостійно)

| Назва                 | Особливість | Чутливість | Точність | Збалансована точність | ROC AUC                                                                               |
|-----------------------|-------------|------------|----------|-----------------------|---------------------------------------------------------------------------------------|
| CYP1A<br>2-Inhibitor  | 0.899       | 0.956      | 0.925    | 0.928                 |  |
| CYP3A<br>4-Inhibitor  | 0.861       | 0.949      | 0.896    | 0.905                 |  |
| CYP3A<br>4-Substrate  | 0.562       | 0.857      | 0.776    | 0.709                 |  |
| CYP2C<br>19-Inhibitor | 0.887       | 0.890      | 0.888    | 0.889                 |  |

Кінець таблиці 3.4

|                  |       |       |       |       |                                                                                     |
|------------------|-------|-------|-------|-------|-------------------------------------------------------------------------------------|
| CYP2C9-Inhibitor | 0.899 | 0.880 | 0.893 | 0.890 |  |
| CYP2C9-Substrate | 0.831 | 0.730 | 0.790 | 0.781 |  |
| CYP2D6-Inhibitor | 0.723 | 0.926 | 0.763 | 0.824 |  |
| CYP2D6-Substrate | 0.853 | 0.783 | 0.819 | 0.818 |  |

Загально, результати моделі показують високий рівень чутливості та особливості для більшості класів прогнозування, що вказує на те, що модель добре розпізнає як позитивні, так і від'ємні випадки. Однак точність моделі для деяких класів може бути не такою високою, що свідчить про можливість помилкових класифікацій. Збалансована точність також показує, що модель є здатною до адекватного розпізнавання класів, враховуючи їхню нерівну кількість у вибірці.

### 3.5.3 Результати експериментів з моделями на основі дескрипторів

У рамках дослідження було застосовано метод опорних векторів (SVM) для розв'язання задачі оцінки метаболізму. Було проведено ряд експериментів, де використовувалися різноманітні дескриптори для побудови класифікаційних моделей та оцінки їхньої ефективності. У таблиці нижче наведено отримані результати (див. табл. 3.5).

Таблиця 3.5 – Результати експериментів з дескрипторами (таблиця виконана самостійно)

| Назва             | Метод         | Дескриптори | Особливість | Чутливість | Точність | Збалансована точність |
|-------------------|---------------|-------------|-------------|------------|----------|-----------------------|
| CYP1A2-Inhibitor  | SVM           | ECFP4       | 0.864       | 0.928      | 0.849    | 0.896                 |
| CYP3A4-Inhibitor  | SVM           | ECFP4       | 0.858       | 0.759      | 0.817    | 0.809                 |
| CYP3A4-Substrate  | Random Forest | ECFP4       | 0.716       | 0.798      | 0.757    | 0.757                 |
| CYP2C19-Inhibitor | SVM           | ECFP2       | 0.819       | 0.826      | 0.822    | 0.822                 |
| CYP2C9-Inhibitor  | SVM           | ECFP4       | 0.898       | 0.719      | 0.837    | 0.809                 |
| CYP2C9-Substrate  | Random Forest | ECFP4       | 0.709       | 0.746      | 0.728    | 0.728                 |
| CYP2D6-Inhibitor  | SVM           | ECFP4       | 0.811       | 0.770      | 0.793    | 0.791                 |
| CYP2D6-Substrate  | Random Forest | ECFP4       | 0.73        | 0.765      | 0.748    | 0.748                 |

Метод класифікації SVM, застосований до набору дескрипторів ECFP4, демонструє досить високі значення особливості, чутливості та точності для більшості ізоферментів, зокрема для CYP1A2-Inhibitor, CYP3A4-Inhibitor, CYP2C19-Inhibitor та CYP2D6-Inhibitor. Однак, для деяких ізоферментів, таких як CYP2C9-Inhibitor, спостерігається низька чутливість, що може вказувати на

нестабільність моделі. Метод Random Forest з використанням дескрипторів ECFP4 також показує деякі добрі результати, особливо для завдань класифікації CYP3A4-Substrate та CYP2C9-Substrate.

### 3.6 Розробка API для інтеграції

Chemprop показав себе краще у порівнянні з іншими моделями, демонструючи високий рівень чутливості та точності для більшості класів прогнозування. Зважаючи на це, ми вирішили удосконалити наші моделі, створені за допомогою Chemprop, шляхом оптимізації гіперпараметрів та розширення вибірки даних.

Для покращення доступності та зручності використання наших вдосконалених моделей, ми розробили API, реалізований на мові програмування Python з використанням фреймворку Flask. Цей API забезпечує надійний інтерфейс для прогнозування властивостей хімічних сполук.

Основний ендпоінт API, «df\_upload/», приймає датафрейм, що містить структури молекул у форматах SMILES та molblocks, а також список властивостей, які необхідно спрогнозувати. На виході API надає спрогнозовані властивості разом з інформацією про точність цих передбачень. Важливо зазначити, що API також оцінює, наскільки дані структури належать до області застосовності (applicability domain) моделі та чи входять вони до тренувальної вибірки. Це критично важливо для забезпечення надійності та точності результатів експериментів.

Для забезпечення зручного розгортання та масштабування, наше рішення підтримує контейнеризацію. Нижче буде представлено ілюстрацію, що демонструє результати прогнозування та приклади вхідних даних, які подаються на вхід API (див. рис. 3.5).

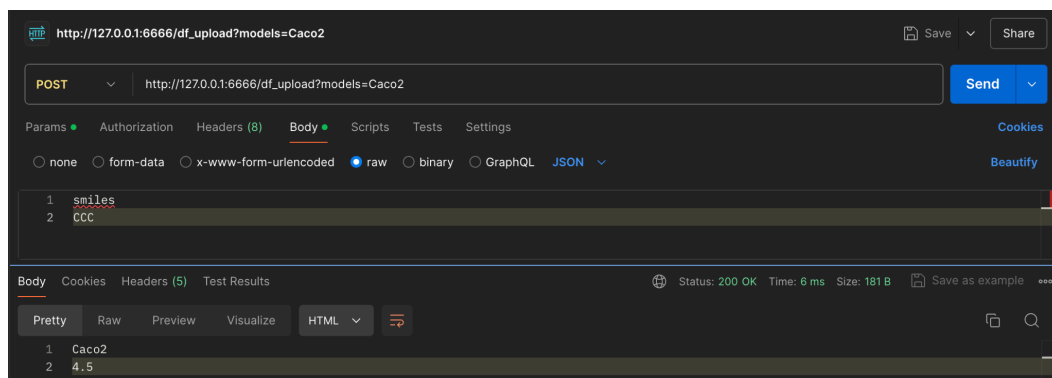


Рисунок 3.5 - Вигляд API (рисунок створено самостійно)

Наступний API вже використовується та інтегрований у платформу Datagrok (див. рис. 3.6). Завдяки API, користувачі можуть завантажувати структури молекул, отримувати точні передбачення та аналізувати результати в рамках єдиної платформи, що забезпечує зручність та ефективність у роботі з хімічними даними.



Рисунок 3.6 - Інтеграція API у платформу Datagrok (рисунок створено самостійно)

## ВИСНОВКИ

В ході виконання кваліфікаційної роботи було визначено актуальність проблеми оцінки властивостей складних хімічних структур та необхідність ефективних методів для їх аналізу. Було проведено всебічний аналіз методів визначення властивостей, включаючи квантово-хімічні розрахунки, молекулярне моделювання та машинне навчання.

Адитивна згорткова модель показала, що найефективнішим методом для прогнозування хімічних властивостей є машинне навчання. Крім того, було детально досліджено та порівняно різні підходи, включаючи Chemprop-RDKit6, фреймворк MGA та моделі на основі фінгерпринтів, визначивши їхні переваги та обмеження.

Ця робота закладає міцну основу для подальших досліджень у галузі прогнозування хімічних властивостей, демонструючи ефективність машинного навчання та можливості його застосування у хімічних дослідженнях. Розробка та інтеграція API у платформу Datagrok є важливим кроком вперед, забезпечуючи зручний доступ до вдосконалених моделей та сприяючи подальшому розвитку методів аналізу хімічних даних.

## ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database / Jie Dong, Ning-Ning Wang, Zhi-Jiang Yao та ін. // J Cheminform. – 2018. – Режим доступу до ресурсу: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020094/>.
2. Evaluation of Free Online ADMET Tools for Academic or Small Biotech Environments / Júlia Dulsat, Blanca López-Nieto, Roger Estrada-Tejedor, José I. Borrell // Molecules. – 2023. – Режим доступу до ресурсу: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9864198/>.
3. Vishwesh Venkatraman. FP-ADMET: a compendium of fingerprint-based ADMET prediction models / Vishwesh Venkatraman // J Cheminform. – 2021. – <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8479898/>.
4. Front Pharmacol. vNN Web Server for ADMET Predictions / Front Pharmacol // Front Pharmacol. – 2017. – Режим доступу до ресурсу: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5722789/>.
5. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties / Guoli Xiong, Zhenxing Wu, Jiakai Yi та ін. // Nucleic Acids Res. – 2021. – Режим доступу до ресурсу: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8262709/>.
6. In silico Prediction of Chemical Ames Mutagenicity / Congying Xu, Feixiong Cheng, Lei Chen та ін. // J Cheminform. – 2012. – Режим доступу до ресурсу: <https://pubs.acs.org/doi/abs/10.1021/ci300400a>.
7. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope / Denis Mulliner, Friedemann Schmidt, Manuela Stolte та ін. // Chem. Res. Toxicol.. – 2016. – Режим доступу до ресурсу: <https://pubs.acs.org/doi/10.1021/acs.chemrestox.5b00465>.
8. Analyzing Learned Molecular Representations for Property Prediction [Електронний ресурс] / Kevin Yang, Kyle Swanson, Wengong Jin та ін.]. – 2019. – Режим доступу до ресурсу: <https://pubs.acs.org/doi/10.1021/acs.jcim.9b00237>.
9. Turuta O. Collection of questionnaire results, received by using the visual

analog scale method, for its further processing in the medical web application [Электронний ресурс] / О. Turuta, Y. Daniil // ScienceRise. – 2017. – Режим доступу до ресурсу: [https://www.researchgate.net/publication/317612987\\_Collection\\_of\\_questionnaire\\_results\\_received\\_by\\_using\\_the\\_visual\\_analog\\_scale\\_method\\_for\\_its\\_further\\_processing\\_in\\_the\\_medical\\_web\\_application](https://www.researchgate.net/publication/317612987_Collection_of_questionnaire_results_received_by_using_the_visual_analog_scale_method_for_its_further_processing_in_the_medical_web_application) (дата звернення: 20.03.2024).

10. Intelligent information system of heterogeneous medical data analysis [Электронний ресурс] / А. Yerokhin, О. Turuta, А. Babii, А. Nechyporenko. – 2017. – Режим доступу до ресурсу: [https://www.researchgate.net/publication/320968847\\_Intelligent\\_information\\_system\\_of\\_heterogeneous\\_medical\\_data\\_analysis](https://www.researchgate.net/publication/320968847_Intelligent_information_system_of_heterogeneous_medical_data_analysis) (дата звернення: 23.03.2024).

## ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

1. Turuta O. Collection of questionnaire results, received by using the visual analog scale method, for its further processing in the medical web application [Електронний ресурс] / O. Turuta, Y. Daniil // ScienceRise. – 2017. – Режим доступу до ресурсу: [https://www.researchgate.net/publication/317612987\\_Collection\\_of\\_questionnaire\\_results\\_received\\_by\\_using\\_the\\_visual\\_analog\\_scale\\_method\\_for\\_its\\_further\\_processing\\_in\\_the\\_medical\\_web\\_application](https://www.researchgate.net/publication/317612987_Collection_of_questionnaire_results_received_by_using_the_visual_analog_scale_method_for_its_further_processing_in_the_medical_web_application) (дата звернення: 20.03.2024).
2. Intelligent information system of heterogeneous medical data analysis [Електронний ресурс] / A.Yerokhin, O. Turuta, A. Babii, A. Nechyporenko. – 2017. – Режим доступу до ресурсу: [https://www.researchgate.net/publication/320968847\\_Intelligent\\_information\\_system\\_of\\_heterogeneous\\_medical\\_data\\_analysis](https://www.researchgate.net/publication/320968847_Intelligent_information_system_of_heterogeneous_medical_data_analysis) (дата звернення: 23.03.2024).