Biomedical Engineering and Technology
Issue 20(5), 2025 ISSN (Online) 2707-8483

УДК 504:57.04

IN SILICO МОДЕЛІ ПРОГНОЗУВАННЯ МУТАГЕННОСТІ ЕЙМСА НА ОСНОВІ ВІДБИТКІВ МОЛЕКУЛЯРНОЇ СТРУКТУРИ КСЕНОБІОТИКІВ

Кисляк Сергій Володимирович
kisluak@ukr.net
Дуган Олексій Мартем`янович
odugan51@gmail.com
Романюк Денис Ігорович
romaniuk.denys@lll.kpi.ua
Яловенко Олена Ігорівна
yalov89@i.ua
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Анотація – *Одна з основних глобальних проблем, з якою стикається людство у 21ст. пов`язана з забрудненням навколишнього середовища. Урбанізація, активний розвиток промисловості та впровадження сучасних технологій у всіх сферах життєдіяльності людини сприяють експоненційному зростанню кількості хімічник сполук, що потрапляють у довкілля. Достатньо велика кількість ксенобіотиків, що представлені у навколишньому середовищі, через механізми прямого або опосередкованого впливу на генетичний апарат людини можуть індукувати розвиток спадкових та/або онкологічних захворювань. Суттєве збільшення кількості зафіксованих випадків онкологічних захворювань у різних країнах світу є основним стимулом для активізації наукової спільноти з метою ефективного виявлення та обліку всіх факторів навколишнього середовища, що можуть проявляти генотоксичні властивості. З урахуванням першочергового вирішення задачі, що пов`язана з підтримкою та збереженням генетичного здоров`я людської популяції, на сьогоднішній день потребують перегляду та удосконалення базові in vitro in vivo методи оцінки генотоксичності впливу факторів навколишнього середовища. У цьому контексті заслуговують на увагу сучасні in silico підходи до оцінки генетичної безпеки факторів навколишнього середовища, що мають достатньо значний, але не повністю реалізований потенціал. Метою роботи є розробка методики створення in silico моделей оцінки мутагенності Еймса (AMES/QSAR), що в якості предикторів використовують різні типи відбитків молекулярної структури (molecular fingerprint) ксенобіотиків. З метою створення ефективних in silico моделей прогнозування мутагенності Еймса було використано базу даних ксенобіотиків, що була сформована шляхом поєднання популярних серед дослідників датасетів: Kazius-Bursi, Hansen та EFSA. Крім того, база даних хімічних сполук, що використовувалась при моделюванні була розширена мікотоксинами. В основі вирішення задачі бінарної класифікації були обрані два ансамблевих метода машинного навчання: метод випадкового лісу та екстремального градієнтного бустінга. Точність бінарних класифікаторів на рівні 80-85%, які були розроблені відповідно до представленої у роботі методики, відповідає відтворюваності теста Еймса в різних лабораторіях. Показано, що орієнтовані на структурні класи бінарні класифікатори є більш ефективними для прогнозування мутагенності Еймса, у порівнянні з AMES/QSAR моделями, що на етапі навчання використовували частину повного необнорідного набору вхідних даних.*

Ключові слова: мутагенність, тест Еймса, QSAR модель, ксенобіотики, молекулярні дескриптори, відбитки молекулярної структури, моделі машинного навчання

I. Вступ

Кількість ксенобіотиків, що мають антропогенне походження та пов`язані з діяльністю людини у різних сферах життя постійно зростає. На початку 2019 року кількість хімічних сполук, що була внесена до реєстрів бази даних CAS (Chemical Abstract Service) налічувала більше ніж 150 млн. сполук [1]. У вересні 2025р. кількість зареєстрованих ксенобіотиків, яким був присвоєний унікальний номер CAS досягла 275 млн. Відповідно, за сім не повних років кількість зареєстрованих речовин збільшилась на 125 млн. Така динаміка збільшення новосинтезованих хімічних сполук пов`язана з активним розвитком різних галузей промисловості, сільського господарства, медицини тощо. Занепокоєння у наукової спільноти та громадськості викликає той факт, що більша частина хімічних сполук може

потрапляти у різні об`єкти навколишнього середовища (атмосферне повітря, питне воду, продукти харчування, косметичні засоби тощо), що може лежати в основі проявів негативних як екологічних так і генетичних наслідків. Крім того, реєстрація достатньо великої кількості ксенобіотиків відбувається без попередньої оцінки їх впливу на здоров`я людини [2]. Взаємодія ксенобіотиків, що мають антропогенне походження, зі спадковим апаратом людини може призводити до появи пошкоджень на рівні ДНК (мутації на рівні соматичних клітин) та розвитку канцерогенезу [3-4]. Вплив ксенобіотиків, що володіють вираженими потенційними генотоксичними властивостями на спадковий матеріал статевих клітин може стати фундаментом для розвитку різноманітних генетичних захворювань [5-6]. Відповідно до нещадавно опублікованої наукової праці [7], вплив хімчних сполук, що представлені у довкіллі на генетичний апарат людини може призводити до аномального перебігу процесу метилювання промоторних ділянок генів білків-гістонів, в результаті чого відбувається формування модифікованого епігеному, який є сприятливим для розвитку нейродегенеративних та онкологічних захворювань. Проблема, що пов`язана з суттєвим збільшенням кількості новонароджених дітей з наявними захворюваннями аутичного спектру також пов`язують з дією ксенобіотиків на спадковий апарат в період раннього онтогенезу [8]. З урахування широкого спектру негативних наслідків впливу ксенобіотиків на здоров`я людської популяції, першочерговим завданням науковців є розробка сучасних методів та підходів, які дозволять ефективно виявляти потенційні генотоксичні сполуки, що представлені у довкіллі.

З метою отримання оцінки генетичної безпеки впливу факторів навколишнього середовища, на сьогоднішній день було розроблено більш ніж 150 *in vitro* та *in vivo* короткострокових тестів. При цьому більша частина методів була представлена для обговорення науковій спільноті понад 30 років тому [9-10]. Для забезпечення об`єктивної генетичної оцінки впливу факторів навколишнього середовища, з урахуванням трьох основних кінцевих результатів пошкодження ДНК [11], що пов`язані з виникненням мутацій, хромосомних аберацій та анеуплоїдії, використовують класичну батарею тест-систем [12-14]. В такій ситуації всебічна токсикологічна оцінка факторів навколишнього середовища може бути здійснена з використанням не однієї, а декількох стандартизованих методик. Така особливість пов`язана з тим, що на сьогоднішній день не існує жодного короткострокового тесту, який дозволив би враховувати всі кінцеві результати пошкодження ДНК [12].

Серед найбільш популярних методів оцінки генотоксичного потенціалу, що відносяться до стандартної батареї короткострокових тестів, особливу увагу заслуговує тест на бактеріальну зворотну мутацію, який був розроблений Брюсом Еймсом ще у далекому 1970 році минулого століття [15,16]. Цей інтерес обумовлений тим, що тест Еймса є достатньо простим з точки зору виконання [10,12]. Тест на зворотню бактеріальну мутацію використовується як базовий *in vitro* метод [17] для оцінки потенційного мутагенного потенціалу та проводиться з використанням бактеріальних штамів Salmonella typhimurium, що є ауксотрофними за гістидином [16,18,19]. Для таких штамів не буде спостерігатись ріст на поживному середовищі, в якому відсутня амінокислота гістидин [12,15]. Вплив ксенобіотиків, що проявляють потенційні генотоксичні властивості може призводити до виникнення зворотних мутацій, що можуть лежати в основі переходу від ауксотрофності по гістидину до прототрофності тест-штамів Salmonella typhimurium [12,15]. Такий процес досить

зручно детектується через активізацію росту ревертантних колоній *Salmonella typhimurium* на поживному середовищі.

Незважаючи на те, що на сьогоднійшній день існують достатньо ефективні *in vitro* та *in vivo* методи оцінки генотоксичності, що були прийняті науковою спільнотою та пройшли процедуру затвердження в рамках таких міжнародних організацій як OECD, ECHA, UK-EMS, US-FDA, EFSA та ін [9,20], для великої кількості ксенобіотиків залишається відсутньою інформація про їх генотоксичний потенціал. Така проблема обумовлена тим, що використання стандартної батареї *in vitro* та *in vivo* тест-систем має недоліки, що пов`язані, у першу чергу, з часовими витратами та вартістю проведення експериментальних досліджень [21-23]. Крім того, відповідно до затвердженої концепції «3R», при проведенні тестувань на генотоксичність необхідно мінімізувати використання піддослідних тварин [24,25]. В такій ситуаціїї вчені зберігають надію щодо подолання вищезазначених проблем сучасної токсикології через розробку та впровадження *in silico* QSAR (Quantitative Structure-Activity Relationship) моделей. Такий підхід дозволяє через пошук взаємозв`язків між наборами вхідних даних, що розраховані відповідно до певної структури досліджуваних ксенобіотиків, отримати інформацію про наявний мутагенний потенціал.

У межах проведеного дослідження нами було зосереджено увагу на розробці ефективних, орієнтованих на основні структурні класи хімічних сполук, *in silico* моделей прогнозування мутагенності Еймса, що використовують в якості предикторів різні типи відбитків молекулярної структури. Вибір тесту Еймса в якості основного метода при проведенні *in silico* моделювання був обумовлений достатньо великою кількістю експериментальних даних, які були отримані в лабораторіях різних країн світу за більше ніж 50 років. Крім того, результати оцінки мутагенності за допомогою *in vitro* таста Еймса знаходяться у відкритому доступі. Підвищений науковий інтерес до відповідного напрямку досліджень простежується у нещодавно опублікованих наукових працях [26-30]

II. МЕТА ДОСЛІДЖЕННЯ

Метою роботи є розробка та тестування орієнтованих на основні структурні класи ксенобіотиків *in silico* моделей прогнозування мутагенності Еймса на основі різних типів відбитків молекулярної структури.

III. МАТЕРІАЛИ ТА МЕТОДИ

3.1 База даних ксенобіотиків

З метою створення ефективних *in silico* моделей прогнозування мутагенності Еймса було використано базу даних ксенобіотиків [29], що була сформована шляхом поєднання популярних серед дослідників датасетів: Kazius-Bursi [31], Hansen [32] та EFSA [33]. Крім того, база даних хімічних сполук, що використовувалась при моделюванні була роширена мікотоксинами [27]. Після видалення хімічних сполук, що мають однакову структурну формулу загальна кількість ксенобіотиків склала 8454. Отриманий набір даних був збережений у *csv*. форматі, в якому для кожної хімічної сполуки була збережена наступна інформація: 1. Ідентифікатор (ID), що відповідає порядковому номеру ксенобіотика; 2. SMILES (Simplified Molecular Input Line Entry System) лінійна нотація – це прийнятий науковою спільнотою текстовий формат даних, який використовується для збереження інформації про структуру хімічних сполук – потенційних мутагенів; 3. Структурний клас, до якого відноситься ксенобіотик, що визначається з урахуванням особливостей будови його молекулярного каркасу; 4. Інформація про наявний або відсутній мутагенний потенціал, що отримана експериментально за допомогою *in vitro*

теста Еймса (позначається 1, якщо хімічна сполука проявляє потенційні мутагенні властивості; 0 відповідає ксенобіотику, для якого відсутні мутагенні властивості. 5. Відбиток молекулярної структури (від анг. molecular fingerprint), що представляє собою бітовий рядок, який дозволяє збірігати інформацію про структуру ксенобіотиків, через позначення наявних (або відсутніх) функціональних груп або підструктур на рівні молекули.

3.2 Розподіл хімічніх сполук за основними структурними класами

На почтаку проведення дослідження нами була сформульована гіпотеза, відповідно до якої, покращення прогностичної здатності розроблених моделей може бути досягнено через створення AMES/QSAR моделей, які на етапі навчання використовують однорідні, з точки зору будови молекулярного каркасу ксенобіотиків, вхідні дані. Відповідно до запропонованого підходу, за допомогою веб-сервіса ClassyFire [34], датасет був розподілений на 9 структурних класів (табл.1).

Таблиця 1. Основні структурні класи ксенобіотиків

№п.п та назва класу	№ групи	Кількість мутагенів	Кількість не мутагенів	Загальна кількість
1. Аліфатичні ациклічні	1	548	774	1322
2. Аліфатичні гетеромоноциклічні	2	189	178	367
3. Аліфатичні гетерополіциклічні		79	141	220
4. Аліфатичні гомомоноциклічні	3	28	101	129
5. Аліфатичні гомополіциклічні		29	128	157
6. Ароматичні гетеромоноциклічні	4	355	675	1030
7. Ароматичні гетерополіциклічні		1248	881	2129
8. Ароматичні гомомоноциклічні	5	871	1176	2047
9. Ароматичні гомополіциклічні		780	273	1053
ЗАГАЛОМ	1-5	4127	4327	8454

Подібність молекулярної структури ксенобіотиків між окремими класами стала базовим критерієм щодо формування 5 груп ксенобіотиків. Перша група ксенобіотиків була представлена окремим класом, до якого відносились ациклічні хімічні сполуки (табл.1). Друга група була отримана шляхом об`єднання ксенобіотиків, що належали до двох структурних класів: аліфатичних гетеромоноциклічних та аліфатичних гетерополіциклічних. Незбалансована кількість ксенобіотиків – потенційних мутагенів по відношенню до не мутагенів, а також не велика кількість мутагенів у порівнянні з мутагенами третьої групи ксенобіотиків може мати негативний вплив на точність, орієнтованих на відповідні структурні класи Ames/QSAR моделей. В такій ситуації клас-орієнтований підхід щодо побудови моделей прогнозування мутагенності Еймса до цієї групи ксенобіотиків не застосовувався. Для отримання *in silico* оцінки мутагенності Еймса хімічних сполук, що відносяться до третьої групи (табл.1), нами було запропоновано застосовувати Ames/QSAR моделі, для яких на етапі навчання використовувалась частина повного датасету, що був представлений 8454 ксенобіотиками. Четверта група була сформована шляхом об`єднання ксенобіотиків, що належать до двох класів: ароматичних гетеромоноциклічних та ароматичних гетерополіциклічних. Ароматичні гомомоноциклічні та ароматичні гомополіциклічні хімічні сполуки представляли об`єднану п`яту групу ксенобіотиків. Існуючий баланс між кількостю мутагенів та не мутагенів для першої, другої, четвертої та п`ятої груп ксенобіотиків дозволив використовувати окремо (на етапі навчання) відповідні набори даних з метою створення ефективних моделей прогнозування мутагенності Еймса.

3.3 Відбитки молекулярної структури (molecular fingerprint)

Загальна класифікація молекулярних дескрипторів враховує їх поділ на три класи, що відповідає 1D (одновимірним), 2D (двовимірним) та 3D (тривимірним) дескрипторам. Найпростіші, з точки зору розрахунків, 1D дескриптори не містять інформацію про зв`язніть атомів на рівні молекули, тому вони не враховують структуру молекули. 2D дескриптори представляють найбільшу групу предикторів. Вони можуть бути розраховані на основі представлення молекули у вигляді графу. 3D дескриптори містять інформацію про розашування у просторі атомів, що формують молекулу. Використання таких дескрипторів має суттєві обмеження, що в першу чергу, пов`язано з необхідністю на етапі розрахунків таких дескрипторів мати достатньо великі ресурси пам`яті та часу. Крім того, прогрмане забезпечення, що дозволяє отримати розрахунки 3D дескрипторів досить часто розповсюджується на платній основі. Тому в науковій літературі при вирішенні задачі прогнозування мутагенності Еймса, зазвичай, в якості предикторів використовують 1D та 2D молекулярні дескриптори. При цьому відбитки молекулярної структури, які відносять до 2D дескрипторів, зазвичай, використовуються дослідниками окремо від інших двовимірних дескрипторів. Така особливість зумовлена специфікою збереження інформації про структуру молекули ксенобіотика, що реалізується за допомогою бітового рядка, в якому кожний біт відповідає за наявність або відсутність певної функціональної групи або підструктури на рівні молекули. Аналіз опублікованих наукових праць [27,36,37,38], в яких дослідники в якості предикторів використовують окремо відбитки молекулярної структури, стали надійним фундаментом для проведення подібних досліджень, але з розробленою, у межах роботи, методикою. Відповідно до запропонованого нами підходу *in silico* оцінки мутагенності Емса факторів навколишнього середовища потребує перевірки гіпотеза щодо можливості покращення ефективності бінарних класифікаторів, які етапі навчання моделей використовували однорідні вхідні дані (відбитків молекулярної структури), які були отримані для чотирьох груп ксенобіотиків (табл.1). Необхідно відмітити, що у науковій праці [39] дослідники розглядають класифікацію молекулярних відбитків структури, що враховує їх поділ на три основні групи, що відповідають класам. До першої групи відносяться, так звані, субструктурні відбитки структри (від анг. «substructure fingerprints»), що представляють собою бітовий рядок певної фіксованої довжини. Друга група дескрипторів відповідає топологічним відбиткам структури (від анг. «topological fingerprint»). Алгоритм розрахунку топологічних відбитків базується на графовому представленні молекули та враховує всі варіанти можливих фрагментів певної довжини, починаючи відлік від кожного атома в структурі молекули. Циркулярні (від анг. «circular fingerprint») відбитки структури, що належать до третьої групи дескрипторів відносяться до хешованих відбитків, які можуть бути розраховані відповідно до молекулярного оточення для кожного атома, що формують молекулу. В такій ситуації необхідним у науковому відношенні є отримання інформації щодо ефективності розроблених моделей прогнозування мутагенності Еймса, що в якості вхідних даних використовують різні відбитки молекулярної структури, кожний з яких відноситься до одного з трьох класів.

У межах роботи було запропоновано в якості предикторів використовувати мелекулярні дескриптори RDkit, MACCS (Molecular Access System) та FCFP (Extended-Connectivity Fingerprints), що відносяться до топологічних,

субструктурних та циркулярних відбитків відповідно та досить часто використовуються у подібних дослідженнях.

3.4 Методи дослідження

Реалізація моделей прогнозування мутагенності Еймса була здійснена на основі двох ансамблевих алгоритмів машинного навчання: методу випадкового лісу (RF) та екстремального градієнтного бустінга. Вибір відповідних методів в якості основних для вирішення задачі бінарної класифікації був обумовлений достатньо великою кількістю наукових праць, в яких розроблені Ames/QSAR моделі на основі методу випадкового лісу [27-29,34] та грдієнтного бустінга [23,27,38,40] демонстрували одну з найкращих точність класифікації.

3.5 Препроцесинг даних

Ефективність розроблених Ames /QSAR моделей на пряму може залежити від того, на скільки якісно було проведено підготовку вхідних даних, що використовуються при моделюванні. Стандартна процедура препроцесингу даних пов`язана з видаленням не інформативних та висококорельнованих ознак, а також досягається через масштабування, що дозволяє розподілити дані у певному визначеному діапазоні [30]. Крім того, для отримання ефективної моделі прогнозування мутагенності Еймса, зазвичай, необхідно вирішити проблему наявних аномальних значень, що можє суттєво знижувати точність бінарних класифікаторів.

Основною перевагою використання відбитків просторової структури, як предкторів полягає у тому, що для таких Ames/QSAR моделей відсутністя необхідність у проведенні базових процедур препроцесингу. Така особливість пов`язана з тим, що вхідні дані, які представлені бітовим рядком можна відразу, без попередньої обробки, використовувати при моделюванні. У рамках проведеного дослідження набори даних (табл.1) випадковим чином були розподілені у співвідношенні 80:20 на навчальну та тестову вибірки. Необхідно зазначити, що подібний підхід щодо розподілу вхідних даних на дві вибірки використовувався як для окремих груп (табл. 1) хімічних сполук, так і для повної бази даних, що налічує 8454 ксенобіотиків. Такий підхід дозволяє отримати оцінку ефективності моделей Ames/QSAR, що використовують на етапі навчання однорідні набори даних, що об`єднані у групи відповідно до подібності будови їх молекулярного каркасу, у порівнянні з бінарними класифікаторами для яких на етапі навчання використовувалась частина повної бази даних, що налічує 8454 ксенобіотика.

Для отримання об`єктивної оцінки ефективності розроблених моделей прогнозування мутагенності Еймса на етапі навчання, була застосована п’ятикратна перехресна перевірка (крос-валідація). Такий підхід також дозволяє здійснювати підбір оптимальних значень гіперпараметрів Ames/QSAR моделей, при яких досягається найбільша точність класифікації.

3.6 Оцінка ефективності Ames/QSAR моделей

Оцінка прогностичної здатності розоблених *in silico* Ames/QSAR моделей здійснювалась за допомогою наступних метрик: загальної точності (*accuracy)*, точності позитивного прогнозу (*precision)*, чутливості (*recall*), специфічності (*specificity*) та F_1-міри ($F_1 - score$), які були отримані з урахуванням матриць помилок відповідно до співвідношень 1-5, де TP, TN, FP, FN – відповідає кількості істинно позитивних, істинно негативних, хибнопозитивних та хибнонегативних результатів класифікації відповідно. Для кожної моделі також була розрахована площа під ROC кривою (AUC), що є достатньо потужним інструментом, що

використовується для оцінки ефективності моделей машинного навчання [41]

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$specificity = \frac{TN}{TN+FP} \quad (4)$$

$$F_1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (5)$$

IV РЕЗУЛЬТАТИ

У таблиці 2 наведено класифікаційні звіти, що були отримані на тестових вибірках для кожної розробленої Ames/QSAR моделі на основі методу випадкового лісу. В якості предикторів для моделей прогнозування мутагенності Еймса використовувались відбитки структури FCFP, RDkit та MACCS.

Відповідно до отриманих значень базових метрик оцінки ефективності можна спостерігати підвищення точності моделей прогнозування мутагенності Еймса для всіх моделей, що були отримані відповідно до однорідних вхідних даних, які відповідають першій, другій, четвертій та п`ятій групі (табл.1) ксенобіотиків. З іншої точки зору, моделі, які на етапі навчання використовували частину повного датасету, що є стандартним підходом у моделюванні, показали зменшення точності в межах від 2 до 11%.

Таблиця 2. Класифікаційний звіт отриманий для тестових вибірок Ames/QSAR моделей, що побудовані на основі методу випадкового лісу (значення метрик для найкращих моделей позначені жирним шрифтом)

Молекулярні дескриптори	Структурні класи ксенобіотиків	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
FCFP	Аліфатичні ациклічні	0,81	0,81	0,71	0,88	0,76	0,85
	Аліфатичні гетеромоно (полі) циклічні	0,81	0,82	0,76	0,86	0,79	0,87
	Ароматичні гетеромоно (полі) циклічні	0,83	0,89	0,77	0,90	0,82	0,91
	Ароматичні гомомоно (полі) циклічні	0,83	0,83	0,86	0,80	0,84	0,90
	Всі	0,78	0,79	0,74	0,82	0,77	0,85
RDkit	Аліфатичні ациклічні	0,81	0,80	0,71	0,88	0,75	0,87
	Аліфатичні гетеромоно (полі) циклічні	0,86	0,88	0,81	0,91	0,85	0,90
	Ароматичні гетеромоно (полі) циклічні	**0,85**	**0,87**	**0,83**	**0,87**	**0,85**	**0,92**
	Ароматичні гомомоно (полі) циклічні	0,84	0,84	0,86	0,81	0,85	0,90
	Всі	0,79	0,80	0,76	0,81	0,78	0,85
MACCS	Аліфатичні ациклічні	**0,82**	**0,82**	**0,73**	**0,88**	**0,77**	**0,88**
	Аліфатичні гетеромоно (полі) циклічні	**0,89**	**0,94**	**0,81**	**0,95**	**0,87**	**0,95**
	Ароматичні гетеромоно (полі) циклічні	0,83	0,85	0,81	0,85	0,83	0,91
	Ароматичні гомомоно (полі) циклічні	**0,85**	**0,85**	**0,87**	**0,82**	**0,86**	**0,90**
	Всі	0,78	0,80	0,75	0,82	0,77	0,86

Результати тестування розроблених бінарних класифікатор дозволяють серед різних, орієнтованих на основні структурні класи Ames/QSAR моделей, обрати найкращу.

Для прогнозування мутагенності Еймса аліфатичних ациклічних хімічних сполук найкращою є модель зі значенням $AUC = 0{,}88$, для якої в якості вхідних даних використовувались відбиткі просторової структури MACCS, що були розраховані для першого класу (табл.1) ксенобіотиків.

In silico оцінку мутагенності Еймса аліфатичних гетеромоноциклічних та аліфатичних гетерополіциклічних хімічних сполук можна отримати за допомогою Ames/QSAR моделі, для якої вхідними даними виступають відбитки молекулярної структури MACCS, що були отримані відповідно до ксенобіотиків, які відносяться до другої групи (табл 1). Необхідно звернути увагу на те, що для даної моделі характерні одні з найкращих значень метрик оцінки прогностичної здатності у порівнянні з іншими, орієнтованими на структурні класи бінарними класифікаторами. При цьому, відповідно до значення $recall = 0{,}81$, дана модель може генерувати приблизно однакову кількість хибнонегативних результатів класифікації у порівнянні з іншими моделями. Очевидно, що при проведенні процедури відбору накращих моделей прогнозування мутагенності Еймса, що здійснюється відповідно до основних метрик бінарної класифікації, необхідно брати за основу ту модель, для якої кількість хибнонегативних результатів буде мінімальною. Якщо модель визначає достатньо велику кількість мутагенних сполук, як не є мутагенами це створює додаткові загрози для генетичного здоров`я людської популяції.

Для *in silico* прогнозування мутагенності Еймса ароматичних гетеромоноциклічних та ароматичних гетерополіциклічних хімічних сполук необхідно використовувати Ames/QSAR модель з $AUC = 0{,}92$, для якої в якості предикторів використовувались відбитки молекулярної структури RDkit, що були розраховані для четвертої групи ксенобіотиків (табл.1). Ames/QSAR модель, що використовує в якості вхідних даних відбитки молекулярної структури MACCS, які були розраховані для п`ятої групи ксенобіотиків (табл.1), дозволяє вирішити задачу бінарної класифікації з найкращим показником точности ($accuracy = 0{,}85$). Значення метрики $recall = 0{,}87$ вказує на те, що в реальних умовах, на нових даних, дана модель Ames/QSAR буде давати найменшу кількість хибнонегативних прогнозів касифікації у порівнянні з іншими бінарними класифікаторами.

У таблиці 3 наведено класифікаційні звіти, що були отримані на тестових вибірках *in silico* Ames/QSAR моделей на основі методу екстрамального градієнтного бустінгу.

Таблиця 3. Класифікаційний звіт отриманий для тестових вибірок Ames/QSAR моделей, що побудовані на основі методу екстремального градієнтного бустінгу (значення метрик для найкращих моделей позначені жирним шрифтом)

Молекулярні дескриптори	**Структурні класи ксенобіотиків**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1 Score**	**AUC**
FCFP	Аліфатичні ациклічні	0,81	0,83	0,69	0,90	0,75	0,88
	Аліфатичні гетеромоно (полі) циклічні	0,78	0,78	0,72	0,83	0,75	0,86
	Ароматичні гетеромоно (полі) циклічні	0,85	0,87	0,82	0,88	0,84	0,92

	Ароматичні гомомоно (полі) циклічні	0,82	0,81	0,86	0,76	0,83	0,89
	Всі	0,82	0,83	0,79	0,85	0,81	0,90
Rdkit	Аліфатичні ациклічні	0,82	0,84	0,71	0,90	0,77	0,90
	Аліфатичні гетеромоно (полі) циклічні	0,83	0,87	0,74	0,91	0,80	0,90
	Ароматичні гетеромоно (полі) циклічні	**0,86**	**0,87**	**0,85**	**0,87**	**0,86**	**0,93**
	Ароматичні гомомоно (полі) циклічні	0,83	0,84	0,85	0,81	0,84	0,91
	Всі	0,85	0,85	0,83	0,86	0,84	0,92
MACCS	Аліфатичні ациклічні	**0,85**	**0,86**	**0,75**	**0,92**	**0,81**	**0,89**
	Аліфатичні гетеромоно (полі) циклічні	**0,84**	**0,91**	**0,72**	**0,94**	**0,80**	**0,93**
	Ароматичні гетеромоно (полі) циклічні	0,82	0,82	0,82	0,82	0,82	0,92
	Ароматичні гомомоно (полі) циклічні	**0,84**	**0,84**	**0,86**	**0,82**	**0,85**	**0,89**
	Всі	0,84	0,84	0,84	0,84	0,84	0,91

Відповідно до отриманіх результатів класифікації, перелік найкращих моделей не змінився (табл. 3) у порівнянні з бінарними класифікаторами, що були отримані за допомогою методу випадкового лісу (табл. 2). Всі моделі з найвищою точністю класифікації побудовані з урахуванням клас-орієнтованого підходу, в основі якого процес навчання моделей відбувався на основі однорідних даних (відбитків молекулярної структури), що були отримані для окремих груп ксенобіотиків (табл.1). Дві моделі Ames/QSAR показали кращу точність класифікації у порівнянні з бінарними класифікаторами, що були побудовані на основі методу випадкового лісу. Відповідно, для *in silico* прогнозування мутагенності Еймса аліфатичних ациклічних хімічних та ароматичних гетеромоно(полі)циклічних сполук необхідно використовувати бінарні класифікатори на основі метода екстремального градієнтного бустінга, що в якості вхідних даних використовують відбитки молекулярної структури MACCS та RDkit відповідно, що були розраховані для першої та четвертої групи ксенобіотиків (табл.1). При цьому, для прогнозування мутагенності Еймса аліфатичних гетеромоно(полі)циклічних та ароматичних гомомоно(полі)циклічних хімічних сполук необхідно використовувати Ames/QSAR моделі, що були побудовані на основі методу випадкового лісу, відповідно до розрахованих відбитків молекулярної структури MACCS, що були розраховані для другої та п’ятої групи (табл. 1) ксенобіотиків.

V ВИСНОВКИ

У роботі запропоновано новий підхід, який лежить в основі покращення прогностичної здатності *in silico* моделей прогнозування мутагенності Еймса, що досягається через формування окремих груп ксенобіотиків, які мають спільні риси будови молекулярної структури та для яких процедура навчання проводиться окремо. Показано, що застосування відбитків молекулярної структури FCFP, RDkit, MACCS в якості предикторів для Ames/QSAR моделей дозволяє отримати *in silico* оцінку мутагенності Еймса, що відповідає варіабельності результатів теста Еймса в різних лабораторіях. Відсутність

застосування стандартних методів препроцесингу даних при створенні таких Ames/QSAR моделей сприяє розробці ефективних бінарних класифікаторів з мінімальними часовими затратами. Використання відбитків молекулярної структури MACCS в якості предикторів дозволяє отримати бінарні класифікатори з найкращою точністю прогнозування мутагенності Еймса трьох структурних класів ксенобіотиків. У такій ситуації молекулярні відбитки структури MACCS необхідно розглядати в якості пріоритетних предикторів при створенні *in silico* Ames/QSAR моделей.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

Згода на публікацію. Усі автори, які мають відношення до рукопису, дали згоду на публікацію цієї наукової праці.

ORCID ID та внесок авторів.
0000-0003-2097-3793(A,B,C,D,E) Sergey Kislyak
0000-0002-5646-917X (A,G,H) Olexii Dugan
0009-0003-2091-1645 (F) Romaniuk Denys
0000-0002-5022-143X (G) Olena Yalovenko
– Концепція роботи та дизайн;
B – Розробка методології покращення точності Ames/QSAR моделей;
C – Формування бази даних ксенобіотиків;
D – Розрахунок молекулярних дескрипторів та розподіл ксенобіотиків за структурними класами;
E – Написання статті;
F – Реалізація моделей;
G – Критичний огляд;
H – Остаточне схвалення статті.

ПЕРЕЛІК ПОСИЛАНЬ

1. Benigni R. In silico assessment of genotoxicity. combinations of sensitive structural alerts minimize false negative predictions for all genotoxicity endpoints and can single out chemicals for which experimentation can be avoided. Regulatory toxicology and pharmacology. 2021. P. 105042.
URL: https://doi.org/10.1016/j.yrtph.2021.105042
2. Honma M. An assessment of mutagenicity of chemical substances by (quantitative) structure–activity relationship. Genes and environment. 2020. Vol. 42, no. 1. URL: https://doi.org/10.1186/s41021-020-00163-1
3. Chatterjee N., Walker G. C. Mechanisms of DNA damage, repair, and mutagenesis. Environmental and molecular mutagenesis. 2017. Vol. 58, no. 5. P. 235–263. URL: https://doi.org/10.1002/em.22087
4. Katerji M., Duerksen-Hughes P. J. DNA damage in cancer development: special implications in viral oncogenesis. American journal of cancer research. 2021. Vol. 11, no. 8.
5. Metallothionein-I/II expression associates with the astrocyte DNA damage response and not Alzheimer-type pathology in the aging brain / R. Waller et al. *Glia*. 2018. Vol. 66, no. 11. P. 2316–2323. URL: https://doi.org/10.1002/glia.23465
6. Activation of the DNA damage response in vivo in synucleinopathy models of Parkinson's disease / C. Milanese et al. Cell death & disease. 2018. Vol. 9, no. 8. URL: https://doi.org/10.1038/s41419-018-0848-7
7. Environmental xenobiotics and epigenetic modifications: implications for human health and disease / A. F. Sobral et al. Journal of xenobiotics. 2025. Vol. 15, no. 4. P. 118. URL: https://doi.org/10.3390/jox15040118
8. Xavier Santos J., Rasga C., Moura Vicente A. Exposure to xenobiotics and gene-environment interactions in autism spectrum disorder: a systematic review. Autism spectrum disorder - profile, heterogeneity, neurobiology and intervention. 2021.
URL: https://doi.org/10.5772/intechopen.95758
9. Search for the optimal genotoxicity assay for routine testing of chemicals: sensitivity and specificity of conventional and new test systems / M. Mišík et al. Mutation research/genetic toxicology and environmental mutagenesis. 2022. Vol. 881. P. 503524. URL: https://doi.org/10.1016/j.mrgentox.2022.503524
10. The various aspects of genetic and epigenetic toxicology: testing methods and clinical applications / N. Ren et al. Journal of translational medicine. 2017. Vol. 15, no. 1. URL: https://doi.org/10.1186/s12967-017-1218-4
11. Turkez H., Arslan M. E., Ozdemir O. Genotoxicity testing: progress and prospects for the next decade. Expert opinion on drug metabolism & toxicology. 2017. Vol. 13, no. 10. P. 1089–1098. URL: https://doi.org/10.1080/17425255.2017.1375097
12. Luan Y., Honma M. Genotoxicity testing and recent advances. Genome instability & disease. 2021. Vol. 3, no. 1. P. 1–21. URL: https://doi.org/10.1007/s42764-021-00058-7
13. Sv R. Genotoxicity: mechanisms, testing guidelines and methods. Global journal of pharmacy & pharmaceutical sciences. 2017. Vol. 1, no. 5. URL: https://doi.org/10.19080/gjpps.2017.01.555575
14. Kislyak S., Dugan O., Yalovenko O. Systems for genetic assessment of the impact of environmental factors. *Innovative biosystems and bioengineering*. 2024. Vol. 8, no. 2. P. 3–27. URL: https://doi.org/10.20535/ibb.2024.8.2.288127

15. Comparison of methods used for evaluation of mutagenicity/genotoxicity of model chemicals - parabens / J. Chrz et al. Physiological research. 2020. P. S661–S679. URL: https://doi.org/10.33549/physiolres.934615
16. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection / B. N. Ames et al. Proceedings of the national academy of sciences. 1973. Vol. 70, no. 8. P. 2281–2285. URL: https://doi.org/10.1073/pnas.70.8.2281
17. Bhagat J. Combinations of genotoxic tests for the evaluation of group 1 IARC carcinogens. Journal of applied toxicology. 2017. Vol. 38, no. 1. P. 81–99. URL: https://doi.org/10.1002/jat.3496
18. Sofuni T. Evolution of genotoxicity test methods in Japan. Genes and environment. 2017. Vol. 39, no. 1. URL: https://doi.org/10.1186/s41021-016-0063-7
19. Genotoxicity evaluation of hospital wastewaters / P. Gupta et al. Ecotoxicology and environmental safety. 2009. Vol. 72, no. 7. P. 1925–1932. URL: https://doi.org/10.1016/j.ecoenv.2009.05.012
20. Overview on genetic toxicology TGs. OECD, 2017. URL: https://doi.org/10.1787/9789264274761-en
21. А. М. Дуган, "Salmonella typhimurium як тест-система для виявлення мутагенної активності забруднювачів навколишнього середовища," Цитологія і генетика, vol. 28, no. 3, pp. 37–41, 1994.
22. І. Р. Бариляк and О. М. Дуган, "Еколого-генетичні дослідження в Україні," Цитологія і генетика, no. 5, pp. 3–10, 2002.
23. C. S. M. Chu et al., "Machine learning – Predicting Ames mutagenicity of small molecules," Journal of Molecular Graphics and Modelling, p. 108011, 2021, doi: 10.1016/j.jmgm.2021.108011.
24. A new 3D model for genotoxicity assessment: EpiSkin™ Micronucleus Assay / L. Chen et al. Mutagenesis. 2020. URL: https://doi.org/10.1093/mutage/geaa003
25. Animal welfare considerations when conducting OECD test guideline inhalation and toxicokinetic studies for nanomaterials / Y. H. Chung et al. Animals. 2022. Vol. 12, no. 23. P. 3305. URL: https://doi.org/10.3390/ani12233305
26. Van Tran T. T., Tayara H., Chong K. T. AMPred-CNN: ames mutagenicity prediction model based on convolutional neural networks. Computers in biology and medicine. 2024. P. 108560. URL: https://doi.org/10.1016/j.compbiomed.2024.108560
27. Chemical rules for optimization of chemical mutagenicity via matched molecular pairs analysis and machine learning methods / C. Lou et al. Journal of cheminformatics. 2023. Vol. 15, no. 1. URL: https://doi.org/10.1186/s13321-023-00707-x
28. Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project / A. Furuhama et al. SAR and QSAR in environmental research. 2023. Vol. 34, no. 12. P. 983–1001.
URL: https://doi.org/10.1080/1062936x.2023.2284902
29. A comparison of nine machine learning mutagenicity models and their application for predicting pyrrolizidine alkaloids / C. Helma et al. Frontiers in pharmacology. 2021. Vol. 12. URL: https://doi.org/10.3389/fphar.2021.708050
30. In silico the ames mutagenicity predictive model of environment / S. Kislyak et al. Innovative biosystems and bioengineering. 2025. Vol. 9, no. 2. P. 42–52. URL: https://doi.org/10.20535/ibb.2025.9.2.316239
31. J. Kazius, R. McGuire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," Journal of Medicinal Chemistry, vol. 48, no. 1, pp. 312–320, 2005, doi: 10.1021/jm040835a.
32. K. Hansen et al., "Benchmark data set for in silico prediction of Ames mutagenicity," Journal of Chemical Information and Modeling, vol. 49, no. 9, pp. 2077–2081, 2009, doi: 10.1021/ci900161g
33. EFSA, "Dietary exposure assessment to pyrrolizidine alkaloids in the European population," EFSA Journal, vol. 14, no. 8, 2016, doi: 10.2903/j.efsa.2016.4572.
34. MicotoXilico: an interactive database to predict mutagenicity, genotoxicity, and carcinogenicity of mycotoxins / J. Tolosa et al. Toxins. 2023. Vol. 15, no. 6. P. 355. URL: https://doi.org/10.3390/toxins15060355
35. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy / Y. Djoumbou Feunang et al. Journal of cheminformatics. 2016. Vol. 8, no. 1. URL: https://doi.org/10.1186/s13321-016-0174-y
36. Machine learning – Predicting Ames mutagenicity of small molecules / C. S. M. Chu et al. Journal of molecular graphics and modelling. 2021. P. 108011. URL: https://doi.org/10.1016/j.jmgm.2021.108011
37. Molecular fingerprint-derived similarity measures for toxicological read-across: recommendations for optimal use / C. L. Mellor et al. Regulatory toxicology and pharmacology. 2019. Vol. 101. P. 121–134. URL: https://doi.org/10.1016/j.yrtph.2018.11.002
38. In silico prediction of chemical genotoxicity using machine learning methods and structural alerts / D. Fan et al. Toxicology research. 2018. Vol. 7, no. 2. P. 211–220. URL: https://doi.org/10.1039/c7tx00259a
39. Molecular fingerprint similarity search in virtual screening / A. Cereto-Massagué et al. Methods. 2015. Vol. 71. P. 58–63. URL: https://doi.org/10.1016/j.ymeth.2014.08.005
40. LightGBM: an effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets / J. Zhang et al. Journal of chemical information and modeling. 2019. Vol. 59, no. 10. P. 4150–4158. URL: https://doi.org/10.1021/acs.jcim.9b00633
41. F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," Korean Journal of Anesthesiology, vol. 75, no. 1, pp. 25–36, 2022, doi: 10.4097/kja.21209

REFERENCES

1. R. Benigni, "In silico assessment of genotoxicity. Combinations of sensitive structural alerts minimize false negative predictions for all genotoxicity endpoints and can single out chemicals for which experimentation can be avoided," Regulatory Toxicology and Pharmacology, vol. 124, p. 105042, 2021. Available: https://doi.org/10.1016/j.yrtph.2021.105042.
2. M. Honma, "An assessment of mutagenicity of chemical substances by (quantitative) structure–activity

relationship," Genes and Environment, vol. 42, no. 1, 2020. Available: https://doi.org/10.1186/s41021-020-00163-1.
3. N. Chatterjee and G. C. Walker, "Mechanisms of DNA damage, repair, and mutagenesis," Environmental and Molecular Mutagenesis, vol. 58, no. 5, pp. 235–263, 2017 . Available: https://doi.org/10.1002/em.22087.
4. M. Katerji and P. J. Duerksen-Hughes, "DNA damage in cancer development: Special implications in viral oncogenesis," American Journal of Cancer Research, vol. 11, no. 8, 2021.
5. R. Waller et al., "Metallothionein-I/II expression associates with the astrocyte DNA damage response and not Alzheimer-type pathology in the aging brain," Glia, vol. 66, no. 11, pp. 2316–2323, 2018. Available: https://doi.org/10.1002/glia.23465.
6. C. Milanese et al., "Activation of the DNA damage response in vivo in synucleinopathy models of Parkinson's disease," Cell Death & Disease, vol. 9, no. 8, 2018. . Available: https://doi.org/10.1038/s41419-018-0848-7.
7. F. Sobral et al., "Environmental xenobiotics and epigenetic modifications: Implications for human health and disease," Journal of Xenobiotics, vol. 15, no. 4, p. 118, 2025. .Available: https://doi.org/10.3390/jox15040118.
8. J. Xavier Santos, C. Rasga, and A. Moura Vicente, "Exposure to xenobiotics and gene-environment interactions in autism spectrum disorder: A systematic review," Autism Spectrum Disorder - Profile, Heterogeneity, Neurobiology, and Intervention, 2021. .Available: https://doi.org/10.5772/intechopen.95758.
9. M. Mišík et al., "Search for the optimal genotoxicity assay for routine testing of chemicals: Sensitivity and specificity of conventional and new test systems," Mutation Research/Genetic Toxicology and Environmental Mutagenesis, vol. 881, p. 503524, 2022. .Available: https://doi.org/10.1016/j.mrgentox.2022.503524.
10. N. Ren et al., "The various aspects of genetic and epigenetic toxicology: Testing methods and clinical applications," Journal of Translational Medicine, vol. 15, no. 1, 2017. .Available: https://doi.org/10.1186/s12967-017-1218-4.
11. H. Turkez, M. E. Arslan, and O. Ozdemir, "Genotoxicity testing: Progress and prospects for the next decade," Expert Opinion on Drug Metabolism & Toxicology, vol. 13, no. 10, pp. 1089–1098, 2017. .Available: https://doi.org/10.1080/17425255.2017.1375097.
12. Y. Luan and M. Honma, "Genotoxicity testing and recent advances," Genome Instability & Disease, vol. 3, no. 1, pp. 1–21, 2021 . Available: https://doi.org/10.1007/s42764-021-00058-7.
13. R. Sv, "Genotoxicity: Mechanisms, testing guidelines and methods," Global Journal of Pharmacy & Pharmaceutical Sciences, vol. 1, no. 5, 2017. .Available: https://doi.org/10.19080/gjpps.2017.01.555575.
14. S. Kislyak, O. Dugan, and O. Yalovenko, "Systems for genetic assessment of the impact of environmental factors," Innovative Biosystems and Bioengineering, vol. 8, no. 2, pp. 3–27, 2024. .Available: https://doi.org/10.20535/ibb.2024.8.2.288127.
15. J. Chrz et al., "Comparison of methods used for evaluation of mutagenicity/genotoxicity of model chemicals - parabens," Physiological Research, p. S661–S679, 2020. .Available: https://doi.org/10.33549/physiolres.934615.
16. B. N. Ames et al., "Carcinogens are mutagens: A simple test system combining liver homogenates for activation and bacteria for detection," Proceedings of the National Academy of Sciences, vol. 70, no. 8, pp. 2281–2285, 1973. . Available: https://doi.org/10.1073/pnas.70.8.2281.
17. J. Bhagat, "Combinations of genotoxic tests for the evaluation of group 1 IARC carcinogens," Journal of Applied Toxicology, vol. 38, no. 1, pp. 81–99, 2017 . Available: https://doi.org/10.1002/jat.3496.
18. T. Sofuni, "Evolution of genotoxicity test methods in Japan," Genes and Environment, vol. 39, no. 1, 2017. . Available: https://doi.org/10.1186/s41021-016-0063-7.
19. P. Gupta et al., "Genotoxicity evaluation of hospital wastewaters," Ecotoxicology and Environmental Safety, vol. 72, no. 7, pp. 1925–1932, 2009. .Available: https://doi.org/10.1016/j.ecoenv.2009.05.012.
20. OECD, "Overview on genetic toxicology TGs," 2017. .Available: https://doi.org/10.1787/9789264274761-en.
21. M. Dugan, "Salmonella typhimurium як тест-система для виявлення мутагенної активності забруднювачів навколишнього середовища," Цитологія і генетика, vol. 28, no. 3, pp. 37–41, 1994.
22. І. Р. Бариляк and О. М. Дуган, "Еколого-генетичні дослідження в Україні," Цитологія і генетика, no. 5, pp. 3–10, 2002.
23. C. S. M. Chu et al., "Machine learning – Predicting Ames mutagenicity of small molecules," Journal of Molecular Graphics and Modelling, p. 108011, 2021. Available: https://doi.org/10.1016/j.jmgm.2021.108011.
24. L. Chen et al., "A new 3D model for genotoxicity assessment: EpiSkin™ Micronucleus Assay," Mutagenesis, 2020. . Available: https://doi.org/10.1093/mutage/geaa003.
25. Y. H. Chung et al., "Animal welfare considerations when conducting OECD test guideline inhalation and toxicokinetic studies for nanomaterials," Animals, vol. 12, no. 23, p. 3305, 2022. Available: https://doi.org/10.3390/ani12233305.
26. T. T. Van Tran, H. Tayara, and K. T. Chong, "AMPred-CNN: Ames mutagenicity prediction model based on convolutional neural networks," Computers in Biology and Medicine, p. 108560, 2024. .Available: https://doi.org/10.1016/j.compbiomed.2024.108560.
27. C. Lou et al., "Chemical rules for optimization of chemical mutagenicity via matched molecular pairs analysis and machine learning methods," Journal of Cheminformatics, vol. 15, no. 1, 2023. .Available: https://doi.org/10.1186/s13321-023-00707-x.
28. Furuhama et al., "Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR

international challenge project," SAR and QSAR in Environmental Research, vol. 34, no. 12, pp. 983–1001, 2023. Available: https://doi.org/10.1080/1062936x.2023.2284902.

29. Helma et al., "A comparison of nine machine learning mutagenicity models and their application for predicting pyrrolizidine alkaloids," Frontiers in Pharmacology, vol. 12, 2021. .Available: https://doi.org/10.3389/fphar.2021.708050.

30. S. Kislyak et al., "In silico the Ames mutagenicity predictive model of environment," Innovative Biosystems and Bioengineering, vol. 9, no. 2, pp. 42–52, 2025. .Available: https://doi.org/10.20535/ibb.2025.9.2.316239.

31. J. Kazius, R. McGuire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," Journal of Medicinal Chemistry, vol. 48, no. 1, pp. 312–320, 2005. .Available: https://doi.org/10.1021/jm040835a.

32. K. Hansen et al., "Benchmark data set for in silico prediction of Ames mutagenicity," Journal of Chemical Information and Modeling, vol. 49, no. 9, pp. 2077–2081, 2009. .Available: https://doi.org/10.1021/ci900161g.

33. EFSA, "Dietary exposure assessment to pyrrolizidine alkaloids in the European population," EFSA Journal, vol. 14, no. 8, 2016. Available: https://doi.org/10.2903/j.efsa.2016.4572.

34. J. Tolosa et al., "MicotoXilico: an interactive database to predict mutagenicity, genotoxicity, and carcinogenicity of mycotoxins," Toxins, vol. 15, no. 6, p. 355, 2023. .Available: https://doi.org/10.3390/toxins15060355.

35. Y. Djoumbou Feunang et al., "ClassyFire: automated chemical classification with a comprehensive, computable taxonomy," Journal of Cheminformatics, vol. 8, no. 1, 2016. Available: https://doi.org/10.1186/s13321-016-0174-y.

36. S. M. Chu et al., "Machine learning – Predicting Ames mutagenicity of small molecules," Journal of Molecular Graphics and Modelling, p. 108011, 2021. .Available: https://doi.org/10.1016/j.jmgm.2021.108011.

37. L. Mellor et al., "Molecular fingerprint-derived similarity measures for toxicological read-across: recommendations for optimal use," Regulatory Toxicology and Pharmacology, vol. 101, pp. 121–134, 2019. Available: https://doi.org/10.1016/j.yrtph.2018.11.002.

38. Fan et al., "In silico prediction of chemical genotoxicity using machine learning methods and structural alerts," Toxicology Research, vol. 7, no. 2, pp. 211–220, 2018. .Available: https://doi.org/10.1039/c7tx00259a.

39. Cereto-Massagué et al., "Molecular fingerprint similarity search in virtual screening," Methods, vol. 71, pp. 58–63, 2015 . Available: https://doi.org/10.1016/j.ymeth.2014.08.005.

40. J. Zhang et al., "LightGBM: an effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets," Journal of Chemical Information and Modeling, vol. 59, no. 10, pp. 4150–4158, 2019. Available: https://doi.org/10.1021/acs.jcim.9b00633.

41. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," Korean Journal of Anesthesiology, vol. 75, no. 1, pp. 25–36, 2022. Available: https://doi.org/10.4097/kja.21209.

Biomedical Engineering and Technology
Issue 20(5), 2025 ISSN (Online) 2707-8483

UDC 504:57.04

IN SILICO MODELS FOR PREDICTING AMES MUTAGENICITY BASED ON MOLECULAR FINGERPRINT OF XENOBIOTICS

Sergey Kislyak
kisluak@ukr.net
Olexii Dugan
odugan51@gmail.com
Romaniuk Denys
romaniuk.denys@lll.kpi.ua
Olena Yalovenko
yalov89@i.ua
National Technical University of Ukraine
"Igor Sikorski Kyiv Polytechnic Institute";
Kyiv, Ukraine

Abstract *– One of the main global problems facing humanity in the 21st century is related to environmental pollution. Urbanization, active industrial development, and the introduction of modern technologies in all spheres of human activity contribute to the exponential growth of the number of chemical compounds entering the environment. A significant number of xenobiotics present in the environment can induce the development of hereditary and/or oncological diseases through mechanisms of direct or indirect influence on the human genetic apparatus. A significant increase in the number of recorded cases of cancer in different countries around the world is the main stimulus for the scientific community to intensify its efforts to effectively identify and record all environmental factors that may exhibit genotoxic properties. Given the priority of maintaining and preserving the genetic health of the human population, the basic in vitro and in vivo methods for assessing the genotoxicity of environmental factors currently need to be reviewed and improved. In this context, modern in silico approaches to assessing the genetic safety of environmental factors deserve attention, as they have significant but not fully realized potential. The paper presents a methodology for developing Ames mutagenicity assessment models (AMES/QSAR) focused on basic structural classes, which used different types of molecular fingerprints of xenobiotics as predictors. Two ensemble machine learning methods were selected to solve the binary classification problem: the random forest method and extreme gradient boosting. The accuracy of binary classifiers at the level of 80-85%, which were developed in accordance with the methodology presented in the paper, corresponds to the reproducibility of the Ames test in different laboratories. It has been shown that structure-oriented binary classifiers are more effective for predicting Ames mutagenicity than AMES/QSAR models, which used part of the complete heterogeneous input data set during the training stage.*

Keywords: mutagenicity, Ames test, QSAR model, xenobiotics, molecular descriptors, molecular fingerprints, machine learning models