УДК 504:57.04

IN SILICO МОДЕЛІ ПРОГНОЗУВАННЯ МУТАГЕННОСТІ ЕЙМСА ОСНОВНИХ СТРУКТУРНИХ КЛАСІВ КСЕНОБІОТИКІВ НА ОСНОВІ МЕТОДУ ВИПАДКОВОГО ЛІСУ

Кисляк Сергій Володимирович
kisluak@ukr.net
Дуган Олексій Мартем`янович
odugan51@gmail.com
Єсипенко Руслана В`ячеславівна
yesypenko.ruslana@lll.kpi.ua
Яловенко Олена Ігорівна
yalov89@i.ua
Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»
Київ, Україна

Анотація – *В умовах активного розвитку промисловості спостерігається щорічне суттєве збільшення кількості хімічних сполук, що можуть потрапляти у навколишнє середовище. Достатньо велика кількість ксенобіотиків, через механізми прямого або опосередкованого впливу на генетичний апарат людини, можуть індукувати спадкові та/або онкологічні захворювання. Зростаюча кількість хімічних речовин, що потрапляють у довкілля та можуть впливати на генетичне здоров`я людської популяції, викликає занепокоєння у громадськості. В такій ситуації критично важливим є ефективний контроль та облік усіх ксенобіотиків, що можуть проявляти потенційні генотоксичні властивості. Для класичних in vitro та in vivo методів, що дозволяють отримати оцінку генотоксичності факторів навколишнього середовища, наявні суттєві недоліки, що пов`язані зі складністю проведення експерименту, значними фінансовими витратами та низькою відтворюваністю результатів у різних лабораторіях. Крім того, при проведенні експериментальних досліджень необхідно враховувати прийняту науковою спільнотою концепцію «3R», що направлена на зменшення, вдосконалення та заміну моделей тварин. Такі обмеження у застосуванні класичних in vitro та in vivo методів генетичної оцінки факторів навколишнього середовища стали фундаментом для формування основного вектору розвитку сучасної комп`ютерної токсикології, в основі якого оцінка генотоксичних та токсичних ефектів досягається за допомогою прогностичних in silico моделей. У цьому контексті заслуговують на увагу QSAR моделі, що відповідно до кількісного зв`язку між структурою ксенобіотиків та активністю дозволяють отримати оцінку їх мутагенного потенціалу. У роботі представлена методологія покращення точності in silico моделей оцінки мутагенності Еймса (Ames/QSAR), на основі ансамблевих алгоритмів машинного навчання. Встановлено, що точність Ames/QSAR моделей можна підвищити через селекцію найбільш впливових ознак, що представлені молекулярними дескрипторами. Крім того, підвищення прогностичної здатності Ames/QSAR моделей може бути досягнуто за рахунок формування однорідних структурних класів ксенобіотиків, що представлені хімічними сполуками, які мають спільну будову молекулярного каркасу. Показано, що обмежений перелік молекулярних дескрипторів може бути використаний для пошуку причинно-наслідкових зв`язків між мутагенністю та фізико-хімічними, електронними, просторовими властивостями ксенобіотиків, що задаються різними наборами молекулярних дескрипторів.*

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

I. ВСТУП

В умовах глобальної індустріалізації та урбанізації відбувається суттєве збільшення кількості хімічних речовин, які можуть бути потенційними забруднювачами навколишнього середовища Кількість синтезованих хімічних сполук антропогенного походження, що можуть потрапляти у довкілля, складає більше ніж 400 на добу [1]. Європейським та американським хімічними товариствами були оприлюднені приблизно 800 тисяч хімічних речовин, для яких на сьогоднішній день відсутня інформація щодо потенційних ризиків для генетичної складової здоров'я людини та негативного впливу на навколишнє середовище [2]. У серпні 2024 року кількість зареєстрованих ксенобіотиків, інформація про які зберігалась на серверах Американського хімічного товариства, склала більше ніж 280 млн. речовин [3]. Крім того, для великої кількості ксенобіотиків процедура реєстрації здійснюється без врахування їх впливу на здоров'я людини взагалі, і зокрема – на спадковий апарат людини [4]. В умовах експоненційного збільшення кількості хімічних сполук, що генерує людство у різних сферах виробництва, особливо актуальною є проблема ефективного виявлення та обліку різноманітних факторів генетичної і канцерогенної небезпеки. Необхідність першочергового вирішення даної проблеми зумовлена генетичними наслідками впливу ксенобіотиків на спадковий апарат людини Взаємодія потенційних генотоксичних сполук з генетичним апаратом людини може стимулювати виникнення різноманітних генетичних захворювань (мутації в статевих клітинах), а також стати відправною точкою у розвитку онкологічних захворювань (мутації у соматичних клітинах)[5-7]. Відповідно до оприлюдненого звіту Всесвітньої організації охорони здоров`я у 2021 році модифікація навколишнього середовища генотоксичними сполуками стала причиною 18,9% усіх зареєстрованих у світі летальних випадків [8]. Мінімізація впливу потенційних генотоксичних сполук на спадковий апарат людини лежить в основі підтримання життєздатності організму на різних етапах онтогенезу та дозволяє забезпечити репродуктивну функцію [9-10].

За останні 50 років, з метою отримання об`єктивної оцінки генетичної безпеки факторів навколишнього середовища, були розроблені різноманітні *in vivo* та *in vitro* методи оцінки генотоксичності, частина з яких була прийнята науковою спільнотою, з подальшим затвердженням відповідними настановами міжнародних організацій (наприклад OECD, ECHA, UK-EMS, US-FDA, EFSA та ін.) [11,12]. Стандартизовані *in vivo* та *in vitro* методики, що формують стандартну батарею короткострокових тест-систем лежать в основі проведення генетичного тестування в країнах Європейського Союзу [11,13].

Застосування класичних методів та підходів для оцінки генотоксичного потенціалу ксенобіотиків мають суттєві недоліки, що пов`язані, у першу чергу, з проблемою відтворюваності результатів тестувань в різних лабораторіях, є складними з точку зору проведення, тривалими у часі та мають достатньо високу вартість [13-16]. Також потребує вирішення проблема, що пов`язана зі зменшенням кількості хибнонегативних та хибнопозитивних результатів тестування потенційних генотоксичних сполук. Крім того, прийнята науковою спільнотою концепція «3R», що направлена на зменшення використання у наукових дослідженнях моделей тварин стала основним стимулом для пошуку та розробки альтернативних методів генетичної оцінки факторів навколишнього середовища [17-18]. Питання щодо необхідності розробки нових підходів та методів генетичної оцінки факторів навколишнього середовища обговорюється у наукових працях [19-22]. З урахуванням активного розвитку інформаційних технологій та систем штучного інтелекту на початку 21 століття, спостерігається активізація наукової спільноти, що за допомогою сучасних *in silico* моделей намагається

отримати об`єктивну оцінку генотоксичного потенціалу факторів навколишнього середовища. У цьому контексті заслуговують на увагу QSAR (Quantitative Structure-Activity Relationship) моделі, що на основі алгоритмів машинного навчання дозволяють вирішити задачу бінарної класифікації з наступним розподілом досліджуваних ксенобіотиків на два класи: мутаген та не мутаген.

ІІ. МЕТА ДОСЛІДЖЕННЯ

Метою роботи є розробка, оптимізація та тестування ефективних *in silico* моделей прогнозування мутагенності Еймса факторів навколишнього середовища на основі методу випадкового лісу.

ІІІ. МАТЕРІАЛИ ТА МЕТОДИ

3.1 База даних ксенобіотиків

У межах проведеного дослідження з метою побудови моделей прогнозування мутагенності Еймса було використано базу даних ксенобіотиків [23], що була сформована шляхом об`єднання широко застосовуваних у наукових дослідженнях датасетів: Kazius-Bursi [24], Hansen [25] та EFSA [26]. Крім того, база даних хімічних сполук – потенційних мутагенів була розширена мікотоксинами [27]. Після видалення однакових за хімічним складом сполук 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 мутагенні властивості; 0 відповідає ксенобіотику, для якого відсутній мутагенний вплив на спадковий апарат людини).

3.2 Розподіл хімічніх сполук за основними структурними класами

З метою підвищення точності розроблених бінарних класифікаторів нами було запропоновано здійснити розподіл датасету на дев`ять однорідних структурних класів ксенобіотиків (таблиця 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

Розподіл хімічних сполук відповідно до особливостей будови молекулярної структури (Molecular Framework) був здійснений за допомогою веб-сервісу ClassyFire [28]. На наступному етапі, з урахуванням подібності між молекулярними структурами окремих класів хімічних сполук, було сформовано 5 груп ксенобіотиків. Відповідно до такого поділу аліфатичні ациклічні хімічні сполуки представляли першу групу хімічних сполук.

Друга група була отримана шляхом об`єднання двох класів ксенобіотиків: аліфатичних гетеромоноциклічних та аліфатичних гетерополіциклічних хімічних сполук. Необхідно зазначити, що наявний дисбаланс щодо кількості хімічних сполук, що відносяться до мутагенів та не мутагенів третьої групи ксенобіотиків, а також невелика кількість мутагенів у порівнянні з не мутагенами, може мати значний негативний вплив на прогностичну здатність орієнтованих на відповідні класи ксенобіотиків Ames/QSAR моделей. В такій ситуації, з метою отримання оцінки мутагенності Еймса ксенобіотиків, що відносяться до аліфатичних гомомоноциклічних та аліфатичних гомополіциклічних хімічних сполук (третя група) нами було прийнято рішення використовувати Ames/QSAR моделі, які на етапі навчання використовували частину повного датасету, що представлений 8454 ксенобіотиками. Четверта група об`єднує ксенобіотики, що належать до ароматичних гетеромоноциклічних та ароматичних гетерополіциклічних хімічних сполук.

Ароматичні гомомоноциклічні та ароматичні гомополіциклічні хімічні сполуки представляли п`яту групу ксенобіотиків. Пропорційний розподіл кількості мутагенів до кількості не мутагенів для першої, другої, четвертої та п`ятої груп ксенобіотиків дозволив використовувати відповідні набори даних з метою створення моделей прогнозування мутагенності Еймса, що орієнтовані на основні структурні класи (табл. 1) хімічних сполук – потенційних мутагенів.

3.3 Молекулярні дескриптори

Молекулярні дескриптори, що виступають у ролі базових предикторів *in silico* Ames/QSAR моделей, можуть бути отримані за допомогою різноманітного програмного забезпечення (наприклад PaDEL, OpenBabel, Chemopy, CDK, RDKit, DRAGON, Mordred та інші). Для розрахунку молекулярних дескрипторів широко використовуваними серед науковців є також такі веб-сервіси, як BioTriangle [29], Galaxy [30], ChemDes [31]. Досить цікавим у науковому відношенні може бути отриманий результат порівняння ефективності орієнтованих на основні структурні класи ксенобіотиків моделей прогнозування мутагенності Еймса з урахуванням різних наборів молекулярних дескрипторів. Тому, у межах роботи, було запропоновано при реалізації Ames/QSAR моделей використовувати три різних наборів молекулярних дескрипторів (PaDel, Mordred та RDkit), що були розраховані за допомогою веб-сервісу Galaxy [30].

Молекулярні дескрипторі PaDEL, RDkit та Mordred що виступають в ролі предикторів при реалізації Ames/QSAR моделей мають різний кількісний та якісний склад. Така особливість, у першу чергу, пов`язана з тим, що розрахунок молекулярних дескрипторів здійснюється за допомогою різноманітних програм та веб-сервісів. При цьому варіабельність між розрахованими дескрипторами, може бути обумовлена різними підходами щодо їх обчислення. У такій ситуації пошук відповіді на питання щодо оцінки ефективності моделей прогнозування мутагенності Еймса, які на етапі навчання використовують різні набори вхідних даних є науково обґрунтованою та потребує вирішення. Кількість розрахованих 1D та 2D дескрипторів PaDEL, що використовувались при реалізації прогностичних моделей, склала 1444, а RDkit – 196. Дескриптори Mordred формують одну з найбільших групу, що представлена 1613 предикторами. Цікавим є той факт, що 30% дескрипторів Mordred відрізняється за якісним складом від

дескрипторів PaDEL. Необхідно звернути увагу на те, що при реалізації Ames/QSAR моделей ми не використовували відбитки просторової структури (molecular fingerprimt), що відносяться до 2D молекулярних дескрипторів. При створенні прогностичних моделей дану групу молекулярних дескрипторів, зазвичай, розглядають окремо

3.4 Методи дослідження

Реалізація моделей прогнозування мутагенності Еймса була здійснена на основі методу випадкового лісу (RF) Відповідний вибір методу, що лежить в основі вирішення задачі бінарної класифікації, був обумовлений достатньо великою кількістю нещодавно опублікованих наукових праць [23,27,32,33], в яких розроблені бінарні класифікатори на основі методу RF демонстрували одну з найкращих точність прогнозування мутагенності Еймса.

Оптимізація моделей машинного навчання була здійснена через зменшення обсягу вхідних даних, які задаються молекулярними дескрипторами, за допомогою алгоритму RFECV (Recursive Feature Elimination with Cross-Validation), що дозволяє здійснювати відбір найбільш впливових дескрипторів на етапі крос-валідації. Такий підхід дозволяє, через селекцію найбільш релевантних ознак, покращити точність розроблених Ames/QSAR моделей. При цьому ознаки, що мали менш вагомий вплив на прогнозовану змінну видалялись. Опубліковані результати досліджень [27,34] дозволяють впевнитись у тому, що оптимізація моделей машинного навчання, що здійснюється через рекурсивне видалення найменш впливових ознак дозволяє отримати більш ефективні класифікатори.

3.5 Препроцесинг даних

Прогностична здатність розроблених моделей оцінки мутагенності Еймса на пряму залежить від того, на скільки якісно була проведена процедура попередньої обробки даних. На початковому етапі препроцесингу було здійснено видалення не інформативних даних, що стосується порядкових номерів ксенобіотиків, SMILES нотації та інформації щодо приналежності хімічних сполук до відповідного структурного класу. Також було проведено видалення ознак, для яких стандартне відхилення дорівнює нулю. Наступний крок був пов`язаний з формуванням вектора $\boldsymbol{y}$ та матриці $\boldsymbol{x}$. Елементи вектору $\boldsymbol{y}$ приймають одне з двох можливих значень цільової змінної (0 – відповідає ксенобіотику для якого, відповідно до *in vitro* тесту Еймса, був отриманий негативний результат тестування; 1 позначає хімічні сполуки з вираженими мутагенними властивостями). Матриця $\boldsymbol{x}$ містить інформацію про молекулярні дескриптори для кожного ксенобіотика.

Проблема мультиколінеарності, що може мати негативний вплив на точність розроблених Ames/QSAR моделей, була вирішена через видалення ознак, що мають достатньо сильну кореляцію ознак. Для цього використовувалась матриця корельованих ознак, кожна комірка якої містила значення коефіцієнтів кореляції Пірсона, що були розраховані між всіма парами молекулярних дескрипторів. Фільтрація вхідних даних здійснювалась з урахуванням граничного значення коефіцієнтів кореляції $r > 0{,}95$.

Наступний етап препроцесингу даних пов`язаний з приведенням значень різних молекулярних дескрипторів до певного масштабу, що в рамках проведеного дослідження було реалізовано за допомогою двох інструментів QuantileTransformer та StandardScaler бібліотеки scikit-learn Pyton. QuantileTransformer з параметром output_distribution='uniform' дозволяє отримати рівномірний розподіл кожної ознаки в межах інтервалу [0,1], що дозволяє мінімізувати негативний вплив аномальних значень (викидів). Стандартизація даних була реалізована за допомогою інструменту StandardScaler, що дозволяє отримати набір ознак з середнім значенням, що дорівнює 0 та стандартним відхиленням 1.

При вирішенні задач бінарної та багатокласової класифікації дослідники

часто стикаються з проблемою, що пов`язана з наявністю дисбалансу класів, що може негативно вплинути на точність побудованих моделей [35]. Не зважаючи на те, що для першої, другої, четвертої та п`ятої груп (таб. 1) ксенобіотиків відсутній критичний дисбаланс з точки зору кількості мутагенів у порівнянні з не мутагенами, нами було прийнято рішення доповнити менш представлені класи згенерованими зразками за допомогою інструменту SMOTE (Synthetic Minority Over-sampling Technique) бібліотеки Imbalanced-learn мови програмування Pyton. Такий підхід дозволяє отримати однакову кількість зразків, які відносяться до двох класів (мутаген/не мутаген), що може мати позитивний вплив на точність розроблених моделей прогнозування мутагенності Еймса.

Наступний етап попередньої обробки даних був пов`язаний з формуванням (у співвідношенні 80:20) двох вибірок – навчальної та тестової. Необхідно зазначити, що подібний підхід відносно розподілу даних на дві вибірки використовувався як для окремих груп (див. табл. 1) хімічних сполук, так і для повної бази даних, що налічує 8454 ксенобіотиків. Важливим з наукової точки зору може бути отриманий результат порівняння ефективності Ames/QSAR моделей, що були отримані відповідно до однорідних наборів даних, для яких спостерігається подібність будови молекулярного каркасу у порівнянні з моделями, для яких на етапі тренування використовувалась частина вхідних даних повного датасету. Результати нещодавно опублікованої наукової праці [36] дозволили нам прийняти рішення щодо зміни стратегії розподілу даних на окремі вибірки. Автори статті наголошують на необхідності формування додаткового альтернативного тестового набору даних, що дозволить отримати об`єктивну оцінку ефективності розроблених класифікаторів особливо у контексті узагальнюючої здатності моделей. Тому дані випадковим чином було розподілено на три вибірки: на тренувальну та дві тестові вибірки, у співвідношенні 80:10:10. З метою отримання об`єктивної оцінки ефективності розроблених класифікаторів на етапі навчання, було запропоновано використовувати п’ятикратну перехресну перевірку(крос-валідацію). Такий підхід дозволяє здійснювати підбір оптимальних гіперпараметрів Ames/QSAR моделей, при яких досягається максимальна точність розроблених бінарних класифікаторів. Остаточну оцінку прогностичної здатності Ames/QSAR моделей було проведено на другій тестовій вибірці.

Слід зазначити, що в науковій літературі терміни «тестовий набір» та «валідаційний набір» іноді вживаються як взаємозамінні [36]. З метою уникнення неоднозначності у трактуванні цих двох термінів, будемо називати валідаційною вибіркою набір даних, що використовується для оцінки ефективності Ames/QSAR моделей та налаштувань гіперпараметрів. При цьому тестова вибірка представляє собою частину вхідних даних, що використовується для остаточної перевірки бінарних класифікаторі за умов вже підібраних оптимальних значень гіперпараметрів Ames/QSAR моделей.

3.6 Оцінка ефективності Ames/QSAR моделей

Ефективність розроблених *in silico* Ames/QSAR моделей оцінювалась за допомогою наступних метрик: загальної точності (*accuracy)*, точності позитивного прогнозу (*precision),* чутливості (*recall*) та F_1-міри ($F_1 - score$), які були отримані з урахуванням матриць помилок відповідно до співвідношень 1-4, де TP, TN, FP, FN – відповідає кількості істинно позитивних, істинно негативних, хибнопозитивних та хибнонегативних результатів класифікації відповідно.

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

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

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

$$F_1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (4)$$

IV РЕЗУЛЬТАТИ

У таблиці 2 наведено інформацію щодо експериментально визначених налаштувань базових гіперпараметрів, при яких точність для п`яти Ames/QSAR моделей на основі методу випадкового лісу була максимальною. На даному етапі проведення досліджень при виконанні процедури навчання Ames/QSAR моделей ми використовували повні набори вхідних даних, що були представлені молекулярними дескрипторами PaDell, RdKit та Mordred.

Таблиця 2. Налаштування базових гіперпараметрів та оцінка точності (крос-валідація) Ames/QSAR моделей, що побудовані на основі повних наборів вхідних даних, без зменшення їх розмірності.

Молекулярні дескриптори	Структурні класи ксенобіотиків	Кількість дерев	Кількість листків	Точність (accuracy)	AUC
PaDell	Аліфатичні ациклічні	500	300	0,8344	0,89
	Аліфатичні гетеромоно (полі)циклчіні	100	200	0,8451	0,92
	Ароматичні гетеромоно (полі)циклчіні	650	200	0,8561	0,93
	Ароматичні гомомоно (полі) циклічні	650	200	0,8429	0,9
	Всі	300	600	0,8462	0,91
RdKit	Аліфатичні ациклічні	200	200	0,8505	0,95
	Аліфатичні гетеромоно (полі)циклчіні	100	100	0,8157	0,9
	Ароматичні гетеромоно (полі) циклчіні	400	300	0,8604	0,93
	Ароматичні гомомоно (полі) циклічні	300	400	0,8417	0,91
	Всі	350	500	0,8462	0,91
Mordred	Аліфатичні ациклічні	250	250	0,8465	0,86
	Аліфатичні гетеромоно (полі) циклчіні	200	200	0,8392	0,89
	Ароматичні гетеромоно (полі) циклчіні	550	350	0,8576	0,93
	Ароматичні гомомоно (полі) циклічні	400	300	0,8432	0,92
	Всі	350	500	0,8466	0,91

Тренування Ames/QSAR моделей здійснювалось окремо для чотирьох груп ксенобіотиків (див. табл.1.). Крім того, на основі трьох наборів молекулярних дескрипторів (PaDell, RdKit та Mordred) був розроблений бінарний класифікатор, що на етапі навчання використовував частину (80%) повної бази даних, що була представлена 8454 ксенобіотиками. Отримані проміжні результати моделювання дозволяють зробити висновки про те, що відповідно до показників точності та площі під ROC-кривою, можна спостерігати тенденцію покращення точності для окремих, орієнтованих на структурні класи (див. табл.2) Ames/QSAR моделей, у порівнянні з бінарними класифікаторами, для яких на етапі навчання використовувався неоднорідний, з точки зору будови молекулярного каркасу набір даних. Разом з

цим, для декількох клас-орієнтованих моделей прогнозування мутагенності Еймса спостерігається незначне зниження показника точності.

Наступний етап дослідження був присвячений перевірці гіпотези щодо можливостей покращення прогностичної здатності *in silico* Ames/QSAR моделей, що може бути здійснено через селекцію релевантних ознак. В якості основного алгоритму, що дозволяє здійснювати відбір найбільш впливових дескрипторів, був обраний RFECV, що поєднаний з п'ятикратною перехресною перевіркою (крос-валідацією).

У таблиці 3 для кожної Ames/QSAR моделі, які на етапі навчання використовували різні набори предикторів, представлена інформація про кількість релевантних дескрипторів, що залишилась після проведення процедури рекурсивного видалення менш впливових ознак.

Таблиця 3 Оцінка точності (крос-валідація) Ames/QSAR моделей, що побудовані на основі найбільш релевантних ознак

Молекулярні дескриптори	Структурні класи ксенобіотиків	Кількість релевантних ознак	Точність (accuracy)	AUC
PaDell	Аліфатичні ациклічні	612	0,8454	0,89
	Аліфатичні гетеромоно (полі) циклічні	49	0,8627	0,93
	Ароматичні гетеромоно (полі) циклічні	649	0,8608	0,93
	Ароматичні гомомоно (полі) циклічні	274	0,847	0,9
	Всі (1444)	289	0,8489	0,91
RdKit	Аліфатичні ациклічні	94	0,8505	0,95
	Аліфатичні гетеромоно (полі) циклічні	85	0,8176	0,9
	Ароматичні гетеромоно (полі) циклічні	139	0,8615	0,93
	Ароматичні гомомоно (полі) циклічні	87	0,8447	0,91
	Всі (196)	145	0,8473	0,91
Mordred	Аліфатичні ациклічні	322	0,8456	0,87
	Аліфатичні гетеромоно (полі) циклічні	218	0,8412	0,9
	Ароматичні гетеромоно (полі) циклічні	675	0,8569	0,93
	Ароматичні гомомоно (полі) циклічні	416	0,8455	0,93
	Всі (1613)	776	0,8466	0,91

Заслуговують на увагу з боку науковців отримані проміжні результати оцінки ефективності орієнтованих на основні структурні класи Ames/QSAR моделі, що використовували в якості предикторів набір релевантних молекулярних дескрипторів. Оцінка точності бінарних класифікаторів була здійснена при проведенні п`ятикратної перехресної перевірки. При цьому у таблиці 3 записані середні значення метрик точності (accuracy) та площі під ROC кривою.

Зменшення кількості вхідних даних у діапазоні від 55% до 97% від початкової кількості дескрипторів (PaDel, Mordred та RDkit) призводило, у більшості випадків, до покращення точності орієнтованих на

структурні класи Ames/QSAR моделей у порівнянні з розробленими бінарними класифікаторами, які на етапі навчання використовували частину (80%) повного набору вхідних даних (див. табл. 3.). Три *in silico* Ames/QSAR моделі, що на етапі навчання використовували в якості предикторів набір релевантних дескрипторів Mordred, показали незначне зниження точності. При цьому, показники AUC для всіх моделей залишились практично без змін, що вказує на збережений баланс бінарних класифікаторів щодо позитивних (мутаген) та негативних (не мутаген) прогнозів. Сформульована, на початку дослідження, гіпотеза щодо покращення точності *in silico* моделей прогнозування мутагенності Еймса отримала своє підтвердження. У такій ситуації наступним необхідним кроком є проведення остаточної перевірки узагальнюючої здатності розроблених Ames/QSAR моделей на даних, що не використовувались у навчанні У таблиці 4 представлені результати тестування моделей що були отримані на двох незалежних тестових вибірках.

Таблиця 4. Класифікаційний звіт на тестовій вибірці для Ames/QSAR моделей з обмеженим переліком релевантних дескрипторів

Молекулярні дескриптори	**Структурні класи ксенобіотиків**	**Точність 1тестова**	**Точність 2 тестова**	**Precision**	**Recall**	**Specificity**	**F1 Score**	**AUC**
PaDell	Аліфатичні ациклічні	0.8133	0,8106	0,8167	0,7778	0.8406	0,7967	0,88
	Аліфатичні гетеромоно (полі) циклічні	0.8359	0,8448	0,871	0,8434	0.8462	0,8571	0,91
	Ароматичні гетеромоно (полі) циклічні	0.8654	0,873	0,9	0,8571	0.8912	0,878	0,93
	Ароматичні гомомоно (полі) циклічні	0.8262	0,8516	0,8758	0,8323	0.8725	0,8535	0,93
	Всі	0.8363	0,8556	0,87	0,846	0.8659	0,8578	0,92
RdKit	Аліфатичні ациклічні	0.8836	0,8947	0,9412	0,8889	0.9048	0,9443	0,96
	Аліфатичні гетеромоно (полі) циклічні	0.8143	0,8103	0,8586	0,8462	0.7812	0,8	0,93
	Ароматичні гетеромоно (полі) циклічні	0.8899	0,8921	0,913	0,8802	0.9054	0,8963	0,94
	Ароматичні гомомоно (полі) циклічні	0.8433	0,8262	0,8188	0,8133	0.8313	0,8161	0,91
	Всі	0.842	0,8568	0,8632	0,8531	0.8606	0,8581	0,93
Mordred	Аліфатичні ациклічні	0.7961	0,7807	0,7451	0,8085	0.806	0,7755	0,85
	Аліфатичні гетеромоно (полі) циклічні	0.8337	0,8966	0,913	0,84	0.9394	0,875	0,95
	Ароматичні гетеромоно (полі) циклічні	0.8563	0,8676	0,8986	0,8158	0.9141	0,8552	0,94

	Ароматичні гомомоно (полі) циклічні	0.849	0,871	0,8741	0,8503	0.8896	0,8621	0,95
	Всі	0.8296	0,8544	0,8665	0,84	0.869	0,853	0,93

Перш ніж перейти до наступного етапу дослідження, що пов`язаний з аналізом класифікаційних звітів розроблених Ames/QSAR моделей, необхідно зазначити, що точність на рівні 85% [23,27,32,37], яка відповідає зареєстрованій варіабельності *in vitro* теста Еймса [38], вважається досить хорошим результатом. Зростання значень основних метрик оцінки ефективності Ames/QSAR моделей навіть у межах 1-2% може мати як наукове, так і практичне значення.

Низька варіативність значень метрики точності (accuracy) на двох тестових вибірках (табл. 4), додає впевненості у тому, що в реальних умовах розроблені бінарні класифікатори проявлятимуть стійкість та стабільність. З метою отримання об`єктивної оцінки узагальнюючої здатності Ames/QSAR моделей на основі другої тестової вибірки були розраховані такі метрики ефективності як: accuracy, precision, recall та F1 Score. Для кожної моделі також була розрахована площа під ROC кривою (AUC), що є достатньо популярним інструментом для оцінки ефективності бінарних класифікаторів [39] та дозволяє зробити оцінку взаємозв`язків між чутливістю (recall) та специфічністю (specificity) у графічному вигляді. За результатами аналізу класифікаційних звітів (табл. 4) можна зробити висновок про те, що більшість Ames/QSAR моделей, що орієнтовані на основні структурні класи ксенобіотиків (табл 1), за умов селекції найбільш впливових ознак, зберігали високі показники класифікації.

Для отримання *in silico* оцінки мутагенності Еймса аліфатичних ациклічних хімчних сполук доцільно використовувати Ames/QSAR модель, яка була отримана, з урахуванням селекції релевантних дескрипторів RdKit, на однорідному, з точки зору будови молекулярного каркасу, наборі даних, що відноситься до першої групи ксенобіотиків. Реалізована модель з $AUC = 0{,}96$ має суттєві переваги у порівнянні з бінарним класифікатором (з $AUC = 0{,}93$), що був отриманий відповідно до сформованої тренувальної вибірки, з урахуванням відбору найбільш впливових ознак, на основі повної бази даних ксенобіотиків.

Для отримання *in silico* оцінки мутагенності Еймса ксенобіотиків, що відносяться до другої групи ксенобіотиків (аліфатичні гетеромоно та аліфатичні поліциклічні хімічні сполуки) необхідно використовувати бінарний класифікатор, для якого на етапі навчання використовувався оптимізований набір даних, що представлений дескрипторами Mordred, які мають найбільший вплив на прогнозовану змінну та були розраховані для другої групи ксенобіотиків. Показник точності (*accuracy*) для даної моделі становив 90%, що є одним з найкращих показників у порівнянні з іншими, розробленими у межах роботи, бінарними класифікаторами. Цікавим є той факт, що точність моделі, яка була отримана відповідно до класичного підходу з використанням дескрипторів повної бази даних та відбором релевантних предикторів, становила тільки 85%.

З метою отримання *in silico* оцінки мутагенності Еймса для ароматичних гетеромоноциклічних та ароматичних гетерополіциклічних хімічних сполук найбільш ефективною є Ames/QSAR модель, що була отримана на основі оптимального набору релевантних дескрипторів RDkit, що були розраховані для четвертої групи ксенобіотиків (див. табл. 1).

Максимальна ефективність щодо розподілу ароматичних гомомоноциклічних та ароматичних гомополіциклічних хімічних сполук між мутагенами та не мутагенами досягається через використання Ames/QSAR

моделі, яка побудована на основі відібраних найбільш впливових дескрипторів Mordred, що були розраховані для однорідної, з точки зори будови молекулярної структури, п`ятої групи (табл. 1) ксенобіотиків.

V ВИСНОВКИ

У роботі запропоновано новий підхід, що лежить в основі покращення прогностичної здатності *in silico* моделей прогнозування мутагенності Еймса, що досягається через зменшення обсягу вхідних даних та селекцію найбільш впливових ознак. Показано, що збільшення ефективності прогностичних моделей може бути реалізовано через формування однорідних груп ксенобіотиків, що мають спільні риси будови молекулярної структури. Представлена у роботі методика може стати фундаментом для пошуку причинно-наслідкових зв`язків між певними властивостями ксенобіотиків, які представлені набором релевантних дескрипторів та проявами мутагенності.

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
0009-0003-4701-7985 (F) Ruslana Yesypenko
0000-0002-5022-143X (G) Olena Yalovenko

A – 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. Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project / M. Honma et al. Mutagenesis. 2018. Vol. 34, no. 1. P. 3–16. URL: https://doi.org/10.1093/mutage/gey031
2. From molecular descriptors to intrinsic fish toxicity of chemicals: an alternative approach to chemical prioritization / S. Samanipour et al. Environmental science & technology. 2022. URL: https://doi.org/10.1021/acs.est.2c07353
3. 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
4. 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
5. The therapeutic potential of DNA damage repair pathways and genomic stability in lung cancer / J. T. Burgess et al. *Frontiers in oncology*. 2020. Vol. 10. URL: https://doi.org/10.3389/fonc.2020.01256
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. 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
8. Clark S. N., Anenberg S. C., Brauer M. Global burden of disease from environmental factors. *Annual review of public health*. 2024. URL: https://doi.org/10.1146/annurev-publhealth-071823-105338
9. Tubbs A., Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. Cell. 2017. Vol. 168, no. 4. P. 644–656. URL: https://doi.org/10.1016/j.cell.2017.01.002
10. USP7 is a master regulator of genome stability / G. J. Valles et al. Frontiers in cell and developmental biology. 2020. Vol. 8. URL: https://doi.org/10.3389/fcell.2020.00717
11. Overview on genetic toxicology TGs. OECD, 2017. URL: https://doi.org/10.1787/9789264274761-en
12. 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
13. 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
14. Дуган, А. М. Salmonella typhimurium як тест-система для виявлення мутагенної активності забруднювачів навколишнього середовища / А. М. Дуган // Цитологія і генетика. – 1994. – Т. 28, № 3. – С. 37–41.
15. Бариляк, І. Р. Еколого-генетичні дослідження в Україні / І. Р. Бариляк, О. М. Дуган // Цитологія і генетика. – 2002. – № 5. – С. 3–10.

16. 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
17. 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
18. Quantitative structure–activity relationship models for genotoxicity prediction based on combination evaluation strategies for toxicological alternative experiments / X. Yang et al. *Scientific reports*. 2021. Vol. 11, no. 1. URL: https://doi.org/10.1038/s41598-021-87035-y
19. A Modern Genotoxicity Testing Paradigm: Integration of the High-Throughput CometChip® and the TGx-DDI Transcriptomic Biomarker in Human HepaRG™ Cell Cultures / J. K. Buick et al. Frontiers in Public Health. 2021. Vol. 9. URL: https://doi.org/10.3389/fpubh.2021.694834
20. A new method to predict genotoxic effects based on serum molecular profile / R. Araújo et al. Spectrochimica acta part A: molecular and biomolecular spectroscopy. 2021. Vol. 255. P. 119680. URL: https://doi.org/10.1016/j.saa.2021.119680
21. Making in silico predictive models for toxicology FAIR / M. T. D. Cronin et al. *Regulatory toxicology and pharmacology*. 2023. P. 105385. URL: https://doi.org/10.1016/j.yrtph.2023.105385
22. Application of a new approach methodology (NAM)-based strategy for genotoxicity assessment of data-poor compounds / A.-M. V. Fortin et al. Frontiers in toxicology. 2023. Vol. 5. URL: https://doi.org/10.3389/ftox.2023.1098432
23. 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
24. Kazius J., McGuire R., Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry. 2005. Vol. 48, no. 1. P. 312–320. URL: https://doi.org/10.1021/jm040835a
25. Benchmark data set for in silico prediction of ames mutagenicity / K. Hansen et al. Journal of chemical information and modeling. 2009. Vol. 49, no. 9. P. 2077–2081. URL: https://doi.org/10.1021/ci900161g
26. Dietary exposure assessment to pyrrolizidine alkaloids in the European population. EFSA journal. 2016. Vol. 14, no. 8. URL: https://doi.org/10.2903/j.efsa.2016.4572
27. 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
28. 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
29. BioTriangle: a web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions / J. Dong et al. Journal of cheminformatics. 2016. Vol. 8, no. 1. URL: https://doi.org/10.1186/s13321-016-0146-2
30. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update / E. Afgan et al. Nucleic acids research. 2022. URL: https://doi.org/10.1093/nar/gkac247
31. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation / J. Dong et al. Journal of cheminformatics. 2015. Vol. 7, no. 1. URL: https://doi.org/10.1186/s13321-015-0109-z
32. 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
33. 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
34. Benchmarking variants of recursive feature elimination: insights from predictive tasks in education and healthcare / O. Bulut et al. Information. 2025. Vol. 16, no. 6. P. 476. URL: https://doi.org/10.3390/info16060476
35. Research on expansion and classification of imbalanced data based on SMOTE algorithm / S. Wang et al. Scientific reports. 2021. Vol. 11, no. 1. URL: https://doi.org/10.1038/s41598-021-03430-5
36. Chicco D., Jurman G. The ABC recommendations for validation of supervised machine learning results in biomedical sciences. Frontiers in big data. 2022. Vol. 5. URL: https://doi.org/10.3389/fdata.2022.979465
37. Shinada, N.K., Koyama, N., Ikemori, M., Nishioka, T., Hitaoka, S., Hakura, A., Asakura, S., Matsuoka, Y., Palaniappan, S.K. Optimizing machine-learning models for mutagenicity prediction through better feature selection // Mutagenesis. – 2022. – Vol. 37, № 3–4. – P. 191–202. URL: https://doi.org/10.1093/mutage/geac010.
38. Piegorsch W. W., Zeiger E. Measuring intra-assay agreement for the ames salmonella assay. Statistical methods in toxicology. Berlin, Heidelberg, 1991. P. 35–41. URL: https://doi.org/10.1007/978-3-642-48736-1_5
39. Nahm F. S. Receiver operating characteristic curve: overview and practical use for clinicians. Korean journal of anesthesiology. 2022. Vol. 75, no. 1. P. 25–36. URL: https://doi.org/10.4097/kja.21209

REFERENCES

1. M. Honma et al., "Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project," Mutagenesis, vol. 34, no. 1, pp. 3–16, 2018, doi: 10.1093/mutage/gey031.
2. S. Samanipour et al., "From molecular descriptors to intrinsic fish toxicity of chemicals: an alternative approach to chemical prioritization," Environmental Science & Technology, 2022, doi: 10.1021/acs.est.2c07353.
3. 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, doi: 10.20535/ibb.2025.9.2.316239.

4. M. Honma, "An assessment of mutagenicity of chemical substances by (quantitative) structure–activity relationship," Genes and Environment, vol. 42, no. 1, 2020, doi: 10.1186/s41021-020-00163-1.
5. J. T. Burgess et al., "The therapeutic potential of DNA damage repair pathways and genomic stability in lung cancer," Frontiers in Oncology, vol. 10, 2020, doi: 10.3389/fonc.2020.01256.
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, doi: 10.1038/s41419-018-0848-7.
7. 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, doi: 10.1002/glia.23465.
8. S. N. Clark, S. C. Anenberg, and M. Brauer, "Global burden of disease from environmental factors," Annual Review of Public Health, vol. 45, 2024. doi: 10.1146/annurev-publhealth-071823-105338
9. A. Tubbs and A. Nussenzweig, "Endogenous DNA damage as a source of genomic instability in cancer," Cell, vol. 168, no. 4, pp. 644–656, 2017, doi: 10.1016/j.cell.2017.01.002.
10. G. J. Valles et al., "USP7 is a master regulator of genome stability," Frontiers in Cell and Developmental Biology, vol. 8, 2020, doi: 10.3389/fcell.2020.00717.
11. OECD, Overview on genetic toxicology TGs, 2017, doi: 10.1787/9789264274761-en.
12. 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, doi: 10.1016/j.mrgentox.2022.503524.
13. 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, doi: 10.20535/ibb.2024.8.2.288127.
14. 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.
15. І. Р. Бариляк and О. М. Дуган, "Еколого-генетичні дослідження в Україні," Цитологія і генетика, no. 5, pp. 3–10, 2002.
16. 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.
17. C. 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, doi: 10.1016/j.yrtph.2018.11.002.
18. X. Yang et al., "Quantitative structure–activity relationship models for genotoxicity prediction based on combination evaluation strategies for toxicological alternative experiments," Scientific Reports, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-87035-y. [Accessed: 14-Jun-2025].
19. J. K. Buick et al., "A Modern Genotoxicity Testing Paradigm: Integration of the High-Throughput CometChip® and the TGx-DDI Transcriptomic Biomarker in Human HepaRG™ Cell Cultures," Frontiers in Public Health, vol. 9, 2021, doi: 10.3389/fpubh.2021.694834.
20. R. Araújo et al., "A new method to predict genotoxic effects based on serum molecular profile," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 255, p. 119680, 2021, doi: 10.1016/j.saa.2021.119680.
21. M. T. D. Cronin et al., "Making in silico predictive models for toxicology FAIR," Regulatory Toxicology and Pharmacology, p. 105385, 2023, doi: 10.1016/j.yrtph.2023.105385.
22. A.-M. V. Fortin et al., "Application of a new approach methodology (NAM)-based strategy for genotoxicity assessment of data-poor compounds," Frontiers in Toxicology, vol. 5, 2023, doi: 10.3389/ftox.2023.1098432.
23. C. Helma et al., "A comparison of nine machine learning mutagenicity models and their application for predicting pyrrolizidine alkaloids," Frontiers in Pharmacology, vol. 12, 2021, doi: 10.3389/fphar.2021.708050.
24. 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.
25. 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.
26. 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.
27. J. Tolosa et al., "MicotoXilico: an interactive database to predict mutagenicity, genotoxicity, and carcinogenicity of mycotoxins," Toxins, vol. 15, no. 6, p. 355, 2023, doi: 10.3390/toxins15060355.
28. Y. Djoumbou Feunang et al., "ClassyFire: automated chemical classification with a comprehensive, computable taxonomy," Journal of Cheminformatics, vol. 8, no. 1, 2016, doi: 10.1186/s13321-016-0174-y.
29. J. Dong et al., "BioTriangle: a web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions," Journal of Cheminformatics, vol. 8, no. 1, 2016, doi: 10.1186/s13321-016-0146-2.
30. E. Afgan et al., "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update," Nucleic Acids Research, 2022, doi: 10.1093/nar/gkac247.
31. J. Dong et al., "ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation," Journal of Cheminformatics, vol. 7, no. 1, 2015, doi: 10.1186/s13321-015-0109-z.
32. 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, doi: 10.1186/s13321-023-00707-x.
33. 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, doi: 10.1080/1062936x.2023.2284902.
34. O. Bulut et al., "Benchmarking variants of recursive feature elimination: insights from predictive tasks in education and healthcare," Information, vol. 16, no. 6, p. 476, 2025, doi: 10.3390/info16060476.
35. S. Wang et al., "Research on expansion and classification of imbalanced data based on SMOTE algorithm," Scientific Reports, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-03430-5.
36. D. Chicco and G. Jurman, "The ABC recommendations for validation of supervised machine learning results in biomedical sciences," Frontiers in Big Data, vol. 5, 2022, doi: 10.3389/fdata.2022.979465.
37. N. K. Shinada et al., "Optimizing machine-learning models for mutagenicity prediction through better feature selection," Mutagenesis, vol. 37, no. 3–4, pp. 191–202, 2022, doi: 10.1093/mutage/geac010.
38. W. W. Piegorsch and E. Zeiger, "Measuring intra-assay agreement for the Ames Salmonella assay," Statistical Methods in Toxicology, Berlin, Heidelberg, pp. 35–41, 1991, doi: 10.1007/978-3-642-48736-1_5.
39. 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

Biomedical Engineering and Technology
Issue 19(4), 2025 ISSN (Online) 2707-8434

UDC 504:57.04

IN SILICO MODELS FOR PREDICTING THE MUTAGENICITY OF MAIN STRUCTURAL CLASSES XENOBIOTICS BASED ON THE RANDOM FOREST ALGORITHM

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

Abstract *– With the rapid development of industry, there has been a significant annual increase in the number of chemical compounds that can enter the environment. A large number of xenobiotics, through direct or indirect effects on the human genetic apparatus, can induce hereditary and/or oncological diseases. The growing number of chemicals entering the environment and potentially affecting the genetic health of the human population is causing public concern. In this situation, it is critically important to effectively control and account for all xenobiotics that may exhibit potential genotoxic properties. Classic in vitro and in vivo methods for assessing the genotoxicity of environmental factors have significant drawbacks related to the complexity of conducting experiments, significant financial costs, and low reproducibility of results in different laboratories. In addition, when conducting experimental studies, it is necessary to take into account the "3R" concept accepted by the scientific community, which aims to reduce, refine, and replace animal models. Such limitations in the application of classical in vitro and in vivo methods of genetic assessment of environmental factors have become the foundation for the formation of the main vector of development of modern computer toxicology, based on which the assessment of genotoxic and toxic effects is achieved using predictive in silico models. In this context, QSAR models deserve attention, as they allow the mutagenic potential of xenobiotics to be assessed based on the quantitative relationship between their structure and activity. The paper presents a methodology for improving the accuracy of in silico Ames mutagenicity assessment models (Ames/QSAR) based on ensemble machine learning algorithms. It has been established that the accuracy of Ames/QSAR models can be improved by selecting the most influential features represented by molecular descriptors. In addition, the predictive power of Ames/QSAR models can be improved by forming homogeneous structural classes of xenobiotics represented by chemical compounds that have a common molecular framework structure. It has been shown that a limited list of molecular descriptors can be used to search for causal relationships between mutagenicity and the physicochemical, electronic, and spatial properties of xenobiotics, which are determined by different sets of molecular descriptors.*

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