

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ  
ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»  
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ  
ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»  
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова  
праця на правах рукопису

**КИСЛЯК СЕРГІЙ ВОЛОДИМИРОВИЧ**

УДК 004.94:577.21:615.9

ДИСЕРТАЦІЯ

***IN SILICO* МОДЕЛІ ПРОГНОЗУВАННЯ МУТАГЕННОСТІ ЕЙМСА  
ОСНОВНИХ СТРУКТУРНИХ КЛАСІВ КСЕНОБІОТИКІВ**

091 Біологія

09 Біологія

Подається на здобуття наукового ступеня доктора філософії. Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

\_\_\_\_\_ Сергій КИСЛЯК

Науковий керівник: Дуган Олексій Мартем'янович, д-р.біол. наук, проф.

Київ – 2026

## Анотація

Кисляк С.В. *In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків. – Кваліфікаційна наукова праця на правах рукопису. Дисертація на здобуття наукового ступеня доктора філософії з галузі знань 09 Біологія за спеціальністю 091 Біологія – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2026.

**Актуальність теми дослідження.** В умовах глобальної індустріалізації у навколишньому середовищі фіксується щорічне значне збільшення кількості хімічних речовин, біологічна дія яких або невідома, або результати досліджень щодо їх генотоксичного потенціалу є відсутніми. Однак, значна частина ксенобіотиків, що потрапляють у довкілля і з якими контактує людська популяція, володіють специфічною біологічною дією: канцерогенною, мутагенною, алергенною. Тому, з метою забезпечення генетичної та екологічної безпеки для людства, потребує негайного вирішення проблема, що пов'язана з оцінкою потенційної генетичної дії різноманітних хімічних сполук, здатних впливати на спадковий матеріал людини і індукувати спадкові і соматичні (онкологічні) захворювання.

Розроблені і широко використовувані у минулі десятиріччя класичні *in vitro* та *in vivo* методи виявлення й оцінки генетичних ефектів факторів навколишнього середовища в короткострокових методах на мікроорганізмах і класичних методах на ссавцях є складними з точки зору їх проведення, є дороговартісними, тривалі в часі, мають проблему відтворюваності результатів експерименту в різних лабораторіях, що не рідко призводять до отримання хибно-позитивних і хибно-негативних результатів. Крім того класичні підходи оцінки генотоксичності факторів навколишнього середовища можуть стикатися з етичними проблемами використання в експериментах теплокровних тварин. Такі обмеження стимулюють наукову спільноту до розробки та впровадження альтернативних сучасних *in silico* методів оцінки потенційної генетичної активності факторів навколишнього середовища, які були б менш

дороговартісними та ефективними, та для яких були б відсутні вищеперераховані недоліки. Стандартна парадигма токсикологічної науки щодо проведення тестування на генотоксичність з використанням прийнятою науковою спільнотою класичної батареї *in vitro* та *in vivo* тест-систем потребує оновлення та розширення переліку ефективних та більш продуктивних методів, особливо з урахуванням концепції «3R», що керується принципами, які направлені на зменшення, вдосконалення та заміну моделей тварин при проведенні тестування на генотоксичність. Проблеми сучасної токсикології можуть бути вирішені через інтеграцію наук, становлення та розвиток яких припадає на кінець 20 ст. В цьому контексті заслуговують на увагу досягнення в області хемоінформатики та комп'ютерних наук. Тому розробка та впровадження сучасних обчислювальних *in silico* прогностичних моделей оцінки генотоксичності факторів навколишнього середовища із застосуванням методів штучного інтелекту можуть розглядатись в якості основного вектору розвитку сучасної обчислювальної токсикології.

**Метою дослідження** є розробка, оптимізація та апробація орієнтованих на основні структурні класи ксенобіотиків *in silico* моделей прогнозування мутагенності Еймса.

#### **Матеріали та методи.**

При створенні *in silico* моделей прогнозування мутагенності Еймса використовувався датасет, що був отриманий шляхом об'єднання трьох загальнодоступних наборів даних: Kazius-Bursi, Hansen та EFSA. Крім того, об'єднаний набір даних було додатково розширено мікотоксинами, дублікати ксенобіотиків були видалені. Відповідно до набору даних, що містив 8454 ксенобіотиків, будь-яка хімічна сполука вважалася мутагенною, якщо при проведенні *in vitro* тесту Еймса на штамах *Salmonella typhimurium* TA97, TA98, TA100, TA102, TA1535, TA1537 і TA1538 був отриманий хоча б один позитивний результат. Для розроблених прогностичних Ames/QSAR моделей використовувався об'єднаний набір даних, розширений мікотоксинами з розрахованими молекулярними дескрипторами PaDel, RDkit та Mordred. На етапі створення ефективних Ames/QSAR моделей, з метою порівняння їх точності, було

запропоновано для кожної хімічної сполуки, що проявляють потенційні генотоксичні властивості, в якості предикторів також використовувати відбитки молекулярної структури (Molecular fingerprint). Відбір найкращого бінарного класифікатора базувався на клас-орієнтованому підході, в основі якого розроблені Ames/QSAR моделі були отримані відповідно до датасету, що був розподілений на п'ять груп ксенобіотиків, що мали спільні риси будови молекулярного каркасу. Для вирішення задачі бінарної класифікації були запропоновані моделі прогнозування мутагенності Еймса на основі наступних методів: логістична регресія (LR-Scikit), логістична регресія на основі стохастичного градієнтного спуску (LR-SGD), метод градієнтного бустінга (Gradient boosting), метод випадкового лісу (Random Forest) та глибинна нейронна мережа.

Заслужують на увагу результати досліджень, в яких науковці намагаються пов'язати прояви мутагенності з наявними функціональними групами або структурами. При проведенні дисертаційного дослідження нами була приділена увага також на іншому підході оцінки генетичної небезпеки ксенобіотиків, що базується на структурних маркерах мутагенності. Згідно з таким підходом було запропоновано отримати візуалізацію багатовимірних даних, що були розподілені на п'ять структурних класів у двовимірному просторі відбитків структури (Molecular fingerprint), відповідно до розрахованих індексів подібності (Танімото та Хемінга). Для вирішення цієї задачі був використаний алгоритм t-розподіленого вкладення стохастичної близькості (t-SNE), що був застосований для візуалізації багатовимірних даних у двовимірному просторі.

**Наукова новизна** отриманих результатів дослідження полягає у наступному:

*вперше:*

- розроблені моделі оцінки мутагенності Еймса на основі різних типів молекулярних дескрипторів (PaDel, RDkit та Mordred), що орієнтовані на основні структурні класи хімічних сполук – потенційних мутагенів. Показано, що *in silico* Ames/QSAR моделі, які побудовані на основі різних наборів релевантних дескрипторів, з урахуванням поділу ксенобіотиків на структурні класи,

дозволяють зменшити кількість хибнонегативних та хибнопозитивних результатів досліджень. Моделі, що були отримані відповідно до набору даних, що відносяться до основних структурних класів ксенобіотиків, дозволяють отримати оцінку мутагенності з високими показниками точності (від 87% до 93%) відповідно до метрики *accuracy*, що перевищує значення метрики загальної точності для *in vitro* тесту Еймса, яка коливається у межах 80-85%.

- розроблені Ames/QSAR моделі прогнозування мутагенності на основі трьох різних типів молекулярних відбитків структури (MACCS, RDkit та FCFP), що орієнтовані на основні структурні класи ксенобіотиків. Показана ефективність використання в якості предикторів відбитків молекулярної структури ксенобіотиків. Точність таких моделей відповідає середньому значенню загальної точності *in vitro* тесту Еймса, яка коливається в межах 80-85%. При цьому основною перевагою даного підходу є спрощена процедура проведення підготовки вхідних даних.
- отримані переліки релевантних молекулярних дескрипторів, використання яких в якості предикторів дає змогу підвищити точність розроблених QSAR моделей, відповідно до метрики загальної точності від 0,1 до 2%.
- розроблений підхід оцінки генетичної активності хімічних сполук дає можливість використовувати алгоритм t-розподіленого вкладення стохастичної близькості (t-SNE) для ефективного пошуку структурних маркерів мутагенності з урахуванням розподілу потенційних генотоксичних сполук на п'ять структурних класів. Такий підхід є ефективним та дозволяє у межах структурного класу здійснювати процедуру відбору схожих за структурою ксенобіотиків. Порівняння структурних формул хімічних сполук, для яких значення метрик відстані (Танімото, Хемінга) будуть мінімальними дозволяє ідентифікувати ті функціональні групи або підструктури, що можуть лежати в основі прояву мутагенності ксенобіотиків.

**Практичне значення результатів дисертаційного дослідження:**

- Сформульовані методологічні основи та принципи, що можуть лежати в основі побудови ефективних, орієнтованих на основні структурні класи, *in silico* моделей прогнозування генетичної активності хімічних сполук.
- Отримані переліки релевантних молекулярних дескрипторів, використання яких в якості предикторів дозволяє підвищити точність (відповідно до метрики загальної точності) від 0,1% до 2% орієнтованих на основні структурні класи Ames/QSAR моделей.
- Розроблено веб-сервіс, який відповідно до запропонованої у межах роботи методики, на основі 1D та 2D молекулярних дескрипторів, дозволяє з мінімальними витратами часу та достатньо високими показниками точності оцінити генетичну активність хімічних сполук.
- Розроблено програмне забезпечення, яке дозволяє, через порівняння схожих за структурою ксенобіотиків, здійснювати пошук структурних маркерів відповідальних за генетичну активність сполуки.
- Результати роботи впроваджено в навчальний процес підготовки фахівців освітньої програми «Біотехнології» магістерського рівня навчання зі спеціальності 162 «Біотехнології та біоінженерія» при вивченні дисциплін «Моделювання молекулярної взаємодії» та «Пакети прикладних програм для задач молекулярної біології» (Акт впровадження від 19.11.2025р.).

**В першому розділі** представлено інформацію щодо тенденцій розвитку сучасної токсикології, акцентовано увагу на проблемах, що потребують вирішення. Проведено огляд літературних даних щодо біологічної складової індукованого впливу ксенобіотиків на генетичний апарат людини, що в першу чергу пов'язано з виникненням генних мутацій, хромосомних аберацій та анеуплоїдії. Приділено увагу особливостям розповсюдження генетично активних хімічних сполук у об'єктах навколишнього середовища. Показана необхідність обліку та отримання генотоксичної оцінки впливу всіх хімічних сполук, що потрапляють у довкілля. Проаналізовано *in vitro* та *in vivo* експериментальні методи оцінки генотоксичності, що формують класичну батарею тест-систем, показані їх переваги та недоліки. Показана необхідність оновлення підходів щодо

отримання оцінки мутагенних ефектів факторів навколишнього середовища. **В другому розділі** детально викладені методи та підходи, що лежать в основі розроблених *in silico* Ames/QSAR прогностичних моделей. Приділена увага питанням класифікації, особливостям розрахунків та застосуванням різних типів 1D та 2D молекулярних дескрипторів, які використовуються в якості предикторів для Ames/QSAR моделей. Висвітлені питання відносно розподілу хімічних сполук на п'ять структурних класів. Представлена інформація щодо особливостей розрахунків основних і метрик оцінки якості *in silico* моделей прогнозування мутагенності Еймса. **В третьому розділі** представлена класична методика побудови Ames/QSAR моделей. Показано, що зменшення розмірності вхідних даних може лежати в основі підвищення ефективності бінарних класифікаторів. Наведено результати моделювання та представлена методологія побудови орієнтованих на основні структурні класи ксенобіотиків Ames/QSAR моделей, що використовують в якості предикторів різні набори 1D та 2D молекулярних дескрипторів. Висвітлена методика покращення точності бінарних класифікаторів з урахуванням розподілу ксенобіотиків за структурними класами та відбором тих молекулярних дескрипторів, що мають вагомий вплив на прогнозовану змінну. Наведено результати моделювання та описані Ames/QSAR моделі оцінки мутагенності, що використовують в якості предикторів відбитки молекулярної структури ксенобіотиків. Встановлено, що якісний склад молекулярних дескрипторів може суттєво впливати на точність моделей прогнозування мутагенності. Проведено аналіз причинно-наслідкових зв'язків між мутагенністю та релевантними дескрипторами основних структурних класів ксенобіотиків

**Особистий внесок здобувача.** Всі основні результати, відображені в дисертаційній роботі, отримано здобувачем особисто. Дисертантом здійснено ґрунтовний аналіз літературних джерел та визначено основний напрям дослідження, що має практичну та наукову цінність. Сформульовано гіпотезу щодо перспектив покращення точності *in silico* Ames/QSAR моделей, з урахуванням розподілу ксенобіотиків на структурні класи та використанням в якості предикторів різних наборів молекулярних дескрипторів. Запропоновано

методологію створення ефективних *in silico* моделей оцінки мутагенності Еймса, покращення прогностичної здатності яких може бути досягнуто через застосування обмеженого набору вхідних даних, що були отримані для окремих структурних класів ксенобіотиків. Показана можливість використання відбитків структури ксенобіотиків для вирішення задачі ідентифікації маркерів мутагенності. Дисертаційна робота виконана на кафедрі промислової біотехнології та біофармації Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», під керівництвом д.б.н., проф. Дугана О.М. Робота є результатом проведення самостійних досліджень Кисляка С. В.

За темою дисертаційного дослідження опубліковано 9 наукових праць, з яких 5 наукових статей, з яких 2 – опубліковані в журналах, що індексуються Scopus, 3 статті опубліковані у фахових періодичних виданнях. Результати досліджень були апробовані на чотирьох міжнародних конференціях.

**Ключові слова:** ДНК, гени, мутації, генотоксичність, канцерогенез, цитотоксичність, тест Еймса, молекулярний дескриптор, класифікатор, машинне навчання, моделювання, нейронна мережа, QSAR модель.



## ABSTRACT

Kislyak S.V. In silico models for predicting the mutagenicity of Ames of the main structural classes of xenobiotics. – Qualification scientific work on the rights of the manuscript. Thesis for the degree of Doctor of Philosophy in the field of knowledge 09 Biology, specialty 091 Biology. – National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, 2026.

Relevance of the research topic. In the context of global industrialization, there has been a significant annual increase in the number of chemicals in the environment whose biological effects are either unknown or for which there are no research results on their genotoxic potential. However, a significant portion of xenobiotics that enter the environment and come into contact with the human population have specific biological effects: carcinogenic, mutagenic, allergenic. Therefore, in order to ensure genetic and environmental safety for humanity, the problem associated with assessing the potential genetic effects of various chemical compounds capable of affecting human genetic material and inducing hereditary and somatic (oncological) diseases needs to be addressed immediately.

The classic in vitro and in vivo methods developed and widely used in recent decades to detect and evaluate the genetic effects of environmental factors in short-term methods on microorganisms and classic methods on mammals are complex in terms of their implementation, expensive, time-consuming, and have problems with the reproducibility of experimental results in different laboratories, which often lead to false positive and false negative results. In addition, classical approaches to assessing the genotoxicity of environmental factors may face ethical problems associated with the use of warm-blooded animals in experiments. Such limitations encourage the scientific community to develop and implement alternative modern in silico methods for assessing the potential genetic activity of environmental factors, which would be less expensive and more effective, and would not have the above-mentioned disadvantages. The standard paradigm of toxicological science regarding genotoxicity testing using the classical battery of in vitro and in vivo test systems accepted by the scientific community needs to be updated and expanded with a list of effective and more

productive methods, especially taking into account the “3R” concept, which is guided by principles aimed at reducing, improving and replacing animal models in genotoxicity testing. The problems of modern toxicology can be solved through the integration of sciences that emerged and developed at the end of the 20th century. In this context, achievements in the field of chemoinformatics and computer science deserve attention. Therefore, the development and implementation of modern computational in silico predictive models for assessing the genotoxicity of environmental factors using artificial intelligence methods can be considered as the main vector for the development of modern computational toxicology.

The aim of the study is to develop, optimize, and test in silico models for predicting Ames mutagenicity based on the main structural classes of xenobiotics.

**Materials and methods.** When creating in silico models for predicting Ames mutagenicity, a dataset was used that was obtained by combining three publicly available datasets: Kazius-Bursi, Hansen, and EFSA. In addition, the combined dataset was further expanded with mycotoxins, and duplicate xenobiotics were removed. According to the dataset containing 8454 xenobiotics, any chemical compound was considered mutagenic if at least one positive result was obtained in the in vitro Ames test on *Salmonella typhimurium* strains TA97, TA98, TA100, TA102, TA1535, TA1537, and TA1538 strains. The combined dataset, expanded with mycotoxins and calculated molecular descriptors PaDel, RDkit, and Mordred, was used for the developed Ames/QSAR predictive models. At the stage of creating effective Ames/QSAR models, in order to compare their accuracy, it was proposed to use molecular fingerprints as predictors for each chemical compound exhibiting potential genotoxic properties. The selection of the best binary classifier was based on a class-oriented approach, in which the developed Ames/QSAR models were obtained according to a dataset divided into five groups of xenobiotics with common features of the molecular framework. To solve the binary classification problem, Ames mutagenicity prediction models were proposed based on the following methods: logistic regression (LR-Scikit), logistic regression based on stochastic gradient descent (LR-SGD), gradient boosting, random forest, and deep neural network.

The results of studies in which scientists attempt to link manifestations of mutagenicity with existing functional groups or structures deserve attention. In conducting our dissertation research, we also focused on another approach to assessing the genetic hazard of xenobiotics, based on structural markers of mutagenicity. According to this approach, it was proposed to visualize multidimensional data, which were divided into five structural classes in a two-dimensional space of molecular fingerprints according to the calculated similarity indices (Tanimoto and Hamming). To solve this problem, the t-distributed stochastic neighbor embedding (t-SNE) algorithm was used, which is applied to visualize multidimensional data in two-dimensional space.

The scientific novelty of the research results lies in the following:

For the first time:

- Models for assessing Ames mutagenicity have been developed based on various types of molecular descriptors (PaDel, RDkit, and Mordred) that focus on the main structural classes of chemical compounds that are potential mutagens. It has been shown that *in silico* Ames/QSAR models, which are built on the basis of various sets of relevant descriptors, taking into account the division of xenobiotics into structural classes, allow reducing the number of false negative and false positive research results. The models obtained in accordance with the data set relating to the main structural classes of xenobiotics allow for the assessment of mutagenicity with high accuracy (from 87% to 93%) according to the accuracy metric, which exceeds the overall accuracy metric for the *in vitro* Ames test, which ranges from 80 to 85%.

- Ames/QSAR models for predicting mutagenicity have been developed based on three different types of molecular structure fingerprints (MACCS, RDkit, and FCFP) that target the main structural classes of xenobiotics. The effectiveness of using molecular structure fingerprints of xenobiotics as predictors has been demonstrated. The accuracy of such models corresponds to the average overall accuracy of the *in vitro* Ames test, which ranges from 80 to 85%. The main advantage of this approach is the simplified procedure for preparing input data.

- Lists of relevant molecular descriptors have been obtained, the use of which as predictors makes it possible to increase the accuracy of the developed QSAR models, according to the overall accuracy metric from 0.1 to 2%.

- The developed approach to assessing the genetic activity of chemical compounds makes it possible to use the t-distributed stochastic neighbor embedding (t-SNE) algorithm to effectively search for structural markers of mutagenicity, taking into account the distribution of potential genotoxic compounds into five structural classes. This approach is effective and allows for the selection of structurally similar xenobiotics within a structural class. Comparing the structural formulas of chemical compounds for which the distance metrics (Tanimoto, Hamming) are minimal allows us to identify those functional groups or substructures that may underlie the mutagenicity of xenobiotics.

Practical significance of the dissertation research results:

- Formulated methodological foundations and principles that can underlie the construction of effective, structure-based, in silico models for predicting the genetic activity of chemical compounds.

- Lists of relevant molecular descriptors have been obtained, the use of which as predictors allows to increase the accuracy (according to the overall accuracy metric) from 0.1% to 2% of Ames/QSAR models focused on basic structural classes.

- A web service has been developed which, in accordance with the methodology proposed in the work, based on 1D and 2D molecular descriptors, allows the genetic activity of chemical compounds to be assessed with minimal time expenditure and sufficiently high accuracy.

- Software has been developed that allows, through comparison of structurally similar xenobiotics, to search for structural markers responsible for the genetic activity of a compound.

- The results of the work have been implemented in the educational process of training specialists in the master's level educational program "Biotechnology" in the specialty 162 "Biotechnology and Bioengineering" when studying the discipline "Modeling of Molecular Interaction" (Act of implementation dated 19.11.2025).

The first chapter presents information on trends in modern toxicology, focusing on problems that need to be addressed. A review of the literature on the biological component of the induced effect of xenobiotics on the human genetic apparatus is presented, which is primarily associated with the occurrence of gene mutations, chromosomal aberrations, and aneuploidy. Attention is paid to the peculiarities of the spread of genetically active chemical compounds in environmental objects. The need to account for and obtain a genotoxic assessment of the impact of all chemical compounds that enter the environment is demonstrated. In vitro and in vivo experimental methods for assessing genotoxicity, which form a classic battery of test systems, are analyzed, and their advantages and disadvantages are shown. The need to update approaches to assessing the mutagenic effects of environmental factors is demonstrated. The second chapter details the methods and approaches underlying the developed in silico Ames/QSAR predictive models. Attention is paid to classification issues, calculation specifics, and the application of various types of 1D and 2D molecular descriptors used as predictors for Ames/QSAR models. Issues related to the division of chemical compounds into five structural classes are highlighted. Information is presented on the features of calculations of the main metrics and quality assessment metrics of in silico Ames mutagenicity prediction models. The third chapter presents the classical method of constructing Ames/QSAR models. It is shown that reducing the dimensionality of input data can be the basis for improving the efficiency of binary classifiers. The results of modeling are presented, and a methodology is presented for constructing Ames/QSAR models oriented toward the main structural classes of xenobiotics, using various sets of 1D and 2D molecular descriptors as predictors. A method for improving the accuracy of binary classifiers is described, taking into account the distribution of xenobiotics by structural classes and the selection of molecular descriptors that have a significant impact on the predicted variable. The results of modeling are presented and Ames/QSAR models for assessing mutagenicity are described, which use molecular structure fingerprints of xenobiotics as predictors. It has been established that the qualitative composition of molecular descriptors can significantly affect the accuracy of mutagenicity prediction models. An analysis of the causal relationships between

mutagenicity and relevant descriptors of the main structural classes of xenobiotics was performed.

Personal contribution of the applicant. All the main results reflected in the dissertation were obtained by the applicant personally. The applicant conducted a thorough analysis of literary sources and determined the main direction of research, which has practical and scientific value. A hypothesis has been formulated regarding the prospects for improving the accuracy of *in silico* Ames/QSAR models, taking into account the distribution of xenobiotics into structural classes and the use of different sets of molecular descriptors as predictors. A methodology for creating effective *in silico* models for assessing Ames mutagenicity has been proposed, the predictive power of which can be improved through the use of a limited set of input data obtained for individual structural classes of xenobiotics. The possibility of using xenobiotic structure fingerprints to solve the problem of identifying mutagenicity markers is demonstrated. The dissertation was completed at the Department of Industrial Biotechnology and Biopharmacy of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” under the supervision of Prof. Dugan O.M., Doctor of Biological Sciences. The work is the result of independent research by Kislyak S.V.

Nine scientific papers have been published on the topic of the dissertation research, including five scientific articles, two of which were published in Scopus-indexed journals, and three articles published in professional periodicals. The research results were tested at four international conferences.

Keywords: DNA, genes, mutations, genotoxicity, carcinogenesis, cytotoxicity, Ames test, molecular descriptor, classifier, machine learning, modeling, neural network, QSAR model.

## СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

### У яких опубліковані основні результати дисертації:

1. 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>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються експериментальні *in vitro* та *in vivo* методи, що використовуються для генетичної оцінки впливу факторів навколишнього середовища, написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Яловенко О.І. – критичний аналіз.

2. Kislyak S., Dugan O., Yesypenko R., Starosyla D., Yalovenko O. In silico the Ames Mutagenicity Predictive Model of Environment. // Innovative Biosystems and Bioengineering. – 2025. –Vol. 9, no. 2. – P. 42–52. URL: <https://doi.org/10.20535/ibb.2025.9.2.316239>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються сучасні *in silico* методи, що використовуються для генетичної оцінки впливу факторів навколишнього середовища. Розробка методології проведення *in silico* моделювання, з урахуванням якісного складу молекулярних дескрипторів. Формування бази даних хімічних сполук. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Єсипенко Р.В. – реалізація моделей; Яловенко О.І. – критичний аналіз.

3. Кисляк С. В., Дуган О. М., Мороз М. О., Яловенко О. І. Система ідентифікації структурних маркерів мутагенності Еймса на основі подібності відбитків структури ксенобіотиків. // Вісник Харківського національного університету ім. В. Н. Каразіна. Серія: Біологія. – 2025. – № 44, вип. 1. – С. 6-14. URL: <https://doi.org/10.26565/2075-5457-2025-44-1>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються *in silico* методи генетичної оцінки впливу факторів навколишнього

середовища, прогностична здатність яких базується на ідентифікації функціональних груп або/і підструктур, що є визначальними з точки зору проявів їх мутагенності. Формування бази даних хімічних сполук. Розподіл хімічних сполук за структурними класами. Розрахунок 2D молекулярних дескрипторів (MACCS, RDkit та FCFP). Розроблений підхід щодо оцінки мутагенного потенціалу, що ґрунтується на структурній подібності між досліджуваними потенційними генотоксичними сполуками. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Мороз М. О. – реалізація моделей; Яловенко О.І. – критичний аналіз статті.

4. Кисляк С. В., Дуган О. М., Єсипенко Р.В., Яловенко О.І. In silico моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу. // Біомедична інженерія і технологія. – 2025. – № 19(4). – С. 48-62 URL: <https://doi.org/10.20535/.2025.19.340327>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються сучасні Ames/QSAR моделі, що використовуються для генетичної оцінки впливу факторів навколишнього середовища. Формування бази даних хімічних сполук. Розрахунок молекулярних дескрипторів та розподіл хімічних сполук за структурними класами. Розробка методології покращення точності *in silico* моделей прогнозування мутагенності Еймса на основі методу випадкового лісу. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Єсипенко Р.В. – реалізація моделей; Яловенко О.І. – критичний аналіз.

5. Кисляк С. В., Дуган О. М., Романюк Д.І, Яловенко О.І. In silico моделі прогнозування мутагенності Еймса на основі відбитків молекулярної структури ксенобіотиків.// Біомедична інженерія і технологія. – 2025. – № 20(5). – С. 1-14. URL: <https://doi.org/10.20535/.2025.20.340837>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються сучасні Ames/QSAR моделі, що використовуються для генетичної оцінки впливу факторів навколишнього середовища. Формування бази даних



хімічних сполук. Розрахунок молекулярних дескрипторів та розподіл хімічних сполук за структурними класами. Розробка методології покращення точності *in silico* моделей прогнозування мутагенності Еймса на основі відбітків молекулярної структури з застосуванням ансамблевих алгоритмів машинного навчання та нейронно-мережевого підходу. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Романюк Д.І. – реалізація моделей; Яловенко О.І. – критичний аналіз

### **Апробація матеріалів дисертації**

6. Кисляк С.В., Есипенко Р.В. *In silico* моделі оцінки генотоксичності впливу факторів навколишнього середовища. //Current challenges of science and education. Proceedings of the 9th International scientific and practical conference. 2024 May 6-8. MDPC Publishing. Berlin, Germany. 2024. P. 50–53.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, підготовка тез до конференції; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Есипенко Р.В. – реалізація моделей.

7. Кисляк С.В., Дуган О.М., Яловенко О.І. *In silico* моделі генетичної оцінки впливу факторів навколишнього середовища.// Science and society: modern trends in a changing world. Proceedings of the 9th International scientific and practical conference. 2024 Aug. 5-7. MDPC Publishing. Vienna, Austria. 2024. P. 25-28.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, підготовка тез до конференції; Дуган О.М., Яловенко О.І. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення.

8. Кисляк С.В., Дуган О.М., Яловенко О.І. Методи оцінки генотоксичних ефектів факторів навколишнього середовища. // European congress of scientific achievements. Proceedings of the 8th International scientific and practical conference. 2024 Aug. 12-14. Barca Academy Publishing. Barcelona, Spain. 2024. P. 9-15.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, підготовка тез до конференції; Дуган О.М., Яловенко О.І. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення.

9. Кисляк С.В., Дуган О.М., Яловенко О.І. Оптимізація *in silico* моделей прогнозування мутагенності Еймса через зменшення розмірності вхідних даних // Current trends in scientific research development. Proceedings of the 11th International scientific and practical conference. 2025 June 5-7. BoScience Publisher. Boston, USA. 2025. P. 31-37.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, розробка методології *in silico* моделювання з урахуванням якісного складу молекулярних дескрипторів; формування бази даних хімічних сполук; аналіз результатів моделювання; підготовка тез до конференції; Дуган О.М., Яловенко О.І. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення.

## ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ ТА ПОЗАЧЕНЬ .....	21
ВСТУП.....	22
РОЗДІЛ 1 ГЕНЕТИЧНІ НАСЛІДКИ ВПЛИВУ ХІМІЧНИХ МУТАГЕНІВ ТА МЕТОДИ ЇХ ОЦІНКИ .....	29
1.1 Поширення мутагенно-активних хімічних сполук в об'єктах навколишнього середовища.....	31
1.2 Молекулярно-біологічні основи впливу мутагенів та їх різновиди.....	35
1.2.1 Методи оцінки генетичних ефектів на основі короткострокових тестів.....	42
1.2.2 Тест-система на бактеріальну зворотну мутацію Еймса (TG 471) .....	45
1.3 Сучасні методи оцінки генотоксичності факторів навколишнього середовища.....	48
Висновки до розділу 1 .....	52
РОЗДІЛ 2 МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕННЯ.....	54
2.1 База даних хімічних сполук – потенційних мутагенів.....	54
2.2 Молекулярні дескриптори, як базові предиктори для <i>in silico</i> моделей прогнозування мутагенності Еймса .....	57
2.2.1 1D та 2D молекулярні дескриптори та особливості їх розрахунків .....	62
2.2.2 Відбитки молекулярної структури та їх класифікація .....	66
2.3 Алгоритми машинного навчання для побудови AMES/QSAR моделей.....	71
2.3.1 Логістична регресія.....	71
2.3.2 Метод випадкового лісу .....	72
2.3.3 Метод екстремального градієнтного бустінгу .....	74
2.3.4 Глибинні нейронні мережі .....	74
2.4 Метрики оцінки ефективності <i>in silico</i> Ames/QSAR моделей.....	77
Висновки до розділу 2 .....	79

РОЗДІЛ 3 РОЗРОБКА ТА ОПТИМІЗАЦІЯ <i>IN SILICO</i> МОДЕЛЕЙ ПРОГНОЗУВАННЯ МУТАГЕННОСТІ НА ОСНОВІ РЕЗУЛЬТАТІВ ТЕСТУ ЕЙМСА.....	80
3.1 Класичні підходи до реалізації Ames/QSAR моделей та шляхи щодо їх оптимізації .....	80
3.1.1 Вплив зменшення набору вхідних даних на ефективність Ames/QSAR моделей .....	85
3.2 Ames/QSAR моделі, орієнтовані на структурні класи ксенобіотиків .....	90
3.2.1 Ames/QSAR моделі на основі методу випадкового лісу .....	95
3.2.2 Ames/QSAR моделі на основі екстремального грідієнтного бустінгу ...	104
3.2.3 Ames/QSAR моделі з використанням нейромережевого підходу .....	112
3.3 Ames/QSAR моделі на основі відбитків молекулярної структури ксенобіотиків .....	121
3.4 Ames/QSAR моделі на основі структурних маркерів мутагенності .....	131
3.5 Причинно-наслідкові зв'язки між мутагенністю та релевантними дескрипторами основних структурних класів ксенобіотиків.....	135
Висновки до розділу 3 .....	147
ЗАГАЛЬНІ ВИСНОВКИ .....	150
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ .....	152
ДОДАТОК А.....	175
ДОДАТОК Б .....	179
ДОДАТОК В .....	183

## ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ ТА ПОЗАЧЕНЬ

<i>Позначення</i>	<i>Зміст скорочення</i>
ACS	American Chemical Society
ADME	Absorption, Distribution, Metabolism, Excretion-Toxicity
CAS	Chemical Abstracts Service
DNN	Deep Neural Network
ECFP	Extended-Connectivity Fingerprints
ECFP	Feature-Class Fingerprint
ECHA	European Chemicals Agency
ecNGS	Error-corrected Next-Generation Sequencing
EFSA	European Food Safety Authority
FN	False Negatives
FP	False Positives
LR	Logistic regression
LR-SGD	Logistic regression using stochastic gradient descent
MACCS	Molecular ACCess System
MDI	Mean decrease impurity
NGS	Next generation sequencing
OECD	Organisation for Economic Co-operation and Development
PFI	Permutation feature importance
QSAR	Quantitative structure–activity relationship
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SMILES	Simplified Molecular Input Line Entry System
TN	True Negatives
TP	True Positives
t-SNE	t-distributed Stochastic Neighbor Embedding
UKEMS	United Kingdom Environmental Mutagen Society
XGBoost	Метод екстремального грідієнтного бустінга
ВООЗ	Всесвітня організація охорони здоров'я
ДНК	Дезоксирибонуклеїнова кислота
ПАВ	Поверхнево-активні речовини

## ВСТУП

**Актуальність теми.** Європейське та Американське хімічні товариства оприлюднили інформацію про більш ніж 800 тисяч хімічних речовин, для яких наслідки взаємодії з генетичним апаратом людини не відомі. На початок 2020 року була доступна інформація про більш ніж 100 000 хімічних речовин, що виробляються промисловістю, які здатні негативно впливати на навколишнє середовище та здоров'я людини. В умовах стрімкого збільшення кількості нових хімічних сполук, що генерує людство в різних сферах виробництва, особливо гострою постає проблема їх ефективного виявлення та обліку. Виходячи з того, що значна частина ксенобіотиків, що потрапляють у навколишнє середовище, так чи інакше контактує з людською популяцією і володіють специфічною біологічною активністю: канцерогенною, мутагенною, алергенною, виникає негайна потреба у забезпеченні генетичної безпеки людства. Це, в деякій мірі, може бути забезпечено наявністю ефективної генетичної оцінки небезпеки різноманітних хімічних сполук шляхом тестування їх активності в методах *in vivo* і *in vitro*. Цей тезис підтверджується тим, що система всебічної токсикологічної оцінки факторів довкілля ґрунтується, в першу чергу, на оцінці їх потенційної генетичної активності. Питання ефективного контролю та виявлення хімічних сполук з потенційними генотоксичними властивостями є надзвичайно чутливими тому, що вплив факторів різної природи на людську популяцію може індукувати появу геномної нестабільності на рівні статевих і соматичних клітин. З іншої сторони, вплив ксенобіотиків довкілля на спадковий апарат людини може стати основою для формування аномального епігенетичного профілю, який буде пов'язаний з розвитком нейродегенеративних та онкологічних захворювань.

Розроблені і широко використовувані у минулі десятиріччя класичні *in vitro* та *in vivo* методи оцінки генетичних ефектів факторів навколишнього середовища є складними з точки зору їх проведення, є дороговартісними, тривалі в часі, мають проблему відтворюваності результатів експерименту в різних лабораторіях

та можуть стикатися з етичними проблемами використання в експериментах теплокровних тварин. Крім того, основний недолік експериментальних методів оцінки генотоксичності пов'язаний з достатньо великою кількістю хибнопозитивних та хибнонегативних результатів прогнозів. Такі обмеження стимулюють наукову спільноту до розробки та впровадження альтернативних сучасних *in silico* методів оцінки потенційної генетичної активності факторів довкілля, які були б менш дорогі та ефективні з точки зору часових витрат. Стандартна парадигма токсикології щодо проведення тестування на генотоксичність з використанням прийнятою науковою спільнотою класичної батареї *in vitro* та *in vivo* тест-систем потребує оновлення та розширення переліку ефективних та більш продуктивних методів, особливо з урахуванням концепції «3R», що керується принципами, які направлені на зменшення, вдосконалення та заміну моделей теплокровних тварин при проведенні досліджень на генотоксичність. Проблеми сучасної токсикології можуть бути вирішені через інтеграцію наук, становлення та розвиток яких припадає на початок 21ст. В цьому контексті заслуговують на увагу досягнення в області хемоінформатики та комп'ютерних наук. Тому розробка та впровадження сучасних обчислювальних *in silico* моделей оцінки генотоксичності факторів навколишнього середовища із застосуванням методів штучного інтелекту можуть розглядатись в якості основного вектору розвитку сучасної токсикології.

За більше ніж 45 років викорисання тесту Еймса в якості базового методу для генетичної оцінки впливу факторів довкілля на спадковий апарат людини, було накопичено великий об'єм експериментальних даних. Така інформація, зазвичай, представлена у вільному доступі та її досить часто використовують науковці для створення *in silico* моделей прогнозування мутагенності. Доступність результатів оцінки мутагенності великої кількості ксенобіотиків, що отримані за допомогою найбільш поширеного *in vitro* метода – теста Еймса, стало основним критерієм для формування основного напрямку дослідження, що пов'язаний з розробкою ефективних *in silico* моделей прогнозування мутагенності Еймса.

**Метою дослідження** є розробка, оптимізація та апробація орієнтованих на основні структурні класи ксенобіотиків *in silico* моделей прогнозування мутагенності Еймса. Для досягнення мети роботи були сформульовані наступні завдання:

- проаналізувати сучасні методи та підходи до оцінки мутагенного потенціалу впливу факторів навколишнього середовища, визначити їх переваги та недоліки;
- проаналізувати існуючі загальнодоступні набори даних, для яких, відповідно до тесту Еймса, була дана оцінка мутагенному потенціалу. На основі відкритих для загального користування наборів сформувати об'єднаний датасет та розширити його мутагенними сполуками-мікотоксинами;
- розподілити хімічні сполуки на п'ять основних структурних класів та розрахувати молекулярні дескриптори для кожної групи ксенобіотиків;
- провести тестування різних алгоритмів машинного навчання та розробити ефективні прогностичні Ames/QSAR моделі для основних структурних класів ксенобіотиків;
- в межах кожного класу ксенобіотиків визначити перелік релевантних молекулярних дескрипторів, які мають вагомий вплив на прогнозовану змінну;
- провести тестування розроблених *in silico* Ames/QSAR моделей з використанням екзанаційної вибірки та оцінити їх прогностичну здатність;
- провести аналіз причинно-наслідкових зв'язків між мутагенністю та релевантними дескрипторами основних структурних класів ксенобіотиків.

**Об'єкт дослідження** – процес реалізації ефективних *in silico* моделей прогнозування мутагенності Еймса факторів навколишнього середовища.

**Предметом дослідження** є розробка ефективних *in silico* Ames/QSAR моделей, покращення точності яких досягається через: формування структурних класів ксенобіотиків; застосування різних типів 1-D та 2-D молекулярних дескрипторів; зменшення розмірності вхідних даних.

**Методи дослідження.** В основі розроблених *in silico* Ames/QSAR моделей були наступні методи: логістична регресія (LR), логістична регресія на основі



стохастичного градієнтного спуску (LR-SGD, метод випадкового лісу, метод градієнтного бустінга (XGBoost) та нейронна мережа. Оптимізація *in silico* Ames/QSAR моделей була реалізована через формування ранжованого переліку молекулярних дескрипторів, який був отриманий відповідно до коефіцієнтів двох моделей регресії (LR-SGD і LR-Scikit) та застосуванням двох підходів оцінки важливості ознак на основі випадкових дерев: критерія середнього зменшення помилок класифікації (MDI) та оцінки важливості ознак шляхом перестановки (PFI). Формування переліків молекулярних дескрипторів, з урахуванням приналежності ксенобіотиків до відповідних структурних класів, була здійснена за допомогою рекурсивного видалення ознак RFECV (Recursive Feature Elimination with Cross-Validation), що поєднаний з крос-валідацією (RFE). Для ідентифікації структурних маркерів мутагенності з урахуванням розподілу потенційних генотоксичних сполук на п'ять класів був використаний алгоритм t-розподіленого вкладення стохастичної близькості (t-SNE).

**Наукова новизна** отриманих результатів дослідження полягає у наступному:

*вперше:*

- розроблені моделі оцінки мутагенності Еймса на основі різних типів молекулярних дескрипторів (PaDel, RDkit та Mordred), що орієнтовані на основні структурні класи хімічних сполук – потенційних мутагенів. Показано, що *in silico* Ames/QSAR моделі, які побудовані на основі різних наборів релевантних дескрипторів, з урахуванням поділу ксенобіотиків на структурні класи, дозволяють зменшити кількість хибнонегативних та хибнопозитивних результатів досліджень. Моделі, що були отримані відповідно до набору даних, що відносяться до основних структурних класів ксенобіотиків, дозволяють отримати оцінку мутагенності з високими показниками точності (від 87% до 93%) відповідно до метрики ассигасу, що перевищує значення метрики загальної точності для *in vitro* тесту Еймса, яка коливається у межах 80-85%.
- розроблені Ames/QSAR моделі прогнозування мутагенності на основі трьох різних типів молекулярних відбитків структури (MACCS, RDkit та FCFP), що

орієнтовані на основні структурні класи ксенобіотиків. Показана ефективність використання в якості предикторів відбитків молекулярної структури ксенобіотиків. Точність таких моделей відповідає середньому значенню загальної точності *in vitro* тесту Еймса. При цьому основною перевагою даного підходу є спрощена процедура проведення підготовки вхідних даних.

- отримані переліки релевантних молекулярних дескрипторів, використання яких в якості предикторів дає змогу підвищити точність розроблених QSAR моделей, відповідно до метрики загальної точності від 0,1 до 2%.
- Розроблений підхід оцінки генетичної активності хімічних сполук дає можливість використовувати алгоритм t-розподіленого вкладення стохастичної близькості (t-SNE) для ефективного пошуку структурних маркерів мутагенності з урахуванням розподілу потенційних генотоксичних сполук на п'ять структурних класів. Такий підхід є ефективним та дозволяє у межах структурного класу здійснювати процедуру відбору схожих за структурою ксенобіотиків. Порівняння структурних формул хімічних сполук, для яких значення метрик відстані (Танімото, Хемінга) будуть мінімальними дозволяє ідентифікувати ті функціональні групи або підструктури, що можуть лежати в основі прояву мутагенності ксенобіотиків.

#### **Практичне значення результатів дисертаційного дослідження:**

- Сформульовані методологічні основи та принципи, що лежать в основі побудови ефективних, орієнтованих на основні структурні класи, *in silico* моделей прогнозування генетичної активності хімічних сполук.
- Отримані переліки релевантних молекулярних дескрипторів, використання яких в якості предикторів дозволяє підвищити точність (відповідно до метрики загальної точності) від 0,1% до 2% орієнтованих на основні структурні класи Ames/QSAR моделей.
- Розроблено веб-сервіс, який відповідно до запропонованої у межах роботи методики, на основі 1D та 2D молекулярних дескрипторів, дозволяє з мінімальними витратами часу та достатньо високими показниками точності оцінити генетичну активність хімічних сполук.

- Розроблено програмне забезпечення, яке дозволяє, через порівняння схожих за структурою ксенобіотиків, здійснювати пошук структурних маркерів відповідальних за генетичну активність сполуки.
- Результати роботи впроваджено в навчальний процес підготовки фахівців освітньої програми «Біотехнології» магістерського рівня навчання зі спеціальності 162 «Біотехнології та біоінженерія» при вивченні дисциплін «Моделювання молекулярної взаємодії» та «Пакети прикладних програм для задач молекулярної біології» (Акт впровадження від 19.11.2025р.).

**Особистий внесок здобувача.** Основні результати проведеного дисертаційного дослідження були отримані здобувачем особисто. Проведено детальний аналіз літературних джерел відповідно до теми та визначено основний напрямок дослідження. Проаналізовані набори даних хімічних сполук, для яких визначений мутагенний потенціал за допомогою тесту Еймса. Об'єднаний набір даних було розширено мікотоксинами. Для кожної хімічної сполуки розширеної бази даних був визначений структурний клас відповідно до особливостей будови молекулярного каркасу ксенобіотиків. На етапі формування розширеного датасету, дисертантом були проведені розрахунки одновимірних (1D) та двовимірних (2D) дескрипторів, а також відбитків молекулярної структури ксенобіотиків. Розроблена методологія підвищення точності існуючих прогностичних Ames/QSAR моделей. З метою створення ефективних бінарних класифікаторів, дисертантом було запропоновано використовувати однорідні набори даних, що відповідають потенційним мутагенам, які розподілені на п'ять структурних класів. Був сформульований і теоретично обґрунтований підхід щодо генетичної оцінки факторів навколишнього середовища з урахуванням різних типів молекулярних дескрипторів. Запропоновано використовувати обмежений набір молекулярних дескрипторів для прогнозування мутагенності хімічних сполук, що відносяться до певного структурного класу ксенобіотиків. Сформульовано підхід до оцінки мутагенної дії факторів навколишнього середовища, що ґрунтується на використанні двовимірних (2D) дескрипторів. Запропонована методика дає змогу, на основі розрахованих індексів подібності

між потенційними мутагенами/не мутагенами, ідентифікувати характерні підструктури та/або функціональні групи, що можуть бути пов'язані з проявами мутагенності досліджуваних ксенобіотиків.

Дисертаційна робота виконана на кафедрі промислової біотехнології та біофармації Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», під керівництвом д.б.н., проф. Дугана О.М. Підготовка публікацій, відповідно до отриманих результатів дослідження, виконана у співавторстві. При цьому вагомий внесок належить дисертанту при відсутності конфлікту інтересів. Усі розділи дисертації написані дисертантом особисто.

**Апробація результатів дисертації.** Апробація отриманих результатів дисертаційного дослідження була здійснена на міжнародних науково-практичних конференціях:

- X Міжнародна науково-практична конференція «Current challenges of science and education», (3-5.06.2024, Берлін, Німеччина);
- VII Міжнародна науково-практична конференція «Science and society: modern trends in a changing world», ( 10-12.06.2024, Відень, Австрія);
- VIII Міжнародна науково-практична конференція «European congress of scientific achievements», (12-14.08.2024, Барселона, Іспанія);
- XI Міжнародна науково-практична конференція «Current trends in scientific research development», (5-7.06.2025, Бостон, США).

### **Структура та обсяг дисертації**

Структура дисертаційної роботи: представлений вступ, три основні розділи та висновки після кожного з них, загальні висновки, список використаних джерел, та два додатки. Загальний обсяг дисертації становить 183 сторінки. Робота містить 31 таблицю та 12 рисунків. Список використаних джерел містить 213 найменувань.

## РОЗДІЛ 1 ГЕНЕТИЧНІ НАСЛІДКИ ВПЛИВУ ХІМІЧНИХ МУТАГЕНІВ ТА МЕТОДИ ЇХ ОЦІНКИ

На початку 20 сторіччя спостерігається стрімкий розвиток природничих наук, що відбувається відповідно до шостого технологічного укладу, який спирається на досягненнях в області сучасних інформаційних технологій, біотехнологій, молекулярної біології, генної інженерії, нанотехнологій та систем штучного інтелекту. За таких умов значну увагу наукова спільнота зосереджує на дослідженнях у галузі наук про життя на молекулярному рівні, розвиток яких став можливим завдяки впровадженню сучасних методів секвенування біологічних послідовностей, що вперше дозволило визначити нуклеотидну послідовність геному людини [1]. Становлення та розвиток обчислювальної молекулярної біології стало надійним фундаментом для формування головного вектору розвитку сучасних медичних та біологічних наук, що спирається на персоналізований індивідуальний підхід, що, у першу чергу, враховує якісний та кількісний склад нуклеотидів геному будь-якого біологічного об'єкта. З точки зору становлення та розвитку генетичної токсикології, персоніфікованої медицини, обчислювальної молекулярної біології досить важливими стали результати досліджень пілотної фази проєкту 1000 геномів, що дозволили дослідити природу генетичних мутацій в різних популяціях еукаріотичних організмів, включно з людиною [2].

У повсякденному житті спадковий апарат людини піддається впливу великої кількості зовнішніх агентів різної природи, що пошкоджують ДНК [3-5]. Серйозною проблемою є те, що новостворені хімічні та біологічні агенти, а також чинники фізичної природи, що володіють потенційними генотоксичними властивостями, здатні індукувати зміни на рівні ДНК, що, в свою чергу, може призвести до появи захворювань спадкової і соматичної природи [6,7]. В цьому контексті цілком логічною є тенденція зростаючої зацікавленості наукової спільноти у розробці ефективних методів та підходів, що дозволили пов'язати

вплив різноманітних чинників довкілля з виникненням мутацій на рівні генів, хромосом і генотипу організмів різного ступеня організації.

У серпні 2024 року кількість зареєстрованих ксенобіотиків, інформація про які зберігається на серверах Американського хімічного товариства, склала більше ніж 280 млн. речовин. На початок 2020 року була доступна інформація про більш ніж 100 000 хімічних речовин, що виробляються промисловістю, які здатні негативно впливати на навколишнє середовище та здоров'я людини і, зокрема, на генетичне здоров'я [8]. На кінець 2022 року Європейське агентство хімічних речовин оприлюднило інформацію про приблизно 800 тисяч хімічних речовин, для яких відсутня повна інформація про прямий або опосередкований вплив на генетичний апарат людини [9]. Індустріалізація спричиняє стрімке зростання концентрації хімічних речовин у довкіллі, при цьому їх генотоксичні властивості або не досліджені, або вивчаються з істотним відставанням від темпів їх розповсюдження. Така тенденція створює нові виклики для людства та стимулює наукову спільноту до розробки, впорядкування та вдосконалення нормативно-правової бази, що повинна супроводжувати проведення відповідних процедур на етапі оцінки, реєстрації, контролю, надання дозволів та заборон щодо використання хімічних речовин, які використовуються в різних сферах життя людини. Ситуацію також ускладнює той факт, що біля 20 % хімічних сполук, які потрапляють у довкілля і з якими контактує людська популяція, володіють специфічною біологічною дією: канцерогенною, мутагенною, алергенною. У контексті підтримки глобальної екологічної та генетичної безпеки пріоритетним завданням науковців є своєчасне отримання інформації щодо генотоксичного потенціалу усіх хімічних сполук, які присутні у навколишньому середовищі та здатні впливати на геном людини. Не зважаючи на те, що питання генетичної та екологічної безпеки перебувають у центрі уваги світової наукової спільноти, одна з ключових проблем сучасної генетичної токсикології пов'язана з необхідністю отримання достовірної генетичної оцінки для усіх хімічних сполук, які можуть потрапляти в навколишнє середовище, інформація про які зберігається в сучасних базах даних, зокрема ChemSpider [10], PubChem [11], SciFinder [12] та інших. На

сьогоднішній день основною перешкодою у вирішенні даної проблеми є відсутність у вільному доступі інформації про понад 50 000 хімічних речовин, оскільки вона вважається конфіденційною та захищена охоронними документами [13]. Крім того, для великої кількості хімічних речовин обмеженими є експериментальні токсикологічні дані [13,14], що не дозволяють використовувати класичні *in vitro* підходи для оцінки їх генотоксичності [15]. Враховуючи негативні наслідки впливу факторів навколишнього середовища на спадковий апарат людини, потребує усунення проблема, що пов'язана з виявленням та обліком різноманітних факторів генетичної і канцерогенної небезпеки. Ефективне вирішення цього завдання потребує від дослідників всебічного аналізу досліджуваних потенційних мутагенів, особливо в контексті їхньої класифікації, механізмів дії, та особливостей поширення у навколишньому середовищі.

### **1.1 Поширення мутагенно-активних хімічних сполук в об'єктах навколишнього середовища**

Пошкодження спадкового апарату людини можуть бути спричинені дією на нього ксенобіотиків, які надходять з різноманітних антропогенних джерел [3,4]. З метою моніторингу та контролю генетичної безпеки факторів навколишнього середовища, науковці намагаються сконцентрувати свої зусилля на тих об'єктах довкілля, з якими постійно контактує людська популяція. Важливим у цьому відношенні є отримання науково обґрунтованої оцінки генетичної безпеки індивідуальних хімічних сполук, що містяться у питній воді, атмосферному повітрі, продуктах харчування, засобах захисту рослин, засобах побутової хімії та лікарських препаратах.

Питна вода, як один з найважливіших складових факторів навколишнього середовища, може виступати в ролі транспортера забруднювачів довкілля в організм людини. Наявність генотоксичних речовин у питній воді має антропогенне походження, що пов'язано з потраплянням у водні екосистеми

фармацевтичних засобів, біоцидів, пестицидів та промислових відходів, в яких якісний склад хімічних сполук-мутагенів, як правило, не є контрольованим [16]. Крім того, навіть вживання людиною хімічно очищеної та знезараженої питної води може мати негативні генетичні наслідки для її здоров'я [17,18]. Виражені мутагенні та канцерогенні ефекти води, в цьому випадку, будуть опосередковані взаємодією побічних продуктів (наприклад тригалометан, галооцтова кислота та інші), що утворюються при її хлоруванні, з геномом людини. Досить цікавим є той факт, що мутагенні хімічні сполуки у воді можуть утворюватись в результаті активізації процесів корозії в металевих трубопроводах [18]. Не оптимістичними, з точки зору підтримки генетичного здоров'я населення України, є результати проведених тестувань на мутагенність поверхневих вод річки Дніпро. Опубліковані у науковій праці [19] результати свідчать про наявний мутагенний потенціал поверхневих вод вздовж всієї акваторії річки Дніпро, що обумовлено присутністю у воді в достатньо великій концентрації генотоксичних хімічних сполук, таких як поверхнево-активні речовини, поліциклічні ароматичні вуглеводні та важкі метали. Така екологічна ситуація, що характерна для основної водної артерії України, є мотивуючим фактором для науковців на шляху до розробки і впровадженню нових методів та підходів, що лежать в основі отримання якісної питної води. В цьому контексті, на сьогоднішній день, необхідним та затребуваним є оновлення методів оцінки мутагенного впливу факторів навколишнього середовища. Крім того, для зниження концентрації генетичних забруднювачів, що потрапляють у водні екосистеми, потребують удосконалення і методи очистки стічних вод.

Ще одним важливим фактором довкілля є атмосферне повітря, якість якого може бути асоційована з розвитком спадкових та генетичних захворювань. Стрімкий розвиток промисловості є основною причиною збільшення концентрації та кількості хімічних сполук, що потрапляють у атмосферне повітря та можуть бути потенційними мутагенами. Сотні мільйонів людей у всьому світі, особливо ті, що проживають у великих містах, піддаються негативному впливу забрудненого повітря, що значно перевищує допустимі порогові значення за



різними показниками, які встановлені ВООЗ [20]. Результати проведених наукових досліджень [21,22] демонструють наявність вираженого зв'язку між частотою захворювань на рак легень та тривалим впливом на людину забрудненого повітря. Занепокоєння наукової спільноти також викликають отримані дані, що свідчать про негативні генетичні наслідки впливу на генетичний апарат людини побічних продуктів згоряння автомобільного палива (особливо дизельного) та вугілля, що потрапляють у атмосферне повітря та формують фракцію завислих твердих мікрочастинок [23,24]. Ще на початку 20 сторіччя було встановлено, що для сажі, як одного з продуктів неповного згоряння вугілля, властиві виражені канцерогенні властивості, що пов'язані з наявністю в ній поліциклічного ароматичного вуглеводня – бенз[а]пірену [25]. Не зважаючи на суттєві негативні наслідки для генетичного здоров'я людської популяції, що можуть бути спричинені побічними продуктами горіння вугілля, воно залишається одним з основних джерел енергії, особливо в країнах з економіками, що розвиваються. Досить цікавий результат з наукової точки зору, та такий, що потребує публічного обговорення, був отриманий при проведенні оцінки впливу повітря на генетичне здоров'я людини у закритих приміщеннях [26]. Виявляється, що повсякденна побутова діяльність людини (наприклад приготування їжі з термічною обробкою) може бути пов'язана з ризиком погіршення генетичного здоров'я, що може бути ініційовано формуванням фракції мікрочастинок розміром 10мкм (газова складова повітря), які потрапляють в організм людини [26]. Результати проведеного дослідження, що опубліковані у науковій праці [27] підтверджують наявний мутагенний потенціал навіть для звичайного пилу, що може накопичуватись у закритих приміщеннях. Виходячи з цього, вологе прибирання у місцях де довгий час можуть перебувати люди є необхідним з точки зору зменшення впливу генотоксичних хімічних сполук на їх геном.

Продукти харчування також є найважливішою складовою навколишнього середовища, через які в організм людини потрапляють ксенобіотики з прямим або опосередкованим вираженим мутагенним потенціалом. Ефективне вирішення

питання контролю та оцінки ризиків впливу таких ксенобіотиків-мутагенів на генетичний матеріал людини, повинен ґрунтуватись на системному підході, в основі якого факт наявності певного мутагена в продуктах харчування (наприклад тваринного походження), повинен враховувати як особливості харчового ланцюга живлення відповідного організму так і етапність складного процесу виробництва продуктів харчування. З урахуванням такої концепції, потрапляння ксенобіотиків, що можуть мати негативний вплив на геном людини, може відбуватись, наприклад, на різних ланках харчового ланцюга. Таке теоретичне обґрунтування можливих шляхів розповсюдження мутагенів дозволяє зробити висновок про те, що всі основні об'єкти навколишнього середовища, з якими контактує людина, можуть бути першопричиною проявів мутагенності харчових продуктів рослинного та тваринного походження. При цьому, засоби захисту рослин, побутової хімії, а також залишки лікарських препаратів, які здатні викликати мутагенні ефекти, можуть потрапляти у водні екосистеми, повітря або ґрунт, а потім поширюватись в інших організмах через харчові ланки вищого рівня. Покращення ситуації щодо збереження генетичного здоров'я людини можливе через усвідомлення рівня загроз на генетичному рівні, з якими стикається людство у випадку неконтрольованих викидів отруйних речовин у атмосферу [28] та водні екосистеми [19], активного застосування пестицидів [29], консервантів та харчових добавок [30] з вираженою мутагенною та канцерогенною дією, використання полімерних органічних сполук [31], що контактують з продуктами харчування тощо. Першим необхідним кроком у вирішенні проблеми, що пов'язана з підтримкою генетичного здоров'я, є адаптація до умов сьогодення нормативних документів та законотворчих актів, що повинні бути направлені на зменшення негативного впливу мутагенів на людську популяцію. Крім того, країнам з економіками, що розвиваються необхідно враховувати позитивний досвід інших країн світу, в яких позитивна динаміка щодо забезпечення генетичного здоров'я ґрунтується відповідно до різноманітних настанов міжнародних організацій, таких як EFSA, OECD, UKEMS та інші.

З метою ефективного виявлення ксенобіотиків, що проявляють потенційні мутагенні властивості, та отримання оцінки рівня заподіяної шкоди генетичному здоров'ю людини необхідно звернути увагу на особливості взаємодії мутагенів з генетичним апаратом людини на молекулярному рівні.

## **1.2 Молекулярно-біологічні основи впливу мутагенів та їх різновиди**

Досить важливим в науковому відношенні, а також у контексті проведеного дослідження, є вивчення особливостей взаємодії та механізмів впливу різноманітних чинників хімічної та фізичної природи на стабільність геному [32]. Ми вже відмічали, що пошкодження ДНК також можуть бути спричинені дією на генетичний апарат факторів навколишнього середовища [3,4]. В залежності від локалізації факторів хімічної природи, що можуть індукувати пошкодження ДНК, розрізняють ендогенні (фізіологічної та метаболічної природи) та екзогенні фактори [4,33,34,35]. Вплив ендогенних та екзогенних факторів на молекулу ДНК може відбуватись шляхом прямої та опосередкованої дії. Прямий механізм впливу детермінований безпосередньою взаємодією ендогенних або екзогенних факторів з молекулою ДНК, що призводить до розривів хімічних зв'язків на рівні ДНК та ініціює зміни в її просторовій структурі [36,37]. Механізм опосередкованої дії екзогенних та ендогенних факторів реалізується через їх метаболізм та активацію проміжних продуктів, взаємодія яких з ДНК лежить в основі її пошкодження [38,39]. Незважаючи на те, що екзогенні та ендогенні чинники мають достатньо великий потенціал щодо модифікації генетичної інформації, відносний внесок внутрішніх та зовнішніх факторів у захворюваність онкологічними захворюваннями залишається поки що не визначеним [40].

Кожна клітина організму людини може зазнавати понад 10 000 пошкоджень ДНК на добу, більшість з яких, як правило, спричинені клітинними метаболічними процесами [41]. Що стосується дії мутагенів на будь яку клітину організму людини, то вони (мутагени) підрозділяються на «прямі» тобто на ті, дія яких обумовлена вихідною хімічною структурою речовини і на «опосередковані»,

дія яких обумовлена їхніми проміжними метаболітами. В даному випадку можна говорити про «клітинні метаболічні процеси». Механізм пошкодження ДНК ендogenous речовинами лежить в основі появи невідповідності заміщення гетероциклічних основ, міжланцюгових та внутрішньоланцюгових зшивок, формування аномальної структури ДНК [4,34,40]. Такі негативні впливи пов'язані з реакціями гідролізу, окиснення, алкілювання, що є результатом перебігу нормальних фізіологічних процесів [40]. Біологічні макромолекули є надзвичайно сприйнятливими до спонтанних хімічних реакцій, що реалізуються переважно через гідроліз та є відповідальними за формування апуринових/апиримідинових сайтів, на рівні яких можуть бути відсутні гетероциклічні азотисті основи [33,34,42]. З реакціями гідролізу також пов'язують індуковане дезамінування азотистих основ нуклеотидів ДНК [4]. Швидкість такого процесу може бути значно підвищена за рахунок впливу ультрафіолетового випромінювання, ДНК інтеркалюючих агентів, азотистої кислоти та бісульфіту натрію [43,44,45,46].

Достатньо велика кількість пошкоджень ДНК хімічними речовинами ендogenous природи виникає за рахунок їх участі у гідролітичних та окислювальних реакціях з водою та активними формами кисню, що є присутніми в клітині [42]. Спонтанні мутації, які притаманні всім без винятку клітинам, під час реплікації ДНК призводять до включення некомпліментарних нуклеотидів у щойно синтезовану молекулу ДНК, створюючи невідповідність спарювання гетероциклічних основ [47] і, як результат – до синтезу білка зі зміненим якісним складом амінокислотних залишків, що може вплинути на виконання білком специфічної функції. Не зважаючи на високорозвинений апарат реплікації, помилки включення піримідинових та пуринових гетероциклічних основ можуть відбуватись з частотою від  $10^{-8}$  до  $10^{-6}$  на клітину за одне покоління [48,49,50,51].

Реакції ДНК з активною формою кисню, за умов його високої концентрації, сприяють розвитку спадкових та спорадичних ракових захворювань [52]. Порушення окисно-відновного балансу через підвищення концентрації активного кисню може призводити до виникнення дисфункцій, що проявляються у вигляді

пошкоджень нуклеїнових кислот, білкових молекул, ліпідів та мембранних структур, які можуть бути пов'язані з розвитком серцево-судинних та нейродегенеративних захворювань [5,34,53]. В той же час, знижена концентрація активної форми кисню може індукувати появу хронічної гранулематозної хвороби та аутоімунних розладів [34]. Активні форми кисню та азоту приймають участь у формуванні понад 70 окиснених різновидів гетероциклічних основ та цукрів, що входять до складу модифікованих мономерних одиниць ДНК та впливають на стабільність генетичної інформації [54].

У межах проведеного дослідження нами було, в першу чергу, акцентовано увагу на вивченні впливу екзогенних факторів хімічної природи, негативну дію яких на генетичний апарат людини можна уникнути, особливо в тій ситуації коли генотоксичність певної хімічної сполуки доведена експериментальними методами або за допомогою сучасних прогностичних *in silico* моделей [55,56,57]. Серед базових екзогенних чинників фізичної природи (факторів навколишнього середовища), що можуть індукувати процеси пошкодження генетичного матеріалу виділяють іонізуюче, ультрафіолетове та інфрачервоне випромінювання, а також хімічні агенти, що можуть проявляти властивості генотоксичності [4,40,58]. В таблиці 1.1, з урахуванням особливостей пошкодження та активованими механізмами репарації, представлена інформація про основні ендо – та екзогенні фактори, що можуть впливати на генетичну стабільність.

Таблиця 1.1

**Ендогенні та екзогенні джерела пошкодження ДНК та можливі механізми репарації [4,5,58,59]**

Ендогенні фактори	Механізм пошкодження	Результат пошкодження	Механізми репарації	Наслідки для ДНК
	Окиснення	Модифікація гетероциклічних азотистих основ	Експізійна репарація азотистих основ	Міссенс мутації
				Локальна зміна просторової структури

## Продовження таблиці 1.1

	Механізм пошкодження	Результат пошкодження	Механізми репарації	Наслідки для ДНК
Ендогенні фактори	Алкілювання	Метилювання гетероциклічних азотистих основ	Пряма репарація; ексцизійна репарація азотистих основ; репарація неспарених пар азотистих основ	Мутації G→A, T→C; Пригнічення реплікації
	Гідроліз	Формування апуринових сайтів	Ексцизійна репарація азотистих основ	Порушення включення гетероциклічних основ при реплікації
			Постреплікативна репарація	Блокування роботи ДНК- та РНК-полімерази
		Дезамінування цитозину та синтез урацилу	Ексцизійна репарація азотистих основ	Міссенс мутації
	Помилки ДНК-полімерази	Заміни, вставки та видалення гетероциклічних азотистих основ	Система репарації невідповідності	Міссенс мутації; колапс реплікативної вилки; хромосомні перебудови
Екзогенні фактори	Іонізуюче випромінювання	Дволанцюгові розриви	Негомологічне з'єднання кінців	Мутагенез; хромосомні транслокації та перебудови.
			Гомологічна рекомбінація	
	Ультрафіолетове випромінювання	Синтез цикlobутанпіримідинових димерів	Ексцизійна репарація нуклеотидів	Транзиції; трансверсії; локальна зміна просторової структури
	Хімічні сполуки (ароматичні аміни, алкілюючі агенти, природні токсини, хіміотерапевтичні засоби )	Пошкодження гетероциклічних азотистих основ; синтез ДНК-адуктів	Ексцизійна репарація нуклеотидів; Пряма репарація; Система репарації невідповідності; Репарація неспарених пар азотистих основ.	Мутації зсуву рамки зчитування; блокування реплікації та транскрипції; локальна зміна просторової структури

Екзогенні джерела порушення генетичної стабільності, такі як іонізуюче випромінювання (рентгенівське випромінювання), космічне та ультрафіолетове випромінювання, а також вплив хімічних речовин-мутагенів сприяють накопиченню пошкоджень ДНК, яким кожна клітина повинна протидіяти щодня [3,40]. Радіаційно-індуковане пошкодження центральної нервової системи

пов'язане з розвитком окислювального стресу, накопиченням вільних радикалів, що лежить в основі молекулярних та клітинних змін, які, включно з пошкодженнями ДНК, можуть призводити до порушень в структурі нейронів, синаптичної пластичності, викликати системне запалення та призводити до загибелі нейронів [60].

Пряма дія інфрачервоного випромінювання викликає хімічні зміни на рівні ДНК, порушує її структуру, що може вплинути на процес реплікації. На цей вид пошкоджень припадає 30-40% індукованих інфрачервоним випромінюванням хімічних модифікацій ДНК [4]. Опосередкований вплив інфрачервоного випромінювання пов'язаний з процесами радіолізу молекул води, що є стимулюючим фактором для накопичення в клітині вільних радикалів, які безпосередньо приймають участь в окислювально-модифікованих пошкодженнях ДНК [61,62].

Ультрафіолетове випромінювання, як один з факторів навколишнього середовища, що здійснює постійний тиск на геномну цілісність організму, є однією з найбільш поширених небезпек для здоров'я людини. Випромінювання з довжинами хвилі від 280 до 315 нм є одним з потужних агентів фізичної природи, що може індукувати різноманітні мутагенні та цитотоксичні порушення [63]. Ультрафіолетове опромінення може ініціювати синтез циклобутанпіримідинових димерів та піримідин-(6,4)-піримідинових фотопродуктів з наступною зміною просторової структури ДНК та блокуванням процесів транскрипції та реплікації [3,64].

Дослідження впливу різноманітних хімічних агентів на генетичний апарат людини є одним з пріоритетних напрямків сучасної генетичної токсикології. Не зважаючи на те, що кількість хімічних речовин, що можуть впливати на генетичну стабільність з кожним роком тільки збільшується, на сьогоднішній день достатньо не погано вивчені механізми пошкодження ДНК для таких найбільш розповсюджених структурних класів екзогенних агентів як: ароматичні аміни, поліциклічні ароматичні вуглеводи, природні токсини, алкілюючі агенти та хіміотерапевтичні засоби [3,4].

Результати досліджень, що опубліковані у роботах [65,66,67,68] дозволяють прослідкувати зв'язок між приналежністю до даного класу хімічних сполук (ароматичних амінів) та вираженою їх мутагенною активністю. Ароматичні аміни є побічними продуктами горіння тютюну, що створює потенційний ризик для здоров'я людини та залишається найпоширенішою причиною смертей від раку легень в усьому світі [48]. Крім того, ароматичні аміни, що є одним з основних забруднювачів навколишнього середовища, можуть використовуватись, як базові компоненти при виробництві косметичної продукції, барвників, пластмас, а також харчових продуктів та пестицидів [3,4,69]. Досить відомими та найбільш дослідженими у науковому відношенні прикладами ароматичних амінів є 2-амінофлуорен та його ацитильований похідний 2-ацетиламінофлуорен, які використовувались як інсектициди, поки для них не була доведена канцерогенна властивість [70]. Амінофлуорени, за участі системи монооксигенази P450 гепатоцитів печінки, трансформуються у канцерогенні складні ефіри та сульфатні алкілюючі агенти, які можуть атакувати восьму позицію карбону гуаніну на рівні молекули ДНК [71]. Якщо утворені ДНК-аддукти не будуть видалені за рахунок ексцизійної репарації нуклеотидів, це може призвести до заміни гетероциклічних основ та подальшого зсуву рамки зчитування [72]. У наукових працях [73,74] досліджені особливості біологічної трансформації ароматичних амінів за участі прокаріотичних організмів, що створює передумови для контролю та підтримки генетично безпечного навколишнього середовища.

Ароматичні аміни широко використовуються як проміжні продукти синтезу діючих компонентів препаратів медичного призначення. В такій ситуації мутагенність домішок створює серйозні перешкоди на шляху дотримання генетичної безпеки та для запобігання випуску фармацевтичної продукції з потенційними генотоксичними властивостями [75,76]. Для отримання оцінки генетичних ефектів базового компоненту певного фармацевтичного препарату, необхідно також враховувати той факт, що ароматичні аміни можуть бути синтезовані як метаболіти за рахунок реакції гідролізу діючих або допоміжних компонентів препарату, що містять у своїй структурі амідні зв'язки [77].



Поліциклічні ароматичні вуглеводні надзвичайно поширені в навколишньому середовищі та є одними з основних забруднювачів атмосфери [78]. Ще у 1983 році Управління з охорони навколишнього середовища США повідомило про шістнадцять поліциклічних ароматичних вуглеводнів, що є основними забруднювачами довкілля [78,79]. Поліциклічні ароматичні вуглеводні є стійкими полютантами, що проявляють властивості токсичності, мутагенності, канцерогенності та імунотоксичності як для прокаріотичних так і для еукаріотичних організмів [80]. Канцерогенність даної групи хімічних сполук, як і для деяких ароматичних амінів, пов'язана з ферментативною активністю системи монооксигенази Р450 гепатоцитів печінки [81,82]. Метаболізм поліциклічних ароматичних вуглеводнів (наприклад, хінонів) відбувається через синтез реакційноздатних проміжних продуктів, які не є достатньо полярними для виведення та можуть викликати пошкодження клітинних мембран, білків та ДНК [4,79,80].

Природні токсини формують клас генотоксичних та канцерогенних хімічних речовин, які використовуються мікроорганізмами або грибами в захисних реакціях. Нитчасті гриби роду *Aspergillus* є основною причиною зараження афлатоксинами зернових і олійних культур, а також молочної продукції [83]. Продуцентами афлатоксинів, є вид *Aspergillus flavus*, які в основному синтезують В-афлатоксини В1 та В2, а також *Aspergillus parasiticus*, що продукують G-афлатоксини G1 та G2 [84,85]. Афлатоксин В1, що є одним з вагомих факторів захворюваності гепатоцелюлярною карциномою в усьому світі Продовольчою та сільськогосподарською організацією ООН був віднесений до канцерогенів [86]. Біотрансформація афлатоксину В1 відбувається за участі ферментної системи монооксигенази Р450 гепатоцитів печінки з утворенням токсичного та канцерогенного продукту афлатоксину В1-8,9-епоксиду, який може взаємодіяти з нітрогеном у сьомій позиції гуаніну з утворенням ДНК-аддукту [3,4,86]. Такий новосинтезований комплекс послаблює глікозильний зв'язок та призводить до депуринізації ДНК [40]. Досліджений також інший шлях хімічної трансформації афлатоксину В1, який враховує додатковий гідроліз ДНК-аддукту з

утворенням афлотоксин-B1-формамідопіримідину, який блокує реплікацію ДНК та має достатньо великий потенціал щодо ініціації мутацій по типу трансверсій [4,86].

З метою пошуку можливих шляхів удосконалення базових методів та пошуку нових *in silico* підходів для ефективної оцінки генотоксичні факторів навколишнього середовища, необхідно розглянути класичні базові (*in vivo* та *in vitro*) та сучасні методи оцінки генетичної безпеки факторів навколишнього середовища.

### **1.2.1 Методи оцінки генетичних ефектів на основі короткострокових тестів**

З метою проведення процедури тестування хімічних речовин на потенційну генетичну активність були розроблені різноманітні *in vitro* та *in vivo* методи з використанням еукаріотичних та прокаріотичних організмів. Відповідно до рекомендацій Організації економічного співробітництва та розвитку (OECD) щодо проведення тестування хімічних речовин [87], існує більше ніж 150 прийнятих науковою спільнотою та регулюючими органами методик, які дають можливість отримати інформацію щодо токсичного та генотоксичного потенціалу факторів навколишнього середовища. OECD надає рекомендації для 20 класичних методик [87], основна частина з яких була розроблена більше ніж 30 років тому [88,89] і на сьогоднішній день не дозволяють в повній мірі адекватно і об'єктивно дати відповідь на питання про те, чи володіє та чи інша хімічна речовина потенційною генетичною активністю. Рекомендації OECD щодо тестування хімічних сполук включають 9 *in vitro* експериментальних моделей з лабораторними тваринами та 11 тест-систем, в яких використовують прокаріотичні і одноклітинні еукаріотичні організми та клітинні лінії ссавців [87,88]. Для отримання достовірної оцінки хімічних речовин щодо їх потенційних генотоксичних властивостей необхідно врахувати три основні кінцеві результати пошкодження ДНК, які в першу чергу пов'язані з виникненням генних мутацій,

хромосомних аберацій та анеуплоїдії [90]. На сьогоднішній день не існує жодної короткострокової тест-системи яка б дозволила врахувати одночасно такі пошкодження спадкової інформації [91]. Тому для комплексної оцінки здатності хімічних речовин викликати пошкодження генетичного матеріалу, з урахуванням трьох кінцевих точок пошкодження, використовують класичну батарею короткострокових тестів [91,92,93]. Хімічні речовини можуть бути оцінені як такі, що не проявляють генотоксичного потенціалу, якщо при використанні *in vitro* методів по всіх кінцевим точкам пошкодження ДНК буде отриманий негативний результат [94]. Не зважаючи на те, що на сьогоднішній день розроблено більше ста методів оцінки генотоксичності, перевага надається тест-системам, в яких стандартизована методика прийнята науковою спільнотою та затверджена відповідними настановами міжнародних організацій (наприклад OECD, ECNA, UK-EMS, US-FDA, EFSA та ін.) [87,88]. В таблиці 2.1 представлена інформація про методи оцінки генотоксичності, що формують класичну батарею короткострокових тест-систем, які рекомендовані OECD в якості базових методів [87,91,92].

Таблиця 1.2

### Стандартна батарея тест-систем для оцінки генотоксичності [91,92]

№ тесту (OECD)	Назва тест-системи	Особливості проведення	Посилання
TG471	Тест на бактеріальну зворотню мутацію Еймса	<i>In vitro</i>	[95]
TG473	Тест на наявність хромосомних аберацій в клітинах ссавців	<i>In vitro</i>	[96]
TG474	Тест на виявлення мікроядер еритроцитів ссавців	<i>In vivo</i>	[97]
TG475	Тест на виявлення хромосомних аберацій в клітинах кісткового мозку ссавців	<i>In vivo</i>	[98]
TG476	Тест на виявлення мутацій в генах HPRT та XPRT в клітинах ссавців	<i>In vivo</i>	[99]
TG478	Тест на виявлення домінантних летальних мутацій	<i>In vivo</i>	[100]
TG483	Тест на виявлення хромосомних аберацій сперматогоніальних зародкових клітинах ссавців	<i>In vivo</i>	[101]
TG485	Тест на виявлення хромосомних транслокацій	<i>In vivo</i>	[102]

## Продовження таблиці 1.2

№ тесту (OECD)	Назва тест-системи	Особливості проведення	Посилання
TG486	Тест репаративного синтезу ДНК	<i>In vivo</i>	[103]
TG487	Тест на виявлення мікроядер у цитоплазмі інтерфазних клітин ссавців	<i>In vitro</i>	[104]
TG489	Тест аналіз ДНК-комет у лужних умовах	<i>In vitro</i>	[105]

У 2012 році Міжнародна рада з гармонізації технічних вимог до реєстрації фармацевтичних препаратів, для використання людиною, затвердила «Керівництво з тестування на генотоксичність та інтерпретацію даних для фармацевтичних препаратів, призначених для використання людиною» ICH S2(R1), відповідно до якої були представлені дві теоретично обґрунтовані схеми застосування батареї *in vitro* та *in vivo* тест-систем для визначення генотоксичності [106]. Перший підхід [91,92,106], що був запропонований для проведення тестування, враховує використання наступних експрес-методів: 1. Теста Еймса *Salmonella-mikrosome* (TG471) (тест на бактеріальну зворотню мутацію); 2. *In vitro* тест на наявність хромосомних аберацій в клітинах кісткового мозку ссавців (TG473); 3. *In vivo* тест для виявлення мікроядер еритроцитів периферичної крові ссавців (TG473). Другий підхід [91,92,106] базується на застосуванні таких тест-систем як: 1. Тест на бактеріальну зворотню мутацію Еймса (TG471); 2. *In vivo* тест для виявлення мікроядер еритроцитів ссавців (TG473); 3. Тест репаративного синтезу ДНК (TG486) або метод ДНК-комет (TG489). До найбільш поширених методів оцінки генотоксичного потенціалу відноситься тест на бактеріальну зворотню мутацію (TG471), тест на виявлення мікроядер у цитоплазмі інтерфазних клітин ссавців (TG487) та тест на наявність хромосомних аберацій в клітинах ссавців (TG473) [15].

На етапі розробки ефективних *in silico* моделей оцінки мутагенності, ми використовували набір даних хімічних сполук, для яких за допомогою тесту Еймса був визначений мутагенний потенціал. Тому в літературному огляді

детально розглянемо методологію проведення тестування за допомогою даного методу.

### 1.2.2 Тест-система на бактеріальну зворотну мутацію Еймса (TG 471)

Базовим якісним методом оцінки генетичної безпеки факторів навколишнього середовища є тест на зворотну мутацію, який був розроблений у 1970-х роках американським молекулярним біологом Брюсом Еймсом [107]. Тест на бактеріальну зворотну мутацію є швидким, недорогим та простим у виконанні методом [89,91]. Тест Еймса використовується як один з основних *in vitro* методів [108] для оцінки мутагенного потенціалу що проводиться з використанням бактеріальних штамів *Salmonella typhimurium*, що є ауксотрофними по гістидину [93,107,109]. Мутації в генах *hisG* та *hisD* на рівні гістидинового оперону штамів *Salmonella typhimurium* є стимулюючими щодо пригнічення процесів росту штамів-тестерів на поживному середовищу, в якому відсутній гістидин [15,91]. Перехід від ауксотрофності по гістидину до прототрофності відбувається під дією хімічних речовин-мутагенів, що здатні індукувати зворотні мутації по типу заміни пар нуклеотидних основ і зсуву рамки зчитування генетичного коду [15,89,91,107]. Мутагенну здатність досліджуваних хімічних речовин оцінюють шляхом підрахунку кількості ревертантних колоній, які культивуються в чашці Петрі у порівнянні з контролем. В результаті оцінюється не частота мутацій, що може бути індукована факторами навколишнього середовища, а кількість ревертантів, що селектуються [89]. Такий підхід представляє собою класичний напівкількісний метод Еймса, який використовується, практично у всіх лабораторіях світу, що займаються тестуванням хімічних сполук на їхню потенційну генетичну активність. Існує модифікація метода Еймса – кількісний метод, який реєструє безпосередньо частоту мутацій, що виникають під дією мутагенів, однак, він не є надто популярним серед дослідників не тільки через його складність, але й проблеми відтворюваності експерименту в різних лабораторіях. Стандартні тестові штами *Salmonella typhimurium*, крім точкових мутацій на рівні гістидинових оперонів, можуть містити додаткові мутації, які обумовлюють підвищену чутливість тест-організму до дії хімічних сполук, що

можуть вплинути на покращення прогностичної здатності такої тест-системи [110]. Наприклад, мутація на рівні гена *rfa* призводить до порушення структури зовнішнього ліпополісахаридного шару збільшуючи тим самим проникність клітинної стінки для потенційних забруднювачів навколишнього середовища [89,91]. Крім того, підвищення чутливості тест-системи досягається за рахунок використання мутантних штамів *Salmonella typhimurium*, в яких відсутній ген *uvrB*, білковий продукт якого є основним компонентом мультиферментної системи ексцизійної репарації нуклеотидів бактерій [107,111]. З метою підвищення прогностичних можливостей тесту Еймса була доведена необхідність введення плізміди *pKM101*, що визначає резистентність штамів *Salmonella typhimurium* до антибіотиків, таких як ампіцилін та карбеніцилін [107]. Крім того, фактор резистентності ( або R-фактор ) несе інформацію про гени *tusA* та *tusB*, які кодують ДНК-полімеразу V, що приймає участь у транслезійному синтезі ДНК [93]. Синтез ДНК через ділянки, що зазнали пошкоджень відіграє ключову роль у підвищенні чутливості тест-системи до ксенобіотиків, що проявляють властивості генотоксичності [93]. Стандартні штами *Salmonella typhimurium*, такі як TA 97, TA 98, TA 100 і TA 102, містять плазмиду *pKM101* [91,110]. Для штамів *Salmonella typhimurium* TA 1538, TA 1535, TA 1537 відсутній фактор резистентності. В таблиці 3 представлена інформація про генотипи основних мутантних штамів *Salmonella typhimurium*, що використовуються для оцінки генетичної безпеки факторів навколишнього середовища. Відповідно до рекомендацій Організації економічного співробітництва та розвитку для оцінки мутагенного потенціалу використовують п'ять штамів *Salmonella typhimurium* (TA1535, TA1537, TA98, TA100, TA102), застосування яких доводить їх прогностичну цінність з точки зору ідентифікації генотоксичних речовин [95]. Заміни, що виникають на рівні пари гетероциклічних основ G та C можуть бути виявлені за допомогою штамів TA100 та TA1535 [95,110,114].

Штами TA98 та TA1537 дозволяють ідентифікувати мутації по типу зсуву рамки зчитування [95,110, 112,115]. В науковій праці [116] була досліджена тест-система *Salmonella typhimurium* TA102, що дозволяє отримати генотоксичну

оцінку мутагенам, які взаємодіють з нуклеотидами А або С. Досить цікавим є той факт, що виявлення деяких мутагенів можливо тільки з використання тест-штаму TA102, тоді як інші штами, що рекомендовані для тестування ОЕСР, дають негативний результат [115].

Таблиця 1.3

**Генотипи мутантних тест-штамів *Salmonella typhimurium* [59,91,110]**

Штами <i>Salmonella typhimurium</i>	Тип реверсії	Мутації на рівні гістидинового оперону	Додаткові мутації		Наявність плазмиди рKM101	Посилання
			<i>rfa</i>	<i>uvr</i>		
TA98	Зсув рамки зчитування	D3052	+	+	+	[112]
TA1538	Зсув рамки зчитування	D3052	+	+	-	[113]
TA100	Заміна основ	G46	+	+	+	[114]
TA1535	Заміна основ	G46	+	+	-	[115]
TA1537	Зсув рамки зчитування	C3076	+	+	-	[116]
TA102	Транзиції/Трансверсії	G428	+	+	+	[117]

Бактеріальний тест на зворотну мутацію використовує в якості тест-системи клітини прокаріотичних організмів, які суттєво відрізняються від клітин ссавців, особливо за такими параметрами як поглинання та метаболізм. Досить суттєві відмінності спостерігаються також і з точки зору молекулярних механізмів репарації, що реалізовані на рівні про- та еукаріотичних організмів. З метою компенсації різниці в особливостях метаболізму еукаріотичних та прокаріотичних організмів для тестування на мутагенність було запропоновано використовувати екзогенне джерело метаболічної активації [95]. Найпоширенішою системою, що використовується для метаболічної активації, є фракція S9, яку отримують з клітин гепатоцитів печінки лабораторних тварин та використовують в *in vitro* тест-системах [94,118]. Хоча система метаболічної активації не дозволяє повністю

врахувати особливості метаболізму вищих еукаріотичних організмів включно з людиною, для такої тест-системи збільшується прогностичний потенціал і значно розширюється спектр хімічних сполук, які виявляють свою активність саме при застосуванні системи метаболічної активації, особливо у випадку дослідження потенційних канцерогенів, які можуть бути проміжними продуктами реакцій за участю ферментної системи монооксигенази P450 [110]. Основною перешкодою на шляху отримання достовірних результатів *in vitro* тестування на мутагенність за допомогою тесту Еймса є висока специфічність метаболічної активації, що залежить від віку, статі, генотипу людини тощо [119]. Використання для тестування на мутагенність сучасної клітинної моделі HepaRG, що демонструє значну активність ферментів системи монооксигенази P450, дозволило частково вирішити проблему не достатнього врахування базових шляхів біотрансформації ксенобіотиків в організмі людини *in vitro* тест-системами [15,120].

### **1.3 Сучасні методи оцінки генотоксичності факторів навколишнього середовища**

Суттєва зміна парадигми щодо проведення тестування на генотоксичність спостерігається після відкриття Сангером у 1977 році методу секвенування біологічних послідовностей, з наступним активним розвитком та становленням біоінформатики. Необхідність перегляду та трансформації базових методів оцінки генетичної безпеки факторів навколишнього середовища, з урахуванням досягнень біоінформатики, системної біології та обчислювальної токсикології, прослідковується у наукових працях [121,122]. Особливу увагу заслуговує активна інтеграція алгоритмів машинного навчання у генетичну токсикологію, що дає надію на вирішення основної проблеми генетичної токсикології, яка пов'язана з відсутньою інформацією про генотоксичний потенціал великої кількості хімічних сполук, які присутні у навколишньому середовищі [123]. Необхідно відмітити, що на сьогоднішній день розроблені сучасні методи та підходи для проведення генетичного тестування, які мають достатньо високі показники



чутливості та специфічності, але при цьому для них відсутні рекомендації OECD [88].

У відповідь на експоненційне збільшення кількості генотоксичних хімічних речовин, що продукує людство, спостерігається активізація наукової спільноти з метою пошуку нових підходів щодо оцінки генетичної безпеки факторів навколишнього середовища. Розвиток сучасних технологій секвенування наступного покоління (NGS) з наступною розробкою нової технології (esNGS), що дозволяє виправляти помилки при отриманні прочитань фрагментів ДНК, продемонстрували перспективні результати щодо виявлення індукованих факторами навколишнього середовища соматичних мутацій, які мають достатньо низьку частоту появи. У науковій праці [124] висвітлені базові принципи дуплексного консенсусного секвенування, що дозволяє оцінити мутагенний потенціал впливу ксенобіотиків на генетичний апарат людини. Методика дозволяє ідентифікувати артефакти створення бібліотек для секвенування на етапі ампліфікації, через порівняння частот появи нуклеотидів у певній позиції великої кількості копій фрагментів ДНК. Мутації, що індуковані факторами навколишнього середовища, відповідно до методики консенсусного дуплексного секвенування, будуть представлені у більшості ампліфікованих фрагментів ДНК [124,125]. Перевагою методу є отримання інформації про генотоксичний потенціал ксенобіотиків з визначеною локалізацією пошкоджень на рівні ДНК та їх якісними характеристиками. Технологія секвенування наступного покоління на основі підходу, що дозволяє ідентифікувати помилково прочитані нуклеотиди, забезпечує отримання деталізованої характеристики індукованих пошкоджень генетичного матеріалу на рівні одного нуклеотиду, що відкриває абсолютно нові можливості для вирішення задачі комплексної оцінки мутагенних ефектів факторів навколишнього середовища з урахуванням дозозалежного генетичного ефекту [126].

Одним з самих найнебезпечніших пошкоджень ДНК є дволанцюгові розриви, що можуть бути ініційовані екзогенними факторами навколишнього середовища фізичної, хімічної або біологічної природи. Репарація таких

пошкоджених ділянок ДНК може спричинити онкогенні перебудови [40]. У відповідь на дволанцюгові розриви відбувається активація ферментів фосфорилювання серин-треонін кіназ ATM, ATR та DNA-РКс [127,128]. У науковій праці [129], після проведеного детального протеомного аналізу, було ідентифіковано біля 900 сайтів фосфорилювання, що охоплює більше ніж 700 білків. Важливим субстратом ATM, ATR та DNA-РКс є білки-гістони H2Ах, які після фосфорилювання серину у 139 позиції перетворюються на  $\gamma$ H2Ах [130]. Досить цікавим, сучасним та не стандартним, з точки зору оцінки кінцевих генетичних ефектів є підхід, що дозволяє отримати оцінку генотоксичності через визначення ступеню прояву процесів фосфорилювання H2Ах у відповідь на вплив хімічних сполук. Така процедура здійснюється за допомогою таких методів, як проточна цитометрія, імуно-флуоресцентна мікроскопія та вестерн-блотінг імуноаналіз [131]. Основними клітинними лініями для даного підходу є В-лімфобласти людини НерG2 і ТК6 [127].

«ToxTracker» представляє собою перспективний інструмент для оцінки генотоксичних ефектів з урахуванням не стандартних точок пошкодження ДНК. Система тестування «ToxTracker» на основі стовбурових клітин ссавців дозволяє виявляти активацію специфічних клітинних сигнальних шляхів, що дають можливість провести генотоксичне профілювання факторів навколишнього середовища з урахуванням їх дозованого впливу. «ToxTracker» використовує панель, що містить шість зелених флуоресцентних білків-репортерів, по одному на кожен клітинну лінію, за допомогою яких оцінюють здатність генотоксичних агентів реагувати з генетичним матеріалом. Результатом такої взаємодії може бути блокування реплікації ДНК, індукування окислювального стресу, активація реакцій, що опосередковані з денатурацією білків або загальні реакції клітинного стресу, залежні від транскрипційного фактору Р5, що виконує функцію супресора пухлинного росту [132,133]. Система «ToxTracker» дозволяє за одне тестування визначити генотоксичний потенціал досліджуваної сполуки, з урахуванням різних кінцевих точок пошкодження ДНК [133].

Тест-система для оцінки генотоксичності факторів навколишнього середовища, що заснована на виявленні мутацій гена *Pig-a* заслуговує на особливу увагу. Не зважаючи на те, що *in vivo* аналіз *Pig-a* продемонстрував свою перспективність щодо оцінки кінцевої точки пошкодження ДНК у вигляді мутацій, ще у далекому 1999 р. [134], рекомендацію OECD методика отримала тільки у середині 2022 р. [135]. Ген *Pig-a* кодує каталітичну субодиницю N-ацетилглюкозамінтрансферази, яка бере участь на ранньому етапі синтезу глікозилфосфатидилінозиту [136], що зв'язує на поверхні гомопоетичних клітин людини та лабораторних ссавців білкові маркери (наприклад продукт гена CD59) [137]. З усіх генів, що асоційовані з глікозилфосфатидилінозитом, лише ген *Pig-a* розташований на X-хромосомі [138]. Відповідно фенотип, для якого характерна відсутність глікозилфосфатидилінозиту буде інформативним з точки зору наявності мутацій на рівні кодуючої ділянки гена *Pig-a*.

У наукових працях [139,140] висвітлені питання оцінки генотоксичності факторів навколишнього середовища за допомогою нової багатообіцяючої моделі, яка використовує методику мікроядерного аналізу з заплідненими курячими яйцями та еритроцитами. До основних переваг методу відноситься можливість оцінки генотоксичних ефектів на рівні моделі *in vitro* з урахуванням параметрів ADME, що є визначальними з точки зору біодоступності хімічної сполуки та пов'язані з її адсорбцією, розподілом, метаболізмом, виділенням та токсичністю. При цьому, відповідно до базових принципів концепції «3R» [141,142], з'являється можливість отримати оцінку генотоксичного потенціалу певного фактора навколишнього середовища без додаткового застосування *in vivo* тест-систем.

Класична схема оцінки генотоксичного потенціалу факторів навколишнього середовища включає використання стандартної батареї *in vitro* та *in vivo* тест-систем, які мають суттєві недоліки з точки зору часових витрат та вартості експериментальних досліджень. Крім того, керуючись базовими принципами концепції «3R» необхідним є зменшення кількості досліджень з піддослідними тваринами. В умовах збільшення кількості хімічних речовин, що можуть

проявляти генотоксичні властивості, особливу увагу вчені приділяють *in silico* методам, що можуть виступати в якості альтернативних підходів для генетичної оцінки факторів навколишнього середовища. Затвердження настанови ІСН М7 «Оцінка та контроль ДНК-реактивних (мутагенних) домішок у фармацевтичних препаратах для обмеження потенційного канцерогенного ризику» є визначальною подією, що стала стимулом на шляху впровадження сучасних *in silico* моделей, які використовують з метою отримання об'єктивної оцінки мутагенної активності факторів навколишнього середовища [6, 77,143] та токсичних ефектів, що можуть бути індуковані ксенобіотиками [144,145]. Обчислювальна токсикологія за допомогою прогностичних *in silico* моделей QSAR у поєднанні з алгоритмами машинного навчання та математичного апарату статистики дозволяє отримати інформацію про мутагенний потенціал, навіть у ситуації, коли для певної хімічної сполуки відсутні експериментальні дані про генотоксичність [146]. Використання *in silico* моделей QSAR є багатообіцяючим перспективним підходом для оцінки мутагенного потенціалу ксенобіотиків. Фундаментом прогностичної здатності таких моделей є набір молекулярних дескрипторів, які представляють фізико-хімічні, просторові, структурні та електронні властивості певного досліджуваного ксенобіотика [91,147]. Необхідність проведення досліджень з використанням *in silico* QSAR моделей для вирішення задач обчислювальної токсикології прослідковується у нещодавно опублікованих дослідженнях [55,147-150].

## Висновки до розділу 1

Стандартні *in vitro* та *in vivo* методи, що відносяться до класичної батареї тест-систем є складними, дорогими, тривалими у часі та мають проблему відтворюваності в різних лабораторіях. Крім того, розвиток сучасної токсикології повинен спиратись на базові принципи концепції «3R», що направлена на зменшення кількості теплокровних тварин, що використовують в експерименті. Такі обмеження щодо використання класичних методів оцінки генотоксичності факторів навколишнього середовища стали відправною точкою на шляху

формування нового вектору розвитку сучасної комп'ютерної токсикології, в основі якого є розробка та впровадження альтернативних сучасних *in silico* методів оцінки генетичної активності факторів навколишнього середовища. Відповідно до такого підходу, базовим фундаментом для отримання оцінки генетичного впливу факторів довкілля на спадковий апарат людини є досягнення в області інформаційних технологій та систем штучного інтелекту, активний розвиток яких припадає на початок 21 століття. Така особливість розвитку сучасної токсикології вимагає від дослідників постійної уваги з метою вирішення питань пов'язаних зі створенням нових та удосконаленням вже існуючих *in silico* моделей прогнозування генотоксичності впливу факторів навколишнього середовища. В цьому контексті, сформульована мета дисертаційного дослідження відповідає загальним тенденціям розвитку сучасної токсикології.

## РОЗДІЛ 2 МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕННЯ

В даному розділі представлена інформація про особливості формування бази даних хімічних сполук, що використовувалась на етапі навчання, валідації та тестування Ames/QSAR моделей прогнозування мутагенності Еймса. Акцентовано увагу на одновимірних та двовимірних молекулярних дескрипторах, що при створенні прогностичних моделей машинного навчання використовувались в якості предикторів. Висвітлені методи та підходи, що лежать в основі розроблених Ames/QSAR моделей оцінки мутагенності факторів навколишнього середовища.

### 2.1 База даних хімічних сполук – потенційних мутагенів

Ідентифікація хімічних сполук з потенційними мутагенними властивостями на основі *in silico* Ames/QSAR моделей полягає у вирішенні задачі бінарної класифікації з наступним розподілом досліджуваних ксенобіотиків на дві групи: мутаген/не мутаген. Кожний етап реалізації моделей прогнозування мутагенності Еймса, що включає навчання, валідацію та тестування *in silico* Ames/QSAR, вимагає від дослідників використання баз даних потенційних мутагенів, для яких експериментально, за допомогою тесту Еймса, отримана інформація про мутагенний потенціал певного ксенобіотика. При проведенні дисертаційного дослідження нами була використана база даних ксенобіотиків [151], яка представляє собою набір даних, що був отриманий шляхом об'єднання трьох, широко використовуваних дослідниками, датасетів: Kazius-Bursi [152], Hansen [153] та EFSA [154]. Крім того, об'єднаний набір даних був доповнений мікотоксинами, які в науковій праці [155] використовувались для побудови *in silico* моделей для отримання оцінки мутагенних та канцерогенних ефектів впливу даної групи генотоксичних сполук на генетичний апарат людини. Після видалення дублікатів ксенобіотиків загальна кількість хімічних сполук в наборі

даних склала 8454. Об'єднаний набір даних, доповнений мікотоксинами, був збережений у csv. форматі, в якому для кожної хімічної сполуки зберігалась наступна інформація: 1. Ідентифікатор (ID), що відповідає порядковому номеру ксенобіотика; 2. SMILES (Simplified Molecular Input Line Entry System) лінійна нотація – це загальноприйнятий 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 мутагенний потенціал ксенобіотиків, що отримана експериментально за допомогою тесту Еймса (позначається 1, якщо хімічна сполука проявляє мутагенні властивості та 0 – не мутаген).

При проведенні дисертаційного дослідження, з метою отримання бінарного класифікатора що має найкращу точність, нами було запропоновано розподілити датасет на дев'ять однорідних структурних класів ксенобіотиків (табл. 2.1)

Таблиця 2.1

### Розподіл ксенобіотиків за структурними класами

№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 ксенобіотиків відповідно до особливостей будови їх молекулярного каркасу (Molecular Framework) був виконаний за допомогою веб-сервіса ClassyFire [156], який на вхід приймає інформацію про ксенобіотик у текстовому форматі SMILES нотації, на виході отримаємо повну класифікацію хімічної

сполуки, включно з даними про особливості будови її молекулярного каркасу, що відповідає структурному класу, до якого відноситься досліджуваний ксенобіотик. Подібність між молекулярними каркасами хімічних сполук на рівні двох різних класів дозволила нам прийняти рішення щодо об'єднання таких ксенобіотиків в окремі групи. З урахуванням такого підходу аліфатичні ациклічні хімічні сполуки сформували першу групу ксенобіотиків (табл 2.1). Друга група була отримана шляхом об'єднання двох класів ксенобіотиків: аліфатичних гетеромоноциклічних та аліфатичних гетерополіциклічних. Заслуговує на увагу третя група ксенобіотиків, в яку входять два класи хімічних сполук: аліфатичні гомомоноциклічні та аліфатичні гомополіциклічні. Не рівномірний розподіл кількості мутагенів до кількості не мутагенів для даної групи ксенобіотиків (табл 2.1), може суттєво вплинути на прогностичну здатність Ames/QSAR моделей. Очевидно, що Ames/QSAR моделі, що будуть орієнтовані на відповідну групу ксенобіотиків (аліфатичні гомомоноциклічні та аліфатичні гомополіциклічні) будуть мати низьку точність. В такій ситуації *in silico* оцінку мутагенного потенціалу для даної групи хімічних сполук було отримано за допомогою моделей, які на етапі навчання, тестування та валідації використовували повний датасет, представлений 8454 ксенобіотиками. Четверта група хімічних сполук сформована з урахуванням двох подібних структурних класів ксенобіотиків: ароматичних гетеромоноциклічних та ароматичних гетерополіциклічних хімічних сполук. П'ята група ксенобіотиків об'єднує два структурних класи: ароматичні гомомоноциклічні та ароматичні гомополіциклічні хімічні сполуки. Рівномірність розподілу хімічних сполук в межах двох класів – мутаген/не мутаген для першої, другої, четвертої та п'ятої груп ксенобіотиків дозволяє використовувати відповідні об'єднані набори даних з метою створення ефективних моделей машинного навчання.

Необхідно відмітити, що точність розроблених *in silico* моделей прогнозування мутагенності Еймса залежить також від кількості хибнонегативних та хибнопозитивних, результатів тестування, що були отримані експериментально за допомогою *in vitro* тесту Еймса. Кількість таких ксенобіотиків в базах даних,



що використовуються для створення Ames/QSAR моделей, повинна бути мінімізована. Крім того, розмір датасету, особливо у випадку невеликої кількості ксенобіотиків з визначеним мутагенним потенціалом, може суттєво вплинути на точність розроблених моделей оцінки мутагенності Еймса. Аналіз нещодавно опублікованих наукових праць [149,151,157], в яких дослідники, з метою створення ефективних *in silico* моделей для прогнозування мутагенності Еймса, використовували ті ж самі набори даних (Kazius-Bursi, Hansen та EFSA) додає впевненості в тому, що початковий крок на шляху отримання ефективного бінарного класифікатора, який пов'язаний з вибором та формуванням бази даних хімічних сполук – потенційних мутагенів, не матиме негативного впливу на точність розроблених Ames/QSAR моделей.

Наступний крок на шляху отримання ефективних *in silico* моделей прогнозування мутагенності Еймса, що може мати суттєвий вплив на кінцеві результати моделювання, пов'язаний з вибором відповідних наборів даних та розрахунком молекулярних дескрипторів, які представляють фізико-хімічні, просторові, структурні та електронні властивості досліджуваних ксенобіотиків [91,147].

## **2.2 Молекулярні дескриптори, як базові предиктори для *in silico* моделей прогнозування мутагенності Еймса**

Молекулярний дескриптор є кінцевим результатом логічної та математичної процедури, яка направлена на трансформацію інформації про хімічну сполуку (наприклад у вигляді SMILES нотації) у числове значення або певний результат деякого стандартизованого експерименту [158]. Молекулярні дескриптори виступають в якості вхідних даних – предикторів, аналіз яких дозволяє отримати прогноз, щодо проявів певної цільової ознаки. Аббревіатура QSAR (Quantitative Structure-Activity Relationship) використовується в науковій літературі, для позначення моделей кількісного співвідношення структура-активність. Фундаментом розроблених сучасних QSAR моделей, що використовуються для

оцінки мутагенного впливу факторів навколишнього є два основних підхода, що відповідають двом типам бінарних класифікаторів: на основі статистичних методів і алгоритмів машинного навчання та на основі правил [8]. Ames/QSAR моделі першого типу використовують інформацію про фізико-хімічні, просторові, електронні властивості досліджуваних ксенобіотиків, що задаються набором молекулярних дескрипторів, для визначення цільової змінної (мутагенності Еймса). Ідентифіковані на рівні молекулярної структури досліджуваних ксенобіотиків певні підструктури та/або функціональні групи, що можуть бути пов'язані з проявами мутагенності (маркери мутагенності), відповідають другому підходу моделювання. З метою вирішення поставлених задач, з урахуванням підвищеного інтересу з боку дослідників до даного напрямку досліджень, нами була зосереджена увага на моделях як першого так і другого типу.

На рис 2.1 представлена схема стандартного робочого процесу побудови Ames/QSAR моделей на основі статистичних методів та алгоритмів машинного навчання [159], що дозволяє отримати генетичну оцінку впливу мутагенів хімічної природи на генетичний апарат людини.

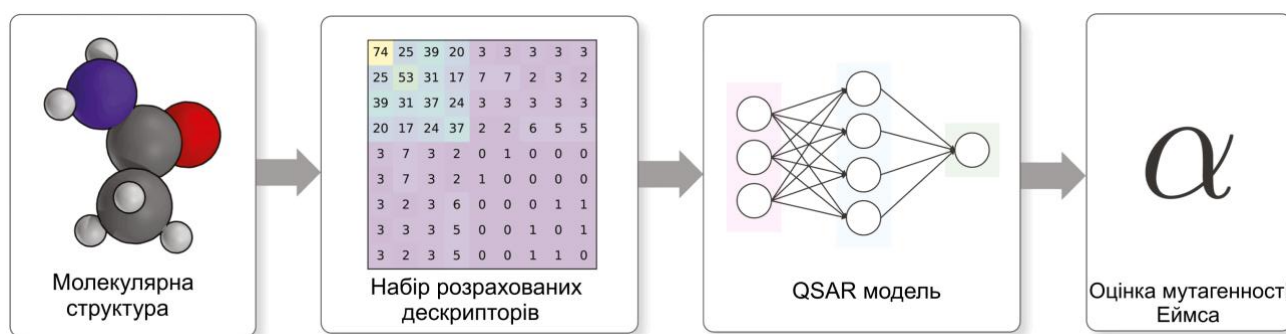


Рис. 2.1. Робочий процес побудови Ames/QSAR моделей з використанням статистичних методів та алгоритмів машинного навчання, адаптовано відповідно до [159]

Відповідно до молекулярної структури досліджуваних ксенобіотиків, що представлені в датасеті, розраховуються молекулярні дескриптори (наприклад PaDel, RDkit, Mordred). Наступний важливий етап моделювання пов'язаний з попередньою обробкою даних, який реалізується, зазвичай, через їх нормалізацію

та видалення аномальних значень. Вибір статистичного методу або певного алгоритму машинного навчання може бути визначальним на шляху отримання ефективного класифікатора, для оцінки мутагенності Еймса факторів навколишнього середовища. Необхідно зазначити, що QSAR моделі мають широке застосування для вирішення великої кількості задач різних галузей сучасної науки. Вони виступають в ролі базового інструменту для вирішення нагальних проблем хемоінформатики, протеоміки, комп'ютерної токсикології, матеріалознавства тощо. QSAR моделі використовуються для оцінки біологічної активності речовин, для яких відсутня повна інформація про їх фізико-хімічні властивості. Проведення віртуального скринінгу потенційних лікарських препаратів на основі структури лігандів також досягається за рахунок застосування QSAR моделей [160]. Крім того, QSAR моделі є досить популярним інструментом для визначення фармакокінетичних параметрів для лігандів, що в науковій літературі об'єднуються аббревіатурою ADME [161].

Молекулярні дескриптори, з урахуванням специфіки отримання їх числових значень, поділяють на обчислювальні дескриптори (теоретичні) та експериментальні. У межах проведеної наукової роботи, для досягнення поставлених задач, нами була акцентована увага на теоретичних дескрипторах, що можуть бути розраховані за допомогою спеціалізованого програмного забезпечення, наприклад PaDel [162], OpenBabel [163] та веб-сервісів Galaxy [164], ChemDes [165]. Крім того, отримати молекулярні дескриптори відповідно до SMILES нотації ксенобіотиків можна за допомогою бібліотеки RDKit мови програмування Python.

Класифікація обчислювальних дескрипторів враховує їх поділ на 3 класи. 1D – об'єднує набір одновимірних молекулярних дескрипторів, в яких не враховується інформація про зв'язність атомів в структурі молекули. 1D дескриптори легко розрахувати, але вони не містять інформації про структурні особливості ксенобіотиків. При моделювання такі дескриптори не використовують окремо, а при поєднанні їх з іншими класами предикторів дозволяють отримати ефективну Ames/QSAR модель. Прикладом 1D

дескрипторів може бути молекулярна маса, кількість донорів та акцепторів водневого зв'язку, коефіцієнт рефрактерності, коефіцієнт ліпофільності тощо. Необхідно зазначити, що в науковій літературі можна зустріти класифікацію молекулярних дескрипторів [166], відповідно до якої 1D дескриптори відносяться до 0D групи. Крім того, автори статті пропонують розглядати додатково 4D клас молекулярних дескрипторів, які можуть бути розраховані за результатами симуляції молекулярної динаміки, що дозволяє досліджувати в часі взаємодію декількох молекул (наприклад ліганда з біологічною мішенню – ферментом, що відбувається в його активному центрі).

Двовимірні 2D молекулярні дескриптори, які ще називають топологічними, формують найбільшу групу предикторів, що можуть бути розраховані відповідно до графового представлення молекули, де атоми формують вершини графа, а його ребра відповідають за зв'язки між атомами. У межах проведеної роботи нами особлива увага була приділена 2D молекулярним відбиткам структури (molecular fingerprint), що представляють собою бітовий рядок, в якому кожний біт відповідає за наявність/відсутність певної функціональної групи або підструктури на рівні молекули. В науковому відношенні достатньо важливим та таким, що потребує вирішення, є пошук відповіді на питання щодо ефективності застосування різних типів відбитків структури, окремо від інших молекулярних дескрипторів, для побудови прогностичних Ames/QSAR моделей.

3D дескриптори дозволяють врахувати інформацію про розташування в просторі атомів, що формують молекулу досліджуваної хімічної сполуки. Розрахунок тривимірних дескрипторів є складним процесом, що потребує достатньо великих обчислювальних ресурсів та часу. Крім того, існують деякі обмеження щодо отримання розрахунків таких дескрипторів, що пов'язано з використанням комерційного програмного забезпечення, що має ліцензійне обмеження та розповсюджується на платній основі. До таких програмних продуктів, наприклад, відноситься Dragon, що дозволяє отримати розрахунки для 4885 дескрипторів включно з 3D дескрипторами.

При проведенні дисертаційного дослідження з метою побудови ефективних *in silico* моделей прогнозування мутагенності Еймса ми використовували тільки 1D та 2D молекулярні дескриптори. Такий вибір, в першу чергу, пов'язаний з тим, що предиктори відповідних класів є найбільш популярними серед науковців, можуть бути достатньо швидко розраховані та досить часто використовуються у подібних дослідженнях. Крім того, досить важливим в науковому відношенні є порівняння ефективності застосування різних наборів 1D та 2D молекулярних дескрипторів, що розраховані за допомогою різного програмного забезпечення, для вирішення задачі прогнозування мутагенності Еймса факторів навколишнього середовища. Крім того, відповідно, до наукової праці [167], якщо кількість ознак, що використовуються при створенні QSAR моделей буде більшою ніж п'ята частина від розміру всієї бази даних, в такій ситуації може бути значно знижена точність таких розроблених моделей. Відсутня пряма кореляція між кількістю предикторів та точністю моделей стає надійним стимулом на шляху реалізації ефективних прогностичних *in silico* моделей, які враховують мінімальний перелік найважливіших ознак, що дозволяє з високими показниками точності для досліджуваних ксенобіотиків передбачити мутагенність Еймса. В науковій праці [155] було отримано багатообіцяючий результат щодо оптимізації моделей машинного навчання через зменшення набору вхідних даних. Автори запропонували використовувати алгоритм RFE (Recursive Feature Elimination), що дозволяє рекурсивно видаляти найменш впливові ознаки, при цьому не втрачаючи на показниках точності розроблених QSAR моделей. Такий позитивний результат було враховано, та на етапі проєктування Ames/QSAR моделей за допомогою модифікованого алгоритму RFECV (Recursive Feature Elimination with Cross-Validation) було отримано мінімальний за кількістю перелік найважливіших дескрипторів для основних структурних класів ксенобіотиків, використання яких може вплинути на покращення точності *in silico* моделей прогнозування мутагенності Еймса. Крім того, формування переліку найбільш релевантних дескрипторів, що відносяться до певного структурного класу ксенобіотиків, може лежати в основі пошуку зв'язків між мутагенністю та фізико-хімічними,

просторовими, електронними тощо властивостями, що можуть задаватися певним числовим значенням відповідного молекулярного дескриптора.

У межах даного розділу дисертаційної роботи заслуговує на увагу висвітлення питань щодо особливостей розрахунку різних наборів молекулярних дескрипторів та їх використання в якості предикторів для Ames/QSAR моделей. Крім того, було приділено особливу увагу відбиткам структури (molecular fingerprint), які використовувались для моделювання двох типів, розроблених при виконанні дисертаційного дослідження, Ames/QSAR моделей: на основі статистичних методів і алгоритмів машинного навчання та на основі правил (фундаментом прогностичної здатності яких виступають ідентифіковані структурні маркери мутагенності).

### **2.2.1 1D та 2D молекулярні дескриптори та особливості їх розрахунків**

Молекулярні дескриптори виступають базовими предикторами *in silico* моделей, що використовуються для прогнозування основних точок пошкодження ДНК, таких як генні мутації та хромосомні аберації. На сьогоднішній день вчені-науковці використовують молекулярні дескриптори, що можуть бути розраховані за допомогою різного програмного забезпечення. При цьому, на етапі розробки таких програмних продуктів, можуть застосовуватись різні підходи щодо їх (дескрипторів) розрахунків. В такій ситуації, досить логічним є пошук відповіді на питання про те, чи можуть певні набори предикторів впливати на точність розроблених моделей прогнозування мутагенності Еймса.

Серед великої кількості програм, що дозволяють отримати молекулярні дескриптори, заслуговують на увагу PaDEL [162], OpenBabel [163], Chemopy [168], CDK[169], RDKit [170], DRAGON [171], Mordred [172] та інші. Досить популярними серед науковців є розроблені для розрахунку молекулярних дескрипторів веб-сервіси BioTriangle [173], Galaxy [164], ChemDes [165]. Зручний веб-інтерфейс, швидкість отримання розрахунків молекулярних дескрипторів та доступність програмного забезпечення були основними критеріями щодо вибору відповідного програмного забезпечення для розрахунку молекулярних дескрипторів. Веб-сервіс Galaxy, що використовується для вирішення

різноманітних задач біоінформатики (включно з хемоінформатикою) відповідає таким вимогам та був обраний нами для проведення розрахунків трьох наборів молекулярних дескрипторів таких як: PaDEL, Mordred та RDKit. Вибір декількох різних наборів молекулярних дескрипторів був пов'язаний з необхідністю перевірки гіпотези щодо ефективності використання певних груп 1D та 2D предикторів при реалізації прогностичних Ames/QSAR моделей. Нетривіальна задача щодо формування таких наборів дескрипторів, з урахуванням сформульованих задач дисертаційного дослідження, була вирішена після опрацювання наукової статті [165], де була представлена схема, що демонструє зв'язок між різними наборами дескрипторів, які можуть бути розраховані за допомогою веб-сервіса ChemDes ( рис 2.2).

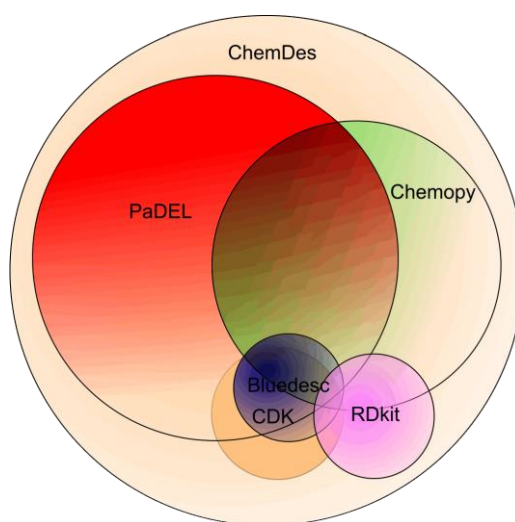
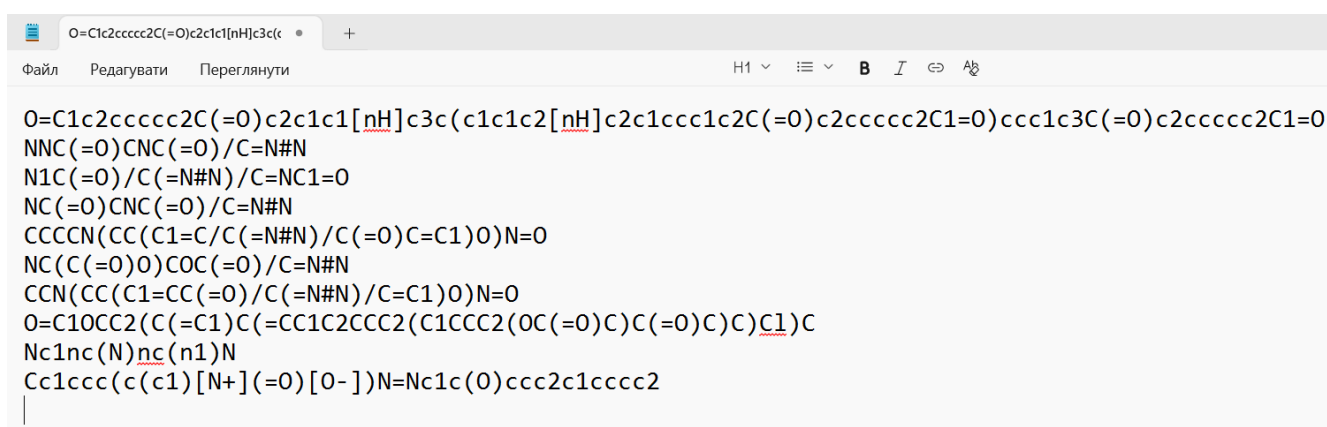


Рис. 2.2. Графічне представлення кількості молекулярних дескрипторів, розрахованих за допомогою веб-сервіса ChemDes, адаптовано відповідно до [165]

На рисунку 2.2. показана схема, де кола, що забарвлені різними кольорами відповідають наборам дескрипторів, що отримані за допомогою різних інструментів. Розмір площі кожного кола пропорційний кількості дескрипторів. Розмір площі перетинів кіл відповідає кількості дескрипторів, які збігаються в різних наборах [165]. Було прийнято рішення обрати набір дескрипторів PaDEL та RDkit, тому що в науковому відношенні достатньо цікавою може бути отримана

відповідь на питання про ефективність Ames/QSAR моделей, що використовують різні за розмірами (кількістю дескрипторів) та якісним складом набори вхідних даних, що представлені молекулярними дескрипторами. Кількість розрахованих 1D та 2D дескрипторів PaDEL, що використовувались при реалізації прогностичних моделей, склала 1444, а RDkit – 196. У зв'язку з тим, що функціонал веб-сервіса Galaxy дозволяє розрахувати додатково дескриптори Mordred, що формують одну з найбільших груп предикторів (1825, з яких 1613 – це 1D та 2D дескриптори), нами було прийняте рішення щодо їх використання при моделюванні. В такій ситуації оцінка ефективності Ames/QSAR моделей, з точки зору прогнозування мутагенності Еймса, може бути зроблена також і з урахуванням 1D та 2D предикторів двох достатньо великих за розміром наборів вхідних даних, для яких приблизно 30% дескрипторів відрізняються за якісним складом.

При розрахунках молекулярних дескрипторів на вхід сервіса Galaxy було подано перелік всіх хімічних сполук – потенційних мутагенів сформованої бази даних у текстовому форматі SMILES-нотації. На рис. 2.3, в якості прикладу, представлена інформація про перші десять ксенобіотиків сформованої бази даних, яка використовувалась при проведенні дослідження.



The image shows a screenshot of a text editor window. The title bar indicates the file path: `O=C1c2cccc2C(=O)c2c1c1[nH]c3c(c1c1c2[nH]c2c1ccc1c2C(=O)c2cccc2C1=O)ccc1c3C(=O)c2cccc2C1=O`. The editor has a menu bar with 'Файл', 'Редагувати', and 'Переглянути'. The toolbar includes icons for font size (H1), list, bold (B), italic (I), undo (↶), and redo (↷). The main text area contains the following SMILES strings, each on a new line:

```
O=C1c2cccc2C(=O)c2c1c1[nH]c3c(c1c1c2[nH]c2c1ccc1c2C(=O)c2cccc2C1=O)ccc1c3C(=O)c2cccc2C1=O
NNC(=O)CNC(=O)/C=N#N
N1C(=O)/C(=N#N)/C=NC1=O
NC(=O)CNC(=O)/C=N#N
CCCCN(CC(C1=C/C(=N#N)/C(=O)C=C1)O)N=O
NC(C(=O)O)COC(=O)/C=N#N
CCN(CC(C1=CC(=O)/C(=N#N)/C=C1)O)N=O
O=C1OCC2(C(=C1)C(=CC1C2CCC2(C1CCC2(OC(=O)C)C(=O)C)C1)C
Nc1nc(N)nc(n1)N
Cc1ccc(c(c1)[N+](=O)[O-])N=Nc1c(O)ccc2c1cccc2
```

Рис 2.3. Вміст текстового файлу, що містить інформацію про ксенобіотики у SMILES нотації



Після запуску відповідних інструментів веб-сервіса Galaxy (Padel, Mordred та RDkit) для кожної хімічної сполуки датасета був отриманий набір дескрипторів, що зберігався у .csv форматі. Такий формат даних легко, за допомогою пакета Excel Microsoft Office, можна конвертувати у звичаний формат .xlsx. На рисунку 2.4 представлений фрагмент бази даних ксенобіотиків, що містить додатково (у порівнянні з інформацією, що викладена на стор. 53) розраховані молекулярні дескриптори (показані значення перших п'яти молекулярних дескрипторів RDkit).

Class mutagene ( Molecular Framework)	Canonical SMILES	mutagenity BalabanJ	BertzCT	Chi0	Chi0n	Chi0v
Aromatic heteropolycyclic compound	<chem>O=C1c2ccccc2C(=O)c2c1c1[nH]c3c(c1c1c2</chem>	0	1.49	3545.358	33.739	25.687
Aliphatic acyclic compound	<chem>NNC(=O)CNC(=O)/C=N#N</chem>	1	3.62	206.528	8.69	5.573
Aliphatic heteromonocyclic compound	<chem>N1C(=O)/C(=N#N)/C=NC1=O</chem>	1	3.037	272.982	7.56	4.735
Aliphatic acyclic compound	<chem>NC(=O)CNC(=O)/C=N#N</chem>	1	3.596	193.434	7.983	5.073
Aromatic homomonocyclic compound	<chem>CCCCN(CC(C1=C/C(=N#N)/C(=O)C=C1)O)N</chem>	1	2.697	463.595	14.251	10.69
Aliphatic acyclic compound	<chem>NC(C(=O)O)COC(=O)/C=N#N</chem>	1	3.755	232.848	9.56	6.006
Aromatic homomonocyclic compound	<chem>CCN(CC(C1=CC(=O)/C(=N#N)/C=C1)O)N=O</chem>	1	2.755	434.832	12.836	9.276
Aliphatic heteropolycyclic compound	<chem>O=C1OCC2(C(=C1)C(=CC1C2CCC2(C1CCC2</chem>	0	1.79	828.784	20.483	16.841
Aromatic heteromonocyclic compound	<chem>Nc1nc(N)nc(n1)N</chem>	0	3.166	169.065	6.853	4.574

Рис. 2.4. Фрагмент бази даних ксенобіотиків з дескрипторами RDkit (значення для 191 молекулярного дескриптора не показані)

Аналіз опублікованих результатів [56,149,174,175], в яких науковці при створенні QSAR моделей використовують в якості 2D дескрипторів, окремо, відбитки молекулярної структури стали визначальними щодо планування подібних досліджень, але з розробленою, при виконанні дисертаційного дослідження, методикою. В цьому контексті, в роботі доцільно висвітлити питання щодо особливостей запису інформації про 2D молекулярну структуру хімічних сполук у вигляді бітового рядку. Крім того, точність розроблених *in silico* моделей прогнозування мутагенності Еймса, з урахуванням структурних класів ксенобіотиків, може залежати також і від того, які види відбитків структур використовувались при моделюванні. Відповідно до вищезазначеного, необхідно приділити увагу також питанням, що присвячені класифікації відбитків просторових структур та способам їх розрахунків.

### 2.2.2 Відбитки молекулярної структури та їх класифікація

Дослідження генетичних наслідків впливу хімічних мутагенів на геном людської популяції повинен враховувати максимальну кількість ксенобіотиків довкілля. При цьому, ефективність передбачення мутагенності Ames/QSAR моделей, у випадку використання 2D дескрипторів, може бути визначена певним типом молекулярного відбитка структури, який використовується при моделюванні та дозволяє у цифровому вигляді зберігати інформацію про молекулярну структуру досліджуваної хімічної сполуки. Серед 2D молекулярних дескрипторів молекулярні відбитки просторової структури формують найбільшу групу двовимірних предикторів, що можуть використовуватись для вирішення різноманітних задач біоінформатики та токсикології. Порівняння молекулярних відбитків, з урахуванням відстані між ними, є фундаментом для проведення віртуального скрінінгу (на основі структури лігандів), що здійснюють з метою пошуку потенційних лікарських препаратів [160,176]. Крім того, використання відбитків структури в якості предикторів для QSAR моделей дозволяє прогнозувати мутагенні та канцерогенні ефекти впливу факторів навколишнього середовища на геном людини [56,174].

В науковій праці [176] дослідники розглядають класифікацію молекулярних відбитків структури, що враховує їх поділ на три основних групи, що відповідають класам. До першої групи відносяться, так звані, субструктурні (від англ. «substructure fingerprints») дескриптори, що представляють собою бітовий рядок певної фіксованої довжини. В науковій літературі дану групу відбитків молекулярної структури також називають структурними ключами. Найбільш широко використовуваними в наукових дослідженнях є відбитки структури MACCS, що можуть мати різну довжину бітового вектора. З урахуванням сформульованих задач дисертаційного дослідження, відповідно до переліку хімічних сполук бази даних ксенобіотиків, нами були отримані розрахунки MACCS 166 (де 166 відповідає кількості біт, необхідних для збереження інформації про молекулярну структуру хімічної сполуки), які використовувались в якості предикторів для розроблених Ames/QSAR моделей. На рисунку 2.5

наведений приклад гіпотетичного 10-бітного молекулярного відбитка структури тирозина, кожний біт якого відповідає за наявність/відсутність певних підструктур або функціональних груп. Біт, що дорівнює 1, відповідає за наявність на рівні молекули певної підструктури. Записане значення, що дорівнює 0, виключає присутність певної функціональної групи.

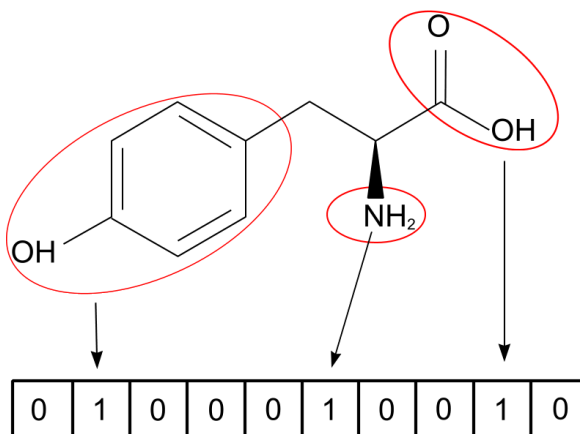


Рис. 2.5. Приклад гіпотетичного 10-бітного субструктурного відбитка MACCS для молекули тирозину (еліпсом червоного кольору позначені підструктури, що наявні в молекулі). Адаптовано відповідно до [176]

До першої групи відбитків також відносяться дескриптори PubChem, що відповідають структурним ключам та досить часто використовуються дослідниками в якості предикторів на етапі розробки QSAR моделей для прогнозування мутагенності Еймса [55,177].

Друга група дескрипторів [176], що відповідає топологічним відбиткам структури (від англ. «topological fingerprint»), мають певні особливості щодо збереження інформації про структуру хімічних сполук. У порівнянні з субструктурними ключами, дескриптори даної групи можуть бути отримані без урахування попередньо визначеної бібліотеки фрагментів, що формують молекулу. Алгоритм розрахунку топологічних відбитків базується на графовому представленні молекули та враховує всі варіанти можливих фрагментів певної довжини, починаючи відлік від кожного атома в структурі молекулі. В результаті, формується набір шляхів (відповідає всім можливим фрагментам), в яких атоми –

вершини графа, а зв'язки формують його ребра. Позиціонування відповідного фрагменту у бітовому рядку відбувається за допомогою хешування. Розрахунок хеш-функції лежить в основі роботи великої кількості алгоритмів, серед яких заслуговує на увагу BLAST (Basic Local Alignment Search Tool), який використовується для пошуку в базах даних схожих біологічних послідовностей та використовується для вирішення великої кількості різноманітних задач біоінформатики [178]. При цьому, максимальна ефективність пошуку досягається через розбиття всієї бази даних біологічних послідовностей та послідовностей-запиту на фрагменти фіксованої довжини (зазвичай 5-7 символів), для яких розраховується хеш-функція. Ідентифікація основної (початкової) ділянки вирівнювання для даного алгоритму здійснюється через збіг значень хеш-функцій. Оцінка ступеня подібності між хімічними сполуками-потенційними мутагенами також використовувалась в межах дисертаційного дослідження для побудови Ames/QSAR моделей на основі правил, фундаментом прогностичної здатності яких виступали ідентифіковані структурні маркери мутагенності.

Загальна концепція щодо збереження інформації про молекулярну структуру відповідно до топологічних відбитків представлена на рисунку 2.6. Починаючи з функціональної групи ОН, рухаючись справа наліво по графу (структурна формула тирозина) враховуємо всі фрагменти, які представлені на рівні даної молекули фіксованої довжини, що має прив'язку до кількості ковалентних зв'язків [179]. У зв'язку з тим, що подібну процедуру необхідно повторити для всіх атомів, реальний розрахунок топологічного відбитка структури буде отриманий з урахуванням значно більшої кількості фрагментів ніж показана на рис. 2.6. Формування гіпотетичного бітового рядку відбувається з урахуванням максимального віддалення від атомів (груп атомів), що для даного прикладу дорівнює п'яти ковалентним зв'язкам. Необхідно зазначити, що основний недолік топологічних відбитків молекулярної структури пов'язаний з виникненням колізій, при яких різні вхідні дані (за якісним складом фрагментів, що представлені на рівні молекули) дають однакові значення хеш-функції.

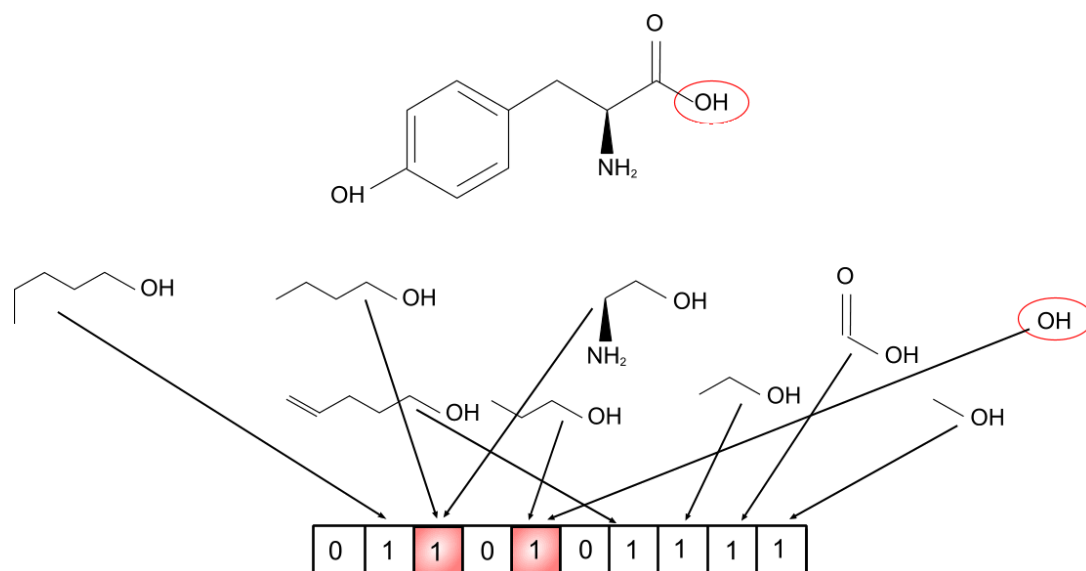


Рис. 2.6. Особливості формування бітового рядку гіпотетичного топологічного відбитка молекулярної структури довжиною 10 біт для молекули тирозину.

Адаптовано відповідно до [176]

В такій ситуації, один і той самий біт може відповідати різним фрагментам (на рисунку 2.6 позначено червоним кольором). Класичними представниками даної групи 2D дескрипторів є RDKit [170,179] та Daylight [180] відбитки молекулярної структури. Оскільки розрахунок RDKit (з довжиною бітового рядка 2048 біт) може бути здійснений відповідно до однойменної бібліотеки, мови програмування Python, що відноситься до відкритого програмного забезпечення, нами було прийнято рішення використовувати відповідні відбитки молекулярної структури в якості предикторів для розроблених в роботі QSAR моделей прогнозування мутагенності. У свою чергу програмне забезпечення, що використовується для розрахунку дескрипторів Daylight розповсюджується на платній основі.

Циркулярні (від англ. «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. При цьому, довжина ковалентного зв'язку виступає в ролі одиниці виміру діаметра молекулярного оточення, який необхідно враховувати для розрахунків циркулярних відбитків [176,179]. Особливість отримання розрахунків таких дескрипторів накладає обмеження щодо їх використання. Вони

не можуть бути застосовані для ідентифікації певних (схожих) фрагментів на рівні декількох досліджуваних молекул, тому що один і той самий атом, з урахуванням алгоритмів розрахунків відбитків структури, може мати різне молекулярне оточення. Серед відбитків просторової структури, що відносяться до третьої групи представленої авторами статті [176] класифікації, найбільш популярними є ECFP (Extended-Connectivity Fingerprints) та FCFP (Feature-Class Fingerprint). Відбитки розширеної зв'язності ECFP дозволяють зберігати інформацію про структуру молекули з урахуванням певного визначеного значення діаметра топологічного оточення атомів відповідно до графового представлення молекули [181]. При цьому важливою особливістю ECFP дескрипторів є те, що їх розрахунок спирається тільки на хімічні властивості атомів, до яких відносяться заряд, кількість зв'язків, атомний номер тощо. У свою чергу відбитки функціональних класів FCFP враховують певні класи ознак (наприклад, ароматичність, наявність донорів/акцепторів водневого зв'язку, кислотні або основні властивості, гідрофільність, гідрофобність тощо), що можуть бути властивими для отриманих фрагментів [181]. З метою реалізації Ames/QSAR моделей, в якості предикторів нами було запропоновано використовувати дескриптори FCFP. Такий вибір був пов'язаний, в першу чергу, з особливостями побудови таких відбитків, які суттєво відрізняються від інших тим, що при формуванні бітового рядка для розрахунків використовується сукупність хімічних властивостей фрагментів, а не тільки хімічні властивості атомів. Відповідно до вищезазначеного, важливою, з наукової точки зору, може бути відповідь на питання про ефективність використання таких відбитків в якості вхідних даних для *in silico* моделей прогнозування мутагенності Еймса факторів навколишнього середовища, з урахуванням розробленої в нашій роботі методики.

Точність *in silico* Ames/QSAR моделей може бути зумовлена також вибором методів, що лежать в основі вирішення задачі бінарної класифікації з розподілом ксенобіотиків на два класи: мутаген/не мутаген. Тому в наступному підрозділі ми приділили увагу базовим методам, що лежать в основі побудови *in silico* моделей оцінки мутагенності Еймса факторів навколишнього середовища.

## 2.3 Алгоритми машинного навчання для побудови AMES/QSAR моделей

Аналіз наукових праць [149,151,155,182] дозволив акцентувати увагу на методах, застосування яких, при реалізації Ames/QSAR моделей, давали не поганий результат з точки зору оцінки точності прогнозування мутагенності Еймса. Для вирішення поставлених завдань дисертаційного дослідження, з урахуванням нещодавно опублікованих результатів *in silico* Ames/QSAR моделювання, нами було обрано наступні методи: логістична регресія, метод випадкового лісу, метод градієнтного бустінгу та нейронна мережа. Вищезазначені методи відносяться до методів машинного навчання з вчителем та використовувались нами для вирішення задачі бінарної класифікації у контексті оцінки мутагенності Еймса.

### 2.3.1 Логістична регресія

Логістична регресія відноситься до одного з класичних статистичних класифікаторів з вчителем, що використовується для вирішення задач бінарної та багатокласової класифікації. Відповідно до набору вхідних даних, що представлені молекулярними дескрипторами логістична регресія дозволяє обчислити імовірність приналежності ксенобіотиків до одного з двох класів (мутаген/не мутаген). Для вирішення задачі, що пов'язана з оцінкою мутагенності Еймса факторів навколишнього середовища, використовувалась біноміальна логістична регресія, яка застосовує логістичну функцію для перетворення незалежних вхідних даних (предикторів) у значення імовірності [183]. Таке перетворення здійснюється відповідно до сигмоподібної функції, що має вигляд  $P = \frac{1}{1+e^{-y}}$ , де  $y$  – стандартне рівняння лінійної регресії  $y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$ ,  $x$  – набір незалежних змінних, що відповідають розрахованим молекулярним дескрипторам [183]. Невідомі значення коефіцієнтів  $a$  та  $b$  рівняння регресії можуть бути визначені за допомогою методу максимальної правдоподібності. Трансформація отриманих значень імовірності, що повертає модель, у відповідний клас (мутаген/не мутаген) відбувається з урахуванням порогового значення, що дорівнює 0,5. В такому випадку цільова змінна буде

приймати тільки два значення: дорівнює одиниці (у випадку якщо імовірність  $P \geq 0,5$ ), що відповідає ксенобіотику з вираженими мутагенними властивостями; приймає нульове значення (при  $P < 0,5$ ), якщо для ксенобіотика не підтверджена *in silico* мутагенність Еймса.

### 2.3.2 Метод випадкового лісу

Метод випадкового лісу та метод градієнтного бустінга, що відносяться до ансамблевих методів машинного навчання, досить часто використовуються для вирішення задачі бінарної класифікації в контексті отримання *in silico* оцінки мутагенності Еймса факторів навколишнього середовища. Специфіка ансамблевого навчання полягає у генерації класифікаторів, кожний з яких навчається окремо [184]. При цьому підвищення точності кінцевої моделі досягається шляхом об'єднання результатів окремих класифікаторів.

Метод випадкового лісу (RF) використовує сукупність незалежних дерев рішень [185], кожні з яких навчаються на випадковій підмножині незалежних вхідних даних (набір розрахованих 1D та 2D молекулярних дескрипторів) та будуються паралельно. Такий підхід відповідає концепції бутстреп-агрегування (в літературі можна зустріти іншу назву – бегінг) який дозволяє створити бінарний класифікатор з високими показниками точності прогнозування мутагенності Еймса та мінімізувати негативні наслідки, що можуть бути пов'язані з перенавчанням моделей. На рис. 2.7 представлена схема формування ансамблю дерев рішень, що лежить в основі метода випадкового лісу. Дерево рішень – є базовою одиницею методу випадкового лісу, що представляє собою набір правил (відповідають шляхам, що ведуть від кореня до листя ациклічного графа), які дозволяють визначити категоріальну змінну (мутаген/не мутаген), відповідно до набору незалежних змінних (дескрипторів). Внутрішні вузли в такому дереві виступають в ролі певного критерія, що лежить в основі перевірки ознак, які використовуються для поділу даних. На рис. 2.7 червоним кольором позначено шляхи від кореня до відповідних листків, які представляють собою набір правил, що дозволяють визначити категоріальну змінну. Розподіл хімічних сполук на два класи – мутаген/не мутаген здійснюється шляхом підрахунку голосів вкладу



кожного дерева рішень (рис 2.7). При вирішенні задачі бінарної класифікації категоріальна змінна, яка в ансамблях випадкових дерев зустрічається частіше буде відповідати прогнозованому класу

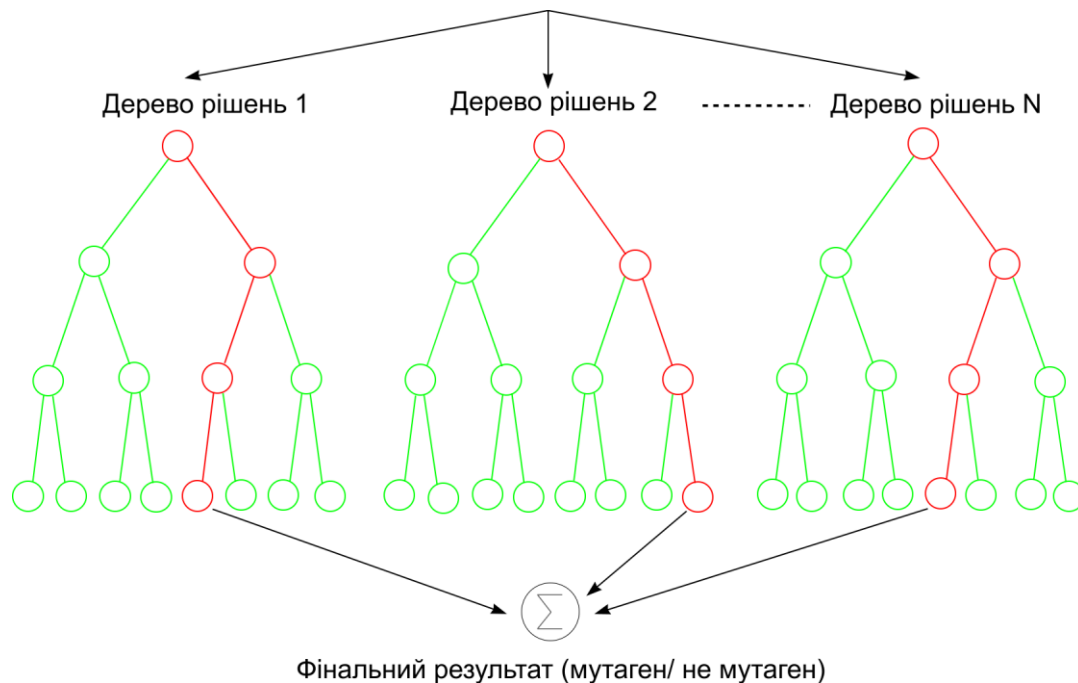


Рис.2.7. Схема роботи бінарного класифікатора на основі випадкового лісу, адаптовано відповідно до [184]

Досить важливою у науковому відношенні є ефективне вирішення задачі класифікації для хімічних сполук-потенційних мутагенів за допомогою прогностичних моделей, з урахуванням обмеженої кількості молекулярних дескрипторів. В межах проведеного нами дослідження метод випадкового лісу використовувався для відбору переліку молекулярних дескрипторів, що мають вагомий вплив на прогнозовану змінну. Відбір ознак (молекулярних дескрипторів) здійснювався відповідно до двох методів з Random Forest: mean decrease impurity та permutation feature importance. Оптимізація моделей, що здійснюється через зменшення набору вхідних даних, лежить в основі вдосконалення Ames/QSAR моделей, дозволяє досягти їх кращої узагальненості та стійкості до перенавчання [186]. Крім того, необхідними в науковому відношенні є дослідження проявів мутагенних ефектів, з урахуванням тільки тих дескрипторів, що мають вагомий

вплив на прогнозовану змінну. В такому випадку оцінка мутагенності Еймса ксенобіотиків може бути дана з урахуванням певного обмеженого набору властивостей (фізико-хімічних, структурних, електронних тощо), що задаються відповідними молекулярними дескрипторами.

### 2.3.3 Метод екстремального градієнтного бустінгу

Серед ансамблевих методів машинного навчання, метод XGBoost, що є модифікованим методом градієнтного бустинга (GBoost), заслуговує на особливу увагу. Такий інтерес пов'язаний з достатньо великою кількістю опублікованих наукових праць [56,149,175,187,188], в яких розроблені QSAR моделі на основі метода XGBoost давали одну з найкращих точностей прогнозування генотоксичних ефектів, у порівнянні з іншими бінарними класифікаторами.

В основі метода XGBoost використовуються дерева рішень, що будуються послідовно. Відповідно, кожна нова побудована модель (дерево рішень) враховує помилки, що були допущені на попередньому етапі. Така концепція побудови прогностичних моделей відповідає принципам бустінгу [189]. Ефективність моделей машинного навчання, що реалізовані відповідно до градієнтного бустинга досягається за рахунок мінімізації значень функції втрат (наприклад середньоквадратичної помилки, крос-ентропії), що розраховуються за допомогою градієнтного спуску, що лежить в основі навчання моделей [190]. Така процедура здійснюється через ітеративне оновлення внутрішніх параметрів моделей (наприклад структури дерева, значення в листках дерев), що направлено на отримання мінімальних значень функції втрат (що відповідає напрямку протилежному градієнту функції втрат). Основна перевага XGBoost у порівнянні з класичним градієнтним бустингом полягає у використанні регуляризації, яка дозволяє завдяки спрощенню моделей запобігти їх перенавчанню та покращити узагальнюючу здатність. Крім того, XGBoost, у порівнянні з GBoost, має перевагу у швидкості.

### 2.3.4 Глибинні нейронні мережі

Серед великої кількості методів, що лежать в основі розроблених сучасних *in silico* Ames/QSAR моделей, нейронні мережі дозволяють розподілити

ксенобіотики на два класи (мутаген/не мутаген) з достатньо високими показниками точності. Впевненості, щодо вибору даного метода в якості базового при виконанні дисертаційного дослідження додають опубліковані в наукових працях [191,192] результати моделювання. У статті [191], серед чотирьох розроблених Ames/QSAR моделей (на основі метода опорних векторів, k-найближчих сусідів та випадкового лісу) бінарний класифікатор на основі глибинної нейронної мережі дозволив отримати найкращу точність. Опублікована наукова праця [192] дозволяє зробити оцінку високому рівню ефективності застосування нейромережового підходу для вирішення задачі *in silico* оцінки мутагенності Еймса з урахуванням різних штамів *Salmonella typhimurium*.

Штучні нейронні мережі представляють собою набір алгоритмів, що лежать в основі симуляції роботи біологічних нейронів. Штучний нейрон, що є функціональною одиницею нейронної мережі, виступає в ролі математичної моделі, яка реалізує функцію суматора, що дозволяє отримати зважену суму вхідних даних з урахуванням вагових коефіцієнтів. Для перетворення та передачі вхідної інформації на наступний нейрон такі моделі можуть використовувати функцію активації як лінійного та і не лінійного типів [193].

В структурі типової нейронної мережі, зазвичай, представлено три шари нейронів: вхідний шар – на який поступає інформація відповідно до набору вхідних даних; прихований шар, який здійснює обчислення зваженої суми вхідних даних, що надходять на штучні нейрони та, після цього, за допомогою функції активації (не лінійного типу) передає отримане вихідне значення на вхід наступного нейрона; вихідний шар, що за допомогою сигмоподібної функції активації (стор. 68) повертає імовірність приналежності ксенобіотиків до одного з двох (мутаген/не мутаген) класів [194]. Кожен шар моделей штучних нейронів зв'язаний з усіма нейронами наступного шару (рис. 2.8). Необхідно зазначити, що навчання нейронної мережі відбувається за рахунок змін, у вузькому діапазоні, вагових коефіцієнтів та значень зсуву для нейронів прихованих шарів, що відносяться до внутрішніх параметрів моделі. Така процедура здійснюється за допомогою алгоритмів оптимізації, серед яких найбільш популярним є Adam

[195]. Крім того, точність класифікації може бути зумовлена налаштуваннями гіперпараметрів (наприклад, кількість шарів, кількість нейронів кожного шару, тип функції активації, кількість епох навчання, вибір оптимізатора), які встановлюються на початку проведення процедури тренування Ames/QSAR моделей та мають суттєвий вплив на перебіг цього процесу.

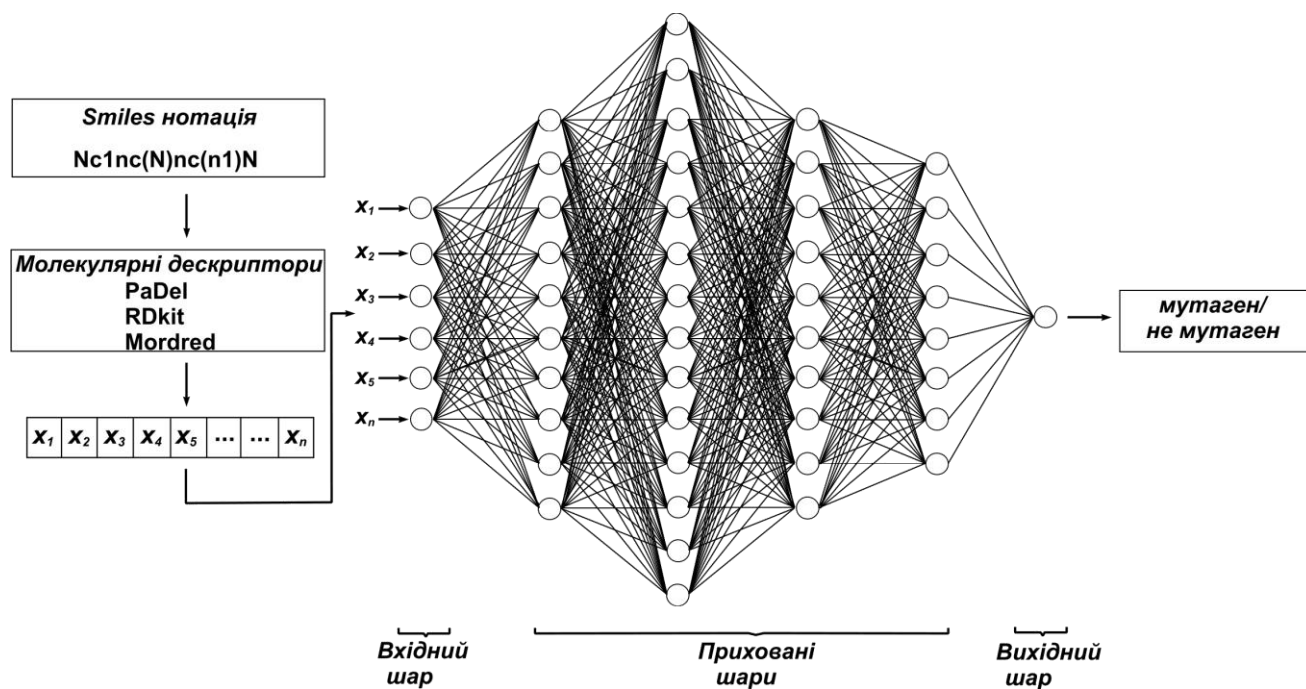


Рис. 2.8. Загальна схема реалізації Ames/QSAR моделей на основі глибинної нейронної мережі

Такий підхід дозволяє вирішити задачу бінарної класифікації з мінімальною кількістю хибнонегативних та хибнопозитивних прогнозів.

Створення ефективних бінарних класифікаторів вимагає наявності бази даних хімічних сполук, для яких, експериментально, за допомогою тесту Еймса отримано інформацію про мутагенність. Кількість експериментальних даних та, особливо, їх якість може мати суттєвий вплив на точність реалізованих Ames/QSAR моделей. В такій ситуації на етапах тренування, валідації та тестування Ames/QSAR моделей необхідно використовувати перевірені набори даних, використання яких, в якості базових, підтверджуються науковими працями в авторитетних рецензованих наукових виданнях.

Розглянемо основні етапи побудови Ames/QSAR моделей (рис.2.8) на основі глибинної нейронної мережі. Відповідно до Smiles-нотації для кожного

ксенобіотика розраховуються 1D та 2D молекулярні дескриптори. На наступному етапі, на вхідний шар нейронної мережі подається вектор ознак, що відповідає молекулярним дескрипторам. Тренування Ames/QSAR моделі здійснюється відповідно до вхідних даних – молекулярних дескрипторів, що є базовими предикторами для розроблених бінарних класифікаторів. В основі навчання лежить мінімізація функції втрат, що дозволяє контролювати та звести до мінімуму відхилення між фактичними (істинними) та прогнозованими значеннями моделі. В цьому процесі приймають участь нейрони прихованих шарів, що через оновлення значень вагових коефіцієнтів за допомогою алгоритмів оптимізації (наприклад Adam) дозволяють отримати прогноз категоріальної залежної змінної (мутаген/не мутаген) з високими показниками точності. Підходи до етапності реалізації бінарних класифікаторів на основі нейронної мережі є спільними для розроблених, в межах роботи, Ames/QSAR моделей. Різниця між ними полягає, в першу чергу, у застосуванні методів, які через пошук закономірностей у вхідних даних дозволяють вирішити задачу бінарної класифікації. Крім того, при проведенні дослідження нами були реалізовані Ames/QSAR моделі з різними наборами молекулярних дескрипторів, що були орієнтовані на основні структурні класи ксенобіотиків.

У межах проведення досліджень, були реалізовані Ames/QSAR моделі, що можуть мати різну прогностичну здатність. У цьому контексті заслуговує на увагу вивчення питань, які пов'язані з загальною характеристикою, розрахунками та особливостями застосування певних метрик, які дозволяють оцінити ефективність розроблених бінарних класифікаторів.

## 2.4 Метрики оцінки ефективності *in silico* Ames/QSAR моделей

Ефективність розроблених прогностичних *in silico* Ames/QSAR моделей оцінювалась за допомогою наступних метрик: загальної точності (*accuracy*), точності позитивного прогнозу (*precision*), чутливості (*recall*), специфічності (*specificity*) та F1-міри ( $F_1$  – *score*), які були отримані з урахуванням матриць

помилки (confusion matrix) відповідно до співвідношень 1-5, де  $TP, TN, FP, FN$  – відповідає кількості істинно позитивних, істинно негативних, хибнопозитивних та хибнонегативних результатів класифікації відповідно.

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

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

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

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

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

Подібні підходи щодо оцінки ефективності QSAR моделей для оцінки мутагенних, а також токсичних ефектів впливу факторів навколишнього середовища прослідковуються у наукових працях [175,196,197,198]. На етапі відбору найкращої Ames/QSAR моделі також використовувалась популярна та достатньо ефективна метрика, що ґрунтується на розрахунку площі під кривою робочої характеристики приймача (ROC) [199]. Необхідно зазначити, що метрика точності розглядається з точки зору оцінки двох параметрів: загальної точності (*accuracy*), що відповідає частині ксенобіотиків, які були правильно розподілені між двома класами та метрики *precision*, що відповідає частині хімічних сполук, які прогностичними Ames/QSAR моделями були правильно віднесені до класу хімічних сполук з вираженими мутагенними властивостями. Метрика *recall* дозволяє визначити частину хімічних сполук, що є мутагенами з урахуванням сумарної кількості істинно позитивних та хибнонегативних результатів. Критерій оцінки ефективності Ames/QSAR моделей *specificity*, подібний до *recall*, але розраховується для визначення частини негативних щодо проявів мутагенного потенціалу ксенобіотиків з урахуванням сумарної кількості істинно негативних та хибнопозитивних результатів класифікації. F1-міра, як критерій оцінки

ефективності розроблених моделей машинного навчання представляє собою гармонічне середнє двох метрик – *precision* та *recall*.

## Висновки до розділу 2

Побудова ефективних Ames/QSAR моделей вимагає наявності бази даних хімічних сполук, в якій для кожного ксенобіотика, експериментально, за допомогою тесту Еймса, отримана інформація про мутагенний потенціал. Такі дані повинні бути отримані з надійних джерел та містити мінімальну кількість хибнонегативних та хибнопозитивних результатів. Точність розроблених моделей прогнозування мутагенності Еймса може залежати від вибору метода, що лежить в основі розподілу ксенобіотиків на два класи (мутаген/не мутаген). Використання різних наборів молекулярних дескрипторів також мають суттєвий вплив на ефективність розроблених бінарних класифікаторів. В цьому контексті заслуговують на увагу різні типи відбитків молекулярної структури, які за допомогою бітового рядка дозволяють записати інформацію про структуру хімічної сполуки. Позитивний результат, з точки зору підвищення прогностичної здатності Ames/QSAR моделей, може бути досягнений через розподіл ксенобіотиків на однорідні групи, що відповідають дев'яти структурним класам ксенобіотиків, для яких прогнозування мутагенності здійснюється з урахуванням обмеженого набору вхідних даних, що задається молекулярними дескрипторами.

## РОЗДІЛ 3 РОЗРОБКА ТА ОПТИМІЗАЦІЯ *IN SILICO* МОДЕЛЕЙ ПРОГНОЗУВАННЯ МУТАГЕННОСТІ НА ОСНОВІ РЕЗУЛЬТАТІВ ТЕСТУ ЕЙМСА

В даному розділі розглядаються питання оптимізації моделей прогнозування мутагенності Еймса, що дозволяє покращити точність, узагальнюючу здатність та стійкість моделей до перенавчання. Наведена інформація щодо особливостей побудови Ames/QSAR моделей на основі різних наборів (PaDel, RDkit, Mordred) одновимірних та двовимірних молекулярних дескрипторів. Представлена методологія покращення прогностичної здатності Ames/QSAR моделей, що реалізується через розподіл ксенобіотиків на окремі групи, які відповідають структурним класам. Приділено увагу питанням отримання переліку молекулярних дескрипторів, з урахуванням структурних класів ксенобіотиків, що мають вагомий вплив на прогнозовану змінну. Висвітлені питання оптимізації Ames/QSAR моделей, що є основою для пошуку причинно-наслідкових зв'язків між мутагенністю та фізико-хімічними, електронними, просторовими властивостями ксенобіотиків, що задаються різними наборами молекулярних дескрипторів.

### 3.1 Класичні підходи до реалізації Ames/QSAR моделей та шляхи щодо їх оптимізації

При створенні *in silico* моделей машинного навчання на початковому етапі дослідження нами використовувався датасет [151], що був отриманий шляхом об'єднання трьох загальнодоступних наборів даних: Kazius-Bursi [152], Hansen [153] та EFSA [154]. Будь-яка хімічна сполука з представленого набору вважалася мутагенно-активною, якщо при проведенні *in vitro* тесту Еймса на штаммах *Salmonella typhimurium* TA97, TA98, TA100, TA102, TA1535, TA1537 і TA1538 був отриманий хоча б один позитивний результат. Для прогностичних Ames/QSAR моделей нами використовувався набір з 1444 дескрипторів PaDel для



кожної хімічної речовини-забруднювача довкілля, що були розраховані відповідно до лінійної SMILES нотації, за допомогою веб-сервіса Galaxy [164]. Необхідно відмітити, що на даному етапі проведення дослідження нами були розраховані 1D та 2D дескриптори PaDel без урахування відбитків молекулярної структури.

Отримані значення молекулярних дескрипторів не можуть бути використані в моделях машинного навчання в якості предикторів без їх попередньої обробки. Підготовка даних для Ames/QSAR моделей машинного навчання була пов'язана з проведенням, спочатку, з – нормалізації, що дозволила видалити аномальні значення предикторів. На наступному етапі було проведено нормування даних, з урахуванням приведення значень предикторів до стандартного діапазону [0,1]. Крім того, було здійснено видалення стовпчика «Canonical SMILES», в якому містилася текстова інформація про структуру молекули, яка не використовувалась для моделювання. Також були видалені стовпчики, в яких були записані однакові значення молекулярних дескрипторів. Для уникнення мультиколінеарності та з метою зменшення розмірності даних нами також було проведено видалення корельованих ознак. У випадку коли дві або більше ознаки мали високу кореляцію (більше 0,95) відповідно до коефіцієнту кореляції Пірсона, одна з них залишалася, а всі інші видалялись. При моделюванні вхідні дані було розподілено на тренувальний та тестовий набір у співвідношенні 75:25 відповідно.

Для вирішення задачі бінарної класифікації нами були запропоновані чотири моделі машинного навчання логістична регресія (LR-Scikit), логістична регресія на основі стахостичного градієнтного спуску (LR-SGD), метод випадкового лісу (Random Forest) та нейронна мережа [200].

Для логістичної регресії (бібліотека Scikit-learn Python) більшість параметрів налаштування моделі використовувалися за замовчуванням. Проблема незбалансованості представлення двох класів ксенобіотиків (мутаген/не мутаген), що може впливати на якість класифікації менш представленого класу, нами було вирішено через встановлення `class_weight='balanced'`. При цьому незбалансованість двох класів (мутаген/не мутаген) компенсувалась ваговими

коефіцієнтами, що дозволило Ames/QSAR моделям враховувати менш представлений клас.

Логістична регресія (LR-SGD) використовує метод стохастичного градієнтного спуску (SGD) для оптимізації параметрів Ames/QSAR моделі. Максимальна ефективність SGD досягалась через оновлення параметрів моделі, що відбувалась для кожного окремого зразка даних або невеликій групі зразків. У розробленій нами Ames/QSAR моделі машинного навчання на основі LR-SGD, з метою оновлення параметрів моделі використовували 64 зразки з навчального набору даних. Модель логістичної регресії була реалізована через архітектуру, що базується на класі *Sequential* бібліотеки *TensorFlow*, що дозволяє створювати моделі, які складаються з послідовності шарів. Для вирішення задачі бінарної класифікації була обрана проста архітектура, яка включала лише один шар (*Dense*) штучних неронів з одним виходом. Після проходження через шар *Dense*, на основі вхідних даних, відбувалось обчислення ваг і зміщень. Модель отримала вхідні дані у вигляді матриці, де кожен рядок представляв спостереження, а кожен стовпчик – ознаку. Функцією активації був обраний *sigmoid*, що перетворював лінійну комбінацію ознак у ймовірність того, що певна хімічна сполука належить до одного з двох класів – мутаген чи не мутаген. Процес ефективного навчання моделі передбачав мінімізацію функції втрат, що дозволило мінімізувати різницю між передбаченим значенням і фактичним результатом оцінки мутагенного потенціалу певного ксенобіотика. Для вирішення задачі, що пов'язана з отриманням оцінки мутагенного потенціалу було запропоновано використовувати функцію втрат *binary\_crossentropy*.

Для реалізації Ames/QSAR моделі випадкового лісу ми використовували клас *RandomForestClassifier* бібліотеки *Scikit-learn* Python. Модель випадкового лісу була представлена 200 деревами, ( $n\_estimators=200$ ), з максимальною кількістю листків, що дорівнювала 600 ( $max\_leaf\_nodes=600$ ).

В структурі нейронної мережі представлений вхідний шар, вихідний шар та 4 приховані шари, які містять 128, 256, 128 та 64 нейрони відповідно (рис. 2.7). В якості функції активації для прихованих шарів була обрана функція *ReLU*, що

мала переваги з точки зору відносної простоти реалізації та дозволяла ефективно вирішити проблему зникаючого градієнту, що негативно впливає на процес навчання глибоких нейронних мереж. На вихідному шарі функцією активації є *Sigmoid* (сигмоподібна функція), що давала змогу отримати значення ймовірності щодо приналежності ксенобіотиків до одного з двох класів (мутаген/не мутаген).

З метою вирішення стандартної для нейронних мереж проблеми, що пов'язана з перенавчанням були використані методи *L1* і *L2* та *Dropout* регуляризації. У створеній нейронній мережі регуляризація *L1* використовувалась на другому прихованому шарі, а на третьому прихованому шарі – регуляризація *L2*. Між всіма шарами мережі наявний *Dropout*, який випадковим чином вимикав 30-60% нейронів під час навчання, що забезпечувало більшу стійкість та покращувала прогностичну здатність моделей на нових вхідних даних. З метою мінімізації функції втрат нами був обраний найбільш ефективний оптимізатор на основі адаптивної оцінки моменту Adam [195]. Навчання нейронної мережі було здійснено протягом 100 епох з розміром партії 64. Ефективність розроблених нами прогностичних *in silico* Ames/QSAR моделей оцінювали за допомогою метрик точності (*accuracy*), чутливості (*recall*), специфічності (*specificity*) та F1-міри ( $F_1$  – *score*).

В таблиці 1 представлені результати класифікації, що були отримані на тестовій вибірці за допомогою логістичної регресії (LR-Scikit), логістичної регресії на основі стахостичного градієнтного спуску (LR-SGD), методу випадкового лісу та нейронної мережі.

Таблиця 3.1

### Результати класифікації для тестової вибірки

	LR-Scikit	LR-SGD	Random Forest	Нейронна мережа
<i>accuracy</i>	0,79	0,79	0,86	0,83
<i>recall</i>	0,81	0,83	0,87	0,82
<i>specificity</i>	0,76	0,76	0,84	0,83
$F_1$ – <i>score</i>	0,80	0,80	0,86	0,83

Відповідно до отриманих класифікаційних звітів, з урахуванням метрик точності (*accuracy*), чутливості (*recall*), специфічності (*specificity*) та F1-міри ( $F_1$  – *score*), серед чотирьох прогностичних Ames/QSAR найкращою, з урахуванням всіх метрик, виявилась Ames/QSAR-метод на основі метода Random Forest з AUC=0,92. Нейронна мережа продемонструвала меншу ефективність з показниками точності 0,83 та чутливості 0,87. Значення площі під кривою ROC = 0,9 для нейронної мережі вказує на цілком прийнятний результат класифікації. Не зважаючи на те, що значення показників класифікаційних звітів для лінійної регресії (LR-Scikit) та лінійної регресії з стохастичним градієнтним спуском (LR-SGD) майже не відрізнявся, аналіз матриць помилок дозволив віддати перевагу останньому методу. На тестовій вибірці модель LR-SGD у порівнянні з LR-Scikit дозволила ідентифікувати більшу кількість істинно позитивних ксенобіотиків, що проявляють властивості генотоксичності.

Отримана оцінка ефективності розроблених Ames/QSAR моделей, відповідно до базових метрик, практично не відрізнялась від таких показників для бінарних класифікаторів, представлених в інших дослідженнях та опублікованих у наступних наукових працях [151,155,182]. Така ситуація стала стимулом до формування основного вектору проведення подальших досліджень, що, в першу чергу, пов'язаний з удосконаленням Ames/QSAR моделей. У цьому контексті, в роботі було запропоновано методику оптимізації моделей прогнозування мутагенності Еймса, яка дозволяє підвищити точність моделей за допомогою наступних підходів: зменшення обсягу вхідних даних, що досягається через видалення нерелевантних предикторів та відбір тих дескрипторів, що мають вагомий вплив на прогнозовану змінну; підбір гіперпараметрів Ames/QSAR моделей; вибір різних наборів молекулярних дескрипторів та різних типів відбитків молекулярної структури (*molecular fingerprint*); поділ бази даних ксенобіотиків на структурні класи. Важливим у науковому відношенні, є відповідь на питання про те, який набір властивостей (фізико-хімічні, електронні, просторові тощо) певного ксенобіотика, що задається молекулярними дескрипторами, є визначальним щодо прояву мутагенності.

### 3.1.1 Вплив зменшення набору вхідних даних на ефективність Ames/QSAR моделей

Оптимізація моделей прогнозування мутагенності Еймса, що досягалась через зменшення розмірності вхідних даних, була реалізована за допомогою сформованого ранжованого переліку молекулярних дескрипторів, який був отриманий відповідно до коефіцієнтів двох моделей регресії (LR-SGD і LR-Scikit) та двох методів з Random Forest: mean decrease impurity та permutation feature importance. В таблиці 3.2 представлений перелік молекулярних дескрипторів, які були визначені відповідно до чотирьох підходів, зустрічались в моделях по декілька раз та мали достатньо вагомий вплив на прогнозовану змінну.

Таблиця 3.2

#### Перелік молекулярних дескрипторів, що мали вагомий вплив на прогнозовану змінну

Random Forest (mean decrease impurity )		Random Forest ( permutation feature importance)	LR-Scikit	LR-SGD
1	<i>MATS1e</i>	<b>SHBint2</b>	<i>AATS2s</i>	<b>GATS1p</b>
2	<b>R_TpiPCTPC</b>	<i>MATS1e</i>	<b>GATS1p</b>	<i>nAtomP</i>
3	<i>nFRing</i>	<b>BCUTp-1h</b>	<b>R_TpiPCTPC</b>	<i>SpMax1_Bhm</i>
4	<i>nAtomP</i>	<i>ATSC2e</i>	<b>BCUTp-1h</b>	<b>R'_TpiPCTPC</b>
5	<i>MLFER_E</i>	<i>SpMin4_Bhm</i>	<i>nHBd</i>	<i>AATSC2i</i>
6	<i>SpMin1_Bhm</i>	<i>SpMin1_Bhm</i>	<i>AATSC0m</i>	<i>nFRing</i>
7	<b>GATS1p</b>	<i>GATS1m</i>	<i>AATSC2i</i>	<i>MATS2c</i>
8	<i>GATS1m</i>	<i>AATSC0m</i>	<i>MATS2c</i>	<i>AATS2s</i>
9	<i>ATSC2e</i>		<i>SpMax1_Bhm</i>	<i>nHBd</i>
10			<b>SHBint2</b>	<b>BCUTp-1h</b>
11			<i>ATS7s</i>	<i>ATS7s</i>
12			<i>MLFER_E</i>	<b>SHBint2</b>
13			<i>SpMin4_Bhm</i>	

Такі параметри моделі є важливими з точки зору оцінки результатів класифікації. Враховуючи велику кількість молекулярних дескрипторів, що використовувались в якості вхідних даних для розроблених класифікаторів, нами

було прийнято рішення щодо формування обмеженого переліку найважливіших дескрипторів (таблиця 3.2) з урахуванням граничного – тридцятого молекулярного дескриптора ранжованого переліку. Молекулярні дескриптори, які були представлені в одній з моделей тільки один раз не записувались у таблицю 3.2. В рамках даного дослідження нами була приділена увага до дескрипторів, які мали вагомий вплив на прогнозовану змінну категоріального типу – мутаген/не мутаген. Ранжований перелік дескрипторів також використовувався нами для вирішення задачі оптимізації моделей машинного навчання, що полягала у визначенні фіксованого базового набору ознак, що обирались з урахуванням переліку дескрипторів, які записані у порядку зменшення вагових коефіцієнтів. При цьому враховувались як дескриптори які повторюються так і ті, що були представлені в моделях один раз. У таблиці 3.2 жирним шрифтом позначені мнемоніки молекулярних дескрипторів, які повторюються в моделях 3-4 рази. Молекулярні дескриптори, що повторювались тільки в двох моделях позначені курсивом. Порядок запису молекулярних дескрипторів для кожного методу відповідав значенням вагових коефіцієнтів (по модулю), що зменшувався при збільшенні порядкового номеру рядка таблиці 3.2.

Досить важливою у науковому відношенні є ефективне вирішення задачі класифікації потенційних мутагенних хімічних сполук за допомогою прогностичних моделей, з урахуванням обмеженої кількості молекулярних дескрипторів. Такий підхід лежить в основі вдосконалення моделей, дозволяє досягти кращої узагальненості та стійкості моделей до перенавчання [201]. З метою оптимізації Ames/QSAR моделей було запропоновано використовувати в якості вхідних даних для кожного з чотирьох класифікаторів набір по 50, 100, 150, 200, 300 та 400 молекулярних дескрипторів, що обирались з кожної моделі. Відбір ознак здійснювався відповідно до ранжованого переліку молекулярних дескрипторів, який був отриманий з урахуванням коефіцієнтів двох моделей регресії (LR-SGD і LR-Scikit) та двох методів з Random Forest: mean decrease impurity та permutation feature importance. При цьому дескриптори, що повторювались, на етапі формування набору вхідних даних, були видалені.

Проведення такої процедури вплинуло на кінцеву кількість предикторів, що використовувались при моделюванні. Кількість ознак, що використовувалась на етапі навчання Ames/QSAR моделей без дублікатів стала 128 (було 200), 276 (було 400), 371 (було 600), 454 (було 800), 589 (було 1200) та 655 (було 1600). При цьому молекулярні дескриптори-дублікати зберігались в окремому файлі для подальшого тестування *in silico* моделей з обмеженим набором даних, що представлені тільки ознаками з найбільшими ваговими коефіцієнтами та мали суттєвий вплив на прогнозовану змінну.

У таблиці 3.3 наведено дані щодо оцінки точності результатів класифікації для чотирьох моделей машинного навчання, що були отримані на тестовій вибірці для фіксованої кількості молекулярних дескрипторів, що не враховує дескриптори-дублікати.

Таблиця 3.3

**Залежність точності (у відсотках) Ames/QSAR моделей від якісного та кількісного складу молекулярних дескрипторів**

Кількість дескрипторів (без повторів)	Точність LR Scikit-learn	Точність LR SGD	Точність RF	Точність NN
128	78	77	84	82
276	80	80	86	84
371	78	78	84	80
454	78	78	83	82
589	79	79	84	82
655	80	79	83	84
1444 (початкова кількість)	79	79	86	83

За результатами проведеного тестування можна помітити, що не має суттєвої кореляції між точністю моделей та якісним і кількісним складом молекулярних дескрипторів. Зменшення кількості молекулярних дескрипторів, зазвичай, призводило до зниження точності розроблених *in silico* Ames/QSAR моделей. Досить цікавим в науковому відношенні є той факт, що використання набору даних з 276 молекулярних дескрипторів призвело до покращення

прогностичної здатності логістичних регресій та нейронної мережі. При цьому точність моделі випадкового лісу не змінилась (рисунок 1.)

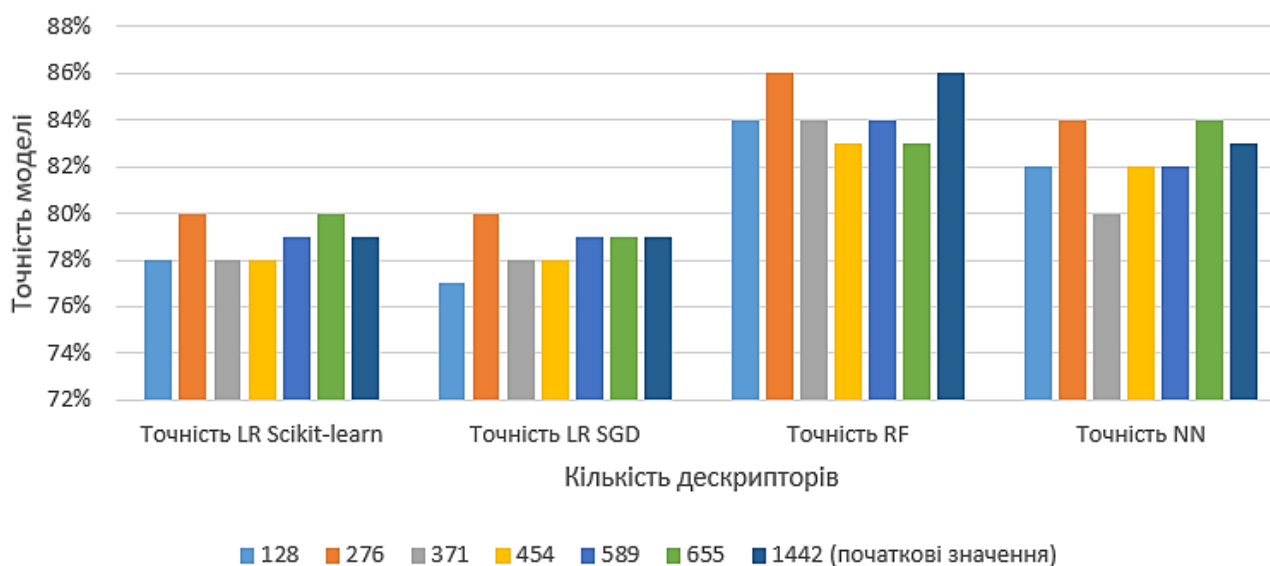


Рис. 3.1. Гістограма залежності точності Ames/QSAR моделей від кількісного складу молекулярних дескрипторів [200]

У таблиці 3.4 представлені результати класифікації, що були отримані на тестовій вибірці за допомогою логістичної регресії (LR-Scikit), логістичної регресії на основі стахостичного градієнтного спуску (LR-SGD), методу випадкового лісу та нейронної мережі з урахуванням обмеженого набору даних з 276 молекулярних дескрипторів.

Таблиця 3.4

**Результати класифікації для тестової вибірки (276 молекулярних дескрипторів)**

	LR-Scikit	LR-SGD	Random Forest	Нейронна мережа
<i>accuracy</i>	0,80	0,80	0,86	0,84
<i>recall</i>	0,80	0,81	0,84	0,85
<i>specificity</i>	0,79	0,80	0,88	0,83
<i>F<sub>1</sub> – score</i>	0,80	0,80	0,86	0,84



Відповідно до отриманого класифікаційного звіту можна відмітити підвищення точності логістичних регресій, яка збільшилась до 80% та нейронної мережі – до 84%. Такий результат є багатообіцяючим та дозволяє зробити висновки про те, що оптимізація Ames/QSAR моделей, яка реалізується через зменшення обсягу вхідних з відбором молекулярних дескрипторів, що мають вагомий вплив на прогнозовану змінну, може підвищити точність *in silico* моделей прогнозування мутагенності Еймса.

Досить важливими у науковому відношенні можуть бути отримані висновки щодо ефективності розроблених класифікаторів, що використовують в якості вхідних даних набір молекулярних дескрипторів-дублікатів, які зустрічались у чотирьох розроблених моделях найбільшу кількість раз та мають вагомий вплив на прогнозовану змінну. Результат класифікації для тестової вибірки з урахуванням 78 молекулярних дескрипторів представлений в таблиці 3.5.

Таблиця 3.5

**Результат класифікації для тестової вибірки з обмеженою кількістю  
молекулярних дескрипторів**

	LR-Scikit	LR-SGD	Random Forest	Нейронна мережа
<i>accuracy</i>	0,74	0,74	0,77	0,78
<i>recall</i>	0,73	0,72	0,80	0,84
<i>specificity</i>	0,75	0,76	0,85	0,72
<i>F<sub>1</sub> – score</i>	0,74	0,74	0,82	0,80

Відповідно до отриманих класифікаційних звітів, з урахуванням метрик точності (*accuracy*), чутливості (*recall*), специфічності (*specificity*) та F1-міри (*F<sub>1</sub> – score*) можемо побачити зниження ефективності розроблених Ames/QSAR моделей у порівнянні з результатами класифікації з повним набором вхідних даних (для 1444 молекулярних дескрипторів). Результат тестування моделей машинного навчання з обмеженим набором молекулярних дескрипторів має певні негативні наслідки з точки зору оцінки їх ефективності. З іншого боку, зменшення кількості молекулярних дескрипторів більш ніж на 95% від початкової кількості (з 1444 до 78) призвело до зниження точності для чотирьох прогностичних моделей

тільки від 3 до 6%. Такий результат може виступати в ролі стимулюючого для пошуку причинно-наслідкових зв'язків між мутагенністю та фізико-хімічними, просторовими, електронними характеристиками певного ксенобіотика, що задаються набором молекулярних дескрипторів-дублікатів, які мають значний вплив на прогнозовану змінну. Відповідно до отриманих результатів, необхідно додатково приділити увагу методам, що дозволяють здійснювати відбір релевантних предикторів та використовуються іншими науковцями для вирішення задач *in silico* оцінки мутагенних ефектів впливу факторів навколишнього середовища.

Наступний підрозділ дисертації присвячений висвітленню питань оптимізації Ames/QSAR моделей, що реалізується шляхом формування дев'яти (табл. 2.1) структурних класів, що об'єднані у 5 груп потенційних мутагенів, застосуванням різних наборів молекулярних дескрипторів та селекції найбільш релевантних дескрипторів.

### **3.2 Ames/QSAR моделі, орієнтовані на структурні класи ксенобіотиків**

На даному етапі досліджень розмір датасету, що використовувався для створення моделей прогнозування мутагенності Еймса, був збільшений (за рахунок мікотоксинів) до 8454 (підрозділ 2.1). З метою покращення прогностичної здатності Ames/QSAR моделей було запропоновано розподілити базу даних хімічних сполук на 5 груп (табл. 2.1), що відповідають основним структурним класам ксенобіотиків. Використання однорідних наборів вхідних даних дозволило отримати перелік найбільш релевантних дескрипторів, які необхідно, в першу чергу, використовувати для створення ефективних, орієнтованих на певний структурний клас *in silico* моделей прогнозування мутагенності. Такий підхід дозволяє отримати *in silico* оцінку мутагенності ксенобіотиків, що відносяться до певного структурного класу, з урахуванням набору властивостей, що характерні для цієї групи.

Точність розроблених бінарних класифікаторів значною мірою може залежати від набору дескрипторів, які використовуються при моделюванні [91,147]. Для порівняння ефективності Ames/QSAR моделей побудованих відповідно до різних наборів вхідних даних, було запропоновано використовувати три набори молекулярних дескрипторів: PaDel, Mordred та RDkit, які були розраховані нами відповідно до Smiles лінійної нотації для кожного ксенобіотика за допомогою веб-сервіса Galaxy [164]. Ефективність моделей прогнозування мутагенності Еймса також може бути детермінована вибором метода, який лежить в основі розподілу ксенобіотиків на дві групи (мутаген/не мутаген). У межах роботи, було обрано три основних метода: випадковий ліс (Random forest), метод градієнтного бустінга (Gradient boosting), та нейронна мережа, які, згідно з опублікованими результатами досліджень [149,151,155,182], дозволяють досягнути найкращої точності *in silico* прогнозування мутагенності Еймса.

Як ми вже відмічали, ефективність Ames/QSAR моделей на пряму залежить від попередньої обробки вхідних даних. Якщо цьому питанню не приділяти достатню увагу, точність розроблених моделей може бути достатньо низькою. Навчання бінарних класифікаторів повинно відбуватись тільки з урахуванням інформативних для моделей ознак, що представлені молекулярними дескрипторами. Не інформативні дані, що стосуються, наприклад, порядкових номерів ксенобіотиків, Smiles лінійної нотації та інформації щодо приналежності хімічних сполук до відповідного класу (з урахуванням особливостей будови їх молекулярного каркасу), були видалені. На етапі препроцесингу було також проведено видалення ознак, для яких відсутня варіативність. Після чого був сформований вектор  $y$ , елементи якого набувають одного з двох можливих значень (0 – сполука проявляє мутагенність за тестом Еймса та 1 – є не мутагеном). Матриця  $x$  зберігає інформацію про розраховані молекулярні дескриптори, що відповідають певному значенню цільової змінної.

Для вирішення проблеми мультиколінеарності використовувалась матриця корельованих ознак, що була отримана відповідно до розрахованих коефіцієнтів кореляції Пірсона між всіма парами молекулярних дескрипторів. Коефіцієнт

кореляції  $r > 0,95$  між парами ознак використовувався в якості основного критерію для проведення фільтрації вхідних даних. При великих значеннях коефіцієнтів кореляції між парами ознак, що наближались до одиниці, одна з них видалялась. Проведення такої процедури сприяло зменшенню розмірності простору вхідних даних, що створює передумови для підвищення стійкості розроблених Ames/QSAR моделей до перенавчання.

Наступний етап препроцесингу пов'язаний з масштабуванням, що був реалізований за допомогою двох інструментів `QuantileTransformer` та `StandardScaler` бібліотеки `scikit-learn` Python. Нормалізація проводилась з метою перетворення набору ознак, що задаються молекулярними дескрипторами у певний фіксований діапазон значень. `QuantileTransformer` з параметром `output_distribution='uniform'` дозволяє отримати рівномірний розподіл кожної ознаки в межах інтервалу  $[0,1]$ , що сприяє зменшенню негативного впливу аномальних значень (викидів). Стандартизація даних була реалізована за допомогою інструменту `StandardScaler`, що дозволила отримати набір ознак з середнім значенням, що дорівнює 0 та стандартним відхиленням 1.

При вирішенні задач бінарної та багатокласової класифікації досить часто виникає проблема незбалансованості класів, що має негативний вплив на ефективність побудованих моделей [202]. У контексті вирішення поставлених задач дисертаційного дослідження, проблему не рівномірного розподілу двох класів (мутаген/не мутаген) було вирішено за допомогою інструменту `SMOTE` (`Synthetic Minority Over-sampling Technique`) бібліотеки `Imbalanced-learn` Python, який дозволяє розширити менш представлений клас новими згенерованими зразками. Таку процедуру було проведено для повної бази даних дескрипторів та, окремо, для чотирьох груп (1,2,4 та 5) ксенобіотиків (табл. 2.1), що були розподілені відповідно до особливостей будови їх молекулярного каркасу. В результаті кількість зразків в різних групах, що відносяться до двох класів стала однаковою.

На наступному етапі вхідні дані було розділено на дві вибірки – навчальну та тестову, у співвідношенні 80:20. Такий підхід щодо розподілу даних на дві

групи ми використовували як для повної бази даних, так і для окремих груп (табл. 2.1) ксенобіотиків, що мають спільні риси будови молекулярного каркасу. Застосування такої методики, у першу чергу, направлено на отримання оцінки ефективності *in silico* моделей прогнозування мутагенності Еймса, що використовували на етапі навчання та валідації різні набори однорідних вхідних даних (відповідно до чотирьох груп ксенобіотиків) у порівнянні з моделями, що побудовані з урахуванням молекулярних дескрипторів що були розраховані для всіх 8454 хімічних сполук. Аналіз наукової роботи [203], в якій розглядаються питання валідації моделей машинного навчання, стала фундаментом для зміни стратегії щодо розподілу даних, які використовувались при моделюванні. Автори статті підкреслюють, що для об'єктивної оцінки ефективності моделей машинного навчання необхідно додатково сформувати альтернативний тестовий набір даних. Тому дані було розподілено випадковим чином на три вибірки: на тренувальну та дві тестові вибірки, у співвідношенні 80:10:10. Перша – тестова вибірка використовувалась для проміжної незалежної перевірки, а друга – для остаточної перевірки узагальнюючої здатності моделей та їх здатності здійснювати прогнози мутагенності Еймса на незалежних даних. Другу тестову вибірку будемо називати екзаменаційною. Слід зауважити, що в різних наукових публікаціях поняття «валідаційний набір» та «тестовий набір» використовуються як взаємозамінні. З метою запобігання плутанини у визначеннях надалі в роботі будемо використовувати термін «валідаційний набір», що позначає частину даних, які використовуються для оцінки ефективності Ames/QSAR моделей з урахуванням налаштувань гіперпараметрів. Термін «тестовий набір» використовували по відношенню до частини даних, які є необхідними для проведення тестування ефективності остаточної Ames/QSAR моделі з підібраними оптимальними значеннями гіперпараметрів.

Оптимізація розроблених нами Ames/QSAR моделей, що полягала у зменшенні розмірності вхідних даних, була реалізована відповідно до алгоритму RFECV (Recursive Feature Elimination with Cross-Validation), що є удосконаленим алгоритмом RFE [155], який поєднує в собі крос-валідацію з рекурсивним

видаленням найменш впливових ознак. Такий підхід дозволив на етапі розподілу вхідних даних не створювати окремо валідаційну вибірку, тому що її роль виконувла частина навчального набору даних у процесі крос-валідації. У роботі було запропоновано використовувати п'ятикратну перехресну перевірку, що дозволяє отримати об'єктивну оцінку ефективності розроблених Ames/QSAR моделей на етапі навчання та запобігти їх перенавчанню. Крім того, крос-валідація дозволяє підібрати оптимальні гіперпараметри моделей, що мають суттєвий вплив на точність бінарних класифікаторів. Такий підхід дозволяє оцінити на скільки ефективним є процес навчання моделей, використовуючи поділ навчальної вибірки на п'ять частин (фолдів). Навчання моделей, відповідно до даної методики, відбувалось спочатку на перших чотирьох фолдах, а п'ятий виступав в ролі валідаційного. Така процедура проводилась п'ять раз (для п'ятикратної перехресної перевірки), поки кожна частина тренувальної вибірки не буде використана як валідаційна. Ефективність застосування такого підходу на етапі валідації Ames/QSAR моделей знайшла підтвердження у наступних наукових публікаціях [56,149,151,155].

Оцінку прогностичної здатності розроблених нами Ames/QSAR моделей було проведено з урахуванням базових метрик: загальної точності (*accuracy*), точності позитивного прогнозу (*precision*), чутливості (*recall*), специфічності (*specificity*) та F1-міри ( $F_1$  – *score*). Порівняння ефективності моделей було здійснено для бінарних класифікаторів, що використовували в якості предикторів, як повний, так і обмежені набори дескрипторів (PaDel, Mordred та RDkit) з урахуванням 4 груп ксенобіотиків (табл. 2.1). Необхідно відмітити, що третя група ксенбіотиків (аліфатичні гомомоноциклічні та аліфатичні гомополіциклічні) містять незначну кількість даних (особливо стосується сполук – мутагенів). Тренування Ames/QSAR моделей на таких даних не дасть позитивних результатів з точки зору точності, навіть після застосування інструменту SMOTE, що дозволяє розширити менш представлені класи. Для даної групи хімічних сполук оцінка мутагенного потенціалу була отримана за допомогою розроблених Ames/QSAR моделей, які на етапі тренування, валідації та тестування

використовували повний набір даних, що містить 8454 ксенобіотиків. Ames/QSAR моделі, що дозволяють отримати оцінку мутагенності Еймса для хімічних сполук, які відносяться до 1,2,4 та 5 групи були побудовані відповідно до однорідних наборів даних, в яких молекулярні дескриптори були розраховані для ксенобіотиків, що мають спільні риси будови молекулярного каркасу. Науковий інтерес становлять результати тестування Ames/QSAR моделей, що реалізовані з урахуванням різних підходів: зниження розмірності вхідних даних, що здійснюється через селекцію найбільш релевантних ознак для окремих структурних класів ксенобіотиків; використання різних наборів молекулярних дескрипторів та методів машинного навчання.

### **3.2.1 Ames/QSAR моделі на основі методу випадкового лісу**

Після проведення препроцесингу даних було розроблено Ames/QSAR моделі на основі методу випадкового лісу, що в якості предикторів використовують різні набори 1D та 2D молекулярних дескрипторів (PaDel, Mordred та RDkit). При цьому, на початковому етапі моделювання використовувався повний набір вхідних даних, що враховував як релевантні ознаки так і ті, що мають мінімальний вплив на прогнозовану змінну. Процедура навчання моделей відбувалась окремо для кожного з чотирьох груп хімічних сполук, що мають спільні риси будови молекулярного каркасу. До першої групи, що відповідає окремому класу, відносяться аліфатичні ациклічні хімічні сполуки. Друга група представляє собою ксенобіотики, що представлені аліфатичними гетеромоноциклічними та аліфатичними гетерополіциклічними сполуками. Ароматичні гетеромоноциклічні та ароматичні гетерополіциклічні сполуки формують третю групу ксенобіотиків. До четвертої групи відносяться ароматичні гомомоноциклічні та ароматичні гомополіциклічні хімічні сполуки. З метою порівняння ефективності різних бінарних класифікаторів, була окремо побудована Ames/QSAR модель, для якої на етапі навчання використовувались 80% даних від загальної кількості хімічних сполук (повної бази даних, що представлена 8454 ксенобіотиками).

Підбір гіперпараметрів моделей, таких як, наприклад, кількість дерев у лісі, максимальна кількість листків у кожному дереві, відбувався емпіричним шляхом. Процедура оцінки якості Ames/QSAR моделей проводилась за результатами п'ятикратної крос-валідації з урахуванням метрики точності *accuracy*. У таблиці 3.6 представлено інформацію про налаштування базових гіперпараметрів, при яких точність (*accuracy*) прогнозування (на етапі крос-валідації) для п'яти розроблених моделей була максимальною.

Достатньо інформативними є мінімальні та максимальні (табл. 3.6) значення показника точності моделей, значення яких залежали від набору вхідних даних, що використовувались на етапі проведення крос-валідації.

Таблиця 3.6

**Налаштування базових гіперпараметрів та оцінка точності (крос-валідація) Ames/QSAR моделей, що побудовані на основі повних наборів вхідних даних, без зменшення їх розмірності [204].**

Набір даних	Структурний клас	Кількість дерев	Кількість листків	Точність	Мін. точність	Макс. точність	AUC
PaDeII	Аліфатичні ациклічні	500	300	0,8344	0,81	0,8492	0,89
	Аліфатичні гетеромоно (полі)циклічні	100	200	0,8451	0,7745	0,8922	0,92
	Ароматичні гетеромоно (полі)циклічні	650	200	0,8561	0,8477	0,8713	0,93
	Ароматичні гомомоно (полі) циклічні	650	200	0,8429	0,8258	0,8752	0,9
	Всі	300	600	0,8462	0,8303	0,8591	0,91
RDKit	Аліфатичні ациклічні	200	200	0,8505	0,8209	0,8756	0,95
	Аліфатичні гетеромоно (полі)циклічні	100	100	0,8157	0,7941	0,8627	0,9
	Ароматичні гетеромоно (полі) циклічні	400	300	0,8604	0,846	0,8694	0,93
	Ароматичні гомомоно (полі) циклічні	300	400	0,8417	0,8295	0,8601	0,91
	Всі	350	500	0,8462	0,8339	0,852	0,91



Продовження таблиці 3.6

Набір даних	Структурний клас	Кількість дерев	Кількість листків	Точність	Мін. точність	Макс. точність	AUC
Mordred	Аліфатичні ациклічні	250	250	0,8465	0,8396	0,8538	0,86
	Аліфатичні гетеромоно (полі) циклічні	200	200	0,8392	0,7843	0,8725	0,89
	Ароматичні гетеромоно (полі) циклічні	550	350	0,8576	0,8379	0,8752	0,93
	Ароматичні гомомоно (полі) циклічні	400	300	0,8432	0,8201	0,8809	0,92
	Всі	350	500	0,8466	0,8354	0,8613	0,91

Достатньо інформативними є мінімальні та максимальні (табл. 3.6) значення показника точності моделей, значення яких залежали від набору вхідних даних, що використовувались на етапі проведення крос-валідації. Спостерігали незначну варіабельність між максимальним та мінімальним значеннями метрики точності (*accuracy*), що свідчило про стабільність та надійність роботи моделі.

ROC – крива, яка є популярним інструментом для оцінки якості моделей машинного навчання [199]. Вона дозволяє отримати графічну візуалізацію взаємозв'язку між чутливістю (*recall*) та специфічністю (*specificity*). Площа під ROC-кривою (Area Under the Curve, AUC) є показником, що використовується для оцінки ефективності бінарних класифікаторів. Значення AUC, що наближаються до 1 відповідає класифікатору, який буде ефективно передбачати як позитивні (мутаген), так і негативні (не мутаген) класи.

Відповідно до отриманих проміжних результатів (табл. 3.6) моделювання, можна спостерігати тенденцію покращення точності моделей, що орієнтовані на основні структурні класи ксенобіотиків, у порівняння з Ames/QSAR моделями, для яких на етапі навчання використовувалась неоднорідна, з точки зору будови молекулярного каркасу ксенобіотиків, навчальна вибірка. Остаточна оцінка ефективності розроблених *in silico* моделей прогнозування мутагенності Еймса буде здійснена на екзаменаційній вибірці.

З наукової точки зору необхідними є дослідження, які дозволяють оцінити вплив зменшення розмірності вхідних даних на точність Ames/QSAR моделей. Потребує вирішення питання оцінки мутагенності Еймса основних структурних класів ксенобіотиків з урахуванням обмеженого набору предикторів, що задаються молекулярними дескрипторами. Тому наступний етап роботи був присвячений розробці *in silico* моделей оцінки мутагенності Еймса, ефективність яких може бути покращена через зменшення розмірності вхідних даних. В якості основного алгоритму, що дозволяє здійснювати відбір релевантних дескрипторів, був обраний RFECV, що поєднаний з крос-валідацією. RFECV дозволяє рекурсивно видаляти ознаки, що мають менш вагомий вплив на прогнозовану змінну. У таблиці 3.7, для кожної з побудованих Ames/QSAR моделей, що використовували на етапі навчання різні набори предикторів (PaDel, RDkit, Mordred), які були розраховані відповідно до чотирьох структурних класів ксенобіотиків, представлена інформація про кількість релевантних дескрипторів, що залишилась після проведення процедури видалення ознак. Оцінка ефективності розроблених класифікаторів була здійснена за допомогою метрики точності (*accuracy*) та площі під ROC- кривою (AUC).

Таблиця 3.7

**Оцінка точності (крос-валідація) Ames/QSAR моделей, що побудовані на основі обмеженого набору релевантних дескрипторів [204]**

Набір даних	Структурний клас	Кількість релевантних ознак	Точність	Мін. точність	Макс. точність	AUC
PaDel	Аліфатичні ациклічні	612	0,8454	0,83	0,8593	0,89
	Аліфатичні гетеромоно (полі) циклічні	49	0,8627	0,8137	0,902	0,93
	Ароматичні гетеромоно (полі) циклічні	649	0,8608	0,8516	0,8772	0,93
	Ароматичні гомомоно (полі) циклічні	274	0,847	0,8314	0,8771	0,9
	Всі (1444)	289	0,8489	0,8375	0,8569	0,91
RDKit	Аліфатичні ациклічні	94	0,8505	0,8209	0,8756	0,95

Продовження таблиці 3.7

Набір даних	Структурний клас	Кількість релевантних ознак	Точність	Мін. точність	Макс. точність	AUC
RDKit	Аліфатичні гетеромоно (полі) циклічні	85	0,8176	0,8039	0,8529	0,9
	Ароматичні гетеромоно (полі) циклічні	139	0,8615	0,848	0,8811	0,93
	Ароматичні гомомоно (полі) циклічні	87	0,8447	0,8239	0,8639	0,91
	Всі (196)	145	0,8473	0,839	0,8569	0,91
Mordred	Аліфатичні ациклічні	322	0,8456	0,831	0,8585	0,87
	Аліфатичні гетеромоно (полі) циклічні	218	0,8412	0,7549	0,9118	0,9
	Ароматичні гетеромоно (полі) циклічні	675	0,8569	0,8418	0,8733	0,93
	Ароматичні гомомоно (полі) циклічні	416	0,8455	0,822	0,8733	0,93
	Всі (1613)	776	0,8466	0,8354	0,8613	0,91

Досить цікавими у науковому відношенні є результати оцінки точності моделей Ames/QSAR, що були нами отримані при проведенні п'ятикратної крос-валідації з урахуванням обмеженого переліку молекулярних дескрипторів. Зменшення кількості вхідних даних в діапазоні від 55% до 97% від початкової кількості дескрипторів (PaDel, Mordred та RDkit) призводило, у більшості випадків, до покращення точності (табл. 3.6) *in silico* моделей прогнозування мутагенності Еймса у порівнянні з такими показниками для розроблених бінарних класифікаторів, які на етапі навчання використовували повний набір вхідних даних. При цьому відсоток збільшення точності (*accuracy*) для орієнтованих на основні структурні класи ксенобіотиків моделей коливався у межах від 0,09% до 1,76%. Одна модель Ames/QSAR, що використовували в якості предикторів набір релевантних дескрипторів Mordred, що були розраховані для четвертої групи ксенобіотиків, показала несуттєве зниження точності. Однак при цьому, показник AUC для цієї моделі залишився без змін, що вказує на збережений баланс бінарних класифікаторів до прогнозування як позитивних (мутаген), так і

негативних класів (не мутаген). Оптимізація Ames/QSAR моделей, що була здійснена через зменшення розмірності вхідних даних, дозволяє покращити точність, стабільність та узагальнюваність моделей, а також знижує ризики їх перенавчання. Зменшення обсягу вхідних даних дозволило покращити інтерпретованість Ames/QSAR моделей. Такий підхід, на нашу думку, може лежати в основі пошуку причинно-наслідкових зв'язків між певними властивостями ксенобіотиків, які представлені набором релевантних дескрипторів та проявами мутагенності.

Отримані результати (на етапі крос-валідації) оцінки ефективності Ames/QSAR моделей, що використовували в якості предикторів як повний, так і обмежений набори молекулярних дескрипторів, дозволили підтвердити сформульовану на початку проведення дослідження гіпотезу, відповідно до якої, покращення якості *in silico* моделей оцінки мутагенності Еймса може бути досягнуто через відбір релевантних дескрипторів, які були обчислені для окремих груп ксенобіотиків. У такій ситуації перехід до наступного кроку, що пов'язаний з остаточною перевіркою моделей на тестовій та екзаменаційній вибірках є логічним та теоретично обґрунтованим та дозволить оцінити ефективність бінарних класифікаторів на нових даних (табл. 3.8), що не використовувались на етапі навчання.

Таблиця 3.8

**Класифікаційний звіт отриманий на екзаменаційній вибірці для моделей Ames/QSAR з обмеженим переліком релевантних дескрипторів [204]**

Набір даних	Структурний клас	Точність (тестова вибірка)	Точність (екзам. вибірка)	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

## Продовження таблиці 3.8

Набір даних	Структурний клас	Точність (тестова вибірка)	Точність (екзам. вибірка)	Precision	Recall	Specificity	F1 Score	AUC
PaDell	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 переходити до детального аналізу отриманих результатів тестування бінарних класифікаторів, необхідно зазначити, що показник точності Ames/QSAR моделей в межах 80 – 85% вважається високим результатом [149,151,155,196], що відповідає зареєстрованій варіабельності *in vitro* тесту Еймса [205]. На сьогоднішній день приділяється особлива увага питанням покращення прогностичної здатності Ames/QSAR моделей. З цієї точки зору заслуговує на увагу ініціатива відділу генетики у та мутагенезу Національного інституту наук про здоров'я Японії, які стали організаторами у 2020-2022 р.

міжнародних змагань науковців з 11 країн світу, що були направлені на удосконалення Ames/QSAR моделей [182]. Покращення навіть на 1-2% значень основних метрик оцінки ефективності Ames/QSAR моделей може мати як наукову, так і практичну цінність.

З метою отримання об'єктивної оцінки узагальнюючої здатності моделей Ames/QSAR, з урахуванням представленої авторами статті [203] методології, тестування розроблених нами моделей проводилось на двох незалежних вибірках. Незначний розкид значень метрики точності (*accuracy*) на тестовій та екзаменаційній вибірках (табл. 3.8) для кожної моделі Ames/QSAR дозволяв зробити припущення про те, що в реальних умовах (на нових даних) розроблені бінарні класифікатори будуть проявляти стабільність. Відповідно до екзаменаційної вибірки, для остаточної перевірки узагальнюючої здатності Ames/QSAR моделей були розраховані базові метрики бінарної класифікації (табл. 3.8). Аналіз класифікаційного звіту дозволив зробити висновок про те, що 7 моделей, які на етапі навчання використовували однорідні оптимізовані набори даних, що були сформовані відповідно до основних структурних класів ксенобіотиків, зберігали високі показники класифікації. При цьому моделі орієнтовані на основні структурні класи ксенобіотиків, які на етапі навчання використовували повні набори даних (молекулярні дескриптори PaDel, RDkit та Mordred), без їх попередньої оптимізації, зазвичай демонстрували менші показники точності. Необхідно зазначити, що прогностична здатність моделей, побудованих відповідно до тренувальної вибірки, що складає 80% від розміру повної бази даних (відповідає класичному підходу до моделювання), була теж підвищена, у межах 1%, через зменшення розмірності вхідних даних.

Для прогнозування мутагенності Еймса аліфатичних ациклічних хімічних сполук необхідно брати за основу модель Ames/QSAR, що на етапі навчання використовувала обмежений набір релевантних дескрипторів RDKit, що був отриманий для першої групи ксенобіотиків (табл. 2.1). Розроблена модель з  $AUC = 0,96$  має переваги у порівнянні з бінарним класифікатором (з  $AUC = 0,93$ ), що в якості вхідних даних використовувала неоднорідний набір найбільш

впливових дескрипторів RDKit, що був отриманий відповідно до повної бази даних ксенобіотиків, яка представлена 8454 хімічними сполуками.

*In silico* оцінка мутагенності Еймса для другої групи ксенобіотиків, що відповідає аліфатичним гетеромоноциклічним та аліфатичним гетерополіциклічним хімічним сполукам може бути здійснена за допомогою бінарного класифікатора, який побудований на основі однорідного оптимізованого набору вхідних даних, що відповідають молекулярним дескрипторам Mordred, які були розраховані для другої групи (табл. 2.1) ксенобіотиків. Точність такої Ames/QSAR моделі, відповідно до метрики *accuracy* становила 90%, що є одним з найбільших показників у порівнянні з іншими моделями. При цьому точність моделі, що була побудована відповідно до стандартного підходу, з використанням дескрипторів повної бази даних та відбором релевантних предикторів, становила 85%. Необхідно відмітити, що показник чутливості (*recall*), що дозволив отримати оцінку ефективності моделі з точки зору ідентифікації мутагенів (позитивний клас), приймає однакове значення (84%) для двох моделей. Такий результат був підставою для висновку про те, що обидві моделі дають однакову кількість хибнонегативних результатів. Показник чутливості є одним з найголовніших метрик, який необхідно враховувати на етапі відбору найкращих Ames/QSAR моделей. Зниження значення цього показника може призвести до того, що модель не буде здатна виявляти більшу частину хімічних сполук з потенційними мутагенними властивостями, що у цілому, може сприяти підвищенню ризиків для генетичного здоров'я людини.

Для оцінки мутагенного потенціалу ксенобіотиків, що відносяться до ароматичних гетеромоноциклічних та ароматичних гетерополіциклічних хімічних сполук найбільш ефективною виявилася Ames/QSAR модель, що в якості предикторів використовувала оптимальний набір дескрипторів RDkit, що були розраховані для четвертої групи ксенобіотиків (табл. 2.1). На тестовій вибірці даний бінарний класифікатор продемонстрував високі результати, зокрема за показником чутливості, та точності позитивного прогнозу (*precision*), що становили відповідно 88% та 91%.

Ефективний розподіл між двома класами (мутаген/не мутаген) для ароматичних гомомоноциклічних та ароматичних гомополіциклічних хімічних сполук, дозволила отримати Ames/QSAR модель, яка побудована на основі релевантного набору вхідних даних, що відповідають молекулярним дескрипторам Mordred, які були розраховані для ксенобіотиків, що належить до п'ятої групи (табл. 2.1).

### 3.2.2 Ames/QSAR моделі на основі екстремального глідієнтного бустінгу

В основі метода XGBoost використовуються дерева рішень, що будуються послідовно. Відповідно, кожна нова побудована модель (дерево рішень) враховує помилки, що були допущені на попередньому етапі. Реалізація Ames/QSAR моделей на основі методу XGBoost передбачає виконання наступних етапів: підготовки даних, що включає в себе їх нормалізацію та балансування класів; навчання моделей на тестовій вибірці, що складає 80% від загальної кількості ксенобіотиків; проведення п'ятикратної крос-валідації, що дозволяє підібрати оптимальні гіперпараметри моделей та дати оцінку їх ефективності на проміжних етапах навчання Ames/QSAR моделей; тестування моделей на двох незалежних вибірках, кожна з яких становила 10% від загальної кількості хімічних сполук.

Для порівняння ефективності бінарних класифікаторів нами було запропоновано, в якості вхідних даних для Ames/QSAR моделей використовувати, як повний набір дескрипторів (PaDel, RDkir, Mordred) так і їх обмежений перелік, що формується з урахуванням найбільш впливових дескрипторів. Відбір релевантних предикторів відбувався на етапі крос-валідації за допомогою алгоритму RFECV. Методологія підготовки даних, що описана у підрозділі 3.2 повністю відповідає вимогам для застосування методу XGBoost при побудові *in silico* моделей прогнозування мутагенності Еймса.

Підбір гіперпараметрів моделей, таких як, кількість дерев, максимальна глибина, швидкість навчання тощо відбувався емпіричним шляхом. Процедура оцінки якості Ames/QSAR моделей проводилась за результатами п'ятикратної крос-валідації з урахуванням метрики *accuracy*. У таблиці 3.9 представлено інформацію про налаштування базових гіперпараметрів моделі на основі



XGBoost, при яких точність прогнозування для розроблених моделей буде найкращою.

Таблиця 3.9

**Налаштування основних гіперпараметрів та оцінка точності (на етапі крос-валідації) Ames/QSAR моделей, що побудовані на основі повних наборів вхідних даних без зменшення їх розмірності**

Набір даних	Структурний клас	Кількість дерев	Макс. глибина	Швидкість навчання	Точність	AUC
PaDell	Аліфатичні ациклічні	200	5	0.1	0.8494	0.9
	Аліфатичні гетеромоно (полі)циклічні	200	7	0.1	0.851	0.9
	Ароматичні гетеромоно (полі)циклічні	200	6	0.1	0.8584	0.92
	Ароматичні гомомоно (полі) циклічні	200	7	0.1	0.85	0.93
	Всі	200	7	0.1	0.8462	0.93
RDKit	Аліфатичні ациклічні	200	7	0.05	0.843	0.94
	Аліфатичні гетеромоно (полі)циклічні	200	5	0.1	0.8196	0.89
	Ароматичні гетеромоно (полі) циклічні	150	7	0.05	0.8557	0.94
	Ароматичні гомомоно (полі) циклічні	200	7	0.1	0.8444	0.91
	Всі	200	7	0.1	0.8411	0.91
Mordred	Аліфатичні ациклічні	200	5	0.1	0.8456	0.87
	Аліфатичні гетеромоно (полі) циклічні	250	7	0.1	0.8412	0.87
	Ароматичні гетеромоно (полі) циклічні	200	5	0.1	0.8541	0.92
	Ароматичні гомомоно (полі) циклічні	200	7	0.1	0.847	0.92
	Всі	200	7	0.1	0.8509	0.92

Під час проведення крос-валідації на кожній ітерації для Ames/QSAR моделей розраховувалась метрика точності (*accuracy*) та площа під ROC-кривою AUC, а її усереднене значення записано у табл. 3.9. Відповідно до отриманих результатів

(табл. 3.9) моделювання, спостерігалась тенденція щодо підвищення точності моделей, орієнтованих на основні структурні класи ксенобіотиків, у порівнянні з Ames/QSAR моделями, для яких на етапі навчання використовувалась неоднорідна, з точки зору будови молекулярного каркасу ксенобіотиків, навчальна вибірка.

У таблиці 3.10, для кожної з Ames/QSAR моделей, для яких на етапі навчання використовували різні набори предикторів (PaDel, RDkit, Mordred), які були розраховані відповідно до чотирьох структурних класів ксенобіотиків, представлена інформація про кількість релевантних дескрипторів, що залишились після видалення найменш впливових ознак. В якості основного алгоритму, що дозволяє здійснювати відбір релевантних дескрипторів, був обраний RFECV, який також використовувався при побудові Ames/QSAR моделей на основі випадкового лісу. Відповідно до отриманих результатів оцінки ефективності Ames/QSAR моделей, що була проведена на етапі крос-валідації (з урахуванням метрик точності (*accuracy*) та площі під ROC-кривою), можна побачити, що при зменшенні кількості вхідних даних у діапазоні від 55% до 90% від початкової кількості дескрипторів (PaDel, Mordred та RDkit) призводило, у більшості випадків, до покращення точності (табл. 3.9) *in silico* моделей прогнозування мутагенності Еймса у порівнянні з такими показниками для розроблених бінарних класифікаторів, які на етапі навчання використовували повний набір вхідних даних.

Таблиця 3.10

**Оцінка точності Ames/QSAR, що побудовані на основі обмеженого набору релевантних дескрипторів**

Набір даних	Структурний клас	Кількість релевантних ознак	Точність	Мін. точність	Макс. точність	AUC
PaDell	Аліфатичні ациклічні	154	0.8616	0.8451	0.8868	0.91
	Аліфатичні гетеромоно (полі) циклічні	345	0.8549	0.7941	0.8922	0.92
	Ароматичні гетеромоно (полі) циклічні	409	0.8611	0.846	0.8713	0.92

Продовження таблиці 3.10

Набір даних	Структурний клас	Кількість релевантних ознак	Точність	Мін. точність	Макс. точність	AUC
PaDel	Ароматичні гомомоно (полі) циклічні	385	0.8538	0.8277	0.885	0.93
	Всі (1444)	204	0.8515	0.8383	0.8699	0.93
RDKit	Аліфатичні ациклічні	74	0.843	0.8037	0.8598	0.95
	Аліфатичні гетеромоно (полі) циклічні	56	0.8235	0.8039	0.8523	0.89
	Ароматичні гетеромоно (полі) циклічні	89	0.8569	0.8421	0.885	0.95
	Ароматичні гомомоно (полі) циклічні	62	0.8466	0.8295	0.8733	0.91
	Всі (196)	100	0.8466	0.8404	0.8526	0.91
Mordred	Аліфатичні ациклічні	465	0.8484	0.8443	0.8685	0.87
	Аліфатичні гетеромоно (полі) циклічні	170	0.8569	0.8039	0.9118	0.87
	Ароматичні гетеромоно (полі) циклічні	397	0.8569	0.8441	0.8752	0.94
	Ароматичні гомомоно (полі) циклічні	374	0.8481	0.822	0.8847	0.92
	Всі (1613)	320	0.8524	0.844	0.8584	0.92

Відповідно до отриманих результатів оцінки ефективності Ames/QSAR моделей, що була проведена на етапі крос-валідації, можна побачити, що при зменшенні кількості вхідних даних у діапазоні від 55% до 90% від початкової кількості дескрипторів (PaDel, Mordred та RDkit) призводило, у більшості випадків, до покращення точності (табл. 3.10) *in silico* моделей прогнозування мутагенності Еймса у порівнянні з такими показниками для розроблених бінарних класифікаторів, які на етапі навчання використовували повний набір вхідних даних. При цьому відсоток збільшення точності (*accuracy*) для орієнтованих на основні структурні класи ксенобіотиків моделей коливався у межах від 0,15% до

1,57%. Для однієї моделі Ames/QSAR (RDkit аліфатичні ациклічні) не було зафіксовано покращення точності.

Застосування алгоритму RFECV для моделей Ames/QSAR на основі методу екстремального грідиєнтного бустінгу дозволило вирішити задачу бінарної класифікації з розподілом хімічних сполук на два класи (мутаген/не мутаген), використовуючи меншу кількість релевантних дескрипторів, у порівнянні з моделями, що побудовані на основі методу випадкового лісу. Такий результат моделювання є позитивним та, можливо, дозволить у подальшому спростити вирішення складної задачі, що пов'язана з аналізом причинно-наслідкових зв'язків між мутагенністю та набором властивостей, що представлені молекулярними дескрипторами.

У таблиці 3.11 наведено значення основних метрик оцінки якості Ames/QSAR моделей на основі методу екстремального грідиєнтного бустінгу, що отримані на тестовій та екзаменаційній вибірках.

Таблиця 3.11

**Класифікаційний звіт на екзаменаційній вибірці для моделей Ames/QSAR з обмеженим переліком релевантних дескрипторів**

Набір даних	Структурний клас	Точність (тестова вибірка)	Точність (екзам. вибірка)	Precision	Recall	Specificity	F1 Score	AUC
PaDell	Аліфатичні ациклічні	0.7961	0.8333	0.8	0.8462	0.8226	0.8224	0.9
	Аліфатичні гетеромоно (полі) циклічні	0.8286	0.8448	0.76	0.8636	0.8333	0.8085	0.88
	Ароматичні гетеромоно (полі) циклічні	<b>0.8685</b>	<b>0.8635</b>	<b>0.8653</b>	<b>0.8718</b>	<b>0.8553</b>	<b>0.8635</b>	<b>0.93</b>
	Ароматичні гомомоно (полі) циклічні	<b>0.849</b>	<b>0.8516</b>	<b>0.8405</b>	<b>0.8726</b>	<b>0.8301</b>	<b>0.8562</b>	<b>0.9</b>
	Всі	0.8521	0.8592	0.8644	0.8624	0.8557	0.8634	0.93
RDKit	Аліфатичні ациклічні	<b>0.8701</b>	<b>0.8923</b>	<b>0.8906</b>	<b>0.9194</b>	<b>0.8654</b>	<b>0.9048</b>	<b>0.96</b>

Продовження таблиці 3.11

Набір даних	Структурний клас	Точність (тестова вибірка)	Точність (екзам. вибірка)	Precision	Recall	Specificity	F1 Score	AUC
RDKit	Аліфатичні гетеромоно (полі) циклічні	0.8286	0.8276	0.7931	0.8519	0.8065	0.8214	0.88
	Ароматичні гетеромоно (полі) циклічні	<b>0.8777</b>	<b>0.8571</b>	<b>0.88</b>	<b>0.8302</b>	<b>0.8846</b>	<b>0.8544</b>	0.92
	Ароматичні гомомоно (полі) циклічні	<b>0.8433</b>	<b>0.8484</b>	<b>0.8247</b>	<b>0.8639</b>	<b>0.8344</b>	<b>0.8439</b>	0.9
	Всі	0.842	0.8651	0.8702	0.8558	0.8723	0.864	0.87
Mordred	Аліфатичні ациклічні	0.8026	0.7895	0.7447	0.7447	0.8209	0.7447	0.87
	Аліфатичні гетеромоно (полі) циклічні	0.8143	0.8103	0.8333	0.8065	0.8148	0.8197	0.9
	Ароматичні гетеромоно (полі) циклічні	<b>0.8746</b>	<b>0.8603</b>	<b>0.8526</b>	<b>0.8636</b>	<b>0.8571</b>	<b>0.8581</b>	<b>0.92</b>
	Ароматичні гомомоно (полі) циклічні	0.8632	0.8419	0.8491	0.8438	0.84	0.8464	0.92
	Всі	0.851	0.8615	0.8716	0.8658	0.8568	0.8687	0.94

Мінімальний розкид значень метрики точності (*accuracy*) на тестовій та екзаменаційній вибірках (табл. 3.11) додає впевненості у тому, що в реальних умовах (на нових даних) розроблені моделі Ames/QSAR будуть проявляти стабільність. Відповідно до екзаменаційної вибірки, для остаточної перевірки узагальнюючої здатності Ames/QSAR моделей, були розраховані базові метрики бінарної класифікації. Аналіз класифікаційного звіту дозволяє зробити висновок про те, що 6 моделей (значення метрик найкращих моделей у табл. 3.11 виділені жирним шрифтом), які на етапі навчання використовували однорідні оптимізовані набори даних, що були сформовані відповідно до основних структурних класів ксенобіотиків, зберігали високі показники класифікації. При цьому моделі,

орієнтовані на основні структурні класи ксенобіотиків, які на етапі навчання використовували повні набори даних (молекулярні дескриптори PaDel, RDkit та Mordred), без їх попередньої оптимізації, зазвичай демонстрували менші показники точності. Необхідно відмітити, що прогностична здатність моделей, які були побудовані відповідно до тренувальної вибірки, що складає 80% від розміру повної бази даних (відповідає класичному підходу до моделювання), була підвищена через зменшення розмірності вхідних даних. Точність таких бінарних класифікаторів була в більшості випадків вищою у порівнянні з орієнтованими на основні структурні класи Ames/QSAR моделей з оптимізацією. Тільки дві орієнтовані на структурні класи (ароматичні гетеромоно(полі) циклічні з дескрипторами PaDell та аліфатичні ациклічні з дескрипторами RDkit) моделі показали кращу ефективність у порівнянні з моделями, що були реалізовані відповідно до стандартної схеми моделювання, але з урахуванням оптимізації, що полягала у зменшенні розміру вхідних даних.

Досить важливими у науковому відношенні є результати порівняння ефективності моделей Ames/QSAR, побудованих на основі двох ансамблевих алгоритмів: випадкового лісу та екстремального градієнтного бустінгу. Для визначення більш ефективних бінарних класифікаторів, розглянемо тільки ті Ames/QSAR моделі, які, відповідно до базових метрик оцінки ефективності, демонстрували найкращі результати класифікації (табл. 3.12).

Таблиця 3.12

**Порівняння ефективності Ames/QSAR моделей на основі методу випадкового лісу та екстремального градієнтного бустінгу**

Набір даних/назва методу	Структурний клас	Точність (accuracy)	Precision	Recall	Specificity	F1 Score	AUC
RDKit/ RF	Аліфатичні ациклічні	0,8947	0,9412	0,8889	0.9048	0,9443	0,96
RDKit/ XGBoost	Аліфатичні ациклічні	<b>0.8923</b>	<b>0.8906</b>	<b>0.9194</b>	<b>0.8654</b>	<b>0.9048</b>	<b>0.96</b>
Mordred / RF	Аліфатичні гетеромоно (полі) циклічні	<b>0,8966</b>	<b>0,913</b>	<b>0,84</b>	<b>0.9394</b>	<b>0,875</b>	<b>0,95</b>

Продовження таблиці 3.12

Набір даних/назва методу	Структурний клас	Точність (accuracy)	Precision	Recall	Specificity	F1 Score	AUC
Mordred / XGBoost	Аліфатичні гетеромоно (полі) циклічні	0.8103	0.8333	0.8065	0.8148	0.8197	0.9
RDKit/ RF	Ароматичні гетеромоно (полі) циклічні	<b>0,8921</b>	<b>0,913</b>	<b>0,8802</b>	<b>0.9054</b>	<b>0,8963</b>	<b>0,94</b>
RDKit/ XGBoost	Ароматичні гетеромоно (полі) циклічні	0.8571	0.88	0.8302	0.8846	0.8544	0.92
Mordred / RF	Ароматичні гетеромоно (полі) циклічні	<b>0,871</b>	<b>0,8741</b>	<b>0,8503</b>	<b>0.8896</b>	<b>0.8621</b>	<b>0,95</b>
Mordred / XGBoost	Ароматичні гетеромоно (полі) циклічні	0.8419	0.8491	0.8438	0.84	0.8464	0.92

У таблиці 3.12 жирним шрифтом позначені метрики оцінки якості, що відповідають найкращим Ames/QSAR моделям. Для прогнозування мутагенності хімічних сполук, що відносяться до другої, четвертої та п'ятої групи ксенобіотиків (табл. 2.1) необхідно використовувати Ames/QSAR моделі, які побудовані на основі методу випадкового лісу з обмеженим набором дескрипторів (RDKit та Mordred), що були розраховані нами для окремих груп ксенобіотиків, які мають спільну будову молекулярного каркасу. Оцінка мутагенного потенціалу для першої групи ксенобіотиків може бути здійснена за допомогою моделей Ames/QSAR на основі екстремального градієнтного бустінгу з дескрипторами RDKit. Відповідно до класифікаційного звіту (табл. 3.12), непоганий результат щодо розподілу аліфатичних ациклічних хімічних сполук на дві групи (мутаген/не мутаген) може бути отриманий, також, за допомогою Ames/QSAR моделі на основі методу випадкового лісу. Але враховуючи той факт, що дана модель може давати більшу кількість хибнонегативних результатів, у порівнянні з бінарним класифікатором на основі екстремального градієнтного бустінгу, її застосування може мати певні обмеження. *In silico* моделі прогнозування мутагенності Еймса

повинні, у першу чергу, зводити до мінімуму кількість хибнонегативних результатів, тому що не ідентифікований мутаген, який буде представлений у навколишньому середовищі, може становити ризик для генетичного здоров'я людської популяції.

### 3.2.3 Ames/QSAR моделі з використанням неймережевого підходу

З метою оцінки мутагенності факторів навколишнього середовища науковці також використовують Ames/QSAR моделі, що побудовані на основі глибинної нейронної мережі [191,192]. Такі бінарні класифікатори демонструють одні з найкращих показників точності класифікації. Крім того, Ames/QSAR моделі на основі нейронних мереж мають значний не реалізований потенціал. У такій ситуації необхідним є проведення наступного етапу дослідження, що пов'язаний з розробкою та тестуванням орієнтованих на основні структурні класи ксенобіотиків Ames/QSAR моделей з використанням неймережевого підходу.

У межах проведеного дослідження нами було прийнято рішення використовувати глибинну нейронну мережу (Deep Neural Network), архітектура якої представлена на рис. 3.2.

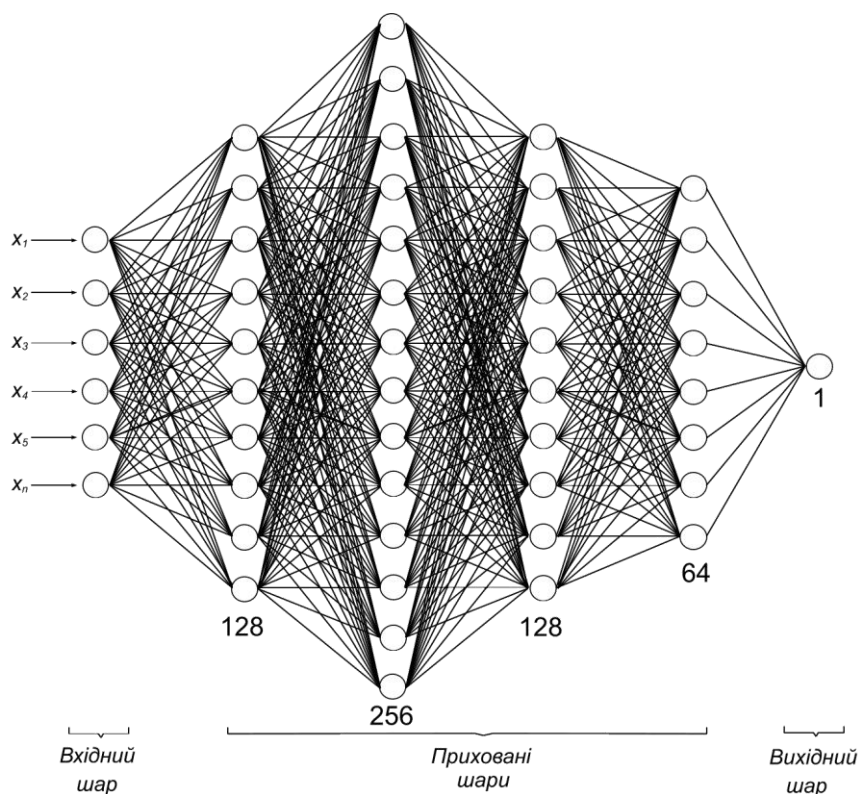


Рис. 3.2. Архітектура глибинної нейронної мережі



Подібна до розробленої у межах представленої роботи архітектура нейронної мережі була запропонована авторами наукової праці [196], що продемонструвала високі результати точності бінарної класифікації. При цьому розроблена *in silico* модель також використовувалась для вирішення задачі оцінки мутагенності ксенобіотиків.

Необхідно зазначити, що на даному етапі роботи ми використовували нейронну мережу, що була розроблена ще на початку проведення дослідження (підрозділ 3.1.). В структурі нейронної мережі (рис. 3.2.) представлений вхідний шар, вихідний шар та 4 приховані шари, які містять 128, 256, 128 та 64 нейрони відповідно. Кількість нейронів першого шару відповідає розмірності простору ознак, що відповідає кількості  $(x_1 \dots x_n)$  одновимірних та двовимірних дескрипторів PaDel, Mordred та RDkit. В якості функції активації для прихованих шарів була обрана функція ReLU, що дозволяє ефективно вирішити проблему зникаючого градієнту, який може негативно впливати на процес навчання глибоких нейронних мереж. На вихідному шарі функцією активації є Sigmoid (сигмоподібна функція), що дає змогу отримати значення ймовірності щодо приналежності ксенобіотиків до одного з двох класів (мутаген/не мутаген). Для вирішення стандартної для нейронних мереж проблеми, що пов'язана з перенавчанням нами були застосовані методи L2 та Dropout-регуляризації. Між всіма внутрішніми шарами мережі задіяний механізм Dropout, який випадковим чином вимикає 30 % нейронів під час навчання, що забезпечує більшу стійкість та покращує прогностичну здатність моделей на нових вхідних даних та запобігає перенавчанню моделей. Бінарна крос-ентропія була обрана в якості функції втрат, що зазвичай використовують для вирішення задачі бінарної класифікації. З метою мінімізації функції втрат був обраний більш ефективний оптимізатор Nadam, що є вдосконаленою версією оптимізатора Adam [195]. Навчання нейронної мережі було здійснено протягом 30 епох з розміром партії 32. Обраний розмір пакета є стандартним та дозволяє досягти балансу між швидкістю та стабільністю.

Ефективність розроблених ефективних *in silico* Ames/QSAR моделей на основі глибинної нейронної мережі оцінювалась на екзаменаційній вибірці за допомогою метрик точності (*accuracy*), точності позитивного прогнозу (*precision*), чутливості (*recall*), специфічності (*specificity*) та F1-міри ( $F_1$  – *score*). Для всіх моделей також була обчислена площа під ROC-кривою (AUC), що є популярним інструментом для оцінки прогностичної здатності бінарних класифікаторів.

Необхідно відмітити, що метрика *accuracy* та площа під ROC-кривою (AUC) використовувалась для оцінки ефективності проміжних моделей, які були отримані з урахуванням етапності розробленої в рамках роботи методики покращення прогностичної здатності Ames/QSAR моделей. Збільшення значень параметру *accuracy* та AUC у такому випадку може виступати в ролі маркеру, що підтверджує правильність сформульованої на початку дослідження гіпотези, відповідно до якої ефективність бінарних класифікаторів можна покращити через зменшення розмірності вхідних даних та формування однорідних груп ксенобіотиків, що мають спільні риси будови молекулярного каркасу. Крім того, параметри *accuracy* та AUC можна використовувати для оцінки вкладу кожної з запропонованих оптимізацій, що лежать в основі покращення Ames/QSAR моделей. Остаточна перевірка прогностичної здатності Ames/QSAR моделей, які були отримані з урахуванням проведених процедур оптимізації, була проведена спочатку на тестовій, а потім на екзаменаційній вибірках, з урахуванням базових метрик оцінки ефективності.

На початковому етапі проведення моделювання необхідно було оцінити точність Ames/QSAR моделей на основі нейронної мережі, що використовують при навчанні повний набір даних (молекулярні дескриптори PaDel, RDkit та Mordred) з урахування розподілу ксенобіотиків на однорідні групи, що мають спільні риси будови молекулярного каркасу. Крім того, необхідно було порівняти точність розроблених орієнтованих на основні структурні класи Ames/QSAR моделей з бінарними класифікаторами, які на етапі навчання використовували різні набори (PaDel, RDkit та Mordred) молекулярних дескрипторів, які були

розраховані для частини (80%) ксенобіотиків повної бази даних, яка налічує 8454 хімічні сполуки.

Процедура оцінки якості Ames/QSAR моделей, проводилась за результатами п'ятикратної перехресної перевірки. При цьому на кожній ітерації, при проведенні крос-валідації, для кожної моделі було розраховано метрику точності (*accuracy*) та AUC (таблиця 3.13). У таблиці 3.12 представлені результати оцінки точності моделей Ames/QSAR, з урахуванням двох метрик, а також записані мінімальні, максимальні та усереднені значення метрики точності (*accuracy*). Декілька моделей показали несуттєве зменшення показника точності (*accuracy*). Невелике зниження значень параметра AUC спостерігалось для однієї моделі. Але загалом, відповідно до отриманих результатів (табл. 3.13) моделювання, спостерігалася закономірність, що пов'язана з підвищенням точності моделей, орієнтованих на основні структурні класи ксенобіотиків, у порівнянні з Ames/QSAR моделями, для яких на етапі навчання використовувалась неоднорідна, з точки зору будови молекулярного каркасу ксенобіотиків, навчальна вибірка. Подібна тенденція спостерігалась також для прогностичних моделей, що були розроблені на основі ансамблевих алгоритмів машинного навчання. Отримані результати є підтвердженням ефективності запропонованої у межах роботи методики, що направлена на покращення точності *in silico* моделей прогнозування мутагенності Еймса.

Таблиця 3.13

**Значення метрик точності (*accuracy*) та площі під ROC-кривою для Ames/QSAR моделей, що побудовані на основі повних наборів вхідних даних, без зменшення їх розмірності**

Набір даних	Структурний клас	Точність ( <i>accuracy</i> )	Мін. точність	Макс. точність	AUC
PaDell	Аліфатичні ациклічні	0.8089	0.7887	0.849	0.93
	Аліфатичні гетеромоно (полі)циклічні	0.8216	0.7745	0.8627	0.95
	Ароматичні гетеромоно (полі)циклічні	0.8342	0.8284	0.8421	0.92
	Ароматичні гомомоно (полі) циклічні	0.8209	0.7985	0.8321	0.9

## Продовження таблиці 3.13

Набір даних	Структурний клас	Точність (accuracy)	Мін. точність	Макс. точність	AUC
	Всі	0.8251	0.8165	0.8466	0.9
RdKit	Аліфатичні ациклічні	0.8551	0.799	0.8925	0.9
	Аліфатичні гетеромоно (полі)циклічні	0.8255	0.7647	0.8922	0.91
	Ароматичні гетеромоно (полі) циклічні	0.8393	0.8246	0.8574	0.91
	Ароматичні гомомоно (полі) циклічні	0.8235	0.8087	0.8447	0.89
	Всі	0.8259	0.8215	0.8374	0.91
Mordred	Аліфатичні ациклічні	0.8253	0.8066	0.8316	0.92
	Аліфатичні гетеромоно (полі) циклічні	0.8237	0.8039	0.8529	0.94
	Ароматичні гетеромоно (полі) циклічні	0.8389	0.8168	0.8594	0.92
	Ароматичні гомомоно (полі) циклічні	0.8258	0.8049	0.839	0.91
	Всі	0.8294	0.8159	0.846	0.9

Наступний етап моделювання пов'язаний з розробкою *in silico* моделей оцінки мутагенності, ефективність яких може бути покращена через зменшення розмірності вхідних даних. При створенні Ames/QSAR моделей на основі глибинної нейронної мережі було запропоновано використовувати обмежений перелік молекулярних дескрипторів, що був отриманий відповідно до моделей на основі методу екстремального градієнтного бустінгу за допомогою алгоритму RFEC. Очевидно, що проведення тренування моделей на обмеженому наборі вхідних даних, що мають суттєвий вплив на прогнозовану змінну (мутаген/не мутаген) може сприяти зниженню кількості хибнонегативних та хибнопозитивних результатів прогнозів. У таблиці 3.14, для кожної з побудованих Ames/QSAR моделей, що використовували на етапі навчання різні набори предикторів, які були розраховані відповідно до чотирьох структурних класів ксенобіотиків, представлена інформація про кількість релевантних дескрипторів. Оцінка

ефективності розроблених класифікаторів була здійснена за допомогою метрики точності (*accuracy*) та площі під ROC- кривою (AUC).

Таблиця 3.14

**Значення метрик точності (*accuracy*) та площі під ROC-кривою (крос-валідація) для Ames/QSAR моделей, що побудовані на основі обмеженого набору релевантних дескрипторів**

Набір даних	Структурний клас	Кількість релевантних ознак	Точність ( <i>accuracy</i> )	Мін. точність	Макс. точність	AUC
PaDel	Аліфатичні ациклічні	154	0.8239	0.8028	0.849	0.95
	Аліфатичні гетеромоно (полі) циклічні	345	0.8278	0.8014	0.8671	0.97
	Ароматичні гетеромоно (полі) циклічні	409	0.8463	0.8382	0.8577	0.82
	Ароматичні гомомоно (полі) циклічні	385	0.8217	0.8011	0.839	0.91
	Всі (1444)	204	0.8335	0.8273	0.844	0.91
RdKit	Аліфатичні ациклічні	74	0.8589	0.8223	0.8878	0.91
	Аліфатичні гетеромоно (полі) циклічні	56	0.8314	0.7647	0.7922	0.92
	Ароматичні гетеромоно (полі) циклічні	89	0.8456	0.8324	0.8613	0.93
	Ароматичні гомомоно (полі) циклічні	62	0.8304	0.8125	0.8372	0.90
	Всі (196)	100	0.8285	0.8217	0.8354	0.91
Mordred	Аліфатичні ациклічні	465	0.8292	0.8075	0.8585	0.93
	Аліфатичні гетеромоно (полі) циклічні	170	0.8296	0.7947	0.8922	0.96
	Ароматичні гетеромоно (полі) циклічні	397	0.8413	0.8285	0.8538	0.92
	Ароматичні гомомоно (полі) циклічні	374	0.8315	0.8163	0.845	0.92
	Всі (1613)	320	0.8306	0.8108	0.849	0.9

Відповідно до отриманих результатів оцінки ефективності моделей на етапі крос-валідації можна побачити, що при зменшенні кількості вхідних даних у діапазоні від 55% до 90% від початкової кількості дескрипторів (PaDel, Mordred та RDkit) призводило, виключно для всіх моделей, до покращення точності (табл. 3.14) *in*

*silico* моделей прогнозування мутагенності хімічних сполук у порівнянні з такими показниками для розроблених бінарних класифікаторів, які на етапі навчання використовували повний набір вхідних даних (табл.3.13). При цьому відсоток збільшення точності (*accuracy*) для орієнтованих на основні структурні класи ксеноіботиків моделей коливався у межах від 0,1% до 2%.

У таблиці 3.15 наведено значення основних метрик оцінки якості Ames/QSAR моделей на основі глибинної нейронної мережі, що були отримані на екзаменаційній вибірці.

Таблиця 3.15

**Класифікаційний звіт на екзаменаційній вибірці для моделей Ames/QSAR з обмеженим переліком релевантних дескрипторів**

Набір даних	Структурний клас	Точність (тестова вибірка)	Точність (екзам. вибірка)	Precision	Recall	Specificity	F1 Score	AUC
PaDell	Аліфатичні ациклічні	<b>0.8881</b>	<b>0.8712</b>	<b>0.9286</b>	<b>0.8</b>	<b>0.9403</b>	<b>0.8595</b>	<b>0.93</b>
	Аліфатичні гетеромоно (полі)	<b>0.9355</b>	<b>0.9365</b>	<b>0.9333</b>	<b>0.9333</b>	<b>0.9394</b>	<b>0.9333</b>	<b>0.96</b>
	Ароматичні гетеромоно (полі)	<b>0.854</b>	<b>0.8656</b>	<b>0.8571</b>	<b>0.8734</b>	<b>0.858</b>	<b>0.8652</b>	<b>0.93</b>
	Ароматичні гомомоно (полі)	<b>0.8364</b>	<b>0.8485</b>	<b>0.8448</b>	<b>0.8647</b>	<b>0.8313</b>	<b>0.8547</b>	<b>0.91</b>
	Всі	0.8335	0.8441	0.8667	0.8217	0.8676	0.8436	0.92
RDKit	Аліфатичні ациклічні	0.8293	0.8496	0.85	0.8226	0.8732	0.8361	0.93
	Аліфатичні гетеромоно (полі)цикліч	<b>0.8615</b>	<b>0.873</b>	<b>0.875</b>	<b>0.875</b>	<b>0.871</b>	<b>0.875</b>	<b>0.93</b>
	Ароматичні гетеромоно (полі)цикліч	<b>0.8727</b>	<b>0.8625</b>	<b>0.8706</b>	<b>0.8706</b>	<b>0.8533</b>	<b>0.8706</b>	<b>0.92</b>
	Ароматичні гомомоно (полі)цикліч	0.8273	0.8308	0.8207	0.8678	0.7898	0.8436	0.89
	Всі	0.8383	0.8462	0.8424	0.8321	0.859	0.8372	0.92

Продовження таблиці 3.15

Набір даних	Структурний клас	Точність (тестова вибірка)	Точність (екзам. вибірка)	Precision	Recall	Specificity	F1 Score	AUC
Mordred	Аліфатичні ациклічні	<b>0.8806</b>	<b>0.8712</b>	<b>0.9091</b>	<b>0.8065</b>	<b>0.9286</b>	<b>0.8547</b>	<b>0.93</b>
	Аліфатичні гетеромоно (полі)	<b>0.9231</b>	<b>0.873</b>	<b>0.8056</b>	<b>0.9667</b>	<b>0.7879</b>	<b>0.8788</b>	<b>0.93</b>
	Ароматичні гетеромоно (полі)	<b>0.8634</b>	<b>0.8875</b>	<b>0.894</b>	<b>0.871</b>	<b>0.903</b>	<b>0.8824</b>	<b>0.93</b>
	Ароматичні гомомоно (полі)	<b>0.8489</b>	<b>0.8636</b>	<b>0.8728</b>	<b>0.8678</b>	<b>0.859</b>	<b>0.8703</b>	<b>0.92</b>
	Всі	0.8403	0.8428	0.8647	0.8304	0.8564	0.8472	0.91

Аналіз класифікаційного звіту (таб. 3.15) дозволяє зробити висновок про те, що 10 з 12 моделей (значення метрик найкращих моделей у табл. 3.15 виділені жирим шрифтом), які на етапі навчання використовували однорідні оптимізовані набори даних, що були сформовані відповідно до основних структурних класів ксенобіотиків, зберігали високі показники класифікації. При цьому моделі, орієнтовані на основні структурні класи ксенобіотиків, які на етапі навчання використовували повні набори даних (молекулярні дескриптори PaDel, RDkit та Mordred), без оптимізації, зазвичай демонстрували зниження ефективності. Крім того, всі моделі, які на етапі тренування використовували дескриптори, розраховані на основі 80% повної бази даних (що відповідає стандартному підходу вирішення задачі бінарної класифікації), також демонстрували гірші результати класифікації. Отримані результати оцінки ефективності Ames/QSAR моделей на основі глибинної мережі підтверджують сформульовану нами гіпотезу, відповідно до якої орієнтовані на основні структурні класи бінарні класифікатори, процедура навчання яких відбувалась з урахуванням обмеженого набору релевантних дескрипторів, можуть демонструвати вищу ефективність у задачах бінарної класифікації.

У таблиці 3.16 наведено перелік Ames/QSAR моделей, що були побудовані на основі методу випадкового лісу, екстремального градієнтного бустінгу та глибинної нейронної мережі, з базовими метриками оцінки якості, що показали найкращі результати класифікації на екзменаційній вибірці.

Таблиця 3.16

**Порівняння прогностичної здатності моделей Ames/QSAR на основі ансамблевих алгоритмів машинного навчання та нейромережевого підходу**

Набір даних/назва методу	Структурний клас	Точність (accuracy)	Precision	Recall	Specificity	F1 Score	AUC
RDKit/ RF	Аліфатичні ациклічні	0,8947	0,9412	0,8889	0.9048	0,9443	0,96
<b>RDKit/ XGBoost</b>	Аліфатичні ациклічні	<b>0.8923</b>	<b>0.8906</b>	<b>0.9194</b>	<b>0.8654</b>	<b>0.9048</b>	<b>0.96</b>
PaDel/DNN	Аліфатичні ациклічні	0.8712	0.9286	0.8	0.9403	0.8595	0.93
Mordred / RF	Аліфатичні гетеромоно (полі) циклічні	0,8966	0,913	0,84	0.9394	0,875	0,95
Mordred / XGBoost	Аліфатичні гетеромоно (полі) циклічні	0.8103	0.8333	0.8065	0.8148	0.8197	0.9
<b>PaDel/DNN</b>	Аліфатичні гетеромоно (полі) циклічні	<b>0.9365</b>	<b>0.9333</b>	<b>0.9333</b>	<b>0.9394</b>	<b>0.9333</b>	<b>0.96</b>
<b>RDKit/ RF</b>	Ароматичні гетеромоно (полі) циклічні	<b>0,8921</b>	<b>0,913</b>	<b>0,8802</b>	<b>0.9054</b>	<b>0,8963</b>	<b>0,94</b>
RDKit/ XGBoost	Ароматичні гетеромоно (полі) циклічні	0.8571	0.88	0.8302	0.8846	0.8544	0.92
Mordred/ DNN	Ароматичні гетеромоно (полі) циклічні	0.8875	0.894	0.871	0.903	0.8824	0.93
<b>Mordred / RF</b>	Ароматичні гомомоно (полі) циклічні	<b>0,871</b>	<b>0,8741</b>	<b>0,8503</b>	<b>0.8896</b>	<b>0.8621</b>	<b>0,95</b>
Mordred / XGBoost	Ароматичні гомомоно (полі) циклічні	0.8419	0.8491	0.8438	0.84	0.8464	0.92
Mordred/ DNN	Ароматичні гомомоно (полі) циклічні	0.8636	0.8728	0.8678	0.859	0.8703	0.92



Для *in silico* прогнозування мутагенної активності хімічних сполук на основі методу Еймса, що належать до першої групи ксенобіотиків (табл. 2.1) необхідно використовувати Ames/QSAR модель, яка побудована відповідно до методу екстремального градієнтного бустінгу з релевантними дескрипторами RDKit. Оцінка мутагенного потенціалу для потенційних мутагенів, що належать до другої групи ксенобіотиків повинна здійснюватися за допомогою розробленого бінарного класифікатора на основі глибинної нейронної мережі з дескрипторами PaDell. Для прогнозування мутагенності Еймса хімічних сполук, що відносяться до четвертої та п'ятої груп ксенобіотиків (табл. 2.1) необхідно використовувати Ames/QSAR моделі, побудовані на основі методу випадкового лісу з обмеженим набором релевантних дескрипторів RDKit та Mordred відповідно.

### **3.3 Ames/QSAR моделі на основі відбитків молекулярної структури ксенобіотиків**

Ефективність *in silico* моделей прогнозування мутагенності Еймса може залежати від багатьох чинників, що обумовлено використанням різних наборів молекулярних дескрипторів, особливістю організації вхідних даних та методів їх обробки. У межах роботи особлива увага була приділена 2D молекулярним відбиткам структури (molecular fingerprint), що представляють собою бітовий рядок, в якому кожний біт відповідає за наявність/відсутність певної функціональної групи або підструктури на рівні молекули. На сьогоднішній день дослідниками у наукових цілях використовуються різні види відбитків молекулярної структури, що відносяться до трьох [176] класів (субструктурні, топологічним та циркулярні). У науковому відношенні достатньо важливою може бути отримана відповідь на питання щодо ефективності Ames/QSAR моделей, які в якості предикторів використовують, різні типи відбитків молекулярної структури. Нещодавно опубліковані результати досліджень [56,149,174,175], в яких науковці при створенні Ames/QSAR моделей використовують відбитки молекулярної структури, стали стимулом для проведення аналогічних досліджень,

але з урахуванням розробленої у межах роботи методики. Для досягнення поставлених завдань дисертаційного дослідження необхідно оцінити ефективність орієнтованих на основні структурні класи ксенобіотиків Ames/QSAR моделей, що в якості вхідних даних, на етапі навчання моделей, використовували різні види відбитків структури ксенобіотиків. Крім того, у межах роботи, необхідно було порівняти прогностичну здатність *in silico* моделей мутагенності Еймса, які на етапі навчання використовували різні набори 1D та 2D дескрипторів (без урахування відбитків молекулярної структури), з бінарними класифікаторами, які, в якості предикторів, використовували відбитки структури ксенобіотиків. Для отримання об'єктивної оцінки ефективності *in silico* моделей прогнозування мутагенності на основі відбитків молекулярної структури перелік основних методів залишився без змін. У якості основних методів для створення бінарних класифікаторів нами використовувались методи випадкового лісу, екстремального градієнтного бустінгу та глибинна нейронна мережа.

Основна перевага Ames/QSAR моделей, які використовують в якості предикторів відбитки молекулярної структури полягає у відсутності проведення повної процедури препроцесингу. Така особливість дозволяє зменшити часові витрати при створенні бінарних класифікаторів та сприяє розробці Ames/QSAR моделей, для яких можливий негативний вплив не якісно проведеної процедури підготовки даних зводиться до мінімуму. Реалізація Ames/QSAR моделей на основі відбитків молекулярної структури відбувалась без зменшення простору вхідних даних, що пов'язано з особливістю збереження інформації про структуру ксенобіотиків за допомогою бітового рядку. Застосування алгоритму RFECV, що дозволяє видаляти менш впливові ознаки на етапі крос-валідації для даних бінарного типу є не ефективним, потребує великих часових витрат та може знижувати ефективність моделей. Негативний вплив такого підходу пов'язаний з тим, що будь-який біт відбитка структури пов'язаний з хімічною структурою ксенобіотика.

В якості вхідних даних були обрані три відбитка молекулярної структури різної довжини: MACCS (166bit), FCFP (1024 bit) та RDkit (2048 bit), що

відносяться до субструктурних, циркулярних та топологічних відбитків відповідно. Розрахунок дескрипторів здійснювався за допомогою бібліотеки RDkit мови програмування Python.

Початковий етап моделювання був пов'язаний з формуванням тренувальної, тестової та екзаменаційної вибірки, що були отримані випадковим чином у співвідношенні 80:10:10 для 5 груп ксенобіотиків (табл. 2.1). Крім того, подібний розподіл даних був здійснений для всіх ксенобіотиків повної бази даних, в якій було представлено 8454 потенційних мутагенів довкілля. Підготовка даних включала в себе тільки видалення не інформативних для моделей даних (Smiles нотація, приналежність до відповідних структурних класів ксенобіотиків, порядковий номер). У таблиці 3.17 представлені результати класифікації Ames/QSAR моделей на основі методу випадкового лісу.

Таблиця 3.17

**Класифікаційний звіт на екзаменаційній вибірці для Ames/QSAR моделей на основі методу випадкового лісу [206]**

Молекулярні дескриптори	Структурний клас	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

Продовження таблиці 3.17

Молекулярні дескриптори	Структурний клас	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
RDkit	Ароматичні гетеромоно (полі) циклічні	<b>0,85</b>	<b>0,87</b>	<b>0,83</b>	<b>0,87</b>	<b>0,85</b>	<b>0,92</b>
	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	<b>0,82</b>	<b>0,82</b>	<b>0,73</b>	<b>0,88</b>	<b>0,77</b>	<b>0,88</b>
	Аліфатичні гетеромоно (полі) циклічні	<b>0,89</b>	<b>0,94</b>	<b>0,81</b>	<b>0,95</b>	<b>0,87</b>	<b>0,95</b>
	Ароматичні гетеромоно (полі) циклічні	0,83	0,85	0,81	0,85	0,83	0,91
	Ароматичні гомомоно (полі) циклічні	<b>0,85</b>	<b>0,85</b>	<b>0,87</b>	<b>0,82</b>	<b>0,86</b>	<b>0,90</b>
	Всі	0,78	0,80	0,75	0,82	0,77	0,86

Реалізація Ames/QSAR моделей на основі методу випадкового лісу була здійснена з урахуванням підбору емпіричним шляхом значень для шести основних гіперпараметрів, таких як: кількість дерев у лісі (*n\_estimators*), мінімальна кількість листків для поділу внутрішнього вузла (*min\_samples\_split*), мінімальна кількість зразків у вузлі листка (*min\_samples\_leaf*), максимальна кількість ознак для поділу вузла (*max\_features*), критерій розділення вузла в кожному дереві (*criterion*), максимальна глибина дерева (*max\_depth*) [206]. При цьому оцінка ефективності Ames/QSAR моделей на кожному кроці підбору оптимальних гіперпараметрів здійснювалась за допомогою п'ятикратної перехресної перевірки.

Відповідно до отриманих результатів класифікації (табл. 3.17) можемо спостерігати покращення точності для всіх орієнтованих на основні структурні класи ксенобіотиків Ames/QSAR моделей у порівнянні з моделями, для яких процедура навчання відбувалась на основі частині (80%) повної бази даних ксенобіотиків. У таблиці 3.16 базові метрики оцінки якості моделей, які

продемонстрували найкращий результат класифікації, позначені жирним шрифтом. Найкращий результат показали три моделі Ames/QSAR, які в якості вхідних даних використовували дескриптори MACCS, з довжиною бітового рядка, що дорівнює 166bit, що були розраховані для першої, другої та п'ятої групи ксенобіотиків (табл. 2.1). Для оцінки мутагенності ароматичних гетеромоно(полі)циклічних хімічних сполук необхідно використовувати модель, що була отримана на основі дескрипторів RDkit, які були розраховані для четвертої об'єднаної групи ксенобіотиків. Дана модель продемонструвала найкращу точність (з  $AUC$  0,92) у порівнянні з іншими бінарними класифікаторами, що в якості вхідних даних використовували дескриптори FCFP та MACCS.

У таблиці 3.18 представлені результати класифікації для Ames/QSAR моделей на основі методу екстремального градієнтного бустінгу.

Таблиця 3.18

**Класифікаційний звіт на экзаменаційній вибірці для Ames/QSAR моделей на основі методу екстремального градієнтного бустінгу [206]**

Молекулярні дескриптори	Структурний клас	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

Продовження таблиці 3.18

Молекулярні дескриптори	Структурний клас	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
RDkit	Ароматичні гетеромоно (полі) циклічні	<b>0,86</b>	<b>0,87</b>	<b>0,85</b>	<b>0,87</b>	<b>0,86</b>	<b>0,93</b>
	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	<b>0,85</b>	<b>0,86</b>	<b>0,75</b>	<b>0,92</b>	<b>0,81</b>	<b>0,89</b>
	Аліфатичні гетеромоно (полі) циклічні	<b>0,84</b>	<b>0,91</b>	<b>0,72</b>	<b>0,94</b>	<b>0,80</b>	<b>0,93</b>
	Ароматичні гетеромоно (полі) циклічні	0,82	0,82	0,82	0,82	0,82	0,92
	Ароматичні гомомоно (полі) циклічні	<b>0,84</b>	<b>0,84</b>	<b>0,86</b>	<b>0,82</b>	<b>0,85</b>	<b>0,89</b>
	Всі	0,84	0,84	0,84	0,84	0,84	0,91

Реалізація Ames/QSAR моделей на основі методу екстремального градієнтного бустінгу була здійснена з урахуванням підбору емпіричним шляхом значень для дев'яти базових гіперпараметрів, таких як: кількість дерев (*n\_estimators*), мінімальна вага вузла (*min\_child\_weight*), максимальна глибина дерева (*max\_depth*), доля даних, що використовується при побудові дерева (*subsample*), частина ознак, що обираються випадковим чином для кожного дерева (*colsample\_bytree*), L1 (*reg\_alpha*) та L2 (*reg\_lambda*) регуляризація, швидкість навчання (*learning\_rate*), мінімальне зменшення функції втрат (*gamma*). Підбір оптимальних значень гіперпараметрів відбувався за допомогою п'ятикратної крос-валідації [206].

Відповідно до отриманих результатів оцінки ефективності розроблених бінарних класифікаторів найкращу прогностичну здатність показали моделі, що орієнтовані на основні структурні класи ксенобіотиків. При цьому перелік

моделей Ames/QSAR, які продемонстрували найкращі показники точності для чотирьох груп ксенобіотиків (табл. 2.1) не змінився у порівнянні з моделями, що були розроблені на основі методу випадкового лісу. Необхідно зазначити, що дві орієнтовані на структурні класи (аліфатичні ациклічні з дескрипторами Masses та ароматичні гетеромоно(полі)циклічні з дескрипторами RDkit) моделі показали збільшення точності відповідно до метрики *accuracy* на 1% та 3% відповідно, у порівнянні з моделями, що були розроблені на основі методу випадкового лісу. При цьому дві моделі (аліфатичні гетеромоно(полі)циклічні та ароматичні гомомоно(полі)циклічні з дескрипторами MACCS) продемонстрували зниження точності (*accuracy*) на 5% та 1% відповідно. Модель яка була отримана на основі тренувальної вибірки, що була сформована з частини ксенобіотиків повної бази даних, показала високий результат класифікації при прогнозуванні ароматичних гомомоно(полі)циклічних хімічних сполук з дескрипторами MACCS.

У таблиці 3.19 представлені результати бінарної класифікації для Ames/QSAR моделей на основі глибинної мережі, що містить чотири приховані шари, у кожному з яких представлено 128, 256, 128 та 64 нейронів відповідно (рис. 3.2). Кількість нейронів вхідного шару відповідає розміру бітового рядка для відбитків молекулярної структури FCFP (166bit), MACCS (1024) та RDkit (2048 bit). В якості функції активації для прихованих шарів була обрана функція ReLU.

Таблиця 3.19

**Класифікаційний звіт на экзаменаційній вибірці для Ames/QSAR моделей на основі глибинної нейронної мережі [206]**

Молекулярні дескриптори	Структурний клас	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
FCFP	Аліфатичні ациклічні	0,82	0,81	0,74	0,88	0,77	0,89
	Аліфатичні гетеромоно(полі) циклічні	<b>0,83</b>	<b>0,83</b>	<b>0,80</b>	<b>0,86</b>	<b>0,81</b>	<b>0,88</b>
	Ароматичні гетеромоно(полі) циклічні	<b>0,87</b>	<b>0,89</b>	<b>0,84</b>	<b>0,89</b>	<b>0,86</b>	<b>0,93</b>

Продовження таблиці 3.19

Молекулярні дескриптори	Структурний клас	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
FCFP	Ароматичні гомомоно (полі) циклічні	0,81	0,83	0,82	0,80	0,82	0,89
	Всі	0,83	0,84	0,82	0,85	0,83	0,89
RDkit	Аліфатичні ациклічні	0,79	0,78	0,68	0,86	0,73	0,85
	Аліфатичні гетеромоно (полі) циклічні	0,80	0,84	0,69	0,89	0,76	0,89
	Ароматичні гетеромоно (полі) циклічні	0,84	0,89	0,79	0,89	0,84	0,92
	Ароматичні гомомоно (полі) циклічні	<b>0,84</b>	<b>0,85</b>	<b>0,85</b>	<b>0,83</b>	<b>0,85</b>	<b>0,89</b>
	Всі	0,84	0,84	0,83	0,85	0,83	0,90
MACCS	Аліфатичні ациклічні	<b>0,83</b>	<b>0,79</b>	<b>0,80</b>	<b>0,85</b>	<b>0,80</b>	<b>0,88</b>
	Аліфатичні гетеромоно (полі) циклічні	0,82	0,79	0,83	0,81	0,81	0,89
	Ароматичні гетеромоно (полі) циклічні	0,84	0,88	0,79	0,89	0,83	0,92
	Ароматичні гомомоно (полі) циклічні	0,80	0,81	0,83	0,78	0,82	0,88
	Всі	0,83	0,83	0,83	0,83	0,83	0,90

На вихідному шарі функцією активації є Sigmoid, що дає змогу отримати значення ймовірності щодо приналежності ксенобіотиків до одного з двох класів (мутаген/не мутаген). Між всіма внутрішніми шарами мережі задіяний механізм Dropout, який випадковим чином вимикає 30 % нейронів під час навчання, що забезпечує більшу стійкість, покращує прогностичну здатність моделей на нових вхідних даних та запобігає перенавчанню моделей. Навчання нейронної мережі було здійснено протягом 100 епох з розміром партії 64. Бінарна крос-ентропія



була обрана в якості функції втрат, що зазвичай використовують для вирішення задачі бінарної класифікації. Мінімізація функції втрат досягалась через застосування класичного оптимізатора Adam.

Серед орієнтованих на основні структурні класи ксенобіотиків Ames/QSAR моделей, що були побудовані нами на основі глибинної нейронної мережі, тільки одна модель (ароматичні гетеромоно(полі) циклічні з дескрипторами FCFP) продемонструвала найвищу точність у порівнянні з моделями, що були розроблені на основі двох ансамблевих алгоритмів машинного навчання (випадкового лісу та екстремального градієнтного бустінгу).

У таблиці 3.20 наведено перелік Ames/QSAR моделей на основі методу випадкового лісу, екстремального градієнтного бустінгу та глибинної нейронної мережі з базовими метриками оцінки якості, що показали найкращі результати класифікації на екзаменаційній вибірці. В якості вхідних даних розроблені моделі використовували класичні набори одновимірних та двовимірних дескрипторів (PaDel, Mordred та RDkit), а також відбитки молекулярної структури (FCFP, RDkt та MACCS), що відносяться до 2D дескрипторів.

Таблиця 3.20

**Порівняння ефективності орієнтованих на основні структурні класи  
ксенобіотиків Ames/QSAR моделей**

Набір даних/назва методу	Структурний клас	Точність (accuracy)	Precision	Recall	Specificity	F1 Score	AUC
RDKit/ RF	Аліфатичні ациклічні	0,8947	0,9412	0,8889	0.9048	0,9443	0,96
<b>RDKit/ XGBoost</b>	Аліфатичні ациклічні	<b>0.8923</b>	<b>0.8906</b>	<b>0.9194</b>	<b>0.8654</b>	<b>0.9048</b>	<b>0.96</b>
PaDel/DNN	Аліфатичні ациклічні	0.8712	0.9286	0.8	0.9403	0.8595	0.93
MACCS/ XGBoost	Аліфатичні ациклічні	0,85	0,86	0,75	0,92	0,81	0,89
Mordred / RF	Аліфатичні гетеромоно (полі) циклічні	0,8966	0,913	0,84	0.9394	0,875	0,95

## Продовження таблиці 3.20

Набір даних/назва методу	Структурний клас	Точність (accuracy)	Precision	Recall	Specificity	F1 Score	AUC
Mordred / XGBoost	Аліфатичні гетеромоно (полі) циклічні	0.8103	0.8333	0.8065	0.8148	0.8197	0.9
<b>PaDel/DNN</b>	Аліфатичні гетеромоно (полі) циклічні	<b>0.9365</b>	<b>0.9333</b>	<b>0.9333</b>	<b>0.9394</b>	<b>0.9333</b>	<b>0.96</b>
MACCS/RF	Аліфатичні гетеромоно (полі) циклічні	0,89	0,94	0,81	0,95	0,87	0,95
<b>RDKit/ RF</b>	Ароматичні гетеромоно (полі) циклічні	<b>0,8921</b>	<b>0,913</b>	<b>0,8802</b>	<b>0.9054</b>	<b>0,8963</b>	<b>0,94</b>
RDKit/ XGBoost	Ароматичні гетеромоно (полі) циклічні	0.8571	0.88	0.8302	0.8846	0.8544	0.92
Mordred/ DNN	Ароматичні гетеромоно (полі) циклічні	0.8875	0.894	0.871	0.903	0.8824	0.93
FCFP/ DNN	Ароматичні гетеромоно (полі) циклічні	0,87	0,89	0,84	0,89	0,86	0,93
<b>Mordred / RF</b>	Ароматичні гомомоно (полі) циклічні	<b>0,871</b>	<b>0,8741</b>	<b>0,8503</b>	<b>0.8896</b>	<b>0.8621</b>	<b>0,95</b>
Mordred / XGBoost	Ароматичні гомомоно (полі) циклічні	0.8419	0.8491	0.8438	0.84	0.8464	0.92
Mordred/ DNN	Ароматичні гомомоно (полі) циклічні	0.8636	0.8728	0.8678	0.859	0.8703	0.92
MACCS/RF	Ароматичні гомомоно (полі) циклічні	0,85	0,85	0,87	0,82	0,86	0,90

Серед розроблених Ames/QSAR моделей прогнозування мутагенності, що в якості предикторів використовували відбитки просторової структури, жодна модель не продемонстрували кращу точність, у порівнянні з бінарними класифікаторами, для яких процедура навчання відбувалась з урахуванням класичних 1D та 2D дескрипторів. При цьому Ames/QSAR моделі на основі дескрипторів FCFP, RDkit та MACCS були ефективними та дозволили з високими показниками точності

визначити мутагенність Еймса. Така оцінка відповідає зареєстрованій варіабельності *in vitro* тесту Еймса. Основною перевагою такого підходу є відсутність проведення класичної процедури препроцесингу даних, що спрощує процес реалізації моделей прогнозування мутагенності та зменшує часові витрати.

### 3.4 Ames/QSAR моделі на основі структурних маркерів мутагенності

Методологія побудови сучасних *in silico* моделей прогнозування мутагенності Еймса факторів навколишнього середовища ґрунтується на двох основних підходах: на основі статистики та на основі правил [8]. Моделі, що відповідають першому підходу побудовані відповідно до набору вхідних даних, що задаються молекулярними дескрипторами. Фундаментом моделей прогнозування мутагенності Еймса на основі правил являються ідентифіковані підструктури або функціональні групи, наявність яких на рівні молекули може виступати у ролі базового критерію, що лежить в основі поділу ксенобіотиків на два класи (мутаген/ не мутаген). Серед Ames/QSAR моделей, що дозволяють отримати генетичну оцінку факторів навколишнього середовища на основі ідентифікованих маркерів мутагенності, найбільш поширеними є методи на основі графів та на основі відбитків молекулярної структури [207]. У рамках дисертаційного дослідження було акцентовано увагу на моделях оцінки мутагенності Еймса, що в якості вхідних даних використовували відбитки молекулярної структури ксенобіотиків. В основі розроблених Ames/QSAR моделей на основі правил була взята за основу сформульована науковою спільнотою концепція хімічної подібності, відповідно до якої сполуки з подібною молекулярною структурою можуть володіти подібними властивостями. Такий підхід дозволяє для двох хімічних сполук з великим показником подібності, одна з яких є мутагеном, а інша – не мутагеном, через порівняння їх структурних формул ідентифікувати функціональну групу(и) або підструктуру(и), що може лежати в основі прояву мутагенності. Оцінка подібності між ксенобіотиками (мутаген/не мутаген) відбувалась за допомогою класичних метрик Танімото та

Хемінга, що достатньо часто використовуються для вирішення подібних задач [151,208]. Очевидно, що спростити складний процес пошуку подібних хімічних сполук між парами (мутаген/не мутаген) можна в тому випадку, коли метрики подібності будуть розраховані для відповідних груп ксенобіотиків (табл.2.1), що мають спільні риси будови молекулярного каркасу. В якості вхідних даних для Ames/QSAR моделей використовувалися відбитки молекулярної структури MACCS, RDkit та FCFP, що відносяться до трьох класів: субструктурні ключі, топологічні та циркулярні відбитки структури [209]. Оцінка попарних відстаней між всіма мутагенами та не мутагенами у межах 5 груп ксенобіотиків зберігалась у вигляді багатовимірної матриці. Такий формат даних є не зручним для роботи, та ускладнює пошук схожих за структурними ознаками ксенобіотиків. Аналіз нещодавно опублікованих наукових праць [151,210,211] дозволив обрати оптимальний алгоритм t-SNE (Т-розподіленого вкладення стохастичної близькості), який використовується для візуалізації багатовимірних даних на площині. Такий підхід є достатньо ефективним та дозволяє серед великої кількості хімічних сполук обрати пари з різними значеннями метрик відстані. Аналіз структурних формул між парами мутаген/не мутаген дозволяє ідентифікувати певні підструктури або/та функціональні групи, що можуть бути маркерами мутагенності. На рисунку 3.3, в якості прикладу, представлена візуалізація t-SNE у просторі дескрипторів FCFP для всіх ксенобіотиків, що відносяться до ароматичних гетерополіциклічних та ароматичних гетеромоноциклічних хімічних сполук (табл. 2.1). Відстань між всіма парами ксенобіотиків була розрахована на основі метрики Танімото. Зеленими точками позначаються ксенобіотики, які, відповідно до тесту Еймса, не проявляли мутагенний потенціал. Червоним кольором – хімічні сполуки з вираженими мутагенними властивостями.

Розроблена у межах проведеного дослідження методологія оцінки мутагенних ефектів може бути достатньо ефективною з урахуванням проведення двох послідовних етапів пошуку маркерів мутагенності. Перший етап враховує порівняння структурних формул двох схожих сполук – мутаген/не мутаген у

межах одного з п'яти структурних класів (табл.2.1). Другий – дозволяє зробити висновок про наявні маркери мутагенності відповідно до результатів порівняння декількох схожих між собою хімічних сполук, що проявляють мутагенні властивості. Вибір схожих за структурою формулою ксенобіотиків здійснювався відповідно до отриманої візуалізації t-SNE з урахуванням найменшої відстані між точками. Відповідно, найбільша схожість буде між тими ксенобіотиками, між якими буде спостерігатися найменша лінійна відстань на площині.

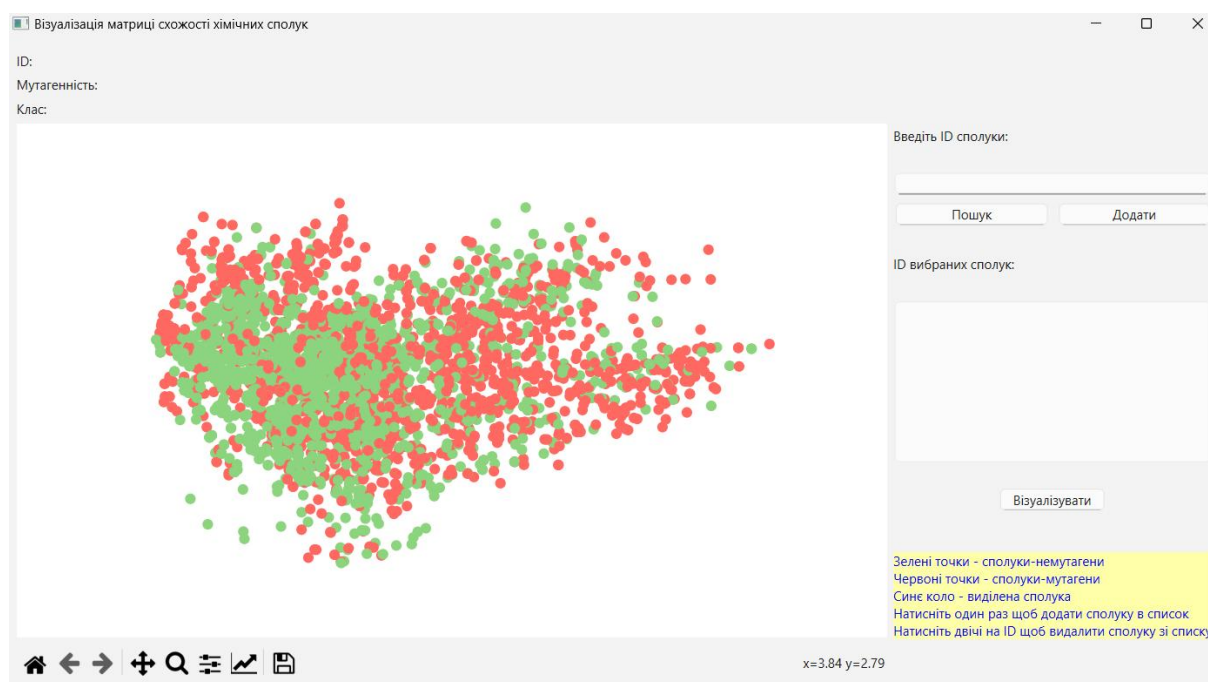


Рис. 3.3. Візуалізація t-SNE даних ароматичних гетерополіциклічних та ароматичних гетеромоноциклічних хімічних сполук у просторі FCFP на основі розрахованих відстаней Танімото [212]

На рисунку 3.4, в якості прикладу, показані пари ксенобіотиків (мутаген/не мутаген) між якими спостерігається високий рівень подібності відповідно до розрахованої метрики Танімото. Аналіз структурних формул пари (мутаген/не мутаген) дозволив сформулювати гіпотезу щодо причини мутагенності хімічної сполуки з ідентифікатором ID2011 (відповідає порядковому номеру ксенобіотика повної бази даних).

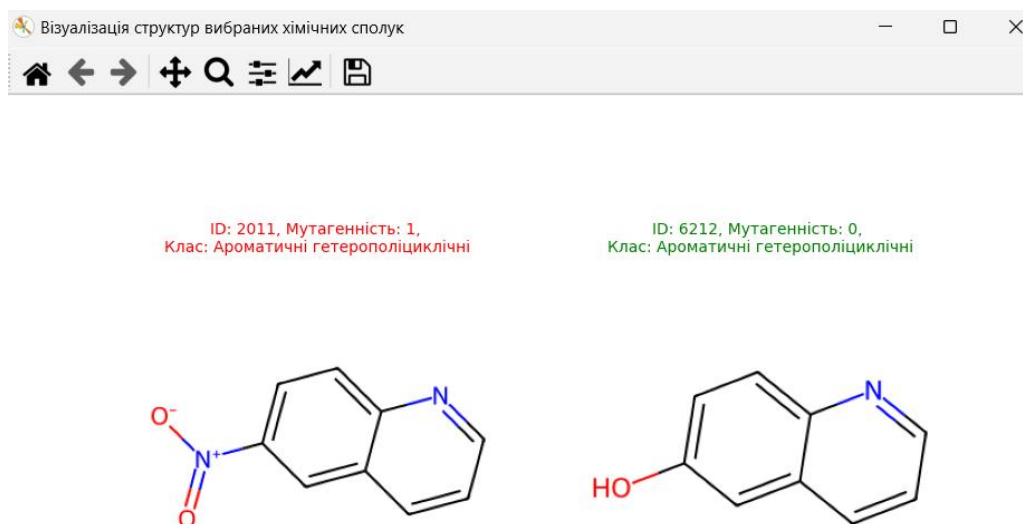


Рис. 3.4. Структурні формули пари мутаген/не мутаген, що відносяться до класу гетеромоноциклічних та гетерополіциклічних сполук [212]

Заміна  $OH$  групи на  $NO_2$  призвела до того, що ксенобіотик з ID6212, який, відповідно до тесту Еймса, не проявляв мутагенні властивості, перетворився на хімічну сполуку з ID2011, що є мутагеном. Досить важливими, з наукової точки зору, є результати порівняння ксенобіотиків-мутагенів (рисунок 3.5) з ID2607, ID 5706, ID5831 та ID3871, між якими, відповідно до розрахованого індекса подібності Танімото, спостерігається мінімальна відстань. Для кожного з чотирьох ксенобіотиків наявна група  $NO_2$ , та всі вони є мутагенами. Група  $NO_2$  для ароматичних хімічних сполук може виступати в ролі маркера мутагенності. Крім того, присутність групи  $NH_2$  в структурі досліджуваних ксенобіотиків також може мати негативний вплив з точки зору прояву генетичних ефектів. Результати опублікованих досліджень [213] підтверджують сформульовану гіпотезу щодо приналежності групи  $NO_2$  та  $NH_2$  до таких, що можуть бути пов'язані з мутагенністю. Остаточний висновок про мутагенність певного ксенобіотика, яка пов'язана з наявною функціональною групою або підструктурою, має бути сформульований з урахуванням великої кількості експериментальних даних, що є надзвичайно складною задачею.

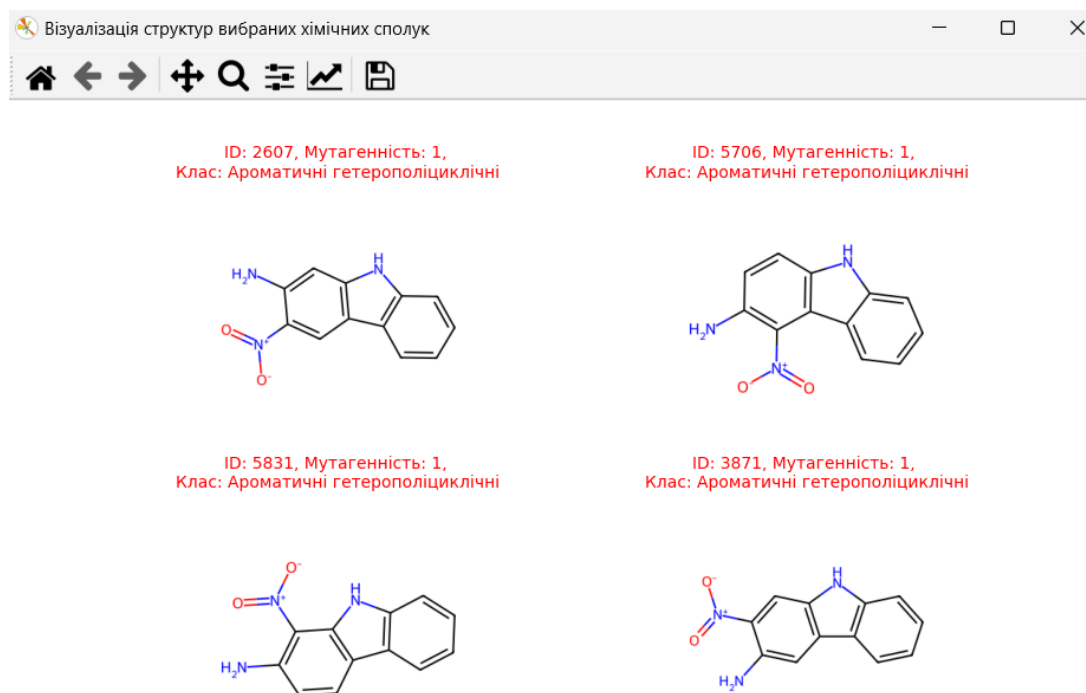


Рис. 3.5 Структурні формули ксенобіотиків – мутагенів, що відносяться до класу гетеромоноциклічних та гетерополіциклічних сполук [212]

Оцінка мутагенності повинна ґрунтуватися на комплексному підході, що враховує фізико-хімічні, просторові, електронні тощо властивості досліджуваних ксенобіотиків. В цьому контексті будь яка група або підструктура, з якою пов'язують мутагенність для відповідного структурного класу ксенобіотиків, може не бути маркером мутагенності для іншого класу хімічних сполук. Запропонована у межах роботи методика може бути застосована для детального аналізу причинно-наслідкових зв'язків між мутагенністю та наявними/відсутніми функціональними групами або підструктурами. Цікавими у науковому відношенні можуть бути висновки щодо мутагенності ксенобіотиків докільля з урахуванням обмеженого переліку релевантних 1D та 2D молекулярних дескрипторів.

### 3.5 Причинно-наслідкові зв'язки між мутагенністю та релевантними дескрипторами основних структурних класів ксенобіотиків

Не зважаючи на те, що наукова спільнота значну увагу приділяє питанням розробки ефективних *in silico* моделей прогнозування мутагенності Еймса, на сьогоднішній день потребує вирішення проблема науково обґрунтованої інтерпретації мутагенної дії ксенобіотиків на спадковий апарат людини. У цьому контексті, мутагенність можна розглядати через призму переліку властивостей, що можуть лежати в основі проявів генотоксичного потенціалу ксенобіотиків. Застосування розроблених у межах роботи Ames/QSAR моделей дозволяє, у першу чергу, вирішити задачу бінарної класифікації з розподілом ксенобіотиків на дві групи (мутаген/не мутаген) та, по друге, отримати набір ключових, з точки зору проявів мутагенності, молекулярних дескрипторів. Дослідження особливостей складних механізмів прямої або опосередкованої дії потенційних мутагенів на генетичний апарат людини, може бути здійснено з урахуванням переліку релевантних дескрипторів, що мають суттєвий вплив на прогнозовану змінну бінарного типу. Відбір найбільш впливових ознак був здійснений за допомогою алгоритму RFECV на основі методу екстремального градієнтного бустінгу та випадкового лісу. У межах проведеного дисертаційного дослідження було запропоновано провести аналіз причинно-наслідкових зв'язків між проявами мутагенності та наборами релевантних молекулярних дескрипторів для чотирьох основних структурних класів ксенобіотиків, що були отримані відповідно до Ames/QSAR моделей на основі екстремального градієнтного бустінгу. Такий вибір методу був пов'язаний у першу чергу з тим, що бінарні класифікатори на основі екстремального градієнтного бустінгу дають можливість отримати менший за кількістю перелік релевантних дескрипторів у порівнянні з моделями на основі випадкового лісу. Предиктори Ames/QSAR моделей RDkit, що представляють собою найменшу за кількістю групу молекулярних дескрипторів, у порівнянні з PaDel та Mordred, досить зручно використовувати для визначення першопричини мутагенності ксенобіотиків, з урахуванням тих ознак, що мають вагомий вплив на прогнозовану змінну. Особливу увагу у роботі було приділено молекулярним дескрипторам RDkit, які були представлені в моделях один раз. Аналіз таких предикторів дозволив отримати інформацію про перелік властивостей, що можуть



лежати в основі прояву мутагенності для окремих однорідних груп ксенобіотиків, що мають спільну будову молекулярного каркасу. Крім того, були отримані найважливіші дескриптори RDkit, які кілька разів (від 2 до 4) зустрічались в моделях, що дає можливість більш поглибленого розуміння причин мутагенності для всіх хімічних сполук, що можуть бути представлені у довкіллі. Використання обмежених наборів релевантних молекулярних дескрипторів, що були отримані у межах дисертаційного дослідження, дозволили отримати орієнтовані на структурні класи Ames/QSAR моделі, для яких властиві високі показники класифікації. Перелік релевантних молекулярних дескрипторів RDkit, отриманих за допомогою Ames/QSAR моделей на основі методу екстремального градієнтного бустінгу представлений у додатку Б.

У таблиці 3.21 представлений перелік релевантних молекулярних дескрипторів RDkit, що був отриманий за допомогою *in silico* моделей прогнозування мутагенності Еймса на основі метода екстремального градієнтного бустінга для аліфатичних ациклічних хімічних сполук. Релевантні дескриптори, які зустрічались в інших орієнтованих на структурі класи моделях, записані у таблиці 3.21 жирним шрифтом. Предиктори, що позначені курсивом є унікальними для даного класу хімічних сполук (аліфатичних ациклічних).

Таблиця 3.21

**Релевантні молекулярні дескриптори RDkit для аліфатичних ациклічних хімічних сполук**

<b>BalabanJ</b>	<b>EState_VSA8</b>	<b>NOCCount</b>	<b>PEOE_VSA5</b>	<b>fr_Al_OH</b>
<b>BertzCT</b>	<b>MaxAbsEStateIndex</b>	<b>NHOHCount</b>	<b>PEOE_VSA7</b>	<b>VSA_EState8</b>
<b>Chi0</b>	<b>MaxAbsPartialCharge</b>	<b>SlogP_VSA10</b>	<b>PEOE_VSA8</b>	<b>VSA_EState9</b>
<i>Chi0v</i>	<i>ExactMolWt</i>	<b>PEOE_VSA9</b>	<b>SlogP_VSA2</b>	<b>SlogP_VSA6</b>
<i>Chi2v</i>	<b>FractionCSP3</b>	<b>SMR_VSA1</b>	<b>SlogP_VSA12</b>	<b>fr_allylic_oxid</b>
<i>Chi3v</i>	<b>HallKierAlpha</b>	<b>NumHeteroatoms</b>	<b>SMR_VSA9</b>	<b>fr_amide</b>
<b>EState_VSA10</b>	<b>Ipc</b>	<b>PEOE_VSA1</b>	<i>SMR_VSA2</i>	<b>fr_halogen</b>
<b>EState_VSA1</b>	<b>Kappa2</b>	<b>PEOE_VSA10</b>	<b>SMR_VSA5</b>	<b>fr_alkyl_halide</b>
<b>EState_VSA4</b>	<i>Kappa3</i>	<b>PEOE_VSA12</b>	<b>SMR_VSA4</b>	<b>fr_aldehyde</b>
<b>EState_VSA5</b>	<b>MinAbsPartialCharge</b>	<b>PEOE_VSA13</b>	<b>SMR_VSA7</b>	<i>fr_unbrch_alkane</i>

## Продовження таблиці 3.21

EState_VSA2	MinPartialCharge	PEOE_VSA4	SMR_VSA6	fr_NH2
EState_VSA3	MaxPartialCharge	PEOE_VSA14	SlogP_VSA3	fr_NH1
EState_VSA6	MinAbsEStateIndex	PEOE_VSA2	SlogP_VSA1	fr_nitroso
EState_VSA7	MolLogP	PEOE_VSA3	fr_N_O	fr_nitro
EState_VSA9	NumHAcceptors	PEOE_VSA6	fr_NH0	

Унікальні для аліфатичних ациклічних хімічних сполук релевантні дескриптори *Chi0v*, *Chi2v*, *Chi3v* враховують об'ємні характеристики та особливості укладки молекули у просторі [165]. Дескриптор *ExactMolWt* відповідає за позначення молекулярної ваги ксенобіотиків. *Kappa3* є одним з топологічних дескрипторів, який розраховується на основі міжатомних взаємодій, що лежить в основі визначення структурної подібності між молекулами. *NumHAcceptors* – це дескриптор, який оцінює кількість акцепторів водневого зв'язку. Цікавим є той факт, що відповідно до правил п'яти Ліпінського даний параметр виступає в якості основного критерію, який необхідно враховувати на етапі відбору лігандів, які можуть бути лікарськими препаратами [160]. Дескриптор *SMR\_VSA2* дозволяє оцінити полярну поверхню молекули ксенобіотика, яка лежить в основі формування слабких Ван-дер-Ваальсових взаємодій, що можуть бути визначальними з точки зору прояву мутагенності. Унікальний для аліфатичних ациклічних сполук дескриптор *fr\_unbrch\_alkane*, який позначає наявність у структурі молекули лінійних алканів. Молекулярний дескриптор *fr\_NH1* є показником наявності у структурі молекули ксенобіотика аміногрупи  $NH_2$  [165]. Нітроген, який входить у склад аміногрупи може виступати у ролі як донору так і акцептору водневого зв'язку, в залежності від того, з якою молекулою він взаємодіє. Така особливість лежить в основі прояву мутагенних ефектів, що можуть бути індуковані аліфатичними ациклічними хімічними сполуками, в структурі яких є група  $NH_2$ . Серед релевантних дескрипторів (табл.3.21) заслуговують на увагу дескриптори, які пов'язані з функціональними групами, що лежать в основі прояву мутагенності та використовувались в якості основних

предикторів для прогнозування мутагенності Еймса інших структурних класів ксенобіотиків. Мутагенна дія на спадковий апарат людини аліфатичних ациклічних хімічних сполук може бути пов'язана з наступними дескрипторами: *fr\_N\_O* – позначає нітроген, який приєднаний до кисню в складі нітро групи ( $\text{NO}_2$ ) або нітросо-групи ( $-\text{N} = \text{O}$ ); *fr\_NH0* – група ( $\text{NH}$ ); *fr\_Al\_OH* – гідроксид алюмінію; *fr\_allylic\_oxid* – позначає алільну групу; *fr\_amide* – позначає амід; *fr\_halogen* – вказує на наявність галогенного атома в структурі ксенобіотика; *fr\_alkyl\_halide* – відповідає алкілгалогеніду; *fr\_aldehyde* – вказує на наявність альдегідної групи; *fr\_NH2* – аміно група ( $\text{NH}_2$ ); *fr\_nitroso* – наявність нітросо групи ( $\text{NO}$ ); *fr\_nitro* – наявність нітро групи ( $\text{NO}_2$ ).

У таблиці 3.22 наведено перелік релевантних молекулярних дескрипторів RDkit, що був отриманий за допомогою *in silico* моделей прогнозування мутагенності Еймса на основі метода екстремального градієнтного бустінга для аліфатичних гетеромоно(полі)циклічних хімічних сполук.

Таблиця 3.22

**Релевантні молекулярні дескриптори RDkit для аліфатичних  
гетеромоно(полі)циклічних хімічних сполук**

BalabanJ	HallKierAlpha	PEOE_VSA14	SMR_VSA10	fr_aldehyde
BertzCT	MaxPartialCharge	PEOE_VSA2	SMR_VSA1	fr_alkyl_hal
EState_VSA1	MaxAbsPartialCharge	PEOE_VSA3	PEOE_VSA9	fr_nitroso
EState_VSA2	Kappa2	PEOE_VSA12	PEOE_VSA8	fr_oxime
EState_VSA4	NumAliphaticCarbocycles	PEOE_VSA1	SMR_VSA5	fr_ketone
EState_VSA3	NOCOUNT	PEOE_VSA10	SMR_VSA3	fr_piperzine
EState_VSA6	NHOHCount	PEOE_VSA11	fr_allylic_oxid	fr_epoxide
EState_VSA5	MinAbsEStateIndex	NumRotatableBond	fr_amide	fr_halogen
EState_VSA9	MinAbsPartialCharge	SlogP_VSA2	fr_ether	
FractionCSP3	MolLogP	SlogP_VSA3	fr_NH0	
EState_VSA7	SMR_VSA7	SlogP_VSA1	fr_Al_OH	
EState_VSA8	PEOE_VSA4	SMR_VSA6	SlogP_VSA5	

Для аліфатичних гетеромоно(полі)циклічних хімічних сполук було отримано тільки два унікальних молекулярних дескриптора *fr\_ether* та *fr\_oxime*,

що відповідають за наявність ефірної ( $R - O - R$ ) та оксимної груп ( $-C = NOH$ ) у структурі молекули ксенобіотика.

Серед релевантних дескрипторів (табл.3.22) заслуговують на увагу дескриптори, які пов'язані з функціональними групами, що лежать в основі прояву мутагенності та використовувались в якості основних предикторів для прогнозування мутагенності Еймса інших структурних класів ксенобіотиків. Мутагенний ефект, що може бути індукований впливом аліфатичних гетеромоно(полі)циклічних хімічних сполук на спадковий апарат людини пов'язаний з наступними дескрипторами: *fr\_allylic\_oxid* – позначає алільну групу; *fr\_amide* – позначає амід; *fr\_aldehyde* – вказує на наявність альдегідної групи; *fr\_alkyl\_halide* – відповідає алкілгалогеніду; *fr\_nitroso* – позначає нітрозно-групу ( $-N = O$ ); *fr\_ketone* – вказує на наявність кетонної групи; *fr\_piperzine* – відповідає за наявність піперазинового кільця ( $C_4H_{10}N_2$ ); *fr\_epoxide* – наявність епоксидної групи; *fr\_halogen* – вказує на наявність галогенного атома в структурі ксенобіотика.

У таблиці 3.23 наведено перелік релевантних молекулярних дескрипторів RDkit, що був отриманий за допомогою *in silico* моделей прогнозування мутагенності Еймса на основі метода екстремального градієнтного бустінга для ароматичних гетеромоно(полі)циклічних хімічних сполук.

Таблиця 3.23

**Релеванті молекулярні дескриптори RDkit для ароматичних  
гетеромоно(полі)циклічних хімічних сполук**

<b>BalabanJ</b>	<i>fr_azide</i>	<i>fr_Ar_NH</i>	<b>NumRotatable Bonds</b>	<b>SlogP_VSA1</b>
<b>BertzCT</b>	<b>fr_aniline</b>	<b>fr_alkyl_halide</b>	<i>NumAromatic Rings</i>	<b>SMR_VSA9</b>
<b>Chi0</b>	<b>fr_amide</b>	<b>fr_ketone</b>	<b>SMR_VSA7</b>	<b>SMR_VSA5</b>
<b>EState_VSA1</b>	<b>SlogP_VSA8</b>	<b>Ipc</b>	<b>SMR_VSA4</b>	<b>SMR_VSA6</b>
<b>EState_VSA10</b>	<b>VSA_EState9</b>	<i>NumAliphaticRings</i>	<b>PEOE_VSA8</b>	<b>RingCount</b>
<b>EState_VSA2</b>	<i>VSA_EState10</i>	<b>MaxAbsEStateIndex</b>	<b>PEOE_VSA9</b>	<b>SMR_VSA1</b>
<b>EState_VSA3</b>	<b>VSA_EState8</b>	<b>MaxAbsPartial Charge</b>	<b>PEOE_VSA1</b>	<b>SMR_VSA10</b>
<b>EState_VSA4</b>	<i>fr_Ndealkylation</i>	<b>MaxPartialCharge</b>	<b>PEOE_VSA10</b>	<b>SMR_VSA3</b>

## Продовження таблиці 3.23

EState_VSA8	fr_NH0	MinAbsEStateIndex	PEOE_VSA12	SlogP_VSA10
EState_VSA5	fr_NH2	MinEStateIndex	PEOE_VSA11	SlogP_VSA7
EState_VSA6	fr_Ar_N	MinPartialCharge	PEOE_VSA2	SlogP_VSA5
EState_VSA7	fr_sulfonamd	MolLogP	PEOE_VSA3	SlogP_VSA6
SlogP_VSA2	fr_piperzine	NumAromatic Heterocycles	PEOE_VSA13	SlogP_VSA4
SlogP_VSA12	fr_pyridine	NHOHCount	PEOE_VSA14	SlogP_VSA3
fr_epoxide	fr_para_ hydroxylation	NOCCount	PEOE_VSA4	
fr_halogen	fr_nitroso	NumAliphatic Carbocycles	PEOE_VSA5	
fr_hdrzine	fr_nitro	NumSaturated Heterocycles	PEOE_VSA7	
fr_bicyclic	fr_thiazole	NumSaturatedRings	PEOE_VSA6	

Опис унікальних для ароматичних гетеромоно(полі)циклічних хімічних сполук релевантних дескрипторів [165], що зустрічались в моделях один раз, наведено у таблиці 3.24. Перелік дескрипторів у таблиці 3.23 дозволяє пов'язати мутагенність ароматичних гетеромоно(полі)циклічних хімічних сполук з наявністю певних функціональних груп або/та структур на рівні досліджуваних ксенобіотиків.

Таблиця 3.24

**Характеристика релевантних унікальних дескрипторів для ароматичних гетеромоно(полі)циклічних хімічних сполук**

	Опис молекулярних дескрипторів
fr_bicyclic	Біциклічна структура
fr_azide	Азидна група ( $N_3$ )
fr_Ar_N	Атом нітрогену в структурі ароматичного кільця
fr_sulfonamd	Сульфонамідна група ( $SO_2NH_2$ )
fr_pyridine	Піридинова група ( $C_5H_5N$ )
fr_thiazole	Тіазолова група ( $C_3H_3NS$ )
fr_Ar_NH	Аміногрупа ( $NH_2$ ) в структурі ароматичного кільця

## Продовження таблиці 3.24

<i>fr_Ndealkylation</i>	Атом нітрогену, що приймає участь у деалкілюванні
<i>NumAromaticHeterocycles</i>	Ароматичні гетероцикли
<i>NumSaturatedHeterocycles</i>	Насичені гетероциклічні структури
<i>NumSaturatedRings</i>	Насичені кільцеві структури
<i>SlogP_VSA4</i>	Оцінка поверхні молекули з вираженими гідрофобними властивостями

При цьому, з урахуванням надзвичайно складних механізмів, що лежать в основі прояву мутагенної дії ксенобіотиків докільля, для отримання їх всебічної генетичної оцінки необхідно застосовувати комплексний підхід, який повинний враховувати фізико-хімічні, просторові, електронні тощо властивості потенційних мутагенів докільля. У цьому контексті, наявність (відсутність) певної функціональної групи або структури (табл. 3.24) може виступати у ролі одного з індикаторів для прогнозування мутагенності, але не є універсальним правилом для її визначення. Така особливість пов'язана з тим, що мутагенний ефект впливу ксенобіотиків на спадковий апарат людини визначається комбінацією достатньо великої кількості факторів (властивостей), які можна враховувати за допомогою різних наборів молекулярних дескрипторів.

Серед релевантних дескрипторів (табл.3.23) заслуговують на увагу дескриптори, які пов'язані з функціональними групами, що лежать в основі прояву мутагенності та використовувались в якості основних предикторів для прогнозування мутагенності Еймса інших структурних класів ксенобіотиків. Предиктори, що мають суттєвий вплив на прояв мутагенності ароматичних гетеромоно(полі)циклічних хімічних сполук: *fr\_aniline* – відповідає за наявність анілінової групи ( $C_6H_5NH_2$ ); *fr\_amide* – позначає амід; *fr\_NH0* – група (NH) *fr\_NH2* – аміногрупа ( $NH_2$ ); *fr\_piperzine* – дає за наявність піперазинового кільця ( $C_4H_{10}N_2$ ); *fr\_para\_hydroxylation* – гідроксильна група у положенні пара-ароматичного кільця; *fr\_nitroso* – наявність нітрузо-групи ( $-N=O$ ); *fr\_nitro* –

наявність нітро групи ( $NO_2$ ); *fr\_alkyl\_halide* – відповідає алкілгалогеніду; *fr\_ketonev* – вказує на наявність кетонної групи.

У таблиці 3.25 представлено перелік релевантних молекулярних дескрипторів RDkit, який був отриманий за допомогою *in silico* моделі прогнозування мутагенності Еймса на основі метода екстремального градієнтного бустінга для ароматичних гомомоно(полі)циклічних хімічних сполук.

Таблиця 3.25

**Релеванті молекулярні дескриптори RDkit для ароматичних  
гомомоно(полі)циклічних хімічних сполук**

EState_VSA1	MaxPartialCharge	PEOE_VSA3	<i>fr_ArN</i>	<b>fr_alkyl_halide</b>
Chi0	MinAbsPartialCharge	PEOE_VSA13	<i>fr_Al_COO</i>	<b>fr_NH0</b>
EState_VSA10	Kappa2	PEOE_VSA11	<b>fr_para_hydroxylation</b>	<b>fr_allylic_oxid</b>
FractionCSP3	MinEStateIndex	PEOE_VSA12	<b>fr_hdrzine</b>	<b>SlogP_VSA3</b>
EState_VSA8	SMR_VSA10	PEOE_VSA1	<i>fr_sulfide</i>	<b>SlogP_VSA7</b>
EState_VSA9	SMR_VSA5	NumHeteroatoms	<b>fr_nitro</b>	<b>VSA_EState9</b>
EState_VSA7	SlogP_VSA10	SlogP_VSA2	<i>fr_nitro_arom_nonortho</i>	<i>fr_Ar_OH</i>
EState_VSA4	PEOE_VSA5	<i>SlogP_VSA11</i>	<i>fr_ketone_Topliiss</i>	<b>SlogP_VSA8</b>
HallKierAlpha	PEOE_VSA8	SMR_VSA7	<b>fr_ketone</b>	<b>SlogP_VSA5</b>
MinPartialCharge	PEOE_VSA6	SMR_VSA6	<b>fr_aniline</b>	<b>SlogP_VSA6</b>
NHOHCount	PEOE_VSA7	SlogP_VSA1	<b>fr_halogen</b>	
<i>NumAromaticCarbocycles</i>	PEOE_VSA14	SMR_VSA9	<i>fr_C_O_noCOO</i>	
MaxAbsPartialCharge	PEOE_VSA2	RingCount	<b>fr_N_O</b>	

Необхідно зазначити, що серед переліку унікальних релевантних дескрипторів, які були отримані окремо для кожної з чотирьох груп ксенобіотиків, більша частина предикторів відповідають за наявність певної специфічної функціональної групи або структури. Такий результат моделювання дозволяє для

окремих груп ксенобіотиків (табл.2.1), що мають спільні риси будови молекулярної структури, сформувати профіль ознак, пов'язаних з мутагенністю, що є унікальними для кожного з чотирьох структурних класів ксенобіотиків.

Опис унікальних релевантних дескрипторів для ароматичних гетеромоно(полі)циклічних хімічних сполук наведено у таблиці 3.26

Таблиця 3.26

**Характеристика релевантних унікальних дескрипторів для ароматичних гетеромоно(полі)циклічних хімічних сполук**

	Опис молекулярних дескрипторів
<i>NumAromaticCarbocycles</i>	Ароматичний цикл, що складається тільки з атомів карбону
<i>fr_ArN</i>	Атом нітрогену є частиною аміногрупи ( $NO_2$ ) на рівні ароматичного кільця
<i>fr_Al_COO</i>	Аліфатичні карбоксильні групи
<i>fr_sulfide</i>	Сульфідний зв'язок
<i>fr_nitro_ arom_ nonortho</i>	Нітро група ( $NH_2$ ) ароматичного кільця знаходиться або в мета- або в пара- позиціях.
<i>fr_ketone_Topliss</i>	Кетонна група ( $R_1 - C = O - R_2$ )
<i>fr_C_O_noCOO</i>	Карбонільна група, що не відноситься до карбонових кислот
<i>fr_Ar_OH</i>	Гідроксильна група на рівні ароматичного кільця

Серед релевантних дескрипторів (табл.3.25) заслуговують на увагу дескриптори, які пов'язані з функціональними групами, що лежать в основі прояву мутагенності та використовувались в якості основних предикторів для прогнозування мутагенності Еймса інших структурних класів ксенобіотиків. Предиктори, що мають суттєвий вплив на прояв мутагенності ароматичних гомомоно(полі)циклічних хімічних сполук: *fr\_para\_hydroxylation* – гідроксильна група у положенні пара- ароматичного кільця; *fr\_hdrzine* – вказує на наявність гідразинового зв'язку ( $-NH - NH_2$ ); *fr\_nitro* – наявність нітро групи ( $NO_2$ ); *fr\_ketone* – вказує на наявність кетонної групи; *fr\_aniline* – відповідає за наявність анілінової групи ( $C_6H_5NH_2$ ); *fr\_halogen* – вказує на наявність галогенного атома в



структурі ксенобіотика; *fr\_N\_O* – позначає нітроген, який приєднаний до кисню в складі нітро групи ( $\text{NO}_2$ ) або нітросо-групи ( $-\text{N} = \text{O}$ ); *fr\_alkyl\_halide* – відповідає алкілгалогеніду; *fr\_NH0* – група ( $\text{NH}$ ); *fr\_allylic\_oxid* – позначає алільну групу.

У таблиці 3.27 наведено перелік універсальних релевантних предикторів, які були представлені у кожній орієнтованій на структурні класи ксенобіотиків Ames/QSAR моделі. Такі молекулярні дескриптори дозволили акцентувати увагу на тих ознаках (властивостях) ксенобіотиків, які є визначальними щодо проявів мутагенності всіх хімічних сполук, що представлені у довкіллі, без урахування їх структурної приналежності. Крім того, отриманий перелік найвпливовіших дескрипторів, дозволив встановити що факт наявної мутагенності для ксенобіотиків, у першу чергу, обумовлений набором електронних властивостей (електронна щільність, полярність, електронегативність, електропозитивність, гібридизація орбіталей, енергетичний рівень орбіталей, дипольний момент тощо), які можуть змінюватись в залежності від умов навколишнього середовища. В такій ситуації стає зрозуміло чому процедура інтерпретації мутагенності, з урахуванням різноманітних властивостей ксенобіотиків є надзвичайно складною задачею.

Таблиця 3.27

**Релевантні універсальні дескриптори, що зустрічались в усіх орієнтованих на основні структурні класи ксенобіотиків моделях [165]**

	Опис молекулярних дескрипторів
<i>EState_VSA1</i> <i>EState_VSA4</i> <i>EState_VSA7</i> <i>EState_VSA8</i> <i>EState_VSA9</i>	Оцінка електростатичних властивостей поверхні молекули. Дескриптори <i>EState</i> розглядають поверхню молекули з різним потенціалом (позитивним, негативним) та враховують ділянки з різною електронною щільністю та дипольним моментом. Розглядають молекулу, як електрично заряджену систему.
<i>FractionCSP3</i>	Кількість атомів вуглецю, для яких характерна $\text{sp}^3$ гібридизація

## Продовження таблиці 3.27

<i>HallKierAlpha</i>	Топологічний дескриптор, що враховує зв'язки між атомами з урахуванням гібридизації електронних орбіталей ( $sp$ , $sp^2$ , $sp^3$ )
<i>MaxAbsPartial Charge</i>	Оцінка максимального значення заряду атому (позитивного або негативного) в молекулі
<i>NHOHCount</i>	Кількість гідроксиламінових груп ( $-NHOH$ )
<i>PEOE_VSA1</i> <i>PEOE_VSA2</i> <i>PEOE_VSA3</i> <i>PEOE_VSA8</i> <i>PEOE_VSA12</i> <i>PEOE_VSA14</i>	Дескриптори <i>PEOE_</i> , мають спільні риси з <i>EState_</i> , але базуються на різних алгоритмах щодо розрахунків. <i>PEOE_</i> дозволяють визначити площу поверхні молекули з частковими електронегативними та електропозитивними зарядами, що необхідно для оцінки здатності атомів до поляризації.
<i>SMR_VSA5</i> <i>SMR_VSA6</i> <i>SMR_VSA7</i>	Оцінюють набір поверхонь молекули, що лежить в основі взаємодії з іншими молекулами на основі полярних та електростатичних взаємодій
<i>SlogP_VSA1</i> <i>SlogP_VSA2</i> <i>SlogP_VSA3</i>	Оцінка поверхні молекули з гідрофобними, гідрофільними та ділянками, які проявляють властивості амфіфільності.
<i>fr_NH0</i>	Наявність в структурі ксенобіотика нітрогену, з яким зв'язаний ковалентним зв'язком водень ( $NH$ )
<i>fr_alkyl_halide</i>	Наявність в структурі ксенобіотика галогеноалканів ( $C_nH_{2n+1}(F, Br, Cl, I)$ )
<i>fr_halogen</i>	В структурі ксенобіотика міститься представник галогенового ряду

Очевидно, що в основі індукованого пошкодження генетичного матеріалу лежить здатність молекули ксенобіотика до формування різних типів взаємодій (електростатичних, Ван-дер-Ваальсових, гідрофобних, водневий зв'язок тощо) з іншими молекулами, зокрема й молекулою ДНК. Тому для оцінки мутагенності факторів навколишнього середовища, незалежно від їх приналежності до структурних класів, у першу чергу, необхідно враховувати електронні

властивості, що пов'язані з розподілом електронів на рівні атомів молекули ксенобітика. Такі властивості як полярність молекули, електронегативність (позитивна та негативна) та електронна щільність, реакційна здатність, формування дипольного моменту є визначальними щодо проявів мутагенних ефектів ксенобіотиків на генетичний апарат людини. Крім того, відповідно до отриманого переліку релевантних універсальних дескрипторів (табл.3.27), які використовуються в усіх орієнтованих на основні структурні класи Ames/QSAR моделях, для оцінки мутагенності необхідно акцентувати увагу також на таких дескрипторах, що позначають функціональні групи такі, як  $NH$ , гідроксиламінові групи  $-NHON$ , а також галогеналканах. Наявність таких груп, пов'язують з мутагенністю факторів навколишнього середовища хімічної природи без урахування їх приналежності до відповідних структурних класів.

Для отримання *in silico* оцінки мутагенності Еймса з високою точністю необхідно використовувати релевантні набори молекулярних дескрипторів, що були отримані нами для основних структурних класів ксенобіотиків.

### Висновки до розділу 3

Ефективність *in silico* моделей прогнозування мутагенності Еймса може залежати від багатьох чинників, що обумовлено використанням різних наборів молекулярних дескрипторів, особливістю організації вхідних даних та методів їх обробки. Отримані результати оцінки ефективності Ames/QSAR моделей, що використовували в якості предикторів, як повний, так і обмежений набори молекулярних дескрипторів, дозволили підтвердити гіпотезу про те, що покращення точності *in silico* моделей оцінки мутагенності Еймса може бути реалізовано через відбір релевантних дескрипторів, які були обчислені для окремих груп ксенобіотиків, що мають спільні риси будови молекулярного каркасу. Оптимізація Ames/QSAR моделей, що була здійснена через зменшення розмірності вхідних даних, дозволяє покращити точність, стабільність та узагальнюваність моделей, а також знижує ризики їх перенавчання. Зменшення

обсягу вхідних даних також дозволяє покращити інтерпретованість Ames/QSAR моделей. Показано, що моделі, які в якості вхідних даних використовували 1D та 2D дескриптори, що були розраховані для основних структурних класів ксенобіотиків, дозволяють вирішити задачу оцінки мутагенності Еймса з високими показниками точності, що навіть перевищує загальну точність *in vitro* тесту Еймса, яка коливається у межах 80-85%. Детальний аналіз нещодавно опублікованих наукових праць [155,157,191,192,196], в яких дослідниками були запропоновані різні підходи щодо побудови ефективних моделей прогнозування мутагенності Еймса, дозволив зробити висновок про те, що розроблена у рамках роботи методика оптимізації Ames/QSAR моделей дозволяє отримати бінарні класифікатори з високими показниками класифікації. При цьому, найкраща *in silico* Ames/QSAR модель з  $AUC = 0,95$ , та точністю  $accuracy = 0,96$  була отримана на основі глибинної нейронної мережі, для якої процедура навчання відбувалась на основі дескрипторів PaDel, що були розраховані для другої групи ксенобіотиків. У нещодавно опублікованих наукових працях [157,191,196] автори також звертають увагу на найвищі показники точності моделей прогнозування мутагенності Еймса, що побудовані на основі нейронних мереж (згорткових та глибинних). Реалізація *in silico* моделей на основі нейромережевого підходу з використанням 1D та 2D дескрипторів показала стабільне покращення ефективності орієнтованих на основні структурні класи ксенобіотиків прогностичних моделей. Більша частина клас-орієнтованих моделей на основі метода випадкового лісу та екстремального градієнтного спуску з дескрипторами Padell, RDkit та PaDell демонстрували також покращення прогностичної здатності, що досягалась за рахунок відбору релевантних дескрипторів та через розподіл ксенобіотиків на окремі структурні класи. Для прогнозування мутагенності Еймса аліфатичних ациклічних хімічних сполук ефективною є модель RDKit/ XGBoost, яка на екзотичній вибірці демонструвала  $AUC = 0,95$  та точність  $accuracy = 0,89$ . Для отримання оцінки мутагенного потенціалу ароматичних гетеромоно(полі)циклічних та ароматичних гомомоно(полі)циклічних хімічних сполук необхідно використовувати

орієнтовані на відповідні структурні класи Ames/QSAR моделі (RDKit/ RF) та (Mordred/RF), що на екзменаційній вибірці показали  $AUC = 0,94$ ,  $accuracy = 0,84$  та  $AUC = 0,95$ ,  $accuracy = 0,87$  відповідно.

Наявність певних функціональних груп або підструктур на рівні молекули може лежати в основі прояву мутагенності основних структурних класів ксенобіотиків. Отриманий перелік унікальних дескрипторів дозволив сформувати профіль ознак пов'язаних з мутагенністю для окремих груп ксенобіотиків, що мають спільні риси будови молекулярного каркасу. Показано, що особливості розподілу електронів на рівні атомів молекул ксенобіотиків, що відносяться до різних структурних класів, може виступити одним з базових критеріїв для оцінки їх здатності до пошкодження спадкового апарату людини.

## ЗАГАЛЬНІ ВИСНОВКИ

1. Проведено детальний аналіз *in vitro* та *in vivo* методів, що відносяться до стандартної батареї тест-систем. Доведена необхідність оновлення методів та підходів до оцінки мутагенності Еймса.
2. Проаналізовані доступні для загального використання набори даних ксенобіотиків хімічної природи, для яких, відповідно до *in vitro* тесту Еймса, була отримана оцінка мутагенному потенціалу.
3. Відповідно до трьох загальнодоступних наборів даних Kazius-Bursi, Hansen та EFSA була сформована об'єднана база даних, яка була додатково розширена мікотоксинами.
4. Запропоновано методику підвищення точності *in silico* моделей прогнозування мутагенності Еймса, яка реалізується через використання однорідних наборів вхідних даних та відбір релевантних ознак. Згідно з представленим підходом до оптимізації *in silico* Ames/QSAR моделей було проведено розподіл хімічних сполук на основні структурні класи, та для кожного ксенобіотика розраховані одновимірні та двовимірні молекулярні дескриптори.
5. Показано, що моделі, які в якості вхідних даних, використовували 1D та 2D релевантні дескриптори (Padel, RDkit та Mordred), що були розраховані для основних структурних класів ксенобіотиків, дозволили підвищити точність таких бінарних класифікаторів. Моделі, що були отримані відповідно до набору даних, що відносяться до основних структурних класів ксенобіотиків, дозволили вирішити задачу оцінки мутагенності з високими показниками точності (від 87% до 93%) відповідно до метрики *accuracy*, що перевищує значення метрики загальної точності для *in vitro* теста Еймса, яка коливається у межах 80-85%. При цьому точність моделей, які на етапі навчання використовували 80% даних повного датасету, що відповідає стандартному підходу до моделювання, була меншою (від 0,1% до 2% для різних моделей), у порівнянні з орієнтованими на основні структурні класи ксенобіотиків моделями. Отримані переліки релевантних дескрипторів для основних структурних класів ксенобіотиків,

дозволяють підвищити точність *in silico* моделей прогнозування мутагенності Еймса.

6. При вирішенні задачі *in silico* оцінки мутагенності підтверджена ефективність використання в якості предикторів відбитків молекулярної структури ксенобіотиків. Точність таких моделей відповідає середньому значенню загальної точності *in vitro* теста Еймса, яка коливається у межах 80-85%, але поступається орієнтованим на основні структурні класи ксенобіотиків моделям прогнозування мутагенності, що в якості предикторів використовували класичні одновимірні та двовимірні дескриптори Padel, RDkit та Mordred.

7. Показана можливість використання алгоритму t-розподіленого вкладення стохастичної близькості (t-SNE), що використовується для візуалізації багатовимірних даних, з метою ідентифікації структурних маркерів мутагенності.

8. Проведений аналіз причинно-наслідкових зв'язків між мутагенністю та набором релевантних ознак, що задаються молекулярними дескрипторами, дозволив зробити висновок про те, що електронні властивості атомів в структурі молекули ксенобіотиків є фундаментом їх негативного генетичного впливу на спадковий апарат людини.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Initial sequencing and analysis of the human genome / Eric S. Lander et al. *Nature*. 2001. Vol. 409, no. 6822. P. 860–921. URL: <https://doi.org/10.1038/35057062>
2. A map of human genome variation from population-scale sequencing / Abecasis G. R. et al. *Nature*. 2010. Vol. 467, no. 7319. P. 1061–1073. URL: <https://doi.org/10.1038/nature09534>
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> (date of access: 12.06.2025).
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. Carusillo A., Mussolino C. DNA damage: from threat to treatment. *Cells*. 2020. Vol. 9, no. 7. P. 1665. URL: <https://doi.org/10.3390/cells9071665> (date of access: 12.06.2025).
6. The potential for chemical mixtures from the environment to enable the cancer hallmark of sustained proliferative signalling / W. Engström et al. *Carcinogenesis*. 2015. Vol. 36, Suppl 1. P. S38–S60. URL: <https://doi.org/10.1093/carcin/bgv030>
7. Mobile phone signal exposure triggers a hormesis-like effect in Atm<sup>+/+</sup> and Atm<sup>-/-</sup> mouse embryonic fibroblasts / C. Sun et al. *Scientific reports*. 2016. Vol. 6, no. 1. URL: <https://doi.org/10.1038/srep37423>
8. 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>
9. 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>



10. ChemSpider - building a foundation for the semantic web by hosting a crowd sourced databasing platform for chemistry / A. J. Williams et al. *Journal of cheminformatics*. 2010. Vol. 2, S1. URL: <https://doi.org/10.1186/1758-2946-2-s1-o16>
11. PubChem 2019 update: improved access to chemical data / S. Kim et al. *Nucleic acids research*. 2018. Vol. 47, no. D1. P. D1102–D1109. URL: <https://doi.org/10.1093/nar/gky1033>
12. Gabrielson S. W. SciFinder. *Journal of the medical library association*. 2018. Vol. 106, no. 4. URL: <https://doi.org/10.5195/jmla.2018.515>
13. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories / Z. Wang et al. *Environmental science & technology*. 2020. Vol. 54, no. 5. P. 2575–2584. URL: <https://doi.org/10.1021/acs.est.9b06379>
14. Challenges and opportunities in the risk assessment of existing substances in Canada: lessons learned from the international community / T. S. B. Maclaren et al. *International journal of risk assessment and management*. 2017. Vol. 20, no. 1/2/3. P. 261. URL: <https://doi.org/10.1504/ijram.2017.082569>
15. Comparison of methods used for evaluation of mutagenicity/genotoxicity of model chemicals - parabens / J. Chrząst et al. *Physiological research*. 2020. P. S661–S679. URL: <https://doi.org/10.33549/physiolres.934615>
16. Genotoxicity detection in drinking water by Ames Test, Zimmermann test and Comet assay / Lah B. et al. *Acta chim slov*. 2005. Vol. 52, no. 341-8.
17. Ubomba-Jaswa E., Fernández-Ibáñez P., McGuigan K. G. A preliminary Ames fluctuation assay assessment of the genotoxicity of drinking water that has been solar disinfected in polyethylene terephthalate (PET) bottles. *Journal of water and health*. 2010. Vol. 8, no. 4. P. 712–719. URL: <https://doi.org/10.2166/wh.2010.136>
18. Agwa O. K., Eze N. J., Okpokwasili G. C. Mutagenic potentials of potable water from ground sources. *The open biotechnology journal*. 2017. Vol. 11, no. 1. P. 81–88. URL: <https://doi.org/10.2174/1874070701711010081>

19. Contaminants, mutagenicity and toxicity in the surface waters of Kyiv, Ukraine / K. T. Ho et al. *Marine pollution bulletin*. 2020. Vol. 155. P. 111153. URL: <https://doi.org/10.1016/j.marpolbul.2020.111153>
20. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Outdoor air pollution. Lyon (FR): International Agency for Research on Cancer; 2016. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, No. 109.) Available from: <https://www.ncbi.nlm.nih.gov/books/NBK368024/>
21. Validation of the harvard six cities study of particulate air pollution and mortality / D. Krewski et al. *New england journal of medicine*. 2004. Vol. 350, no. 2. P. 198–199. URL: <https://doi.org/10.1056/nejm200401083500225>
22. An association between long-term exposure to ambient air pollution and mortality from lung cancer and respiratory diseases in japan / K. Katanoda et al. *Journal of epidemiology*. 2011. Vol. 21, no. 2. P. 132–143. URL: <https://doi.org/10.2188/jea.je20100098>
23. Aoki Y. Evaluation of in vivo mutagenesis for assessing the health risk of air pollutants. *Genes and environment*. 2017. Vol. 39, no. 1. URL: <https://doi.org/10.1186/s41021-016-0064-6>
24. Characteristics, sources and health risks of toxic species (PCDD/Fs, PAHs and heavy metals) in PM<sub>2.5</sub> during fall and winter in an industrial area / C. Bi et al. *Chemosphere*. 2020. Vol. 238. P. 124620. URL: <https://doi.org/10.1016/j.chemosphere.2019.124620>
25. Luch A. Nature and nurture – lessons from chemical carcinogenesis. *Nature reviews cancer*. 2005. Vol. 5, no. 2. P. 113–125. URL: <https://doi.org/10.1038/nrc1546>
26. Toxicological and mutagenic effects of particulate matter from domestic activities / D. Figueiredo et al. *Toxics*. 2023. Vol. 11, no. 6. P. 505. URL: <https://doi.org/10.3390/toxics11060505>
27. Mutagenic and carcinogenic hazards of settled house dust II: salmonella mutagenicity / R. M. Maertens et al. *Environmental science & technology*. 2008. Vol. 42, no. 5. P. 1754–1760. URL: <https://doi.org/10.1021/es702448x>

28. DeMarini D. M., Linak W. P. Mutagenicity and carcinogenicity of combustion emissions are impacted more by combustor technology than by fuel composition: a brief review. *Environmental and molecular mutagenesis*. 2022. URL: <https://doi.org/10.1002/em.22475>
29. Comprehensive investigation of the mutagenic potential of six pesticides classified by IARC as probably carcinogenic to humans / R. Martinek et al. *Chemosphere*. 2024. P. 142700. URL: <https://doi.org/10.1016/j.chemosphere.2024.142700>
30. Chauhan S. S., Garg P., Parthasarathi R. Computational framework for identifying and evaluating mutagenic and xenoestrogenic potential of food additives. *Journal of hazardous materials*. 2024. Vol. 470. P. 134233. URL: <https://doi.org/10.1016/j.jhazmat.2024.134233>
31. Safety assessment of recycled plastics from post-consumer waste with a combination of a miniaturized ames test and chromatographic analysis / E. Mayrhofer et al. *Recycling*. 2023. Vol. 8, no. 6. P. 87. URL: <https://doi.org/10.3390/recycling8060087>
32. Owiti N. A., Nagel Z. D., Engelward B. P. Fluorescence sheds light on DNA damage, DNA repair, and mutations. *Trends in cancer*. 2020. URL: <https://doi.org/10.1016/j.trecan.2020.10.006>
33. Hoeijmakers J. H. J. DNA damage, aging, and cancer. *New england journal of medicine*. 2009. Vol. 361, no. 15. P. 1475–1485. URL: <https://doi.org/10.1056/nejmra0804615>
34. Reactive oxygen species: from health to disease / K. Brieger et al. *Swiss medical weekly*. 2012. URL: <https://doi.org/10.4414/smw.2012.13659>
35. ROS and the DNA damage response in cancer / U. S. Srinivas et al. *Redox biology*. 2019. Vol. 25. P. 101084. URL: <https://doi.org/10.1016/j.redox.2018.101084>
36. Epigenetic modification and a role for the E3 ligase RNF40 in cancer development and metastasis / J. Fu et al. *Oncogene*. 2020. URL: <https://doi.org/10.1038/s41388-020-01556-w>

37. PRIMPOL ready, set, reprime! / S. Tirman et al. Critical reviews in biochemistry and molecular biology. 2020. P. 1–14.  
URL: <https://doi.org/10.1080/10409238.2020.1841089>
38. DNA binding and antiradical potential of ethyl pyruvate: Key to the DNA radioprotection / D. Sharma et al. Chemico-Biological interactions. 2020. Vol. 332. P. 109313. URL: <https://doi.org/10.1016/j.cbi.2020.109313>
39. Significant improvement in rat kidney cold storage using UW organ preservation solution supplemented with the immediate-acting prc-210 free radical scavenger / B. M. Verhoven et al. Transplantation direct. 2020. Vol. 6, no. 8. P. e578. URL: <https://doi.org/10.1097/txd.0000000000001032>
40. 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>
41. Lindahl T., Barnes D. E. Repair of endogenous DNA damage. Cold spring harbor symposia on quantitative biology. 2000. Vol. 65. P. 127–134. URL: <https://doi.org/10.1101/sqb.2000.65.127>
42. Accelerated deamination of cytosine residues in UV-induced cyclobutane pyrimidine dimers leads. Current opinion in toxicology. 2018. Vol. 7. P. 9–16. URL: <https://doi.org/10.1016/j.cotox.2017.10.009>
43. Peng W., Shaw B. R. Accelerated deamination of cytosine residues in uv-induced cyclobutane pyrimidine dimers leads to CC→TT transitions†. *Biochemistry*. 1996. Vol. 35, no. 31. P. 10172–10181. URL: <https://doi.org/10.1021/bi960001x>
44. Echinomycin, a bis-intercalating agent, induces C→T mutations via cytosine deamination / R. Moyer et al. Mutation research/fundamental and molecular mechanisms of mutagenesis. Vol. 288, no. 2. P. 291–300. URL: [https://doi.org/10.1016/0027-5107\(93\)90097-y](https://doi.org/10.1016/0027-5107(93)90097-y)
45. Hayatsu H. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis – A personal account. Proceedings of the japan academy, series B. 2008. Vol. 84, no. 8. P. 321–330. URL: <https://doi.org/10.2183/pjab.84.321>

46. Frankel A. D., Duncan B. K., Hartman P. E. Nitrous acid damage to duplex deoxyribonucleic acid: distinction between deamination of cytosine residues and a novel mutational lesion. *Journal of bacteriology*. 1980. Vol. 142, no. 1. P. 335–338. URL: <https://doi.org/10.1128/jb.142.1.335-338.1980>
47. Ganai R. A., Johansson E. DNA replication—a matter of fidelity. *Molecular cell*. 2016. Vol. 62, no. 5. P. 745–755. URL: <https://doi.org/10.1016/j.molcel.2016.05.003>
48. 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>
49. McCulloch S. D., Kunkel T. A. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell research*. 2008. Vol. 18, no. 1. P. 148–161. URL: <https://doi.org/10.1038/cr.2008.4>
50. Kunkel T. A. Evolving views of DNA replication (in)fidelity. *Cold spring harbor symposia on quantitative biology*. 2009. Vol. 74. P. 91–101. URL: <https://doi.org/10.1101/sqb.2009.74.027>
51. Kunkel T. A., Bebenek K. DNA replication fidelity. *Annual review of biochemistry*. 2000. Vol. 69, no. 1. P. 497–529. URL: <https://doi.org/10.1146/annurev.biochem.69.1.497>
52. Oxidative stress in cancer-prone genetic diseases in pediatric age: the role of mitochondrial dysfunction / S. Perrone et al. *Oxidative medicine and cellular longevity*. 2016. Vol. 2016. P. 1–7. URL: <https://doi.org/10.1155/2016/4782426>
53. Martin L. J. DNA damage and repair. *Journal of neuropathology & experimental neurology*. 2008. Vol. 67, no. 5. P. 377–387. URL: <https://doi.org/10.1097/nen.0b013e31816ff780>
54. Proceedings of a workshop on DNA adducts: Biological significance and applications to risk assessment Washington, DC, April 13–14, 2004 / M. Sander et al. *Toxicology and applied pharmacology*. 2005. Vol. 208, no. 1. P. 1–20. URL: <https://doi.org/10.1016/j.taap.2004.12.012>
55. Quantitative structure–activity relationship models for genotoxicity prediction based on combination evaluation strategies for toxicological alternative experiments /

- X. Willett et al. Scientific reports. 2021. Vol. 11, no. 1.  
URL: <https://doi.org/10.1038/s41598-021-87035-y>
56. 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>
57. Computational approaches to identify structural alerts and their applications in environmental toxicology and drug discovery / H. Yang et al. Chemical research in toxicology. 2020. Vol. 33, no. 6. P. 1312–1322.  
URL: <https://doi.org/10.1021/acs.chemrestox.0c00006>
58. Huang R., Zhou P.-K. DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. Signal transduction and targeted therapy. 2021. Vol. 6, no. 1. URL: <https://doi.org/10.1038/s41392-021-00648-7>
59. 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>
60. Ionizing radiation-induced risks to the central nervous system and countermeasures in cellular and rodent models / E. Pariset et al. International journal of radiation biology. 2020. P. 1–19.  
URL: <https://doi.org/10.1080/09553002.2020.1820598>
61. Desouky O., Ding N., Zhou G. Targeted and non-targeted effects of ionizing radiation. Journal of radiation research and applied sciences. 2015. Vol. 8, no. 2. P. 247–254. URL: <https://doi.org/10.1016/j.jrras.2015.03.003>
62. Azzam E. I., Jay-Gerin J.-P., Pain D. Ionizing radiation-induced metabolic oxidative stress and prolonged cell injury. Cancer letters. 2012. Vol. 327, no. 1-2. P. 48–60. URL: <https://doi.org/10.1016/j.canlet.2011.12.012>
63. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair / R. P. Rastogi et al. Journal of nucleic acids. 2010. Vol. 2010. P. 1–32.  
URL: <https://doi.org/10.4061/2010/592980>

64. Deciphering uv-induced DNA damage responses to prevent and treat skin cancer / J. W. Lee et al. *Photochemistry and photobiology*. 2020. Vol. 96, no. 3. P. 478–499. URL: <https://doi.org/10.1111/php.13245>
65. Benigni R., Bossa C. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. *Chemical reviews*. 2011. Vol. 111, no. 4. P. 2507–2536. URL: <https://doi.org/10.1021/cr100222q>
66. Explanation for main features of structure–genotoxicity relationships of aromatic amines by theoretical studies of their activation pathways in CYP1A2 / I. Shamovsky et al. *Journal of the american chemical society*. 2011. Vol. 133, no. 40. P. 16168–16185. URL: <https://doi.org/10.1021/ja206427u>
67. Identification of mutagenic aromatic amines in river samples with industrial wastewater impact / M. Muz et al. *Environmental science & technology*. 2017. Vol. 51, no. 8. P. 4681–4688. URL: <https://doi.org/10.1021/acs.est.7b00426>
68. Recent advances in heterocyclic aromatic amines: an update on food safety and hazardous control from food processing to dietary intake / X. Chen et al. *Comprehensive reviews in food science and food safety*. 2019. Vol. 19, no. 1. P. 124–148. URL: <https://doi.org/10.1111/1541-4337.12511>
69. Norinder U., Myatt G., Ahlberg E. Predicting aromatic amine mutagenicity with confidence: a case study using conformal prediction. *Biomolecules*. 2018. Vol. 8, no. 3. P. 85. URL: <https://doi.org/10.3390/biom8030085>
70. Kriek E. Fifty years of research on N-acetyl-2-aminofluorene, one of the most versatile compounds in experimental cancer research. *Journal of cancer research and clinical oncology*. 1992. Vol. 118, no. 7. P. 481–489. URL: <https://doi.org/10.1007/bf01225261>
71. Plant activation of aromatic amines mediated by cytochromes P450 and flavin-containing monooxygenases / C. Chiapella et al. *Mutation research/genetic toxicology and environmental mutagenesis*. 2000. Vol. 470, no. 2. P. 155–160. URL: [https://doi.org/10.1016/s1383-5718\(00\)00098-x](https://doi.org/10.1016/s1383-5718(00)00098-x)
72. Influence of flanking sequence context on the mutagenicity of acetylaminofluorene-derived DNA adducts in mammalian cells/ S. Shibutani et

- al. *Biochemistry*. 2001. Vol. 40, no. 12. P. 3717–3722.  
URL: <https://doi.org/10.1021/bi0027581>
73. Arora P. K. Bacterial degradation of monocyclic aromatic amines. *Frontiers in microbiology*. 2015. Vol. 6. URL: <https://doi.org/10.3389/fmicb.2015.00820>
74. Application of novel bacterial consortium for biodegradation of aromatic amine 2-ABS using response surface methodology / M. Fatima et al. *Journal of microbiological methods*. 2020. Vol. 174. P. 105941.  
URL: <https://doi.org/10.1016/j.mimet.2020.105941>
75. Extending (Q)SARs to incorporate proprietary knowledge for regulatory purposes: a case study using aromatic amine mutagenicity / E. Ahlberg et al. *Regulatory toxicology and pharmacology*. 2016. Vol. 77. P. 1–12.  
URL: <https://doi.org/10.1016/j.yrtph.2016.02.003>
76. A pharma-wide approach to address the genotoxicity prediction of primary aromatic amines / M. Patel et al. *Computational toxicology*. 2018. Vol. 7. P. 27–35.  
URL: <https://doi.org/10.1016/j.comtox.2018.06.002>
77. A local QSAR model based on the stability of nitrenium ions to support the ICH M7 expert review on the mutagenicity of primary aromatic amines / A. Furukawa et al. *Genes and environment*. 2022. Vol. 44, no. 1. URL: <https://doi.org/10.1186/s41021-022-00238-1>
78. Comprehensive review of polycyclic aromatic hydrocarbons in water sources, their effects and treatments / A. Mojiri et al. *Science of the total environment*. 2019. Vol. 696. P. 133971. URL: <https://doi.org/10.1016/j.scitotenv.2019.133971>
79. Biomass burning contributed most to the human cancer risk exposed to the soil-bound PAHs from Chengdu Economic Region, western China / H. Zheng et al. *Ecotoxicology and environmental safety*. 2018. Vol. 159. P. 63–70.  
URL: <https://doi.org/10.1016/j.ecoenv.2018.04.065>
80. Polycyclic aromatic hydrocarbons: sources, toxicity, and remediation approaches / A. B. Patel et al. *Frontiers in microbiology*. 2020. Vol. 11. URL: <https://doi.org/10.3389/fmicb.2020.562813>



81. Ewa B., Danuta M.-Š. Polycyclic aromatic hydrocarbons and PAH-related DNA adducts. *Journal of applied genetics*. 2016. Vol. 58, no. 3. P. 321–330. URL: <https://doi.org/10.1007/s13353-016-0380-3>
82. Abdel-Shafy H. I., Mansour M. S. M. A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation. *Egyptian journal of petroleum*. 2016. Vol. 25, no. 1. P. 107–123. URL: <https://doi.org/10.1016/j.ejpe.2015.03.011>
83. Aflatoxins: a global concern for food safety, human health and their management / P. Kumar et al. *Frontiers in microbiology*. 2017. Vol. 07. URL: <https://doi.org/10.3389/fmicb.2016.02170>
84. Taxonomy of *Aspergillus* section *Flavi* and their production of aflatoxins, ochratoxins and other mycotoxins / J. C. Frisvad et al. *Studies in mycology*. 2019. Vol. 93. P. 1–63. URL: <https://doi.org/10.1016/j.simyco.2018.06.001>
85. Identification of *Aspergillus* species in Central Europe able to produce G-type aflatoxins / N. Baranyi et al. *Acta biologica hungarica*. 2015. Vol. 66, no. 3. P. 339–347. URL: <https://doi.org/10.1556/018.66.2015.3.9>
86. Priesterjahn E.-M., Geisen R., Schmidt-Heydt M. Influence of light and water activity on growth and mycotoxin formation of selected isolates of *aspergillus flavus* and *aspergillus parasiticus*. *Microorganisms*. 2020. Vol. 8, no. 12. P. 2000. URL: <https://doi.org/10.3390/microorganisms8122000>
87. Overview on genetic toxicology TGs. OECD, 2017. URL: <https://doi.org/10.1787/9789264274761-en>
88. 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>
89. 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>

90. 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>
91. 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>
92. 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>
93. 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>
94. Scientific opinion on genotoxicity testing strategies applicable to food and feed safety assessment. *EFSA journal*. 2011. Vol. 9, no. 9. URL: <https://doi.org/10.2903/j.efsa.2011.2379>
95. Test no. 471: bacterial reverse mutation test. OECD, 2020. URL: <https://doi.org/10.1787/9789264071247-en>
96. Test no. 473: in vitro mammalian chromosomal aberration test. OECD Publishing, 2014. URL: <https://doi.org/10.1787/9789264224223-en>
97. Test no. 474: mammalian erythrocyte micronucleus test. OECD, 2014. URL: <https://doi.org/10.1787/9789264224292-en>
98. Test no. 475: mammalian bone marrow chromosomal aberration test. OECD Publishing, 2014. URL: <https://doi.org/10.1787/9789264224407-en>
99. Test no. 476: in vitro mammalian cell gene mutation tests using the hprt and xprt genes. OECD, 2016. URL: <https://doi.org/10.1787/9789264264809-en>
100. Test no. 478: rodent dominant lethal test. OECD, 2016. URL: <https://doi.org/10.1787/9789264264823-en>
101. Test no. 483: mammalian spermatogonial chromosomal aberration test. OECD, 2016. URL: <https://doi.org/10.1787/9789264264847-en>
102. Test no. 485: genetic toxicology, mouse heritable translocation assay. OECD, 1986. URL: <https://doi.org/10.1787/9789264071506-en>

103. Test no. 486: unscheduled DNA synthesis (UDS) test with mammalian liver cells in vivo. OECD, 1997. URL: <https://doi.org/10.1787/9789264071520-en>
104. Test no. 487: in vitro mammalian cell micronucleus test. OECD Publishing, 2014. URL: <https://doi.org/10.1787/9789264224438-en>
105. Test no. 489: in vivo mammalian alkaline comet assay. OECD Publishing, 2014. URL: <https://doi.org/10.1787/9789264224179-en>
106. S2(R1) genotoxicity testing and data interpretation for pharmaceuticals intended for human use / ed. by I. C. o. Harmonisation. Rockville, MD] : U.S. Dept. of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, 2008. 37 p.
107. 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>
108. 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>
109. 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>
110. Microbial mutagenicity assay: ames test / U. Vijay et al. Bio-protocol. 2018. Vol. 8, no. 6. URL: <https://doi.org/10.21769/bioprotoc.2763>
111. The nucleotide excision repair protein UvrB, a helicase-like enzyme with a catch / K. Theis et al. Mutation research/dna repair. 2000. Vol. 460, no. 3-4. P. 277–300. URL: [https://doi.org/10.1016/s0921-8777\(00\)00032-x](https://doi.org/10.1016/s0921-8777(00)00032-x)
112. Mutagenic activity of chemicals identified in drinking water/ V. F. Simmon et al. Mutation research/environmental mutagenesis and related subjects. 1978. Vol. 53, no. 2. P. 262. URL: [https://doi.org/10.1016/0165-1161\(78\)90337-0](https://doi.org/10.1016/0165-1161(78)90337-0)
113. Ames B. N., McCann J., Yamasaki E. Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. Mutation

- research/environmental mutagenesis and related subjects. 1975. Vol. 31, no. 6. P. 347–363. URL: [https://doi.org/10.1016/0165-1161\(75\)90046-1](https://doi.org/10.1016/0165-1161(75)90046-1)
114. Maron D. M., Ames B. N. Revised methods for the Salmonella mutagenicity test. Mutation research/environmental mutagenesis and related subjects. 1983. Vol. 113, no. 3-4. P. 173–215. URL: [https://doi.org/10.1016/0165-1161\(83\)90010-9](https://doi.org/10.1016/0165-1161(83)90010-9)
115. S. E. Zeiger et al. Environmental and molecular mutagenesis. 1992. Vol. 19, S21. P. 2–141. URL: <https://doi.org/10.1002/em.2850190603>
116. Comparison of Salmonella typhimurium TA102 with Escherichia coli WP2 tester strains / P. Wilcox et al. Mutagenesis. 1990. Vol. 5, no. 3. P. 285–292. URL: <https://doi.org/10.1093/mutage/5.3.285>
117. Nessler F. The current limitations of in vitro genotoxicity testing and their relevance to the in vivo situation. Food and chemical toxicology. 2017. Vol. 106. P. 609–615. URL: <https://doi.org/10.1016/j.fct.2016.08.035>
118. Assessment of the predictive capacity of the optimized in vitro comet assay using HepG2 cells / Y.-h. Hong et al. Mutation research/genetic toxicology and environmental mutagenesis. 2018. Vol. 827. P. 59–67. URL: <https://doi.org/10.1016/j.mrgentox.2018.01.010>
119. The utility of metabolic activation mixtures containing human hepatic post-mitochondrial supernatant (S9) for in vitro genetic toxicity assessment / J. A. Cox et al. Mutagenesis. 2015. Vol. 31, no. 2. P. 117–130. URL: <https://doi.org/10.1093/mutage/gev082>
120. Cytochrome P450 1A1/2, 2B6 and 3A4 hepatic cell-based biosensors to monitor hepatocyte differentiation, drug metabolism and toxicity / M. Vlach et al. Sensors. 2019. Vol. 19, no. 10. P. 2245. URL: <https://doi.org/10.3390/s19102245>
121. From classical toxicology to tox21: some critical conceptual and technological advances in the molecular understanding of the toxic response beginning from the last quarter of the 20th century / S. Choudhuri et al. *Toxicological sciences*. 2017. Vol. 161, no. 1. P. 5–22. URL: <https://doi.org/10.1093/toxsci/kfx186>

122. Toxicity testing in the 21st century: progress in the past decade and future perspectives / D. Krewski et al. *Archives of toxicology*. 2019. Vol. 94, no. 1. P. 1–58. URL: <https://doi.org/10.1007/s00204-019-02613-4>
123. Combining machine learning models of in vitro and in vivo bioassays improves rat carcinogenicity prediction / D. Guan et al. *Regulatory toxicology and pharmacology*. 2018. Vol. 94. P. 8–15. URL: <https://doi.org/10.1016/j.yrtph.2018.01.008>
124. Direct quantification of in vivo mutagenesis and carcinogenesis using duplex sequencing / C. C. Valentine et al. *Proceedings of the national academy of sciences*. 2020. Vol. 117, no. 52. P. 33414–33425. URL: <https://doi.org/10.1073/pnas.2013724117>
125. Salk J. J., Schmitt M. W., Loeb L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature reviews genetics*. 2018. Vol. 19, no. 5. P. 269–285. URL: <https://doi.org/10.1038/nrg.2017.117>
126. Genotoxicity assessment: opportunities, challenges and perspectives for quantitative evaluations of dose–response data / J. Menz et al. *Archives of toxicology*. 2023. URL: <https://doi.org/10.1007/s00204-023-03553-w>
127. Rahmanian N., Shokrzadeh M., Eskandani M. Recent advances in  $\gamma$ H2AX biomarker-based genotoxicity assays: A marker of DNA damage and repair. *DNA repair*. 2021. T. 108. C. 103243. URL: <https://doi.org/10.1016/j.dnarep.2021.103243>
128. DNA double-strand breaks induce H2Ax phosphorylation domains in a contact-dependent manner / P. L. Collins et al. *Nature communications*. 2020. Vol. 11, no. 1. URL: <https://doi.org/10.1038/s41467-020-16926-x>
129. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage / S. Matsuoka et al. *Science*. 2007. Vol. 316, no. 5828. P. 1160–1166. URL: <https://doi.org/10.1126/science.1140321>
130. Fragkos M., Choleza M., Papadopoulou P. The role of  $\gamma$ h2ax in replication stress-induced carcinogenesis: possible links and recent developments. *Cancer diagnosis & prognosis*. 2023. T. 3, № 6. C. 639–648. URL: <https://doi.org/10.21873/cdp.10266>

131. Reddig A., Roggenbuck D., Reinhold D. Comparison of different immunoassays for  $\gamma$ H2AX quantification. *Journal of laboratory and precision medicine*. 2018. T. 3. C. 80. URL: <https://doi.org/10.21037/jlpm.2018.09.01>
132. Quantitative interpretation of toxtracker dose response data for potency comparisons and mode-of-action determination / L. Boisvert et al. *Environmental and molecular mutagenesis*. 2023. URL: <https://doi.org/10.1002/em.22525>
133. The extended toxtracker assay discriminates between induction of DNA damage, oxidative stress, and protein misfolding / G. Hendriks et al. *Toxicological sciences*. 2015. Vol. 150, no. 1. P. 190–203. URL: <https://doi.org/10.1093/toxsci/kfv323>
134. Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals / D. J. Araten et al. *Proceedings of the national academy of sciences*. 1999. Vol. 96, no. 9. P. 5209–5214. URL: <https://doi.org/10.1073/pnas.96.9.5209>
135. Test no. 470: mammalian erythrocyte pig-a gene mutation assay. OECD, 2022. URL: <https://doi.org/10.1787/4faea90e-en>
136. Brodsky R. A., Hu R. PIG-A mutations in paroxysmal nocturnal hemoglobinuria and in normal hematopoiesis. *Leukemia & lymphoma*. 2006. Vol. 47, no. 7. P. 1215–1221. URL: <https://doi.org/10.1080/10428190600555520>
137. Cross G. A. M. Glycolipid anchoring of plasma membrane proteins. *Annual review of cell biology*. 1990. Vol. 6, no. 1. P. 1–39. URL: <https://doi.org/10.1146/annurev.cb.06.110190.000245>
138. Molecular cloning of murine pig-a, a gene for gpi-anchor biosynthesis, and demonstration of interspecies conservation of its structure, function, and genetic locus / K. Kawagoe et al. *Genomics*. 1994. Vol. 23, no. 3. P. 566–574. URL: <https://doi.org/10.1006/geno.1994.1544>
139. Hen's egg test for micronucleus induction (HET-MN) / K. Reisinger et al. *Methods in molecular biology*. New York, NY, 2019. P. 195–208. URL: [https://doi.org/10.1007/978-1-4939-9646-9\\_10](https://doi.org/10.1007/978-1-4939-9646-9_10)
140. The hen's egg test for micronucleus induction (HET-MN): validation data set / K. Reisinger et al. *Mutagenesis*. 2021. URL: <https://doi.org/10.1093/mutage/geab016>

141. A new 3D model for genotoxicity assessment: EpiSkin™ Micronucleus Assay / L. Chen et al. *Mutagenesis*. 2020. URL: <https://doi.org/10.1093/mutage/geaa003>
142. 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>
143. Wichard J. D. In silico prediction of genotoxicity. *Food and chemical toxicology*. 2017. Vol. 106. P. 595–599. URL: <https://doi.org/10.1016/j.fct.2016.12.013>
144. Cavasotto C. N., Scardino V. Machine learning toxicity prediction: latest advances by toxicity end point. *ACS omega*. 2022. URL: <https://doi.org/10.1021/acsomega.2c05693>
145. Toxicity prediction method based on multi-channel convolutional neural network / Yuan et al. *Molecules*. 2019. Vol. 24, no. 18. P. 3383. URL: <https://doi.org/10.3390/molecules24183383>
146. Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses / A. Amberg et al. *Regulatory toxicology and pharmacology*. 2016. Vol. 77. P. 13–24. URL: <https://doi.org/10.1016/j.yrtph.2016.02.004>
147. The (re)-evolution of quantitative structure–activity relationship (QSAR) studies propelled by the surge of machine learning methods / T. A. Soares et al. *Journal of chemical information and modeling*. 2022. Vol. 62, no. 22. P. 5317–5320. URL: <https://doi.org/10.1021/acs.jcim.2c01422>
148. Keyvanpour M. R., Shirzad M. B. An analysis of QSAR research based on machine learning concepts. *Current drug discovery technologies*. 2020. Vol. 17. URL: <https://doi.org/10.2174/1570163817666200316104404>
149. 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>
150. Local QSAR based on quantum chemistry calculations for the stability of nitrenium ions to reduce false positive outcomes from standard QSAR systems for the mutagenicity of primary aromatic amines / S. Muto et al. *Genes and environment*. 2024. Vol. 46, no. 1. URL: <https://doi.org/10.1186/s41021-024-00318-4>

151. 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>
152. 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>
153. 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>
154. 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>
155. 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>
156. 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>
157. 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.compbimed.2024.108560>
158. *Molecular descriptors for chemoinformatics*. Weinheim, Germany : WILEY-VCH, 2009. 1220 p.
159. DScibe: library of descriptors for machine learning in materials science / L. Himanen et al. *Computer physics communications*. 2020. Vol. 247. P. 106949. URL: <https://doi.org/10.1016/j.cpc.2019.106949>
160. Кисляк С. В., Голуб Н. Б., Дуган О. М., Аверьянова О. А. Моделювання молекулярної взаємодії [Електронний ресурс] : підручник для здобувачів ступеня магістра за спеціальністю 162 «Біотехнології та біоінженерія» / С. В. Кисляк, Н. Б. Голуб, О. М. Дуган, О. А. Аверьянова; КПП ім. Ігоря Сікорського. – Електронні



текстові дані (1 файл, 26 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2023. – 203 с.  
– Назва з екрана.

161. Talevi A. In Silico ADME: QSPR/QSAR. *The ADME Encyclopedia*. Cham, 2022. P. 525–531. URL: [https://doi.org/10.1007/978-3-030-84860-6\\_149](https://doi.org/10.1007/978-3-030-84860-6_149)

162. Yap C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*. 2010. Vol. 32, no. 7. P. 1466–1474. URL: <https://doi.org/10.1002/jcc.21707>

163. Open Babel: an open chemical toolbox / N. M. O'Boyle et al. *Journal of cheminformatics*. 2011. Vol. 3, no. 1. URL: <https://doi.org/10.1186/1758-2946-3-33>

164. 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>

165. 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>

166. Molecular descriptors for structure–activity applications: a hands-on approach / F. Grisoni et al. *Methods in molecular biology*. New York, NY, 2018. P. 3–53. URL: [https://doi.org/10.1007/978-1-4939-7899-1\\_1](https://doi.org/10.1007/978-1-4939-7899-1_1)

167. Topliss J. G., Edwards R. P. Chance factors in studies of quantitative structure–activity relationships. *Journal of medicinal chemistry*. 1979. Vol. 22, no. 10. P. 1238–1244. URL: <https://doi.org/10.1021/jm00196a017>

168. ChemoPy: freely available python package for computational biology and chemoinformatics / D.-S. Cao et al. *Bioinformatics*. 2013. Vol. 29, no. 8. P. 1092–1094. URL: <https://doi.org/10.1093/bioinformatics/btt105>

169. Chemistry Development Kit (CDK): [Електронний ресурс]. – Режим доступу: <http://sourceforge.net/projects/cdk>

170. RDKit: [Електронний ресурс]. – Режим доступу: <http://sourceforge.net/projects/rdkit/>

171. Dragon (Software for Molecular Descriptor Calculation): [Електронний ресурс]. – Режим доступу: [http://www.taletе.mi.it/products/dragon\\_description.htm](http://www.taletе.mi.it/products/dragon_description.htm)

172. Mordred: a molecular descriptor calculator / H. Moriwaki et al. Journal of cheminformatics. 2018. Vol. 10, no.1. URL: <https://doi.org/10.1186/s13321-018-0258-y>
173. 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>
174. 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>
175. 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>
176. 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>
177. In silico prediction of chemical ames mutagenicity / C. Xu et al. Journal of chemical information and modeling. 2012. Vol. 52, no. 11. P. 2840–2847. URL: <https://doi.org/10.1021/ci300400a>
178. Durbin R., Eddy S., Krogh A., Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. — Cambridge: Cambridge University Press, 1998. — 371 c. — ISBN 978-0-521-62971-3.
179. Venkatraman V. FP-MAP: an extensive library of fingerprint-based molecular activity prediction tools. Frontiers in chemistry. 2023. Vol. 11. URL: <https://doi.org/10.3389/fchem.2023.1239467>
180. Daylight Chemical Information Systems [Электронный ресурс]. – Режим доступа: <http://www.daylight.com>
181. Effectiveness of molecular fingerprints for exploring the chemical space of natural products / D. Boldini et al. Journal of cheminformatics. 2024. Vol. 16, no. 1. URL: <https://doi.org/10.1186/s13321-024-00830-3>

182. Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project / A. Furuhashi 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>
183. Read this paper if you want to learn logistic regression / A. A. T. Fernandes et al. Revista de sociologia e política. 2020. Vol. 28, no. 74. URL: <https://doi.org/10.1590/1678-987320287406en>
184. IBM. What is random forest? [Электронный ресурс] // IBM. – Режим доступа: <https://www.ibm.com/think/topics/random-forest>.
185. Random forest: A classification and regression tool for compound classification and QSAR modeling / V. Svetnik et al. Journal of chemical information and computer sciences. 2003. Vol. 43, no. 6. P. 1947–1958. URL: <https://doi.org/10.1021/ci034160g>
186. Guyon I., Elisseeff A. An introduction to variable and feature selection // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 1157–1182.
187. 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>
188. Hong J., Kwon H. Multimodal deep learning for chemical toxicity prediction and management. Scientific reports. 2025. Vol. 15, no. 1. URL: <https://doi.org/10.1038/s41598-025-95720-5>
189. Machine Learning – XGBoost [Электронный ресурс] // GeeksforGeeks. – Режим доступа: <https://www.geeksforgeeks.org/machine-learning/xgboost/>
190. Chen T., Guestrin C. XGBoost. KDD '16: the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco California USA. New York, NY, USA, 2016. URL: <https://doi.org/10.1145/2939672.2939785>
191. A deep neural network–based approach for prediction of mutagenicity of compounds / R. Kumar et al. Environmental science and pollution research. 2021. Vol. 28, no. 34. P. 47641–47650. URL: <https://doi.org/10.1007/s11356-021-14028-9>

192. Multitask deep neural networks for ames mutagenicity prediction / M. J. Martínez et al. *Journal of chemical information and modeling*. 2022. URL: <https://doi.org/10.1021/acs.jcim.2c00532>
193. Review of machine learning and deep learning models for toxicity prediction / W. Guo et al. *Experimental biology and medicine*. 2023. URL: <https://doi.org/10.1177/15353702231209421>
194. Artificial: understanding the basic concepts without mathematics / S.-H. Han et al. *Dementia and neurocognitive disorders*. 2018. Vol. 17, no. 3. P. 83. URL: <https://doi.org/10.12779/dnd.2018.17.3.83>
195. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, Banff, 14-16 April 2014.
196. 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: [10.1093/mutage/geac010](https://doi.org/10.1093/mutage/geac010).
197. DeepAmes: A deep learning-powered Ames test predictive model with potential for regulatory application / T. Li et al. *Regulatory toxicology and pharmacology*. 2023. P. 105486. URL: <https://doi.org/10.1016/j.yrtph.2023.105486>
198. Predicting drug-induced liver injury using convolutional neural network and molecular fingerprint-embedded features / T.-H. Nguyen-Vo et al. *ACS omega*. 2020. Vol. 5, no. 39. P. 25432–25439. URL: <https://doi.org/10.1021/acsomega.0c03866>
199. 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>
200. 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>

201. Guyon I., Elisseeff A. An introduction to variable and feature selection // *Journal of Machine Learning Research*. — 2003. — Vol. 3. — P. 1157–1182. — DOI: 10.1162/153244303322753616.
202. 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>
203. 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>
204. *In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу / С. Кисляк та ін. *Біомедична інженерія і технологія*. 2025. Т. 4, № 19. С. 48–62. URL: <https://doi.org/10.20535/.2025.19.340327>
205. 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](https://doi.org/10.1007/978-3-642-48736-1_5)
206. *In silico* моделі прогнозування мутагенності Еймса на основі відбитків молекулярної структури ксенобіотиків / С. Кисляк та ін. *Біомедична інженерія і технологія*. 2025. Т. 5, № 20. С. 1–14. URL: <https://doi.org/10.20535/.2025.20.340837>
207. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts / H. Yang et al. *Frontiers in chemistry*. 2018. Vol. 6. URL: <https://doi.org/10.3389/fchem.2018.00030>
208. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity / S. J. Swamidass et al. *Bioinformatics*. 2005. Vol. 21, Suppl 1. P. i359–i368. URL: <https://doi.org/10.1093/bioinformatics/bti1055>
209. Willett P. The calculation of molecular structural similarity: principles and practice. *Molecular informatics*. 2014. Vol. 33, no. 6-7. P. 403–413. URL: <https://doi.org/10.1002/minf.201400024>

210. Machine learning on drug-specific data to predict small molecule teratogenicity / A. P. Challa et al. *Reproductive toxicology*. 2020. Vol. 95. P. 148–158. URL: <https://doi.org/10.1016/j.reprotox.2020.05.004>
211. From high dimensions to human insight: exploring dimensionality reduction for chemical space visualization / A. A. Orlov et al. *Molecular informatics*. 2024. URL: <https://doi.org/10.1002/minf.202400265>
212. Система ідентифікації структурних маркерів мутагенності Еймса на основі подібності відбитків структури ксенобіотиків/ Кисляк С.В. та ін. Вісник Харківського національного університету імені В.Н. Каразіна. Серія «Біологія», 2025, 44, с. 6-14. URL: <https://doi.org/10.26565/2075-5457-2025-44-1>
213. A rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess potential for genotoxicity / L. Müller et al. *Regulatory toxicology and pharmacology*. 2006. Vol. 44, no. 3. P. 198–211. URL: <https://doi.org/10.1016/j.yrtph.2005.12.001>

## ДОДАТОК А

### СПИСОК ПУБЛІКАЦІЙ ТА ВІДОМОСТІ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

1. 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>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються експериментальні *in vitro* та *in vivo* методи, що використовуються для генетичної оцінки впливу факторів навколишнього середовища, написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Яловенко О.І. – критичний аналіз.

2. Kislyak S., Dugan O., Yesypenko R., Starosyla D., Yalovenko O. In silico the Ames Mutagenicity Predictive Model of Environment. // Innovative Biosystems and Bioengineering. – 2025. –Vol. 9, no. 2. – P. 42–52. URL: <https://doi.org/10.20535/ibb.2025.9.2.316239>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються сучасні *in silico* методи, що використовуються для генетичної оцінки впливу факторів навколишнього середовища. Розробка методології проведення *in silico* моделювання, з урахуванням якісного складу молекулярних дескрипторів. Формування бази даних хімічних сполук. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Єсипенко Р.В. – реалізація моделей; Яловенко О.І. – критичний аналіз.

3. Кисляк С. В., Дуган О. М., Мороз М. О., Яловенко О. І. Система ідентифікації структурних маркерів мутагенності Еймса на основі подібності відбитків структури ксенобіотиків. // Вісник Харківського національного університету ім. В. Н. Каразіна. Серія: Біологія. – 2025. – № 44, вип. 1. – С. 6-14. URL: <https://doi.org/10.26565/2075-5457-2025-44-1>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються *in silico* методи генетичної оцінки впливу факторів навколишнього

середовища, прогностична здатність яких базується на ідентифікації функціональних груп або/і підструктур, що є визначальними з точки зору проявів їх мутагенності. Формування бази даних хімічних сполук. Розподіл хімічних сполук за структурними класами. Розрахунок 2D молекулярних дескрипторів (MACCS, RDkit та FCFP). Розроблений підхід щодо оцінки мутагенного потенціалу, що ґрунтується на структурній подібності між досліджуваними потенційними генотоксичними сполуками. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Мороз М. О. – реалізація моделей; Яловенко О.І. – критичний аналіз статті.

4. Кисляк С. В., Дуган О. М., Єсипенко Р.В., Яловенко О.І. In silico моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків на основі методу випадкового лісу. // Біомедична інженерія і технологія. – 2025. – № 19(4). – С. 48-62 URL: <https://doi.org/10.20535/.2025.19.340327>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються сучасні Ames/QSAR моделі, що використовуються для генетичної оцінки впливу факторів навколишнього середовища. Формування бази даних хімічних сполук. Розрахунок молекулярних дескрипторів та розподіл хімічних сполук за структурними класами. Розробка методології покращення точності *in silico* моделей прогнозування мутагенності Еймса на основі методу випадкового лісу. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Єсипенко Р.В. – реалізація моделей; Яловенко О.І. – критичний аналіз.

5. Кисляк С. В., Дуган О. М., Романюк Д.І, Яловенко О.І. In silico моделі прогнозування мутагенності Еймса на основі відбитків молекулярної структури ксенобіотиків.// Біомедична інженерія і технологія. – 2025. – № 20(5). – С. 1-14. URL: <https://doi.org/10.20535/.2025.20.340837>

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, в яких розглядаються сучасні Ames/QSAR моделі, що використовуються для генетичної оцінки впливу факторів навколишнього середовища. Формування бази даних



хімічних сполук. Розрахунок молекулярних дескрипторів та розподіл хімічних сполук за структурними класами. Розробка методології покращення точності *in silico* моделей прогнозування мутагенності Еймса на основі відбітків молекулярної структури з застосуванням ансамблевих алгоритмів машинного навчання та нейронно-мережевого підходу. Аналіз результатів моделювання. Написання статті; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Романюк Д.І. – реалізація моделей; Яловенко О.І. – критичний аналіз

### **Апробація матеріалів дисертації**

6. Кисляк С.В., Есипенко Р.В. *In silico* моделі оцінки генотоксичності впливу факторів навколишнього середовища. //Current challenges of science and education. Proceedings of the 9th International scientific and practical conference. 2024 May 6-8. MDPC Publishing. Berlin, Germany. 2024. P. 50–53.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, підготовка тез до конференції; Дуган О.М. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення; Есипенко Р.В. – реалізація моделей.

7. Кисляк С.В., Дуган О.М., Яловенко О.І. *In silico* моделі генетичної оцінки впливу факторів навколишнього середовища.// Science and society: modern trends in a changing world. Proceedings of the 9th International scientific and practical conference. 2024 Aug. 5-7. MDPC Publishing. Vienna, Austria. 2024. P. 25-28.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, підготовка тез до конференції; Дуган О.М., Яловенко О.І. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення.

8. Кисляк С.В., Дуган О.М., Яловенко О.І. Методи оцінки генотоксичних ефектів факторів навколишнього середовища. // European congress of scientific achievements. Proceedings of the 8th International scientific and practical conference. 2024 Aug. 12-14. Barca Academy Publishing. Barcelona, Spain. 2024. P. 9-15.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, підготовка тез до конференції; Дуган О.М., Яловенко О.І. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення.

9. Кисляк С.В., Дуган О.М., Яловенко О.І. Оптимізація *in silico* моделей прогнозування мутагенності Еймса через зменшення розмірності вхідних даних // Current trends in scientific research development. Proceedings of the 11th International scientific and practical conference. 2025 June 5-7. BoScience Publisher. Boston, USA. 2025. P. 31-37.

Внесок авторів: Кисляк С.В – опрацювання літературних джерел, розробка методології *in silico* моделювання з урахуванням якісного складу молекулярних дескрипторів; формування бази даних хімічних сполук; аналіз результатів моделювання; підготовка тез до конференції; Дуган О.М., Яловенко О.І. – концепція роботи та дизайн, критичний аналіз та остаточне схвалення.

**ДОДАТОК Б**  
**Перелік релевантних 1D та 2D дескрипторів RDkit основних структурних класів хімічних сполук**

Таблиця Б.1

**Релеванті молекулярні дескриптори RDkit для аліфатичних ациклічних хімічних сполук**

<b>BalabanJ</b>	<b>EState_VSA8</b>	<b>NOCCount</b>	<b>PEOE_VSA5</b>	<b>fr_Al_OH</b>
<b>BertzCT</b>	<b>MaxAbsEStateIndex</b>	<b>NHOHCount</b>	<b>PEOE_VSA7</b>	<b>VSA_EState8</b>
<b>Chi0</b>	<b>MaxAbsPartialCharge</b>	<b>SlogP_VSA10</b>	<b>PEOE_VSA8</b>	<b>VSA_EState9</b>
<i>Chi0v</i>	<i>ExactMolWt</i>	<b>PEOE_VSA9</b>	<b>SlogP_VSA2</b>	<b>SlogP_VSA6</b>
<i>Chi2v</i>	<b>FractionCSP3</b>	<b>SMR_VSA1</b>	<b>SlogP_VSA12</b>	<b>fr_allylic_oxid</b>
<i>Chi3v</i>	<b>HallKierAlpha</b>	<b>NumHeteroatoms</b>	<b>SMR_VSA9</b>	<b>fr_amide</b>
<b>EState_VSA10</b>	<b>Ipc</b>	<b>PEOE_VSA1</b>	<i>SMR_VSA2</i>	<b>fr_halogen</b>
<b>EState_VSA1</b>	<b>Kappa2</b>	<b>PEOE_VSA10</b>	<b>SMR_VSA5</b>	<b>fr_alkyl_halide</b>
<b>EState_VSA4</b>	<i>Kappa3</i>	<b>PEOE_VSA12</b>	<b>SMR_VSA4</b>	<b>fr_aldehyde</b>
<b>EState_VSA5</b>	<b>MinAbsPartialCharge</b>	<b>PEOE_VSA13</b>	<b>SMR_VSA7</b>	<i>fr_unbrch_alkane</i>
<b>EState_VSA2</b>	<b>MinPartialCharge</b>	<b>PEOE_VSA4</b>	<b>SMR_VSA6</b>	<b>fr_NH2</b>
<b>EState_VSA3</b>	<b>MaxPartialCharge</b>	<b>PEOE_VSA14</b>	<b>SlogP_VSA3</b>	<i>fr_NH1</i>
<b>EState_VSA6</b>	<b>MinAbsEStateIndex</b>	<b>PEOE_VSA2</b>	<b>SlogP_VSA1</b>	<b>fr_nitroso</b>
<b>EState_VSA7</b>	<b>MolLogP</b>	<b>PEOE_VSA3</b>	<b>fr_N_O</b>	<b>fr_nitro</b>
<b>EState_VSA9</b>	<i>NumHAcceptors</i>	<b>PEOE_VSA6</b>	<b>fr_NH0</b>	

Таблиця Б.2

**Релевантні молекулярні дескриптори RDkit для аліфатичних  
гетеромоно(полі)циклічних хімічних сполук**

<b>BalabanJ</b>	<b>HallKierAlpha</b>	<b>PEOE_VSA14</b>	<b>SMR_VSA10</b>	<b>fr_aldehyde</b>
<b>BertzCT</b>	<b>MaxPartialCharge</b>	<b>PEOE_VSA2</b>	<b>SMR_VSA1</b>	<b>fr_alkyl_halide</b>
<b>EState_VSA1</b>	<b>MaxAbsPartialCharge</b>	<b>PEOE_VSA3</b>	<b>PEOE_VSA9</b>	<b>fr_nitroso</b>
<b>EState_VSA2</b>	<b>Kappa2</b>	<b>PEOE_VSA12</b>	<b>PEOE_VSA8</b>	<i>fr_oxime</i>
<b>EState_VSA4</b>	<b>NumAliphaticCarbocycles</b>	<b>PEOE_VSA1</b>	<b>SMR_VSA5</b>	<b>fr_ketone</b>
<b>EState_VSA3</b>	<b>NOCOUNT</b>	<b>PEOE_VSA10</b>	<b>SMR_VSA3</b>	<b>fr_piperzine</b>
<b>EState_VSA6</b>	<b>NHOHCount</b>	<b>PEOE_VSA11</b>	<b>fr_allylic_oxid</b>	<b>fr_epoxide</b>
<b>EState_VSA5</b>	<b>MinAbsEStateIndex</b>	<b>NumRotatableBond</b>	<b>fr_amide</b>	<b>fr_halogen</b>
<b>EState_VSA9</b>	<b>MinAbsPartialCharge</b>	<b>SlogP_VSA2</b>	<i>fr_ether</i>	
<b>FractionCSP3</b>	<b>MolLogP</b>	<b>SlogP_VSA3</b>	<b>fr_NH0</b>	
<b>EState_VSA7</b>	<b>SMR_VSA7</b>	<b>SlogP_VSA1</b>	<b>fr_Al_OH</b>	
<b>EState_VSA8</b>	<b>PEOE_VSA4</b>	<b>SMR_VSA6</b>	<b>SlogP_VSA5</b>	

**Релевантні молекулярні дескриптори RDkit для ароматичних  
гетеромоно(полі)циклічних хімічних сполук**

<b>BalabanJ</b>	<i>fr_azide</i>	<i>fr_Ar_NH</i>	<b>NumRotatable Bonds</b>	<b>SlogP_VSA1</b>
<b>BertzCT</b>	<b>fr_aniline</b>	<b>fr_alkyl_halide</b>	<i>NumAromatic Rings</i>	<b>SMR_VSA9</b>
<b>Chi0</b>	<b>fr_amide</b>	<b>fr_ketone</b>	<b>SMR_VSA7</b>	<b>SMR_VSA5</b>
<b>EState_VSA1</b>	<b>SlogP_VSA8</b>	<b>Ipc</b>	<b>SMR_VSA4</b>	<b>SMR_VSA6</b>
<b>EState_VSA10</b>	<b>VSA_EState9</b>	<i>NumAliphaticRings</i>	<b>PEOE_VSA8</b>	<b>RingCount</b>
<b>EState_VSA2</b>	<i>VSA_EState10</i>	<b>MaxAbsEStateIndex</b>	<b>PEOE_VSA9</b>	<b>SMR_VSA1</b>
<b>EState_VSA3</b>	<b>VSA_EState8</b>	<b>MaxAbsPartial Charge</b>	<b>PEOE_VSA1</b>	<b>SMR_VSA10</b>
<b>EState_VSA4</b>	<i>fr_Ndealkylation1</i>	<b>MaxPartialCharge</b>	<b>PEOE_VSA10</b>	<b>SMR_VSA3</b>
<b>EState_VSA8</b>	<b>fr_NH0</b>	<b>MinAbsEStateIndex</b>	<b>PEOE_VSA12</b>	<b>SlogP_VSA10</b>
<b>EState_VSA5</b>	<b>fr_NH2</b>	<b>MinEStateIndex</b>	<b>PEOE_VSA11</b>	<b>SlogP_VSA7</b>
<b>EState_VSA6</b>	<i>fr_Ar_N</i>	<b>MinPartialCharge</b>	<b>PEOE_VSA2</b>	<b>SlogP_VSA5</b>
<b>EState_VSA7</b>	<i>fr_sulfonamd</i>	<b>MolLogP</b>	<b>PEOE_VSA3</b>	<b>SlogP_VSA6</b>
<b>SlogP_VSA2</b>	<b>fr_piperzine</b>	<i>NumAromatic Heterocycles</i>	<b>PEOE_VSA13</b>	<i>SlogP_VSA4</i>
<b>SlogP_VSA12</b>	<i>fr_pyridine</i>	<b>NHOHCount</b>	<b>PEOE_VSA14</b>	<b>SlogP_VSA3</b>
<b>fr_epoxide</b>	<b>fr_para_hydroxylation</b>	<b>NOCCount</b>	<b>PEOE_VSA4</b>	
<b>fr_halogen</b>	<b>fr_nitroso</b>	<b>NumAliphatic Carbocycles</b>	<b>PEOE_VSA5</b>	
<b>fr_hdrzine</b>	<b>fr_nitro</b>	<i>NumSaturated Heterocycles</i>	<b>PEOE_VSA7</b>	
<i>fr_bicyclic</i>	<i>fr_thiazole</i>	<i>NumSaturatedRings</i>	<b>PEOE_VSA6</b>	

Таблиця Б.4

**Релеванті молекулярні дескриптори RDkit для ароматичних  
гомомоно(полі)циклічних хімічних сполук**

<b>EState_VSA1</b>	<b>MaxPartialCharge</b>	<b>PEOE_VSA3</b>	<i>fr_ArN</i>	<b>fr_alkyl_halide</b>
<b>Chi0</b>	<b>MinAbsPartialCharge</b>	<b>PEOE_VSA13</b>	<i>fr_Al_COO</i>	<b>fr_NH0</b>
<b>EState_VSA10</b>	<b>Kappa2</b>	<b>PEOE_VSA11</b>	<b>fr_para_hydroxylation</b>	<b>fr_allylic_oxid</b>
<b>FractionCSP3</b>	<b>MinEStateIndex</b>	<b>PEOE_VSA12</b>	<b>fr_hdrzine</b>	<b>SlogP_VSA3</b>
<b>EState_VSA8</b>	<b>SMR_VSA10</b>	<b>PEOE_VSA1</b>	<i>fr_sulfide</i>	<b>SlogP_VSA7</b>
<b>EState_VSA9</b>	<b>SMR_VSA5</b>	<b>NumHeteroatoms</b>	<b>fr_nitro</b>	<b>VSA_EState9</b>
<b>EState_VSA7</b>	<b>SlogP_VSA10</b>	<b>SlogP_VSA2</b>	<i>fr_nitro_arom_nonortho</i>	<i>fr_Ar_OH</i>
<b>EState_VSA4</b>	<b>PEOE_VSA5</b>	<i>SlogP_VSA11</i>	<i>fr_ketone_Topliss</i>	<b>SlogP_VSA8</b>
<b>HallKierAlpha</b>	<b>PEOE_VSA8</b>	<b>SMR_VSA7</b>	<b>fr_ketone</b>	<b>SlogP_VSA5</b>
<b>MinPartialCharge</b>	<b>PEOE_VSA6</b>	<b>SMR_VSA6</b>	<b>fr_aniline</b>	<b>SlogP_VSA6</b>
<b>NHOHCount</b>	<b>PEOE_VSA7</b>	<b>SlogP_VSA1</b>	<b>fr_halogen</b>	
<i>NumAromaticCarbocycles</i>	<b>PEOE_VSA14</b>	<b>SMR_VSA9</b>	<i>fr_C_O_noCOO</i>	
<b>MaxAbsPartialCharge</b>	<b>PEOE_VSA2</b>	<b>RingCount</b>	<b>fr_N_O</b>	

## ДОДАТОК В

### АКТ ВПРОВАДЖЕННЯ

**ЗАТВЕРДЖУЮ**

Проректор з навчальної роботи

Національного технічного університету  
України «Київський політехнічний  
інститут імені Ігоря Сікорського»



Тетяна ЖЕЛЯСКОВА

«\_\_\_» \_\_\_\_\_ 2025 р.

### АКТ

впровадження результатів дисертаційного дослідження аспіранта Кисляка Сергія Володимировича на тему: «*In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків»

Комісія у складі:

голова – завідувачка кафедри біоенергетики, біоінформатики та екобіотехнології КПІ ім. Ігоря Сікорського, д.т.н., доц. Голуб Н.Б.;

члени комісії – проф., д.т.н., проф. Саблій Л.А.;  
доц., к.т.н., доц. Козар М.Ю.

цим Актом засвідчує, що результати дисертаційного дослідження Кисляка Сергія Володимировича на тему: «*In silico* моделі прогнозування мутагенності Еймса основних структурних класів ксенобіотиків» впроваджено у викладанні дисциплін вибіркового блоку «Моделювання молекулярної взаємодії» (2022/2023 н.р.) та «Пакети прикладних програм для задач молекулярної біології» (2024/2025 н.р.) для здобувачів другого (магістерського) рівня вищої освіти спеціальності 162 Біотехнології та біоінженерія.

В навчальний процес було впроваджено:

1. Методику розробки ефективних Ames/QSAR моделей орієнтованих на основні структурні класи органічних сполук
2. Методику оцінки мутагенних ефектів на основі відбитків молекулярної структури органічних сполук.
3. Методику пошуку причинно-наслідкових зв'язків між мутагенністю та релевантними дескрипторами основних структурних класів органічних сполук.

**Голова комісії**

Завідувачка кафедри д.т.н., доц.

Наталія ГОЛУБ

**Члени комісії**

Проф., д.т.н., проф.

Лариса САБЛІЙ

Доц., к.т.н., доц.

Марина КОЗАР