

Список використаних джерел

1. H.-Y. Li, W. Zeng, F.-S. Liu, Z.-T. Liu, Z.-X. Bai, Q.-J. Liu. Theoretical study of friction sensitivity of energetic materials. *J. Phys. Chem. Solids*, 2025, 202, 112692
2. S. V. Bondarchuk. Friction sensitivity of nitramine energetic materials: a prediction based on genetic function approximation. *Fire Phys Chem* 2022, 2 (3), 272-278.
3. H. Gao, J. M. Shreeve. Azole-based energetic salts. *Chem. Rev.* 2011, 111 (11), 7377–7436.
4. H. Gao, Q. Zhang, J. M. Shreeve. Fused heterocycle-based energetic materials (2012–2019). *J. Mater. Chem.* A2020, 8, 4193–4216.
5. Z. Zhang L. Wen T. Yu, C. Chen, S. V. Bondarchuk, Y. Liu. “Unraveling the Tangle” of calculated solid-state enthalpies of formation of C–H–N–O energetic materials with a high-precision approach based on the predicted asymmetric cell total energy evaluation. *Fuel* 2026, 404, 136307.
6. Materials Studio 2023; Dassault Systèmes BIOVIA, San Diego, CA, 2023.

ПРОГНОЗУВАННЯ ТОКСИЧНОСТІ ОРГАНІЧНИХ РЕЧОВИН ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Анастасія Комкова, Сергій Бондарчук

Черкаський національний університет імені Богдана Хмельницького
e-mail: komkova.anastasiia2222@vu.cdu.edu.ua, bondchem@cdu.edu.ua

Оцінка токсичності нових хімічних сполук є обов'язковою для багатьох галузей (фармацевтика, харчова, сільськогосподарська, промислова тощо). Комп'ютерні методи є сучасною альтернативою класичним тестам, оскільки вони замінюють дослідження на тваринах, здешевлюють процес і прискорюють пошук безпечних речовин [1].

Моделювання базується на принципі QSAR – математичному прогнозуванні біологічної/токсичної дії невідомих молекул на основі аналізу їхньої хімічної структури та властивостей.

Основою дослідження стали дані, наведені у статті [2], присвяченій прогнозуванню гострої пероральної токсичності пестицидів для американської перепілки.

З архівів EFSA Open Food Tox, EPA ECOTOX Knowledgebase та PPDB узято значення токсичності з параметрами: пероральне введення, гострий характер впливу та тривалість спостереження не менше 14 днів. Попередня фільтрація виключила неорганічні речовини, полімери, суміші, дублікати та сполуки з неповними даними, що забезпечило вибірку тільки з органічних пестицидів – 199 сполук. Гетерогенність вибірки була високою й охоплювала всі категорії токсичності за GHS [3].

Для побудови моделі сформовано вибірку зі 199 органічних пестицидів, закодованих у форматі SMILES. Перед побудовою моделі провели статистичний аналіз набору даних. За допомогою інструментів Excel перевірили вибірку на промахи і візуалізували розподіл даних через коробковий графік (рис. 1).

За результатами графічного аналізу та додаткової перевірки методом стандартизованої оцінки z-балами, зробила висновок, що речовини, які мають значення $LD_{50} = 5000$ мг/кг маси тіла (2 молекули); $LD_{50} = 7000$ мг/кг маси тіла (1 молекула) й $LD_{50} = 4400$ мг/кг маси тіла (1 молекула) є промахами. Такі дані вважаються нерепрезентативними для загального набору даних, але ми вирішили перенести їх у зовнішній тестовий набір, щоб перевірити перспективність в розрахунку для структурно нетипових молекул.

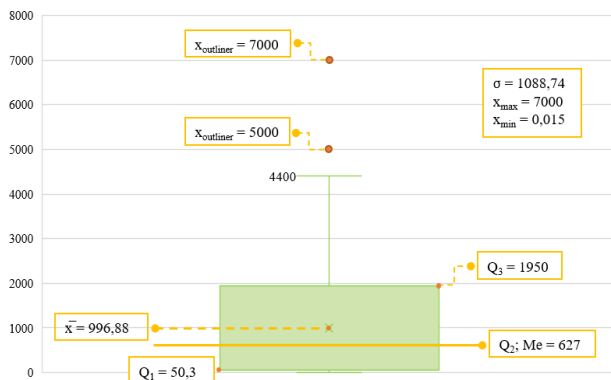


Рис. 1. Статистичний розподіл значень LD_{50} для вибірки зі 199 сполук

Для побудови моделі використовували метод апроксимації генетичної функції (GFA), який нещодавно продемонстрував свою ефективність у передбаченні водної токсичності високоазотних сполук до *Aliivibriofischeri* [4]. Масив даних був розбитий на тренувальний і тестовий набори (150/49) й всі налаштовані параметри GFA наведені у таблиці 1:

Таблиця 1.

Початкові параметри регресійного аналізу методом GFA

Розмір популяції	70
Максимальна кількість поколінь	100000
Початкові члени рівняння	20
Максимальна довжина рівняння	35
Стала довжина рівняння	Ні
Ймовірність мутації	0.1
Лінійні сплайни	Так
Функція ранжування	Fridman LOF

У ході регресійного аналізу отримано рівняння:

$$LD_{50} = -28.663 \cdot S_{ssO} - 217.088 \cdot \langle AlogP + 0.84 \rangle - 242.56 \cdot \langle Balaban\ index\ JY - 2.15 \rangle + 366.526 \cdot \langle \chi(3):cluster - 1.0868 \rangle + 1955.17 \cdot \langle S_{aaaC} - 1.567 \rangle - 242.42 \cdot \langle 13.785 - \chi(1) \rangle + 212.461 \cdot \langle 9.118 - \chi(3):path \rangle + 2120.366$$

де S_{ssO} – атоми Оксигену, які з'єднані з молекулою двома одинарними зв'язками; $AlogP$ – коефіцієнт розподілу октанол/вода (показник ліпофільності); $Balabanindex\ JY$ – індекс ступеня розгалуженості вуглецевого скелета молекули; $\chi(3):cluster$ – індекс молекулярної зв'язності 3-го порядку (3 зв'язки сходяться в один вузол); S_{aaaC} – карбону, які знаходяться всередині ароматичного кільця і сполучені трьома ароматичними зв'язками; $\chi(1)$ – індекс одинарних зв'язків між неконденсованими атомами й показує розмір та площу поверхні молекули; $\chi(3):path$ – індекс молекулярної зв'язності 3-го порядку, який оцінює лінійність ланцюжків сформованих чотирма послідовно зв'язаними атомами.

Якість побудованої моделі оцінювалася порівнянням коефіцієнтів (R^2). Близькість

значень коефіцієнтів детермінації ($R^2 = 0,46$), перехресної валідації ($R^2_{\text{кв}} = 0,42$) та скоригованого коефіцієнта ($R^2_{\text{скуп}} = 0,43$) свідчить про те, що модель не перенавчена і надійна. Також це підтверджується критерієм Фішера, який значно вищий ($F = 17,12$) за порогове значення, яке дорівнює 2,08 при довірчій ймовірності 95 %. Тобто рівняння регресії є статистично значущим й придатне для прогнозування.

Перший графік було навмисно побудовано з викидами й коефіцієнт детермінації для прогнозу становив 0,11 відносно експериментального ряду для якого $R^2 = 0,84$. Напрямок лінії тренду підтверджує, що у ході аналізу, модель змогла визначити ключові закономірності й задовільно прогнозує загальну тенденцію як низьких, так і високих значень токсичності. Коефіцієнти ймовірно можна збільшити видаливши викиди, які погіршують загальне значення показників. Відповідність між експериментальними й прогнозованими показниками LD_{50} з виключенням промахів візуалізовано на графіку:



Рис. 2. Порівняльний аналіз експериментальних та прогнозованих значень LD_{50} пестицидів за отриманою QSAR-моделлю без показу викидів

Математичні параметри моделі значно покращилися після вилучення нетипових значень, які порушували критерії загальної хімічної однорідності вибірки. На графіку (рис. 2) коефіцієнт детермінації для прогнозу збільшився до 0,72 при експериментальному $R^2 = 0,84$. Також, можна зробити висновок, що модель найкраще описує речовини значення LD_{50} яких знаходиться приблизно в діапазоні 0–2000 мг/кг, що не є дивним, бо загальна вибірка включала найбільше сполук зі значеннями саме у цьому інтервалі токсичності. Значення 4400, 5000 й 7000, які класифіковані як викиди були розраховані не коректно через те, що у моделі було недостатньо матеріалу для навчання з цього проміжку.

Це говорить про те, що рівняння, яке було отримано у ході регресійного аналізу можна вільно використовувати для попереднього скринінгу або грубої експрес-оцінки й доводить те, що з вибірки варто виключати викиди, щоб результати були точніші й мали кращу відтворюваність.

Список використаних джерел

1. Tonoyan L., Siraki A. G.. Machine Learning in Toxicological Sciences: Opportunities for Assessing Drug Toxicity. Front. Drug Discov. – 2024. – Vol. 4. – Article: 1336025.
2. Iovine N., Roncagliani A., Benfenati E. Predicting Acute Oral Toxicity in Bobwhite Quail: Development of QSAR Models for LD_{50} . Environments – 2025. – Vol. 12, № 2. – Article: 56.
3. Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Tenth revised edition. ST/SG/AC.10/30/Rev.10. New York and Geneva : United Nations, 2023. 592 p. URL: <https://unece.org/transport/dangerous-goods/ghs-rev10-2023>
4. Bondarchuk S. V.. Prediction of aquatic toxicity of energetic materials using genetic function approximation. Fire Phys Chem – 2023. – Vol. 3, № 1. – P. 23-28.