

Особливості формування спеціалізованого набору даних для прогнозування хімічних властивостей фторвмісних сполук

Вінницький національний технічний університет

Анотація

У роботі представлено особливості та результати формування спеціалізованого набору даних для прогнозування фізико-хімічних властивостей насичених фторвмісних сполук. Пропонований датасет містить експериментальні значення pK_a та $\log P$ для 183 сполук і охоплює кілька структурних класів, що мають значення для медичної хімії. Проведено попередню обробку даних, включно з видаленням викидів, стратифікованим поділом та контролем структурної подібності молекул. Первинний аналіз виявив нелінійні залежності між кількістю та положенням атомів фтору і властивостями сполук, що обґрунтовує доцільність застосування методів машинного навчання для подальшого моделювання.

Ключові слова: фторвмісні сполуки, набір даних, pK_a , $\log P$, попередня обробка даних, машинне навчання, прогнозування властивостей.

Abstract

The paper presents the features and results of forming a specialized dataset for predicting the physicochemical properties of saturated fluorine-containing compounds. The proposed dataset includes experimental pK_a and $\log P$ values for 183 compounds and covers several structural classes relevant to medicinal chemistry. Data preprocessing was performed, including outlier removal, stratified splitting, and molecular similarity control. Preliminary analysis revealed nonlinear relationships between the number and position of fluorine atoms and the properties of the compounds, which justifies the use of machine learning methods for further modeling.

Keywords: fluorine-containing compounds, dataset, pK_a , $\log P$, data preprocessing, machine learning, property prediction.

Вступ

Фтор є унікальним елементом у фармацевтичній хімії завдяки своїм малим розмірам, високій електронегативності та здатності істотно змінювати фізико-хімічні властивості органічних сполук. Введення атома фтору в молекулу може значно впливати на її кислотність/основність, ліпофільність, стабільність і біологічну активність [1]. Такі зміни зумовлені як індуктивним, так і мезомерним ефектами, що змінюють електронну густину в реакційно-активних центрах молекули. Саме тому передбачення властивостей фторвмісних сполук потребує застосування моделей, здатних враховувати як електронні, так і просторові ефекти фторування.

Попри значну кількість досліджень, присвячених прогнозуванню фізико-хімічних параметрів органічних сполук (зокрема $\log P$ та pK_a [2-5]), систематичні роботи, спрямовані на насичені фторвмісні сполуки, залишаються поодинокими та маловідомими. Більшість існуючих моделей створювалися на основі великих, але загальних наборів даних, що не відображають специфіку цього класу речовин [6-8]. Це призводить до зниження точності передбачення таких важливих властивостей, як ліпофільність та кислотність/основність. Основними чинниками, що ускладнюють прогнозування властивостей фторованих сполук, є:

- нелінійні ефекти – вплив фтору на молекулярні властивості часто має нелінійний характер, через що прості регресійні моделі не забезпечують адекватного опису;
- залежність від положення атомів фтору – хімічні та фізичні властивості сильно залежать від того, де саме в структурі розташований атом або група CF_x ;
- кооперативні ефекти – у поліфторованих сполуках спостерігається взаємодія кількох атомів фтору, вплив яких не є простою сумою індивідуальних ефектів.

Як наслідок, стандартні алгоритми оцінки $\log P$ і pK_a демонструють низьку точність для насичених фторвмісних похідних, що у свою чергу обумовлює необхідність створення спеціалізованого набору даних, який охоплює фторовані та нефторовані аміни й карбонові кислоти [9].

Результати дослідження

Для проведення дослідження було сформовано унікальний спеціалізований датасет, що містить експериментальні значення pK_a та $\log P$ для 183 насичених фторорганічних сполук. Ці дані були отримано з попередніх опублікованих досліджень, що гарантує узгодженість методології вимірювань та високу точність визначення значень. Датасет охоплює три основні класи сполук, кожен з яких представляє важливий структурний тип для медичної хімії.

Для дослідження кислотно-основних властивостей (pK_a) датасет включає 59 первинних екзоциклічних амінів (клас 1), 65 вторинних гетероциклічних амінів (клас 2) та 54 карбонових кислот (клас 3) (рис. 1). Оскільки експериментальні обмеження методу вимірювання $\log P$ вимагали використання модельних сполук, для дослідження ліпофільності було використано відповідні похідні: N-бензоїльні похідні 63 первинних (клас 4) та 65 вторинних амінів (клас 5), а також 55 анілідів карбонових кислот (клас 6) (рис. 2). Така модифікація дозволила коректно визначити коефіцієнт розподілу октанол-вода, зберігаючи при цьому структурні особливості вихідних молекул.

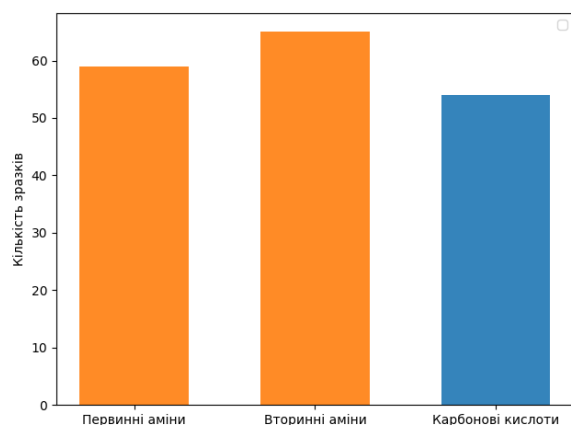


Рисунок 1 – Розподіл кількості молекул за типом із розрахованим pK_a

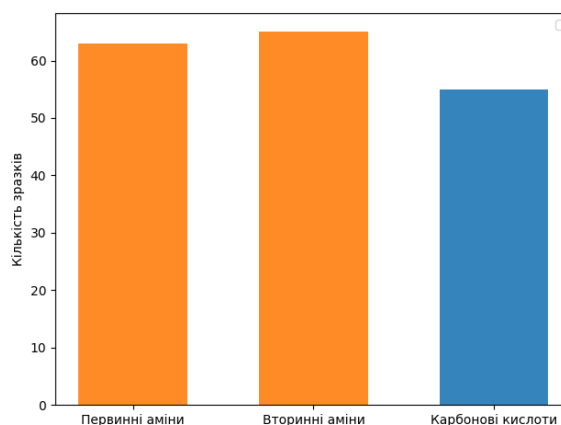


Рисунок 2 – Розподіл кількості молекул за типом із розрахованим $\log P$

Датасет демонструє широкий спектр структурних особливостей, що дозволяє систематично вивчати вплив різних факторів на фізико-хімічні властивості. Включені сполуки охоплюють моноциклічні системи від тричленних до семичленних циклоalkanів, а також більш складні біциклічні структури, включаючи спіроциклічні, мостикові та конденсовані системи. Окремо представлені ациклічні похідні, що дозволяє оцінити вплив циклічної структури як такої. Особлива увага приділена різноманітності фторованих замісників (рис. 3): Монофтор (CHF), Монофторметил (CH_2F), Дифторметил (CHF_2), Трифторметил (CF_3), gem-Дифтор (CF_2)

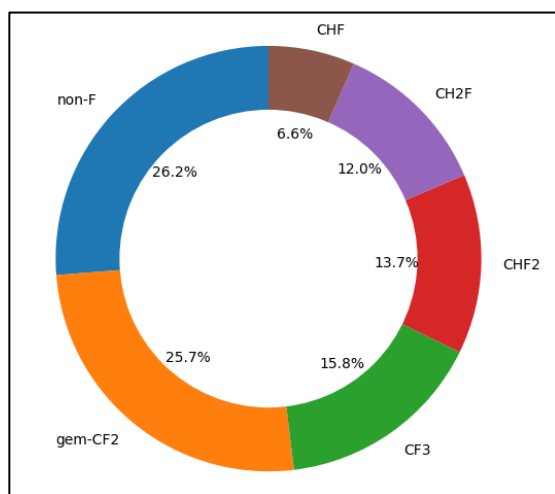


Рисунок 3 – Структурний розподіл молекул за фтором

	Smiles	F group	LogP	pKa
1	FC1(F)CCC(C(O)=O)CC1	gem-CF2	2.63	4.32
2	OC(C1CCCCC1)=O	non-F	2.88	4.85
3	OC([C@@H]1C[C@H]2C[C@H]2C1)=O	non-F	2.87	4.35
4	OC([C@H]1C[C@H]2C[C@H]2C1)=O	non-F	2.92	4.24
5	OC([C@H]1C[C@H]2C(F)(F)[C@H]2C1)=O	gem-CF2	2.37	3.80
...	...	...	...	...
179	FC([H])([C@H]1C[C@H]1N)[H]	CH2F	1.41	NaN
180	FC([H])([C@H]1C[C@H]1N)[H]	CH2F	1.12	NaN
181	NC1C2(CC2)C1	non-F	2.08	9.41
182	NC1C2(CC2)C1(F)F	gem-CF2	2.14	NaN
183	NC1C2(C(F)(F)C2)C1	gem-CF2	2.04	7.44

183 rows × 4 columns

Рисунок 4 – Скріншот структури датасету фторвмісних сполук

Перед використанням датасету (рис. 4) для навчання моделей була проведена також попередня обробка даних. Важливим етапом стало виявлення та видалення викидів, які можуть негативно вплинути на якість моделей. Для цього використовувались різні статистичні методи залежно від типу розподілу даних. Спочатку за допомогою тесту Шапіро-Вілکا визначався характер розподілу кожної ознаки. Для нормально розподілених ознак застосовувався z-score критерій, а в інших випадках – метод інтерквартильного розподілу. В результаті цього процесу було видалено чотирнадцять викидів для задачі прогнозування рKa та десять викидів для задачі прогнозування LogP.

Наступним критичним кроком стало розділення даних на навчальну та тестову вибірки. Було обрано співвідношення 80 до 20, де більша частина використовується для крос-валідації під час навчання, а менша зберігається як незалежний тестовий набір для фінальної оцінки моделей. Щоб забезпечити подібність розподілу цільової змінної в обох підмножинах, застосовувався метод стратифікованого поділу (stratified split). Це особливо важливо, враховуючи що датасет містить різні класи сполук з різними діапазонами значень рKa та LogP.

Особливу увагу було приділено запобіганню витоку даних між навчальною та тестовою вибірками. Оскільки деякі молекули в датасеті є структурно дуже подібними, існував ризик, що інформація про тестові зразки опосередковано потрапить у навчальну вибірку. Для вирішення цієї проблеми була розрахована міра подібності Tanimoto між усіма парами молекул. Молекули з коефіцієнтом подібності вище 90 відсотків були ідентифіковані та переміщені в одну підмножину, зберігаючи при цьому цільові пропорції поділу. Такий підхід гарантує, що модель не "бачить" схожі структури під час навчання і тестування, що забезпечує точнішу оцінку її узагальнюючої здатності.

Попередній аналіз експериментальних даних виявив чіткі закономірності, які підтверджують якість датасету та відповідність хімічній інтуїції. Для вторинних амінів спостерігається систематичне зменшення рKa зі збільшенням кількості атомів фтору та їх наближенням до основного центру (рис. 5). Для ліпофільності закономірності виявилися складнішими та менш передбачуваними. Загальний тренд показує зростання LogP зі збільшенням кількості атомів фтору, причому трифторметильна група демонструє найбільший вплив, значно перевищуючи ефект дифторметильної та монофторметильної груп (рис. 6). Однак вплив інших факторів, таких як розмір циклу, положення замісника та стереохімія, виявився більш складним і нелінійним. Це підкреслює багатофакторну природу ліпофільності та необхідність використання складних моделей для її точного прогнозування.

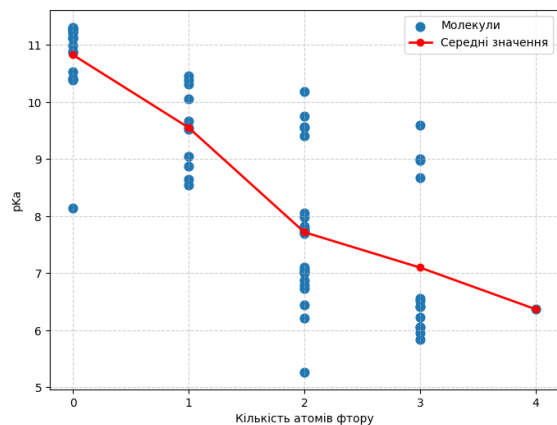


Рисунок 5 – Залежність рKa від кількості атомів фтору у вторинних амінах

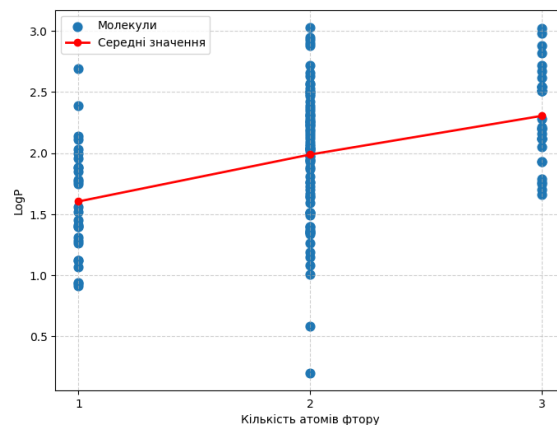


Рисунок 6 – Залежність LogP від кількості атомів фтору

Таким чином, хоч і розмір датасету (рис. 4) обмежений ста вісімдесятьма трьома сполуками, він є достатнім для навчання та валідації моделей завдяки високій якості даних та структурній організації. Відносна структурна подібність молекул дозволяє моделям ефективно навчатися на виявленні тонких відмінностей, що критично важливо для точного прогнозування. Водночас цей малий розмір вимагає ретельного підходу до вибору архітектури моделей та стратегії навчання, щоб уникнути перенавчання.

Висновки

У результаті дослідження створено спеціалізований набір даних для прогнозування кислотно-основних та ліпофільних властивостей насичених фторвмісних сполук. Проведено повний цикл

підготовки даних, включно з очищенням, стратифікованим поділом і контролем структурної подібності, що забезпечує коректність подальшого машинного аналізу. Первинна аналітика підтвердила наявність нелінійних і багатофакторних залежностей між структурою молекул і їхніми властивостями, що визначає доцільність використання методів штучного інтелекту для моделювання таких взаємозв'язків. Отримані результати формують базу для розроблення інтелектуальних систем прогнозування фізико-хімічних параметрів фторвмісних сполук.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Benaiges, D., Pedro-Botet, J., & Climent, E. (2021). Hydrophilic or lipophilic statins? *Frontiers in Cardiovascular Medicine*. <https://doi.org/10.3389/fcvm.2021.687585>
2. Baltruschat, M., & Czodrowski, P. (2020). Machine learning meets pKa. *F1000Research*. <https://doi.org/10.12688/f1000research.22090.2>
3. Caine, B. (2019). On the development and application of AIBL-pKa, a pKa predictor based on equilibrium bond lengths of a single protonation state.
4. Yang, Q., Li, Y., Liu, Y., Zhang, L., Luo, S., & Cheng, J. (2020). Holistic prediction of the pKa in diverse solvents based on a machine-learning approach. *Angewandte Chemie International Edition*, 43. <https://doi.org/10.1002/anie.202008528>
5. Xiong, J., Li, Z., Wang, G., Fu, Z., Zhong, F., Xu, T., Liu, X., Huang, Z., Liu, X., Chen, K., Jiang, H., & Zheng, M. (2022). Multi-instance learning of graph neural networks for aqueous pKa prediction. *Bioinformatics*, 38(3). <https://doi.org/10.1093/bioinformatics/btab714>
6. Li, M., Zhou, J., Hu, J., Fan, W., Zhang, Y., Gu, Y., & Karypis, G. (2017). DGL-LifeSci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega*, 6(41). <https://doi.org/10.1021/acsomega.1c04017>
7. Landrum, G. (2023, September). RDKit: Open-source cheminformatics [Online]. Retrieved from <https://www.rdkit.org>
8. Raddi, R. M., & Voelz, V. A. (2022). Stacking Gaussian processes to improve pKa predictions in the SAMPL7 challenge. *Journal of Computer-Aided Molecular Design*, 35. <https://doi.org/10.1007/s10822-021-00411-8>
9. Gurbych, O., Pavliuk, P., Krasnienkov, D., Liashuk, O., Melnykov, K., Grygorenko, O. (2025). Filling the gap in LogP and pKa evaluation for saturated fluorine-containing derivatives with machine learning. *Journal of Computational Chemistry*, 46(2). <https://doi.org/10.1002/jcc.70002>

Шевчук Олександр Федорович – доцент кафедри комп'ютерних наук, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: shevchuk@vntu.edu.ua

Павлюк Петро Петрович – студент факультету інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: p4tro.pavlyk@gmail.com;

Shevchuk Oleksandr F. – Associate Professor of the Department of Computer Sciences, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: shevchuk@vntu.edu.ua

Pavliuk Petro P. – student, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: p4tro.pavlyk@gmail.com