

ОПТИМІЗАЦІЯ QSAR-МОДЕЛЕЙ ДЛЯ ПЕРЕДБАЧЕННЯ БІОЛОГІЧНОЇ АКТИВНОСТІ МОЛЕКУЛ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Д. В. Маслов <https://orcid.org/0009-0002-1376-1450>

О. А. Голуб <https://orcid.org/0000-0003-1823-2523>

Національний університет «Кієво-Могилянська академія», Україна

Вул. Сковороди 2, 04070 Київ, Україна

e-mail: dv.maslov@ukma.edu.ua

Молекулярне моделювання є важливим інструментом сучасної обчислювальної хімії, яке широко використовують на ранніх етапах розроблення лікарських засобів для передбачення біологічної активності потенційних кандидатів. Дослідження проведено на датасеті з 3782 молекул, описаних 3291 молекулярним дескриптором і значеннями активності рChEMBL у діапазоні 5,01–8,52, який містив 733 унікальні молекулярні структури. Порівняння різних підходів до розділення вибірки показало перевагу стратифікованого scaffold-орієнтованого розподілу, який забезпечив реалістичну оцінку якості моделей із R^2 до 0,72 при MAE = 0,41. Отримано оптимізовану QSAR-модель, яка є придатною для раннього віртуального скринінгу і яку можна використовувати для пріоритизації сполук у процесі розроблення знеболювальних препаратів, спрямованих на рецептор TRPV1.

Ключові слова: QSAR-моделювання, машинне навчання, TRPV1, молекулярні дескриптори.

ВСТУП. Молекулярне моделювання займає провідне місце серед сучасних напрямів обчислювальної хімії. При розробленні нових фармацевтичних засобів дослідники стикаються з необхідністю швидко та ефективно передбачувати біологічну активність великої кількості потенційних кандидатів. Для цього широко використовують моделі кількісних співвідношень структура-активність (QSAR – Quantitative Structure-Activity Relationship).

Традиційні QSAR-моделі базуються на розрахунку молекулярних дескрипторів, які характеризують структурні та фізико-хімічні властивості молекул. Однак якість таких моделей часто переоцінюють через неправильне розділення даних на тренувальний та тестовий набори [1].

Основна проблема, яку розглянуто у цій роботі, що слідує з парадигми машинного навчання [2], полягає в наступному: якщо тестовий набір містить молекулярні

структури (scaffolds), яких немає у тренувальному наборі, то модель не зможе адекватно передбачувати їхню активність, навіть якщо показує високу точність на звичайному випадковому розділенні.

Мета дослідження – розробити та протестувати методи оптимізації QSAR-моделей для потенційних знеболювальних препаратів за концентраціями інгібування (IC50) рецептора болю TRPV1 [3] з урахуванням молекулярної різноманітності набору характеристичних даних – датасету.

ЕКСПЕРИМЕНТ І ОБГОВОРЕННЯ РЕЗУЛЬТАТІВ. Дослідження проводили на датасеті (наборі даних) із 3782 молекул (взяті з бази даних ChEMBL – Chemical European Molecular Biology Laboratory) з обчисленими за допомогою бібліотек (RDKit, Mordred, Morgan fingerprints) 3291 молекулярними дескрипторами. Цільовою змінною була інгібувальна активність до TRPV1 (від’ємний десятковий логарифм IC50 – pChEMBL Value), яка варіювалася від 5.01 до 8.52.

За допомогою випадкової вибірки датасет було розділено на тренувальну (80%) та тестову (20%) вибірки [4]. Одержали датасет, ключовою особливістю якого є наявність 733 унікальних молекулярних структур (scaffolds), при цьому 72 з’явилися лише у тестовому наборі при звичайному випадковому розділенні, що свідчить про значний потенціал для втрати інформації [2].

Зокрема для оцінки втрати інформаційного потенціалу було порівняно три підходи до розділення датасету [5]:

1. Звичайний K-Fold (крос-валідація), або випадкове розділення між наборами без урахування молекулярних структур. У результаті отримали коефіцієнт детермінації $R^2 = 0.54$, яке є

оптимістичною оцінкою через втрату інформації про структури.

2. Груповий K-Fold (Group K-Fold), або розділення так, щоб усі молекули з однаковою (схожою) молекулярною структурою потрапляли лише в один набір. У результаті маємо $R^2 = 0.31$, що показує реальну складність завдання на невидимих структурах.
3. Стратифіковане розділення (Stratified Split) 80/20, яке застосовували у цьому дослідженні з урахуванням розподілу активності та молекулярних структур. У результаті обчислень отримали $R^2 = 0.646-0.720$ залежно від методу моделювання, що показує оптимальність підходу.

У ході роботи було протестовано декілька моделей машинного навчання, зокрема: випадковий ліс (Random Forest, 500 дерев), градієнтне підсилення (Gradient Boosting, GB; 300 дерев, швидкість навчання (learning rate) = 0.05) та нейронні мережі (Neural Networks) із малими та середніми архітектурами.

Для представлення молекулярних структур використовували морганівські відбитки (Morgan fingerprints), а також застосували метод головних компонент (Principal Component Analysis, PCA) для зменшення розмірності дескрипторного простору.

Градієнтне підсилення показало найкращі результати і було обрано для подальших оптимізацій.

Для збільшення передбачуваної сили моделі було застосовано наступні методи оптимізації [6]:

1. Розширення тренувального набору (Data augmentation – Aug) у 2 та в 3 рази шляхом додавання до значень

дескрипторів малих випадкових збурень ($\sigma=0.02$), яке моделює природну варіативність в експериментальних даних.

2. Подвоєння та потроєння важливих молекулярних дескрипторів на основі їхньої важливості, розрахованої з базової моделі, що надає моделі більший внесок (сигнал) від важливих ознак.
3. Комбінування попередніх методів із застосуванням їх водночас.

У ході обчислень було знайдено топ-20 найважливіших дескрипторів, які визначають близько 26% від загальної важливості. На чолі списку є наступні: n10FaRing_2D та n10FaRing_3D (індекси для 10-членних ароматичних кілець), BCUTre-1l та BCUTse-1l (BCUT-дескриптори, що характеризують розподіл електронної густини), fr_NH1 (частота первинних амінів), MolLogP (ліпофільність). Позначення дескрипторів взяті

з відповідних бібліотек. Саме ці дескриптори відображають ключові структурні та фізико-хімічні властивості, які впливають на IC50 для TRPV1.

Але таке число дескрипторів, навіть найважливіших, не дозволяє отримати реальну картину. Тому відбір дескрипторів для подальшої оптимізації здійснювали на основі їхньої важливості (feature importance), визначеної за допомогою базової моделі. Було протестовано різні розміри підмножин (20, 50, 100, 150, 200, 250 дескрипторів). Встановлено, що менші набори призводять до втрати інформативності, тоді як збільшення їхньої кількості понад 100 не додає прецизійності, а спричиняє розмивання внеску найбільш значущих ознак. Таким чином, вибір топ-100 дескрипторів є емпірично обґрунтованим компромісом між точністю та стабільністю моделі.

Таблиця

Table.

Порівняння результатів обчислень різними методами оптимізації

Comparison of calculation results using different optimization methods.

Метод	R ²	MAE*	RMSE**	Покращення
Базова GB	0.7080	0.425	0.560	–
GB + Aug	0.7112	0.421	0.555	+0.45%
GB + Aug + Подвоєння	0.7126	0.418	0.553	+0.66%
GB + Aug + Потроєння	0.7201	0.410	0.547	+1.71%

* MAE – Mean Absolute Error (середня абсолютна похибка);

** RMSE – Root Mean Squared Error (середньоквадратична похибка).

Як видно з таблиці, найкращий результат було отримано при поєднанні Data augmentation із потроєнням топ-100 де-

скрипторів, вибраних за результатами машинного навчання.

ВИСНОВКИ. Таким чином модель машинного навчання, яка дозволяє досягнути значення коефіцієнта детермінації $R^2 = 0.7201$, можна вважати придатною для скринінгу молекул на ранніх етапах розроблення лікарських засобів [7]. Значення MAE = 0.41 свідчить, що в середньому передбачена активність відрізняється від експериментальної всього на ± 0.41 логарифмічної одиниці. Така точність є прийнятною для QSAR-моделей і її можна використовувати у поєднанні з хімічною інтуїцією, застосуванням інструментів ШІ та додатковими методами валідації, зокрема молекулярним докінгом або експериментальними випробуваннями [8–10].

Критично важливим є правильне розділення даних, зокрема Group K-Fold показав, що реальна якість моделі на невидимих структурах є значно нижчою, ніж оціночні значення. Окрім цього, вельми ефективною є Data augmentation або розширення датасету в 2–3 рази шляхом додавання малих збурень, що поліпшує узагальнюваність моделей.

Важливо відзначити, що комбіновані методи перевершують окремі, але найбільше поліпшення ($R^2=0.7201$, +1.71%) досягнуто за одночасного застосування обох методів (Aug та потроєння топ-100 дескрипторів).

Також слід підкреслити, що не всі підходи є універсальними. Зокрема, використання морганівських відбитків, глибоких нейронних мереж (Deep Neural Networks), а також попереднє оброблення даних за допомогою методу головних компонент (Principal Component Analysis, PCA) не

призвели до покращення якості моделей для зазначеного датасету.

Це підкреслює важливість емпіричного тестування різних підходів, а також необхідність адаптації вибору дескрипторів і моделей до конкретної задачі та властивостей набору даних.

Таким чином розроблена модель з $R^2 = 0.7201$ є практично корисною для розширеного скринінгу молекулярних бібліотек на ранніх етапах розроблення нових лікарських засобів.

ДЕТАЛІЗАЦІЯ ВКЛАДУ АВТОРІВ У ПІДГОТОВКУ РУКОПІСУ. Автори роботи зробили рівнозначний внесок у розроблення концепції та дизайну дослідження, збір, систематизацію, аналіз та інтерпретацію отриманих даних. Автори брали рівновелику участь у підготовці, редагуванні та доопрацюванні статті. Усі автори ознайомилися з результатами дослідження та схвалили остаточну версію статті.

КОНФЛІКТ ІНТЕРЕСІВ. Автори заявляють про відсутність конфлікту інтересів.



ПОДЯКА.

Цю роботу було підтримано грантом від Міжнародного фонду Саймонса [SFI-PD-Ukraine-00014577, O.G.] (This work was supported by a grant from the Simons Foundation International [SFI-PD-Ukraine-00014577, O.G.]). Державний реєстраційний номер: 0121U100174.

OPTIMIZATION OF QSAR MODELS FOR PREDICTION OF BIOLOGICAL ACTIVITY MOLECULES USING MACHINE LEARNING METHODS.**D. V. Maslov**<https://orcid.org/0009-0002-1376-1450>**O. A. Golub**<https://orcid.org/0000-0003-1823-2523>*National University "Kyiv-Mohyla Academy",
Ukraine**2 Skovoroda st., 04070 Kyiv, Ukraine**e-mail: dv.maslov@ukma.edu.ua*

Molecular modeling plays a central role in modern computational chemistry, particularly in the early stages of drug discovery, where researchers must rapidly and reliably predict the biological activity of large sets of potential candidates. Quantitative Structure–Activity Relationship (QSAR) models are widely used for this purpose; however, their true predictive performance is often overestimated due to improper data splitting strategies. A key challenge arises when test sets contain molecular scaffolds absent from the training data, resulting in models that appear accurate under random splits but fail to generalize to unseen chemical space.

This study investigates optimization strategies for QSAR modeling while explicitly accounting for molecular diversity. A dataset of 3,782 molecules with 3,291 computed descriptors and pChEMBL anesthetic activity values (5.01–8.52) for receptor TRPV1 was analyzed. The dataset contained 733 unique scaffolds, and 72 occurred exclusively in the test set under random 80/20 splitting, revealing

substantial information leakage. Three splitting strategies were compared: standard K-Fold ($R^2 = 0.54$), scaffold-based Group K-Fold ($R^2 = 0.31$), and stratified scaffold-aware splitting ($R^2 = 0.646\text{--}0.7201$), the latter demonstrating the most realistic and stable performance.

Multiple machine-learning approaches were evaluated, with Gradient Boosting achieving the best baseline accuracy. Optimization techniques included descriptor-level data augmentation ($\sigma = 0.02$), descriptor weighting by duplicating the most important features, and combined methods. The best model ($R^2 = 0.7201$, MAE = 0.41) was obtained by integrating augmentation with triple duplication of top-ranking descriptors. Several commonly used approaches—Morgan fingerprints, deep neural networks, PCA—yielded significantly weaker performance, highlighting the superior informativeness of physicochemical descriptors for this dataset.

The resulting model demonstrates practical utility for early-stage virtual screening and prioritization of candidate molecules, providing a reliable tool for guiding medicinal chemistry decisions.

Keywords: QSAR modeling; machine learning; TRPV1; molecular descriptors.

ЛІТЕРАТУРА

1. Golbraikh A., Tropsha A. Beware of q^2 !. *Journal of Molecular Graphics and Modelling*. 2002. **20**(4). 269–276. doi: [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1).
2. Yang K., Swanson K., Jin W. et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*. 2019. **59**(8). 3370–3388. doi: <https://doi.org/10.1021/acs.jcim.9b00237>.

- Caterina M. J., Schumacher M. A., Tomimaga M., Rosen T. A., Levine J. D., Julius D. The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature*. 1997. **389**(6653). 816–824. doi: <https://doi.org/10.1038/39807>.
- Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer. 2009. doi: <https://doi.org/10.1007/978-0-387-84858-7>.
- Petrov K. P., Bender A. An open-source implementation of scaffold identification. *ChemRxiv* (preprint). 2024. doi: <https://doi.org/10.26434/chemrxiv-2024-84r9x>.
- Lange J. J., Strickfaden S., Klein R., Hinselmann G. Comparative analysis of chemical descriptors by machine learning. *Molecular Pharmaceutics*. 2024. **21**(5). 1874–1888. doi: <https://doi.org/10.1021/acs.molpharmaceut.4c00080>.
- Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*. 2010. **29**(6–7): 476–488. doi: <https://doi.org/10.1002/minf.201000061>.
- Cherkasov A., Muratov E. N., Fourches D. et al. QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*. 2014. **57**(12). 4977–5010. doi: <https://doi.org/10.1021/jm4004285>.
- Roy K., Kar S., Das R. N. A primer on QSAR/QSPR modeling. Springer. 2015. doi: <https://doi.org/10.1007/978-3-319-17281-1>.
- Gramatica P. On the development and validation of QSAR models. *Methods in Molecular Biology*. 2013. **930**: 499–526. doi: https://doi.org/10.1007/978-1-62703-059-5_21.
- learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*. 2019. **59**(8): P. 3370–3388. doi: <https://doi.org/10.1021/acs.jcim.9b00237>.
- Caterina M. J., Schumacher M. A., Tomimaga M., Rosen T. A., Levine J. D., Julius D. The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature*. 1997. **389**(6653): P. 816–824. doi: <https://doi.org/10.1038/39807>.
- Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer. 2009. doi: <https://doi.org/10.1007/978-0-387-84858-7>.
- Petrov K. P., Bender A. An open-source implementation of scaffold identification. *ChemRxiv* (preprint). 2024. doi: <https://doi.org/10.26434/chemrxiv-2024-84r9x>.
- Lange J. J., Strickfaden S., Klein R., Hinselmann G. Comparative analysis of chemical descriptors by machine learning. *Molecular Pharmaceutics*. 2024. **21**(5): P. 1874–1888. doi: <https://doi.org/10.1021/acs.molpharmaceut.4c00080>.
- Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*. 2010. **29**(6–7): 476–488. doi: <https://doi.org/10.1002/minf.201000061>.
- Cherkasov A., Muratov E. N., Fourches D. et al. QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*. 2014. **57**(12): P. 4977–5010. doi: <https://doi.org/10.1021/jm4004285>.
- Roy K., Kar S., Das R. N. A primer on QSAR/QSPR modeling. Springer. 2015. doi: <https://doi.org/10.1007/978-3-319-17281-1>.
- Gramatica P. On the development and validation of QSAR models. *Methods in Molecular Biology*. 2013. **930**: P. 499–526. doi: https://doi.org/10.1007/978-1-62703-059-5_21.

REFERENCES

- Golbraikh A., Tropsha A. Beware of q^2 !. *Journal of Molecular Graphics and Modelling*. 2002. **20**(4): P. 269–276. doi: [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1).
- Yang K., Swanson K., Jin W. et al. Analyzing

Стаття надійшла: 25.03.2026.

Статтю прийнято до друку: 11.04.2026.

Статтю опубліковано: 30.04.2026.