УДК 004.85 DOI: https://doi.org/10.30837/IYF.IIS.2024.442

ДОСЛІДЖЕННЯ МЕТОДІВ ПРОГРАМНОЇ РЕАЛІЗАЦІЇ ОЦІНКИ ВЛАСТИВОСТЕЙ СКЛАДНИХ СТРУКТУР

Сергієнко О. С.
Науковий керівник – к.т.н., доцент Турута О. П.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
e-mail: oleksandra.serhiienko@nure.ua

This work is dedicated to validating QSAR models, which are pivotal for predicting molecular activity based on structural attributes, with a particular focus on ADMET. It meticulously examines the performance of two prominent methodologies: chemprop, employing graph neural networks for biological activity prediction, and a SVM classification model utilizing molecular fingerprints. Diverse mathematical approaches are rigorously assessed, employing metrics such as balanced accuracy (BA) and area under the ROC curve (AUC). The ultimate objective of this research is to enhance the efficiency of evaluating molecular properties, offering promising implications in the realm of drug discovery.

Актуальність дослідження ADMET властивостей складних структур у галузі медицини та фармацевтики пояснюється наявністю постійної потреби у нових лікарських засобах, особливо для подолання пандемій, таких як COVID-19. Використання сучасних комп'ютерних методів аналізу даних дозволяє ефективно оцінювати властивості молекул, що значно спрощує процес досліджень, зменшує ризики невдалих експериментів та дозволяє оперативно реагувати на зміни в галузі громадського здоров'я. Такі дослідження є ключовими для розвитку сучасної фармацевтичної науки та сприяють створенню більш ефективних та безпечних лікарських препаратів.

Метод (Q)SAR відіграє важливу роль у сучасній медичній хімії та фармацевтиці, особливо в контексті пошуку нових лікарських засобів. Цей метод відображає залежність між характеристиками молекули X та її активністю або властивістю Y. Для оцінки активності молекули за її характеристиками використовуються статистичні оцінки та методи машинного навчання. Вибір характеристик молекули X для створення (Q)SAR моделі є важливим завданням. Один з підходів – це представлення молекули у вигляді вектора чисел – дескрипторів, що описують її хімічну структуру та властивості. Інші підходи можуть використовувати 3D дескриптори або графові нейронні мережі для аналізу молекулярної структури.

Для побудови QSAR моделей використовуються різні математичні підходи, такі як метод опорних векторів, логістична регресія, випадковий ліс та штучні нейронні мережі. Ці методи відрізняються за кількістю

використовуваних змінних, інтерпретованістю та прогностичною здатністю.

У даному досліджені використовувалися дані з бази даних Tox21, яка містить інформацію про токсичність хімічних сполук. Дані Tox21 представляють собою масиви даних, які є не збалансованими та містять різні співвідношення активних та неактивних сполук.

У ході експерименту використовувалися різні моделі, включаючи моделі на основі глибокого навчання, такі як chemprop, і класичні класифікаційні моделі на основі молекулярних фінгерпринтів.

Модель chemprop є методом для передбачення біологічної активності молекул на основі їх хімічної структури. Вона використовує графові нейронні мережі для аналізу молекулярної структури та передбачення їх властивостей.

Класифікаційна модель на основі молекулярних відбитків є методом, заснованим на представленні молекули у вигляді бінарного вектора, де кожен елемент відповідає наявності або відсутності певного хімічного фрагмента. Цей метод використовується для передбачення біологічної активності молекул на основі їх структури.

Для порівняння і валідації моделей використовувалися різні критерії оцінки точності, включаючи збалансовану точність та площу під ROC-кривою при крос-валідації. Ці критерії були обрані для оцінки прогностичної здатності моделей на перехресному контролі.

Збалансована точність при крос-валідації визначається як середнє арифметичне точності для кожного класу:

$$BA = \frac{1}{n_{classes}} \sum_{i=1}^{n_{classes}} \frac{TP_i}{TP_i + FN_i},$$

де:

TP_i – це кількість вірно передбачених позитивних прикладів для класу i,

FN_i – це кількість хибно відкинутих прикладів для класу i,

Площа під ROC-кривою при крос-валідації обчислюється як інтеграл від ROC-кривої:

$$AUC = \int_0^1 TPR\left(FPR^{-1}(t)\right) dt,$$

де:

TPR – чутливість,

FPR – специфічність,

$FPR^{-1}(t)$ – обернена функція до функції, що повертає значення FPR при заданому порозі t.

Результати порівняння за наведеними вище критеріями наведені у таблиці.

Таблиця 1 – Результати валідації моделей

Модель	Збалансована точність (BA)	Площа під ROC-кривою (AUC)
Chemprop	0.85	0.92
Класифікаційна модель (SVM)	0.78	0.86

Ці результати свідчать про вищу прогностичну здатність моделі chemprop у порівнянні з класифікаційною моделлю на основі молекулярних відбитків. Для подальшого розвитку дослідження, рекомендується акцентувати увагу на оптимізації параметрів та розширенні набору даних для навчання моделі chemprop. Також, варто розглянути можливість поєднання chemprop з іншими методами машинного навчання для досягнення ще кращих результатів.

Список використаних джерел:

1. In silico Prediction of Chemical Ames Mutagenicity / Congying Xu, Feixiong Cheng, Lei Chen та ін. // J Cheminform. 2012. URL: https://pubs.acs.org/doi/abs/10.1021/ci300400a (дата звернення: 20.01.2024).

2. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope / Denis Mulliner, Friedemann Schmidt, Manuela Stolte та ін. // Chem. Res. Toxicol.. 2016. URL: https://pubs.acs.org/doi/10.1021/acs.chemrestox.5b00465 (дата звернення: 13.12.2023).

3. Turuta O. Collection of questionnaire results, received by using the visual analog scale method, for its further processing in the medical web application / O. Turuta, Y. Daniiel // ScienceRise. 2017. URL: https://www.researchgate.net/publication/317612987_Collection_of_questionnaire_results_received_by_using_the_visual_analog_scale_method_for_its_further_processing_in_the_medical_web_application (дата звернення: 20.03.2024).

4. Intelligent information system of heterogeneous medical data analysis / A.Yerokhin, O. Turuta, A. Babii, A. Nechyporenko. 2017. URL: https://www.researchgate.net/publication/320968847_Intelligent_information_system_of_heterogeneous_medical_data_analysis (дата звернення: 23.03.2024).