DOI 10.31891/2307-5732-2022-307-2-14-24
УДК 004.89

ГУРБИЧ О. В.
Національний університет "Львівська політехніка"
https://orcid.org/0000-0002-6821-3390
e-mail: oleksandr.v.hurbych@lpnu.ua

МЕТОД МЕТА-НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ МОЛЕКУЛЯРНОЇ СПОРІДНЕНОСТІ

У цій роботі застосований принцип мета-навчання для передбачення молекулярної спорідненості між рецептором (велика біомолекула) та лігандами (малі органічні молекули). Мета-моделі вивчають оптимальну комбінацію окремих базових моделей у двох послідовних ансамблях: класифікаційному та регресійному. Кожен із ансамблів містить по шість моделей машинного навчання, які поєднуються методом стекінгу. Базові моделі включають в себе метод опорних векторів, випадковий ліс, градієнтний бустинг, графові нейронні мережі та нейронні мережі прямого поширення, а також трансформери. Перший ансамбль прогнозує імовірність зв'язування та класифікує усі молекули-кандидати до обраного рецептору на активні та неактивні. Ліганди, які перший ансамбль визнав активними, подаються у другий ансамбль, який передбачає ступінь їхню спорідненості до рецептору у вигляді коефіцієнту енгібіювання (Ki). Особливістю методу є відмова від використання координат атомів окремих молекул та їхніх комплексів – у такий спосіб нівелюються експериментальні похибки під час підготовки зразків та вимірювання координат атомів, а також уможливлюється застосування методу для визначення спорідненості біомолекул із невідомими просторовими конфігураціями. Показано, що мета-навчання збільшує відгук (Recall) класифікаційного ансамблю на 34,9% та коефіцієнт детермінації (R^2) регресійного ансамблю на 21% у порівнянні із середніми значеннями. У цій роботі показано, що ансамбль з мета-стекінгом є асимптотично оптимальною системою для навчання. Розглядається особливість Super Learning'у для використання перехресної перевірки (k-fold cross-validation) для формування передбачень «першого рівня», на яких виконується навчання моделей другого рівня або мета-моделей, які комбінують моделі першого рівня оптимальним чином. Досліджується здатність передбачати молекулярну спорідненість шести моделей машинного навчання, а також покращення ефективності внаслідок поєднання моделей у ансамблі методом стекінгу. Показані моделі, які поєднані у два послідовні ансамблі.

Ключові слова: мета-навчання, машинне навчання, методи ансамблювання, молекулярна спорідненість, трансформери, бустинг, стекінг, коефіцієнт енгібування.

OLEKSANDR GURBYCH
Lviv Polytechnic National University

METHOD SUPER LEARNING FOR DETERMINATION OF MOLECULAR RELATIONSHIP

This paper uses the Super Learning principle to predict the molecular affinity between the receptor (large biomolecule) and ligands (small organic molecules). Meta-models study the optimal combination of individual basic models in two consecutive ensembles - classification and regression. Each costume contains six models of machine learning, which are combined by stacking. Base models include the reference vector method, random forest, gradient boosting, neural graph networks, direct propagation, and transformers. The first ensemble predicts binding probability and classifies all candidate molecules to the selected receptor into active and inactive. Ligands recognized as involved by the first ensemble are fed to the second ensemble, which assumes the degree of their affinity for the receptor in the form of an inhibition factor (K_i). A feature of the method is the rejection of the use of atomic coordinates of individual molecules and their complexes - thus eliminating experimental errors in sample preparation and measurement of nuclear coordinates and the method to determine the affinity of biomolecules with unknown spatial configurations. It is shown that meta-learning increases the response (Recall) of the classification ensemble by 34.9% and the coefficient of determination (R^2) of the regression ensemble by 21% compared to the average values. This paper shows that an ensemble with meta-stacking is an asymptotically optimal system for learning. The feature of Super Learning is to use k-fold cross-validation to form first-level predictions that teach second-level models — or meta-models — that combine first-level models optimally. The ability to predict the molecular affinity of six machine learning models is studied, and the efficiency improvement is due to the combination of models in the ensemble by the stacking method. Models that are combined into two consecutive ensembles are shown.

Keywords: Super Learning, machine learning, ensemble methods, molecular affinity, transformers, boosting, stacking, inhibition coefficient.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Початкові етапи розробки лікарських речовин вимагають локалізації рецептору, що відповідає за розвиток хвороби, вивчення молекулярного механізму і наступний дизайн та випробовування молекул-кандидатів. Після того як молекула-рецептор була ідентифікована, створюється список лігандів-кандидатів. Кращі молекули зі списку обираються за найвищою спорідненістю кандидату до цільового рецептора. Спорідненість характеризує силу взаємодії між цілями та лігандами. Вона може бути кількісно описана константою інгібіювання *Ki*. Чим менше значення *Ki*, тим сильніше ліганд зменшує біологічну активність рецептору, і тим сильніший лікувальний ефект може бути досягнутий меншою дозою ліків. Існує багато методів експериментального вимірювання *Ki*, але вони є затратними за часом, ресурсами та зусиллями. Тому перед експериментальними застосовуються обчислювальні підходи для зменшення набору лабораторних випробувань шляхом ранжування та виключення кандидатів з низькою розрахунковою афінністю. Одним з таких підходів є молекулярний докінг [1, 5] у якому досліджуються можливі просторові

взаєморозташування рецептору і ліганіду та оптимізується конформація зчеплення структур. Розрахована енергія не є величиною молекулярної спорідненості, оскільки конформації з низькою оціночною енергією зв'язування не завжди співпадають з експериментальними [2, 6]. Важливо відмітити, що докінг може постраждати від неточного виявлення лігандових точок/положень, або взагалі зчепити зовсім неактивні компоненти [3, 8]. Разом із класичними симуляціями силових полів і докінгом, технології машинного навчання стали потужним інструментом в галузі віртуального відсіювання.

Аналіз досліджень та публікацій

Одне із перших досліджень у машинному навчанні для молекулярної спорідненості, яке опубліковане [5], де ефективність нейронних мереж прямого поширення, методу К-найближчих сусідів і дерев рішень порівнювалися на 200 лігандах та двох рецепторах. Список дескрипторів лігандів включає молекулярну масу, гнучкість, полярність, поляризацію, кількість донорів/акцепторів електронів, тощо. [9, 12] оцінили афінність лігандів, використавши метод опорних векторів. Їхній датасет складався з кількох сотень пар ліганд-рецептор. Було згенеровано більш ніж 500 молекулярних дескрипторів, проте лише 50 з них виявилися інформативними. Важливість виділення правильних параметрів також висвітлюється у дослідженні [4, 7], у якому вивчається спорідненість білків з білками.

Концепція усередненого простору взаємодій, представлена у дослідженні [10, 13]. Для багаторецепторної бінарної класифікації активних/неактивних лігандів запропоновано набір з 439 ознак: 107 описують рецептори, 166 параметрів є MACCS відбитками [12, 9], що описують ліганди, решта 166 є відбитками, усередненими за лігандами до одного рецептора. В даному випадку було обрано метод Баєсових адитивних регресійних дерев. У наступних підходах використано метод опорних векторів, випадковий ліс, дерева рішень та логістичну регресію. Автори досягли точності в близько 95% для бінарної класифікації активних і неактивних лігандів. Схожа концепція простору, усередненого за взаємодією, також використовується у дослідженні [15, 16]. Автори збудували декілька регресійних моделей для передбачення молекулярної афінності використовуючи вектори параметризованих лігандів, що усереднюються для кожного рецептора.

Також варто згадати ряд досліджень [1, 17], що перевірялись на однакових наборах даних [5, 14], зокрема кіназах. Молекулярна спорідненість у KronRLS [12] визначається через добуток Кронекера між матрицями подібностей пар ліганд-ліганд та рецептор-рецептор. Маючи подібності лігандів та подібності рецепторів, [13] використовує градієнтний бустинг для передбачення молекулярної спорідненості та повідомляє про середньоквадратичну похибку в 0.28 для датасету Дейвіса [10, 11]. Глибока нейромережа DeepDTA [15–17] використовує FASTA рецепторів [1, 3] та SMILES лігандів [2, 7], кодує і доповнює відсутні символи нулями. Кодування проводиться двома окремими згортковими мережами, виходи з яких конкатенуються і передаються до повноз'єднаних шарів нейронної мережі, яка передбачає молекулярну афінність. DeepDTA досягає середньоквадратичної похибки 0.26-0.66 (залежно від набору кодування) на датасеті Дейвіса. WideDTA [14, 17] використовує чотири текстові вводи: білкова послідовність, SMILES лігандів, білкові домени і мотиви, та максимальну спільну підструктуру слів для передбачення афінності. Ці вводи передаються у чотири окремі рукави згорткової нейронної мережі, конкатенуються і передаються через набір повноз'єднаних шарів для передбачення спорідненості. Модель демонструє середньоквадратичну похибку 0.26 на датасеті Дейвіса. Підхід GraphDTA [5, 8] також покладається на роздільні входи – один для ліганда і один для рецептора. Виводи рукавів також конкатенуються і регресуються. Ліганди представляються як графи з атомами в якості вузлів та зв'язками в якості ребер. Чотири реалізації графової нейромережі перевіряються для лігандової частини, у той час як рецептори кодуються з FASTA нотації і обробляються набором конволюційних мереж. Отримана похибка MSE коливається в районі від 0.23 до 0.25 pK_d на датасеті Дейвіса. Подібно до KronRLS, модель SimCNN-DTA, розроблена [2, 9], базується на хімічних подібностях. Зокрема, для пари ліганд-рецептор обчислюються два вектори: перший складається з подібностей Танімото між молекулярними відбитками лігандів, другий – з подібностей Сміта-Вотермана між FASTA послідовностями рецепторів. Зовнішній добуток цих двох векторів являє собою двовимірну матрицю, яка слугує вхідними даними для двовимірної конволюційної мережі для передбачення молекулярних спорідненостей. SimCNN-DTA перевірялась на датасетах Дейвіса і KIBA, і показала однакову та кращу продуктивність, ніж інші дослідження у цьому абзаці.

Дослідження [14] порівнює продуктивність низки методів машинного навчання (випадковий ліс, метод опорних векторів, Гаусовий процес, нейромережа прямого поширення) для передбачення афінності на датасеті PDBbind (v.2015). Для всіх розглянутих підходів рецептори і ліганди описувались набором структурних і фізико-хімічних параметрів, що формують єдиний вхідний вектор для кожної пари рецептор-ліганд. Передбачуваними вихідними величинами були значення K_i та K_d. Найкращих результатів вдалось досягнути за допомогою випадкового лісу.

Підхід DGraphDTA [5, 9] використовує структурну інформацію молекул і білків. Будуються два графи: один для білкового рецептора, інший – для молекули ліганда, на яких проводиться регресія для передбачення афінності. Цікаво, що білкові графи будуються з контактних матриць білків, які передбачаються з послідовностей FASTA предиктором PconsC4 [3, 7].

Jimenez та співавтори [14, 15] запропонували Kdeep – тривимірну конволюційну нейромережу (3D-CNN) для передбачення молекулярної афінності. Кожна пара білку з лігандом описується вокселізованим

об'ємом у 24 Ангстреми представленням сайту зв'язування. Кожен воксель описаний вісьмома фармакофорними властивостями. Автори досягнули RMSE в 1.27 *pK* між експериментальним і передбаченим значенням на стандартному PDBbind (v.2016) базовому тестовому наборі. DeepAtom [12, 17] – ще один 3D-CNN фреймворк, що кодує властивості, пов'язані зі зв'язуванням, за допомогою вокселізованих представлень. З датасетом Astex Diverse [3, 4] в якості навчальної вибірки, DeepAtom досягає RMSE в 1.23 в одиницях pK на тестовій вибірці PDBbind (v.2016).

Дослідження проведене [8,10] обговорює підхід для передбачення спорідненості на основі ансамблів 3D-CNNів. Ансамбль має найменші метрики похибок з MAE значенням в 1.01 ккал/моль та RMSE в 1.29 ккал/моль. Використання ансамблю мереж покращило якість передбачень на 0.1 ккал/моль порівняно з продуктивністю поодиноких мереж. Ансамбль моделей випадкового лісу, AdaBoost-Regressor, Gradient Boosting Regressor та мережі прямого поширення, розроблений [16, 17] передбачає афінність пептидних лігандів до кількох білків пов'язаних з пухлинами. [11] розповідає про доволі високі коефіцієнти детермінації R^2 в 0.81/0.9 на навчальній/тестовій вибірці і, в якості додаткового кроку валідації, виконує ряд комп'ютерних симуляцій для обраних пар білків-лігандів. Проте, вибрані ліганди не показали очікувану сильну зв'язаність. Інтегрований підхід, що використовує зчеплення лігандів на кількох структурних ансамблях для відображення рецепторної гнучкості пропонується [8, 10]. Підхід поєднує етап лігандного зчеплення з етапом передбачення афінності для комплексу ліганд-рецептор, застосовуючи модель випадкового лісу. Для того, щоб покращити рішення, перевіряються різні набори молекулярних дескрипторів.

З часу своєї появи нейромережі-трансформери були успішно застосовані для відображення хімічних символів у представлення хімічних функцій і властивостей. У дослідженні [12, 15] моделі self-attention використовуються для передбачення продуктів хімічних реакцій, сформульованих як проблема машинного перекладу SMILES-рядків реактивів та реагентів на продукти. [17] аналізує застосування BERT-моделі для вивчення контекстних представлень хімічних сполук для передбачення токсичності, розчинності, подібності до ліків та синтетичної доступності. У роботі [4, 8] кодування білкових послідовностей вивчаються BERT-моделлю з широкомасштабного нерозміченого датасету і ретельно вивчається на предмет кодування різноманітних білкових аспектів. Модель MT-DTI [1, 3] поєднує здібності моделювати послідовності двох моделей – CNN для FASTA і self-attention механізми для SMILES, для задачі передбачення афінності. Таке поєднання нейронних архітектур дозволило досягнути неперевершених результатів на вище згаданих датасетах KIBA і Дейвіса.

[8, 9] описує пошук ефективних інгібіторів проти рецепторів, пов'язаних з SARS-CoV-2, за допомогою методів машинного навчання. [3, 5] пропонує градієнтний бустинг для пошуку інгібіторів до протеази SARS-CoV-2 3CL [1, 2] серед вже затверджених FDA (Управління з продовольства і медикаментів США) ліків. Так як протеаза є єдиним рецептором у цьому дослідженні, немає потреби у його параметризації. З навчальною вибіркою у 314 інгібіторів, ліганди параметризуються усередненням трьох видів молекулярних відбитків. Дослідження [4, 8] застосовує багатоетапний конвеєр, що поєднує класифікацію активності, фільтрацію лікувальної подібності, передбачення зчеплення і зв'язної спорідненості для пошуку інгібіторів протеази. Оскільки розглядався підхід з одним рецептором, тільки ліганди були описані набором дескрипторів і як вхідні дані. Обрані інгібітори перевірялися у розрахунковому експерименті за методом молекулярної динаміки. [7, 10] виконали скринінг бази даних ChEMBL [1, 11] рекурентною нейромережею для пошуку інгібіторів проти протеази. Молекули, які були визначені як активні, були в подальшому проаналізовані з використанням молекулярного докінгу. [4, 12] запропонували конвеєр для визначення ліків-кандидатів проти кількох рецепторів, пов'язаних з SARS-CoV-2, із особливою увагою на виборі параметрів лігандів. Натренований конвеєр в подальшому використовувався для ранжування списку тисяч відомих ліків та мільйонів доступних до синтезу молекул за зв'язною спорідненістю, токсичністю і леткістю. Модель MT-DTI від [8] застосовувалася для передбачення афінності відомих противірусних ліків для шести рецепторів, пов'язаних з SARS-CoV-2 [3]. У дослідженні від [15], відомі ліки та доступні до синтезу молекули перевіряються на здатність взаємодіяти з трьома рецепторами (шипи, капсиди і трансферази) засобами AutoDock Vina і підходами машинного навчання.

Огляд Ellingson [14] освітлює популярні підходи машинного навчання і перешкоди, що постають у вивченні молекулярної афінності. Окрім типових проблем з представленням, які згадуються вище, також обговорюються складності з доступністю та узгодженістю даних. Якість передбачень спорідненості може постраждати від неузгоджених даних, на які вплинув експериментальний шум. Важливо, що деякі з обговорених досліджень фокусуються на одному рецепторі, тому в них не потрібно параметризувати рецептори. Крім того, поєднання моделей у ансамблі допомагає розширити область передбачень через поєднання рішень для вужчого кола завдань.

Мета-навчання, яке також називають мета-стекінгом, Super Learning [5, 8] або Stacked Regression [7, 9], — це клас алгоритмів, у якому навчається мета-модель другого рівня для визначення оптимальної комбінації моделей першого рівня. На відміну від пакетування (bagging) та бустингу (boosting), мета-стекінг полягає в тому, щоб об'єднати сильні, різноманітні групи базових моделей. Концепція мета-стекінгу була створена на початку 1992-го року [1, 5], та теоретично обґрунтована у 2007-му році [4]. У цій роботі було показано, що ансамбль з мета-стекінгом є асимптотично оптимальною системою для навчання. Визначальною особливістю Super Learning'у є використання перехресної перевірки (k-fold cross-validation)

для формування передбачень «першого рівня», на яких виконується навчання моделей другого рівня або мета-моделей, які комбінують моделі першого рівня оптимальним чином.

Формулювання цілей статті

Метою роботи був аналіз дослідження у машинному навчанні для молекулярної спорідненості для використання перехресної перевірки, на яких виконується навчання моделей другого рівня або мета-моделей.

Виклад основного матеріалу

У роботі досліджується здатність передбачати молекулярну спорідненість шести моделей машинного навчання, а також покращення ефективності унаслідок поєднання моделей у ансамблі методом стекінгу. Моделі поєднані у два послідовні ансамблі. Ми використовували наступні алгоритми: метод опорних векторів (SVM) [3, 5], випадковий ліс (RF) [6], градієнтний бустинг в реалізації бібліотеки CatBoost (CB) [10, 11], повнозв'язну нейронну мережу прямого поширення (FNN), графова нейронна мережа (GNN) [13], і двоспрямовані кодувальні представлення з трансформерів (BERT) [1]. Перші чотири моделі – SVM, RF, CB, FNN – використовують молекулярні відбитки ECFP4 [17] як вхідні дані. П'ята модель – GNN – використовує графове представлення лігандів, розглядаючи атоми як вузли, а зв'язки між атомами - як ребра графа. Нами було обрано графи із параметризацією як вузлів, так і ребер [10] з метою розширення здатності представлення оригінального формалізму GNN [9]. На рис. 1 показаний спосіб утворення графового представлення на прикладі молекули хлороформу ($CHCl_3$). На відміну від згаданих вище підходів, що мають справу лише з фізико-хімічними властивостями, BERT працює безпосередньо зі стрічковими представленнями лігандів і, таким чином, відкидає потребу в інженерії ознак.

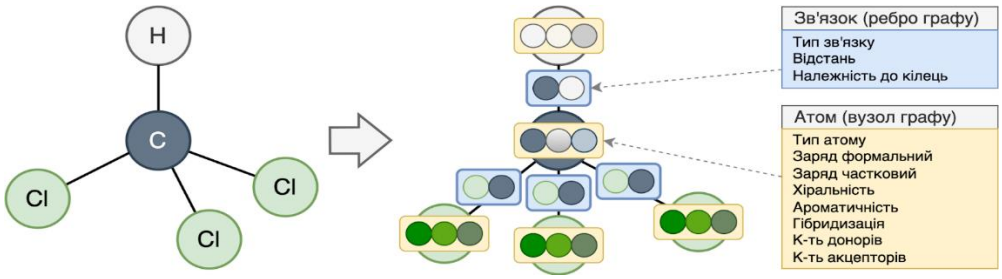

Рис. 1. Хлороформ як приклад ліганду, представленого у вигляді молекулярного графу із вузлами (атомами) та ребрами (хімічними зв'язками між атомами). Кожен вузол та ребро описуєтся набором параметрів, що вказані на малюнку

Зазначені моделі та їх ансамблі застосовуються для прогнозування афінності органічних молекул до людського тромбіну. Тромбін був обраний через доступність значної кількості публічних даних для навчання моделей, а також перспективи перевірки результатів у лабораторії високопродуктивного скринінгу. Приклад комплексу альфа-тромбіну та гірудину показано на рис. 2. Однорецепторна парадигма є поширеним підходом у застосуванні машинного навчання для хемінформатики. Усі моделі можуть бути перенавчені для інших рецепторів.

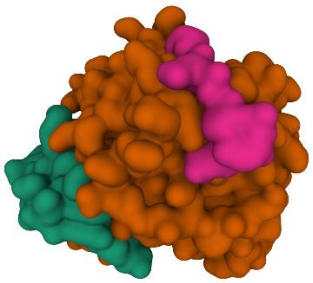

Рис. 2. Комплекс людський альфа-тромбін (велика субодиниця забарвлена коричневим, мала - зеленим), лігандом - пептид гірудин (рожевий)

Принципова схема методу показана на рис. 3. Метод складається із двох поєднаних послідовно ансамблів: перший класифікує ліганди-кандидати на активні та неактивні; другий прогнозує значення константи інгібування (*Ki*) для активних лігандів.

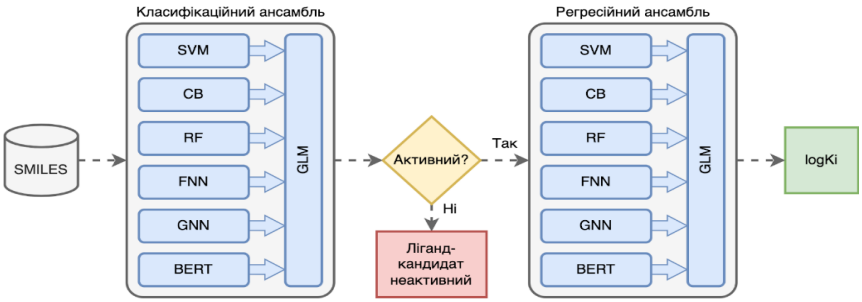

Рис. 3. Схема методу з послідовними класифікаційним та регресійним ансамблями. Неактивні ліганди-кандидати відсіваються першим ансамблем; для активних лігандів другий ансамбль передбачає *Ki*

Ансамблі – як класифікаційний, так і регресійний – утворюються із окремих моделей за методом мета-навчання (див. рис. 4). Мета-навчання відноситься до класу алгоритмів машинного навчання із вчителем [2]. Його суть полягає у тренуванні моделей другого рівня для виявлення оптимальної комбінації моделей першого рівня та поєднання їх у ансамбль із вищою предиктивною здатністю [7]. У якості алгоритму мета-навчання нами був обраний набір лінійних моделей [1, 5] у реалізації h2o.ai. На додаток до гаусівського (нормального) розподілу, до GLM належать розподіли Пуассона, біноміальні та гамма-розподіли. Кожен з них служить різним цілям і залежно від вибору функції розподілу та цільового значення може використовуватися або для регресії, або для класифікації [2]. Нами були обрані гаусова регресія та біноміальна логістична регресія для відповідних задач регресії (logKi) та класифікації (активний чи неактивний). Детально схема підготовки даних та тренування мета-моделей показана на рисунку 4. Щоб знизити або повністю виключити вплив малоінформативних моделей, було застосовано L1 регуляризацію. Кількість перехресних валідацій - 5, функція втрат - Binary LogLoss. Застосовувалася рання зупинка через 10 епох в разі відсутності зменшення значення функцій втрат на валідаційному датасеті. Солвер - градієнтний спуск.

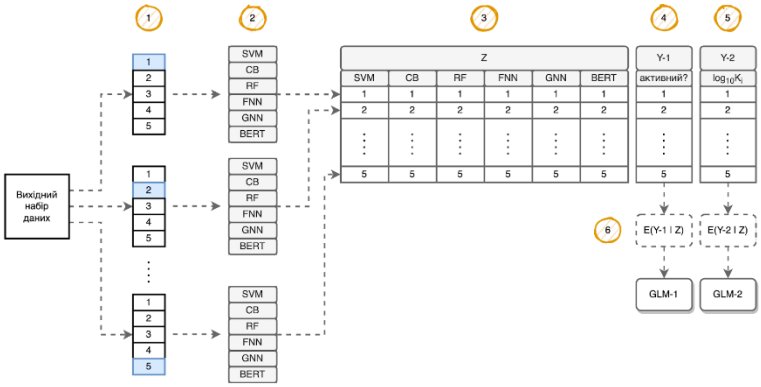

Рис. 4. Блок-схема навчання мета-моделі. 1 - крос-валідаційні набори даних. Синім кольором позначено блок даних для валідації. 2 - тренування шести моделей першого рівня. 3 - результати передбачень моделей першого рівня на валідаційних наборах даних на кожному розбитті (1-5). 4 - справжні мітки активності лігандів. 5 - справжні значення $log_{10}K_i$. 6 - тренування моделей другого рівня із використанням передбачень моделей першого рівня (Z) як вхідних даних та справжніх міток (Y-1, Y-2) як цільових значень

Ми використовуємо метод опорних векторів (SVM) із Гаусівською радіальною базисною функцією (RBF), яка використовується для трансформування простору ознак. Моделі SVM задіяні як у класифікаційних, так і в регресійних ансамблях. В обох випадках були використані реалізації бібліотеки Scikit-learn [5, 9]. Завдяки використанню ядра $RBF\,k\left(x_i, x_j\right) = \exp(-\frac{1}{2\gamma^2} x_i - x_j^{\,2})$ на векторах ознак x_i і x_j з двійковими компонентами, отримана модель порівнювала вхідні молекули за схожістю їх відбитків ECFP4 на бітовому рівні. Параметр регуляризації C моделі було обрано базуючись на результатах сіткового пошуку серед значень 0,01, 0,1, 0,2, 0,5, 1,0, 2,0, 5,0, 10,0 і 100,0. Пошук показав, що значення C = 1 є оптимальним вибором з точки зору найвищої метрики точності. Під час навчання моделі було встановлено режим class_weight = «balanced», який використовує значення цільових міток для автоматичного регулювання ваг, обернено пропорційних частотам класів у вхідних даних. Для регресії була використана модель Epsilon-Support Vector Regression, яка також реалізована в бібліотеці Scikit-learn.

Для випадкового лісу (RF) ми також використовуємо реалізацію від Scikit-learn. RF також використовує параметр class_weight = «balanced», що автоматично регулює баланс між кількістю активних та неактивних зразків у задачі класифікації. На початковому етапі ми застосували процедуру пошуку найкращих гіпер-параметрів методом сітки. Під час цієї оптимізації були перевірені всі комбінації наступних параметрів: кількість дерев в ансамблі (n_estimators) – 200, 500, 1000, 2000, максимальна глибина дерева (max_depth) – 2, 5, 7, 8, 10, мінімальна кількість екземплярів, необхідних для поділу вузла (min_samples_split) – 1, 2, 4, 8, 10, 20. Перехресна валідація на п’яти поділах була застосована, щоб знайти комбінацію гіпер-параметрів, які дають найкращу усереднену точність. Оптимальною виявилася комбінація таких гіпер-параметрів: n_estimators = 200, max_depth = 10, min_samples_split = 10.

Пошук гіпер-параметрів для регресії був подібним, але цього разу в рамках п’ятиразової перехресної валідації була мінімізована середня квадратична похибка моделі регресії. В результаті, оптимальними для регресії були наступні значення гіпер-параметрів: n_estimators = 2000, max_depth = 10, min_samples_split = 2.

У цьому розділі описується модель градієнтного бустингу з деревами рішень. Ми використали реалізацію CB [15] цього алгоритму. Моделі були натреновані на бінарних ECFP4 відбитках лігандів. Гіпер-параметри відповідних класифікаційних і регресійних моделей спочатку були налаштовані шляхом пошуку по такій сітці: швидкість навчання – 0,1, 0,03, 0,01; L2 регуляризація – 1, 3, 5, 7, 9; максимальна глибина дерев рішень – 6, 8, 10. Показниками для вибору оптимальних гіпер-параметрів були точність для класифікації та MSE для регресії, обчислені на тестовій вибірці. В результаті обраний набір гіпер-параметрів для завдання класифікації включає швидкість навчання = 0,1; параметр регуляризації листків L2 = 3; максимальну глибину дерев = 10. Той самий набір гіпер-параметрів, за винятком регуляризації L2 = 1, використовується для завдання регресії. Валідаційні метрики зазвичай припиняли покращуватися після 200 епох навчання для класифікації та 500 епох для регресії. Під час навчання було застосовано алгоритм ранньої зупинки з параметром patience, встановленим на 100 епох.

У цьому підрозділі описується підхід, заснований на формалізмі нейронних мереж прямого поширення (FNN). Подібно до наведених вище методів, вхідними даними FNN є відбитки ECFP4. Розмір вхідного шару становить 2048 нейронів. Вхідні дані надалі надходять у набір повнозв'язних шарів. Під час оптимізації гіпер-параметрів ми випробували різні архітектури FNN, у яких змінювали ширину та глибину мережі, спробували рекурентні та пропускові з'єднання, такі функції активації як ReLU, Softplus та PReLU. Незважаючи на наші очікування, рекурентні та пропускові з'єднання не покращили продуктивність як для класифікації, так і для регресії. У випадку класифікації, остаточна архітектура складалася з п'яти шарів: вхідний шар розміром 2048, три прихованих шари з 512, 256 і 64 нейронами відповідно та вихідний шар із 2 нейронами. Функція активації на першому, другому та третьому рівнях — ReLU. У випадку завдання регресії архітектура складалася також з п'яти шарів: вхідний шар має розмір 2048, приховані шари – 1024, 256 і 64 нейронів відповідно, а вихідний шар – з 1 нейроном. Функція активації для прихованих шарів – Softplus. Навчання проводилося за допомогою технології стохастичного градієнтного спуску з оптимізатором Adam та batch size рівним 32. Під час навчання ми також використовували планувальник швидкості навчання, який зменшував початкову швидкість навчання в 0,001 кожні 50 епох навчання в 0,9 рази. У задачі класифікації навчена модель була оптимізована за втратою перехресної ентропії, тоді як найкраща модель була обрана відповідно до кращого показника точності в тестовій підмножині. MSE використовувався як функція оптимізації втрат у задачі регресії. Була обрано модель із найнижчим MSE на тестовій підмножині.

Як вже зазначалося, представлення лігандів у рамках підходу GNN будуються у вигляді молекулярних графів. У роботі використовується модель AttentiveFP GNN [1, 17] у реалізації DGL-LifeSci (https://lifesci.dgl.ai/) як для завдань класифікації, так і для регресії. AttentiveFP передає повідомлення між вузлами та вивчає нелокальні ефекти завдяки механізму уваги. Механізм уваги полягає у тому, що додаткові шари нейронів навчаються на вагах вузлів. Це дозволяє атомам агрегувати ознаки стану сусідів та поширювати власні ознаки назад до сусідів. Таким чином, окремі атоми роблять внесок у вектор стану усієї молекули. Результати Attentive FP передаються на повнозв'язний шар із двома (класифікація) або одним (регресія) нейроном. Перерахуємо деякі параметри архітектури: кількості параметрів вузлів і ребер становили 27 і 12. Векторні представлення для атомів (вузлів) та хімічних зв'язків (ребер) були згенеровані за допомогою WeaveAtomFeaturizer і WeaveEdgeFeaturizer; кількість шарів в Attentive FP – 2, розмір ознак графу – 200, кількість рекурентних кроків – 2, коефіцієнт дропауту – 0,2. Вся мережа навчалася наскрізним способом за допомогою оптимізатора Adam з розміром батчу рівним 30, коефіцієнтом регуляризації L2 - 0,0002 та затримкою ранньої зупинки рівною 40 епох. Початкова швидкість навчання 0,001 поступово зменшувалася в 0,9 разів кожні 10 епох протягом процесу навчання. У задачі класифікації ми використали фокальну функцію втрат [3], а у задачі регресії - MSE як для оптимізації гіпер-параметрів, так і для вибору найкращої моделі.

У цьому підрозділі описується методологія навчання, підготовки даних та оптимізації гіпер-параметрів моделі-трансформера. Ми обрали двоспрямовану трансформерну модель BERT [5, 6], оскільки вона добре враховує як лівий, так і правий контексти під час побудови векторних представлень, що має вирішальне значення для повного охоплення складних взаємозалежностей у хімічних структурах. Навчання BERT зазвичай складається з двох етапів. Перший етап – попереднє навчання на великому нерозміченому датасеті з метою прогнозування випадково замаскованого символу у послідовності. Другий етап – донавчання моделі для конкретного завдання з відповідно розміченим (невеликим) датасетом.

Для завдання класифікації архітектура BERT виглядає наступним чином: 4 шари, 12 внутрішніх нейронних мереж механізму уваги, прихований розмір 768. Для кожного символа з рядка SMILES створюється ембедінг шляхом додавання позиційного та контекстного кодувань. Попередньо натреновані ваги BERT були взяті з Wolf et al, де вони були оптимізовані на приблизно 155 000 послідовностях SMILES з бази даних PubChem [2, 12]. Максимальна довжина рядка SMILES була 128. Словник створювався за допомогою алгоритму BytePair; його розмір становить приблизно 52 000 токенів. На етапі донавчання ми експериментували з кількома функціями втрат. Найкраща точність і precision/recall були досягнуті за допомогою зваженої фокальної функції втрат із параметрами гама 2 і альфа 0,81. Щоб уникнути перенавчання та покращити узагальнення під час навчання, у BERT було додано ймовірності дропауту 0,3 до вхідних шарів, шарів уваги та прихованих станів, що виводяться кожним шаром моделі.

Регресійна модель BERT була меншою: 3 прихованих шари, 6 self-attention шарів, розміром 768 та розміром словника 2229 токенів. Попередньо навчені ваги для такої конфігурації також були взяті з Wolf et al., які були оптимізовані на 400 000 послідовностях SMILES з бази даних ChEMBL [5, 8].

Навчання обох моделей проводилося протягом 8 та 14 епох відповідно, з batch size рівним 32. Початкова швидкість навчання становила $9 \cdot 10^{-6}$ та $1 \cdot 10^{-4}$ для класифікації та регресії відповідно та лінійно знижувалася протягом навчання. Ваги моделей оновлювалися оптимізатором Adam.

У публічному доступі є небагато наборів даних які містять експериментальні константи інгібування K_i та для класифікації активних/неактивних лігандів до різних рецепторів. K_i - це концентрація ліганду в розчині, необхідна для інгібування функції рецептора-мішені наполовину. K_i відображає, наскільки сильно молекула-інгібітор пригнічує біологічну функцію молекули-рецептора: чим нижче K_i, тим сильнішим (більш активним) є інгібітор, і навпаки. Не існує строгого порогу для константи інгібування, яка розрізняє активні та неактивні ліганди, однак K_i в 10 000 нмоль часто використовується як такий роздільник [16, 17].

Оскільки досліджуваний метод складається з ансамблів класифікації та регресії, ми підготували два набори даних для навчання відповідних моделей. Класифікаційні та регресійні набори даних були об'єднані з трьох баз даних: BindingDB [1], DUD-E [1] і ChEMBL [9]. Ліганди були представлені за допомогою рядка SMILES, під час попередньої обробки вони були канонізовані за допомогою RDKit (http://www.rdkit.org) із видаленням інформації про ізометрію молекул.

Початковий розмір необробленого датасету для класифікації становив 27497 записів. Було канонізовано SMILES, видалено дублікати та записи з помилками. Ліганд вважався неактивним, якщо K_i для нього мав значення >10 000 нмоль. Якщо для одного ліганду в різних базах зустрічалися суперечливі записи, клас зв'язування визначався за більшістю одностайних записів. Якщо більшість неможливо було встановити, ліганд вилучався з набору даних. Під час навчання ми використовували п'ятиразову валідацію моделей класифікації та зберігали сталий розподіл активних та неактивних лігандів у кожній навчальній і тестовій вибірці.

Для видалення аномалій була розрахована молекулярна маса усіх лігандів. Видалення аномалій проводилося за методом Д. Тюки, який заснований на розрахунку міжквартильної відстані (IQR) за формулою 1 і наступного визначення меж м'яких викидів за формулами 2 та 3:

$$IQR=Q3-Q1 \quad (1)$$

$$LIF=Q1-1,5*IQR \quad (2)$$

$$UIF=Q3-1,5*IQR, \quad (3)$$

де *LIF* – нижня внутрішня межа для виявлення м'яких викидів; *UIF* – верхня внутрішня межа для виявлення м'яких викидів; *Q1* – значення першого квартиля; *Q3* – значення третього квартиля; *IQR* – міжквартильна відстань.

Початковий розмір датасету для регресії становив 4161 записів. До набору даних було включено лише записи з точними значеннями K_i. Якщо для одного ліганду було кілька різних значень K_i, ми обирали найбільше значення - як негативний сценарій. Додаткова умова при створенні вибірки даних для регресії випливає з конструкції конвеєра: регресійний ансамбль передбачає K_i, лише якщо класифікаційний стек моделей визначив активний ліганд. Тому для задачі регресії ми обирали значення K_i лише в межах «активного» діапазону та опускали зразки з високими K_i. В результаті, мітки в наборі даних для регресії знаходяться у діапазоні $K_i \in [0 : 30\,000$ нМ]. Невелике перевищення діапазону K_i поза 10 000 нмоль мало на меті дати змогу регресійним моделям не обмежуватись лише обмеженою вибіркою найактивніших лігандів, а й передбачати K_i для слабких інгібіторів (10 000 - 30 000 нМ).

Оскільки похибка вимірювання *Ki* зростає прямопропорційно її значенню, ми працювали з десятковим логарифмом константи - *log10Ki*. Таке перетворення допомогло збалансувати внесок помилок у функцію втрат у різних діапазонах концентрації на наблизити розподіл цільових значень до нормального. Гістограма, що показує розподіл зразків із значеннями *log10Ki*, показана на рис. 5. Набір даних регресії також був розділений на п'ять частин для перехресної валідації. Так само як і для класифікаційного набору, з даних було видалено дублікати. Аномалії за методом Д. Тюки із використанням двох параметрів: молекулярної маси та значення *log10Ki*.

Після підготовки, набір даних для *класифікації* складався з 26 808 лігандів, що відносились до двох класів активності по відношенню до людського тромбіну: 3565 активних (13,3%) та 23243 не активних (86,7%). Набір даних для *регресії* складався із 3940 унікальних лігандів тромбіну із відповідними значеннями констант інгібування (K_i). Відповідно до описаних вище правил, 3225 (81,6%) з них були активними та 726 (18,4%) – неактивними. Розподіли активних та неактивних лігандів можна бачити на рис. 4. Відзначимо, що набори даних для класифікації та регресії мають подібну до нормального розподілу форму та близькі середні значення молекулярних мас – 443 та 481 відповідно. Розподіл класифікаційних даних в основному локалізовано в межах від 300 до 600 а.о.м., тоді як розподіл даних для регресії демонструє більш широкий розкид від 300 до 700. Середнє значення $log_{10}K_i$ становило 3 та розкид від -2 до 6.

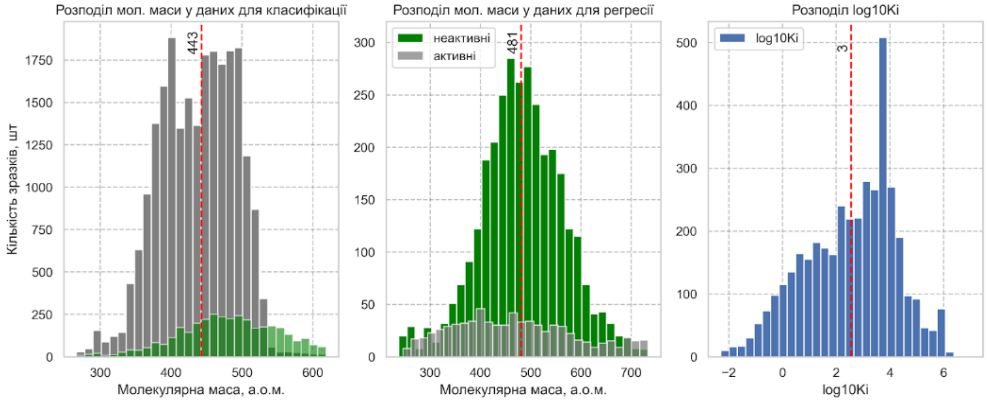

Рис. 5. Розподіли молекулярної маси лігандів у наборах даних для класифікації (зліва) та регресії (посередині) після видалення дублікатів та аномальних точок. Активні ліганди показано зеленим, неактивні - сірим кольором. Розподіл значень $log_{10}K_i$ у для регресії показано справа. Червоними вертикальними лініями показані середні значення величин, відкладених по горизонтальній осі координат

Для кращого розуміння даних було проаналізовано частоту зустрічань спільних молекулярних фрагментів серед активних та неактивних молекул. Для цього ми скористалися методом [11], що визначає різноманітні функціональні групи, а також хімічні групи, утворені лише атомами вуглецю, гетероцикли, ароматичні структури та поодинокі атоми. Аналіз мав на меті встановлення можливих структурних причин, що визначають міру активності ліганду, із точку зору хімічної будови речовини та розподілу даних.

У класифікаційному наборі даних розраховані частоти зустрічань молекулярних фрагментів у активних лігандах та співставленні із такими у неактивних молекулах. Результат порівняння для 20 фрагментів, що зустрічаються серед активних лігандів найбільш часто, можна бачити на рис. 6. Активні ліганди відзначаються більшим вмістом вторинного вуглецю (C020), бензольних кілець (c1ccccc1) та пептидних груп (NC=O). Неактивні ж молекули мають більше метильних груп (C010), етильних фрагментів (CC) та загальним вмістом атому Оксигену (O).

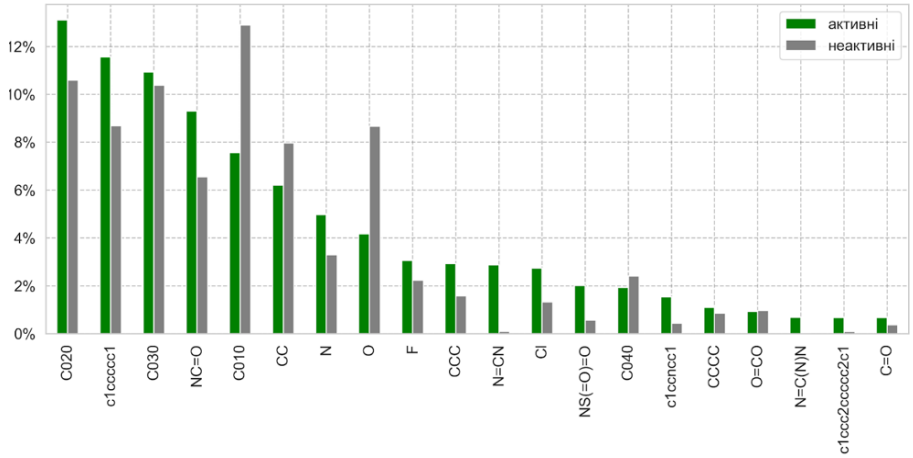

Рис. 6. Нормалізована частота зустрічань молекулярних фрагментів у активних та неактивних молекулах, що використовувалися для тренування класифікаційних моделей

У регресійному наборі даних спочатку було виділено терцилі за цільовою метрикою ($log_{10}K_i$). Концептуально, терцилі відповідали уявному класу активності лігандів: q_1 - "найактивніші", q_2 - "активні", q_3 - "помірно активні". Розраховані межі терцилів становили -2.3, 1.9, 3.6, 7. Частоти зустрічання фрагментів були розраховані для усіх молекул. Потім, 20 найчастіших фрагментів першого терцилю ("найактивніші" ліганди) були співставлені із частотами цих фрагментів у двох інших терцилях. Результат можна побачити на рис. 6. Легко побачити, що найактивніші ліганди вирізняються високим вмістом вторинного (C020) та третинного (C030) вуглецю, а також довшими нерозгалуженими алкільними елементами (CCC) - на противагу більшому вмісту коротших (CC) у менш активних лігандах. Цікаво, що фрагменти бензолу (c1ccccc1) сприяють зменшенню активності лігандів - їх вміст у q_2 та q_3 більший, ніж у q_1. Також варто відзначити, що у регресійному наборі даних прослідковується та ж тенденція, що і у класифікаційному: сполуки із вищим вмістом Оксигену (O) виявляють нижчу активність.

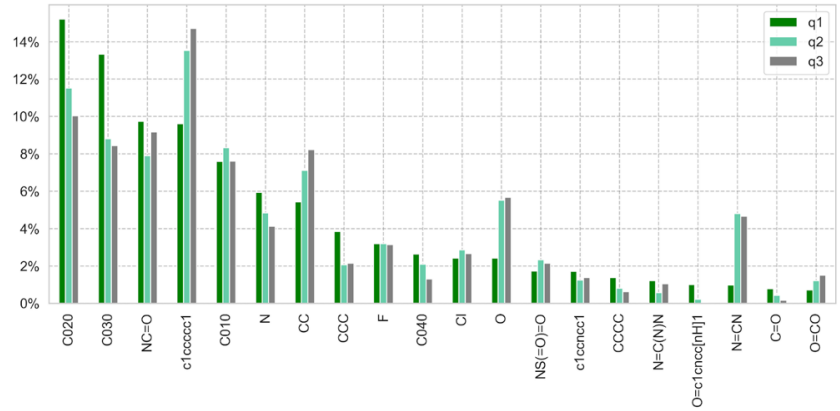

Рис. 7. Нормалізована частота зустрічань молекулярних фрагментів у сполуках, що належать до трьох терцилів регресійних даних: q1 - K_i до 100 нмоль, q2 - K_i від 100 до 4000 нмоль, q3 - K_i 4000 нмоль та більше

В даному дослідженні ліганди кодуються трьома способами. Перший спосіб – використання відбитків ECFP4 – використовується в підходах SVM, RF, CB та нейронних мережах прямого поширення. Ми генерували ці відбитки за допомогою бібліотеки RDKit. Другий спосіб – представлення лігандів через молекулярні графи, які використовуються в підході з графовими нейронними мережами. Третій спосіб - токенізація текстових позначень SMILES при використанні BERT. Жоден із цих підходів не використовує

просторові координати лігандів, що є важливою перевагою у порівнянні із підходами, що використовують координати молекул, оскільки, по-перше, дозволяє передбачати афінність для нових лігандів із невідомими просторовими структурами; по-друге, у такий спосіб із даних виключається неминуча експериментальна похибка вимірювань координат, що залежить від методу та приладу, яким проводиться вимірювання.

Перш ніж перейти до наступного розділу, варто більш детально обговорити молекулярні графи. Очікується, що визначальні взаємодії між атомами в молекулі можуть бути змоделювані за допомогою графу, і така математична модель може описати функції та властивості молекули. Щоб створити молекулярний граф, необхідно закодувати ознаки вузлів (атоми) та ребер (зв'язки). Для цієї мети ми використовували Weave atom and bond featurizer [15, 18], який генерує дев'ять атомних параметрів та три параметри зв'язку для характеристики атомів та їх оточення. До характеристик атома відноситься тип атома ('H', 'C', 'N', 'O', 'F', 'P', 'S', 'Cl', 'Br', 'I', 'інший'), формальні та часткові заряди, хіральність, ароматичність, тип sp-гібридизації, кількість донорів та/або акцепторів водневого зв'язку та розмір кільця. До характеристик зв'язку відноситься тип зв'язку («одинарний», «подвійний», «потрійний» або «ароматичний»), довжина та належність до одного кільця. Більшість цих ознак кодуються підходом one-hot, за винятком зарядів і кількості кілець, до яких належить атом, - ці ознаки є цілими числами через їх адитивну природу.

Результати

У цій частині ми аналізуємо ефективність моделей класифікації за такими метриками як accuracy, precision, recall, та AUC. Оцінка виконувалась за методом п'ятиразової перехресної валідації: кожна з моделей була натренована та провалідована п'ять разів - щоразу на унікальному розбитті вихідного набору даних на тренувальну та тестову підмножини. Тобто загалом було отримано п'ять моделей і п'ять наборів метрик. Усереднені значення класифікаційних метрик показані в Таблиці 1. Найкращі результати за AUC серед класифікаційних моделей показали SVM (0.98), CB (0,97) та RF (0,97).

Також наведено результати тренування регресійних моделей: усереднені середньоквадратичні (MSE), абсолютні похибки (MAE) та коефіцієнти детермінації (R^2). Серед окремих регресійних моделей, найменші похибки та найвищі R^2 продемонструвала модель SVM. Варт відзначити, що складні моделі засновані на нейронних мережах (FNN, GNN, BERT) показали гірші результати за лінійні моделі (SVM) та моделі засновані на деревах рішень (CB, RF). Це можна пояснити порівняно невеликим набором даних для класифікації та малим – для регресії.

Ансамблювання моделей виявляється результативним, якщо частота та розподіл помилок між методами різняться: зменшення кількості "спільних" помилок моделей призводить до зниження кількості FP ансамблю. Щоб перевірити цю гіпотезу, потрібно дослідити частоту перекривань помилок окремих моделей, що входять до ансамблю. Це можна зробити, підрахувавши площу перетину об'єднання (IoU) хибно-негативних (FN) та хибно-позитивних (FP) результатів усіх методів. Відповідні коефіцієнти IoU представлені на рис. 8. Найбільший внесок у виключення хибно-позитивних помилок зробила пара FNN-GNN (перекриття FP IoU рівне 0.18); найменший внесок зробили концептуально різні підходи SVM і RF (0.72). Найбільше перекриття хибно-негативних значень спостерігається у пари SVM-FNN (0.92), тоді як найнижчий - 0,59–0,61 – для комбінацій GNN з іншими методами. Такий аналіз виконаний лише для класифікаційних моделей, оскільки він є простішим.

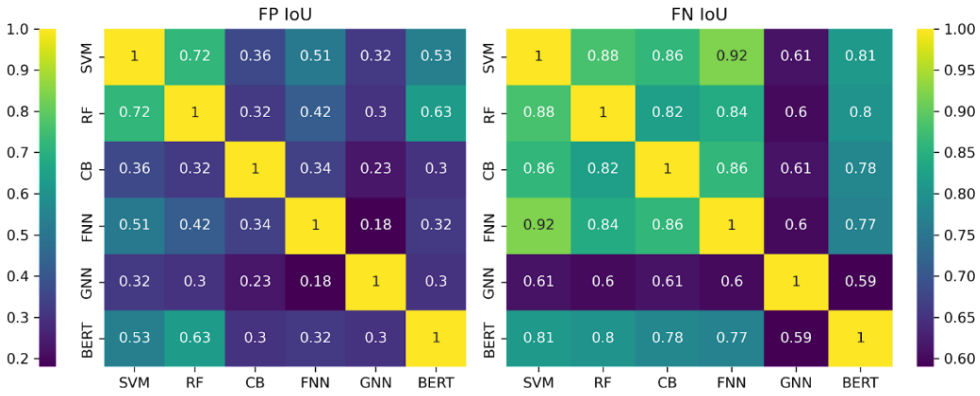

Рис. 8. Матриці коефіцієнтів перетину по об'єднанню (IoU) хибно-позитивних (FP IoU, зліва) та справжніх позитивних (FN IoU, справа) передбачень усіх моделей класифікаційного ансамблю

Для покращення ефективності передбачень, ми поєднали окремі моделі у ансамблі методом мета-стекінгу. Навчання мета-моделей проводилося на перехресних передбаченнях базових моделей (5 підвибірок), які поєднувалися у матриці розміром N x L, де N – кількість рядків набору даних, L – кількість базових моделей, яка у нашому випадку становила 6. Тренування та валідація моделей другого рівня (мета-моделей) також виконувалася із 5-разовою перехресною валідацією. Таке комбінування слабших моделей у ансамблі із мета-моделями дозволило покращити відгук (Recall) класифікаційного ансамблю на 34,9% та коефіцієнт детермінації (R^2) регресійного ансамблю на 21% у порівнянні із середніми значеннями окремих моделей. Усі результати можна бачити у зведеній таблиці 1.

Таблиця 1

Зведена таблиця показників регресійних та класифікаційних моделей розрахованих за методом п'ятиразової перехресної валідації

Тип	Класифікація				Регресія		
Модель	Accuracy	Precision	Recall	AUC	R^2	MAE	MSE
SVM	0.95	0.84	0.91	0.98	0.74	0.55	0.56
RF	0.93	0.80	0.82	0.97	0.66	0.65	0.71
CB	0.94	0.87	0.77	0.97	0.71	0.59	0.62
FNN	0.95	0.85	0.86	0.91	0.69	0.60	0.66
GNN	0.88	0.80	0.45	0.94	0.69	0.59	0.66
BERT	0.87	0.69	0.55	0.90	0.61	0.66	0.82
Середнє	0.92	0.808	0.727	0.945	0.68	0.61	0.67
Ансамблі	0.99	0.94	0.98	0.998	0.827	0.49	0.49
Зміна	+7.6%	+16.3%	+34.9%	+5.6%	+21%	-19.2%	-27%

Для мета-стекінгу використовувалися лінійні моделі (GLM). Коефіцієнти цих лінійних моделей розраховувалися шляхом максимізації оцінки максимальної правдоподібності (maximum likelihood estimation), отож абсолютні значення знайдених ваг моделей другого рівня можна розглядати як мірило інформаційного внеску певної моделі-учасника ансамблю до загального результату. Абсолютні величини коефіцієнтів були нормалізовані та виражені у відсотковому вигляді. Результат можна бачити на рисунку 9.

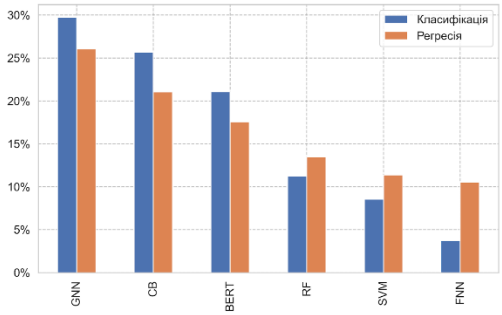

Рис. 9. Нормалізовані коефіцієнти класифікаційних (синім) та регресійних (помаранчевим) мета-моделей

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У цій роботі представлено новий метод високопродуктивного віртуального скринінгу для передбачення молекулярної спорідненості лігандів до одного рецептора. Основою методу є використання методу мета-стекінгу та різноманітних за своєю природою моделей машинного навчання: методу опорних векторів, випадкового лісу, градієнтного бустингу, повнозв'язної нейронної мережі прямого поширення, графової нейронної мережі і двоспрямованого кодувального представлення з трансформерів. Реалізовано два послідовні ансамблі: класифікаційний та регресійний. Класифікаційний ансамбль передбачає імовірність зв'язування рецептора та ліганда. Ліганди, які були класифіковані як активні, надходять до регресійного ансамблю, який передбачає афінність ліганду до рецептору кількісно - у вигляді константи інгібування K_i.

Таким чином, запропонований метод дає змогу передбачити активність молекул-кандидатів не лише якісно (активний або неактивний), але й кількісно (значення K_i). У роботі показано, що поєднання моделей методом мета-стекінгу суттєво знижує кількість помилок, підвищує загальну точність та статистичну достовірність результатів; дозволяє виключити (у випадку класифікації) або компенсувати (у випадку регресії) помилки, допущені іншими моделями ансамблю.

References

1. Beck D., Haffari, G., Cohn, T. Graph-to-sequence learning using gated graph neural networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 1. Association for Computational Linguistics, Melbourne, Australia. 2018. P. 273–283. https://doi.org/10.18653/v1/P18- 1026.

2. Beck B., Shin B., Choi Y., Park S., Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. Comput. Struct. Biotechnol. J. 2020. P. 784–790. https://doi.org/10.1016/j.csbj.2020.03.025.

3. Breiman L. Random forests. Mach. Learn. 45 (1). 2010. P. 5–32. https://doi.org/10.1023/ A:1010933404324.

4. Chen Y.-C. Beware of docking! Trends Pharmacol. Sci. 36 (2). 2015. P. 78–95. https://doi. org/10.1016/j.tips.2014.12.001.

5. Chen J.-Q., Chen H.-Y., Dai W.-j., Lv Q.-J., Chen C.-C. Artificial intelligence approach to find lead compounds for treating tumors. J. Phys. Chem. Lett. 10 (15). 2019. P. 4382–4400. https://doi.org/10.1021/acs.jpclett.9b01426.

6. Chupakhin V., Marcou G., Baskin I., Varnek A., Rognan D. Predicting ligand binding modes from neural networks trained on protein-ligand interaction fingerprints. 53 (4). 2016. P. 763–772. https://doi.org/10.1021/ ci300200r.

7. Davis M., Hunt J., Herrgard S., Ciceri P., Wodicka L., Pallares G., Hocker M., Treiber D., Zarrinkar P. Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29. 2010. P. 1046–1051. https://doi.org/10.1007/978-1-4939-9752-7.

8. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding.

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Association for Computational Linguistics, Minneapolis, Minnesota. 2019. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

9. Ellingson S., Davis B., Allen J. Machine learning and ligand binding predictions: a review of data, methods, and obstacles. Biochim. Biophys. Acta (BBA) – General Subj. 1864 (6), 2020. P. 129-545. https://doi.org/10.1016/j.bbagen.2020.129545.

10. Gao K., Nguyen D., Chen J., Wang R., Wei G.-W. Repositioning of 8565 existing drugs for COVID-19. 11 (13). 2020. P. 5373–5382. https://doi.org/ 10.1021/acs.jpclett.0c01579.

11. Hartshorn M., Verdonk M., Chessari G., Brewerton S., Mooij W., Mortenson P., Murray C. Diverse, high-quality test set for the validation of protein-ligand docking performance. J. Med. Chem. 50 (4). 2007. P. 726–741. https://doi.org/10.1021/ jm061277y.

12. He T., Heidemeyer M., Ban F., Cherkasov A., Ester M. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. 9. 2017. P. 24. https://doi.org/10.1186/s13321-017-0209-z.

13. Heck G., Pintro V., Pereira R., de Avila M., Levin N., de Azevedo Jr. W. Supervised machine learning methods applied to predict ligand-binding affinity. Curr. Med. Chem. 24 (23). 2017. P. 2459–2470. https://doi.org/10.2174/ 092986732466617062309250 3.

14. Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q., Shoemaker B., Thiessen P., Yu B., Zaslavsky L., Zhang J., Bolton E. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019. P. 1102–1109.

15. Kowalewski J., Ray A. Predicting novel drugs for sars-cov-2 using machine learning from a >10 million chemical space. Helion 6, e04639. 2020. https://doi.org/ 10.1016/j.heliyon.2020.e04639.

16. Kundu I., Paul G., Banerjee R. A machine learning approach towards the prediction of protein-ligand binding affinity based on fundamental molecular properties. RSC Adv. 8. 2018. P. 12127–12137. https://doi.org/10.1039/C8RA00003D.

17. Kwon Y., Shin W.-H., Ko J., Lee J. Ak-score: accurate protein-ligand binding affinity prediction using an ensemble of 3d-convolutional neural networks. Int. J. Mol. Sci. 21 (22). 2020. P. 8424. https://doi.org/10.3390/ijms21228424.