

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри молекулярної біотехнології та біоінформатики

к.б.н. Нипорко Олексій Юрійович

Протокол №_____ засідання кафедри

від “_____” _____20____ р.

**ПРОЄКТУВАННЯ PROTAC-СПОЛУК ДЛЯ ДЕГРАДАЦІЇ BRD4 ІЗ
ЗАСТОСУВАННЯМ ГЕНЕРАТИВНОГО МАШИННОГО НАВЧАННЯ
ТА *IN SILICO* ОПТИМІЗАЦІЇ ЛІНКЕРІВ**

Випускна кваліфікаційна робота магістра
студента спеціальності 091 Біологія та біохімія
ОП «Біоінформатика та структурна біологія»

Мельник Варвари Володимирівни

Науковий керівник від кафедри
доцент кафедри молекулярної біотехнології та біоінформатики
к.ф.-м.н. **Войтешенко Іван Сергійович**

Науковий керівник
молодший науковий співробітник,
Науково-виробничого хіміко-біологічного центру
Київського національного університету імені Тараса Шевченка
к.б.н. **Протопопов Микола Васильович**

Оцінка захисту роботи

Київ — 2026 р.

АНОТАЦІЯ

Мельник В. В. Проектування PROTAC-сполук для деградації BRD4 із застосуванням генеративного машинного навчання та *in silico* оптимізації лінкерів. — Випускна кваліфікаційна робота магістра. — Київський національний університет імені Тараса Шевченка, Навчально-науковий інститут високих технологій, кафедра молекулярної біотехнології та біоінформатики. — Київ, 2026.

Ключові слова: PROTAC, BRD4, BET-інгібітори, таргетна деградація білків, генеративне машинне навчання, навчання з підкріпленням, Link-INVENT, ансамблеві моделі, молекулярний докінг, фармакокінетика, дизайн лінкерів.

Робота присвячена розробці та валідації інтегрованої обчислювальної методології генерації й ранжування PROTAC-молекул, спрямованих на таргетну деградацію епігенетичного регулятора BRD4 — ключової онкологічної мішені, гіперекспресія якої асоційована з гематологічними злоякісними новоутвореннями та низкою твердих пухлин.

Дослідження реалізовано у два концептуально пов'язані етапи. На першому етапі розроблено гетерогенну ансамблеву модель прогнозування біологічної активності PROTAC. Ансамбль інтегрує чотири базові алгоритми (XGBoost, Random Forest, SVM та KNN), кожен з яких працює у власному просторі ознак, сформованому на основі фармакофорних відбитків, відбитків Моргана, 2D-дескрипторів RDKit, ембедингів ChemBERTa та PCA-редукованих білкових представлень. Агрегацію прогнозів реалізовано за допомогою мета-класифікатора на основі логістичної регресії, що дозволяє ефективно інтегрувати різномірні типи даних. На незалежній тестовій вибірці модель досягла високих показників якості ($MCC = 0.75$, $ROC-AUC = 0.92$, специфічність $= 0.97$), що підтверджує її придатність як надійної оціночної функції. Валідність моделі додатково обґрунтовано за допомогою Y-рандомізації та аналізу епістемічної невизначеності.

На другому етапі розроблену модель інтегровано у генеративний конвеєр молекулярного дизайну на платформі REINVENT4 з використанням модуля Link-INVENT та алгоритму навчання з підкріпленням DAP. Генерацію лінкерів здійснено для чотирьох комбінацій лігандів (2 для BRD4 і 2 для CRBN). Отримані RL-оптимізовані сполуки демонструють статистично значуще підвищення прогнозованої біологічної активності порівняно з апіорною вибіркою при збереженні фізико-хімічних параметрів у межах bRo5-простору ($MW \approx 843$ Да, $\log P \approx 5.03$, $TPSA 200 - 250 \text{ \AA}^2$).

Подальший структурний відбір здійснено за допомогою молекулярного докінгу (5 незалежних запусків на сполуку). Серед топ-100 кандидатів ідентифіковано три пріоритетні PROTAC (P34, P87 та P96), які поєднують високий афінітет до BRD4 та CRBN (середня оцінка від -32.44 до -35.90) із високою відтворюваністю результатів ($CV < 10\%$), що свідчить про стабільність передбачуваних комплексів.

Якісний аналіз взаємодій підтвердив, що всі три пріоритетні кандидати відтворюють канонічну просторову організацію трикомпонентного комплексу BRD4–CRBN: зберігаються характерні контакти з WPF-шельфом у бромодомені BD1 та триптофановим кластером CRBN. Сконструйовані лінкерні фрагменти не є інертними просторовими розділювачами, а є активними елементами інтерфейсу зв'язування, формуючи додаткові стабілізуючі контакти вздовж міжбілкового інтерфейсу.

Запропонована методологія формує узгоджений та масштабований обчислювальний підхід — від побудови предиктивної моделі до структурно обґрунтованого відбору кандидатів — і має транслятивний потенціал для інших систем POI/E3-ліганд у межах раціонального дизайну PROTAC.

ABSTRACT

Melnyk V. V. Design of PROTAC compounds for BRD4 degradation using generative machine learning and *in silico* linker optimization. — Master's thesis. — Taras Shevchenko National University of Kyiv, Educational and Scientific Institute of High Technologies, Department of Molecular Biotechnology and Bioinformatics. — Kyiv, 2026.

Keywords: PROTAC, BRD4, BET inhibitors, targeted protein degradation, generative machine learning, reinforcement learning, Link-INVENT, ensemble models, molecular docking, pharmacokinetics, linker design.

This thesis is devoted to the development and validation of an integrated computational methodology for the generation and evaluation of PROTAC molecules aimed at the targeted degradation of the epigenetic regulator BRD4, a key oncological target whose overexpression is associated with hematological malignancies and a range of solid tumors.

The study was carried out in two conceptually interconnected stages. At the first stage, a heterogeneous ensemble model for predicting PROTAC biological activity was developed. The ensemble integrates four base classifiers (XGBoost, Random Forest, SVM, and KNN), each operating on its own feature representation derived from pharmacophore descriptors, Morgan fingerprints, RDKit 2D physicochemical descriptors, ChemBERTa embeddings, and PCA-reduced protein embeddings. Prediction aggregation is performed using a logistic regression-based meta-classifier, enabling effective integration of heterogeneous feature modalities. On an independent test set, the model achieved high performance (MCC = 0.75, ROC-AUC = 0.92, specificity = 0.97), demonstrating its suitability as a reliable scoring function. Model validity was further confirmed through Y-randomization and epistemic uncertainty analysis.

At the second stage, the developed model was integrated into a generative molecular design framework based on the REINVENT4 platform using the Link-INVENT module and the DAP reinforcement learning algorithm. Linker generation was performed for four warhead combinations (2 for BRD4 and 2 for CRBN). The RL-optimized compounds demonstrated a statistically significant

increase in predicted biological activity compared to the prior distribution while maintaining physicochemical properties within the range characteristic of the bRo5 chemical space ($MW \approx 843$ Da, $\log P \approx 5.03$, $TPSA\ 200 - 250\ \text{\AA}^2$).

Subsequent structural prioritization was performed using molecular docking (5 independent runs per compound). Among the top 100 candidates, three priority PROTACs (P34, P87, and P96) were identified, combining strong binding affinity to BRD4 and CRBN (mean docking score from -32.44 to -35.90) with high reproducibility ($CV < 10\%$), indicating the stability of the predicted complexes.

The qualitative interaction analysis confirmed that all three prioritized candidates reproduce the canonical spatial organization of the BRD4–CRBN ternary complex: key contacts with the WPF shelf in the BD1 bromodomain and the tryptophan cluster of CRBN are preserved. The designed linker fragments act not as inert spacers but as active elements of the binding interface, forming additional stabilizing interactions along the protein–protein interface.

The proposed methodology establishes a coherent and scalable computational workflow, spanning from predictive model development to structure-based candidate prioritization, and demonstrates translational potential for other POI/E3 ligand pairs within the PROTAC design paradigm.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

- УПС — убіквітин-протеасомна система;
- ADMET (*absorption, distribution, metabolism, excretion and toxicity*) — абсорбція, розподіл, метаболізм, виведення, токсичність;
- BD1 / BD2 — перший та другий бромодомени;
- BET (*Bromodomain and Extra-Terminal*) — родина білків із бромодоменом та позатермінальним доменом;
- BRD4 (*Bromodomain-containing protein 4*) — бромодомен-вмісний білок 4, білок родини BET, епігенетичний регулятор транскрипції;
- bRo5 (*beyond Rule of Five*) — поза правилами Ліпінського;
- CRBN (*Cereblon*) — E3-убіквітин-лігаза;
- CTD (*C-terminal domain*) — С-кінцевий домен;
- CV (*Coefficient of Variation*) — коефіцієнт варіації;
- DAP (*Dual Action Policy*) — політика подвійної дії, тобто стратегія агента, яка передбачає використання двох типів дій або двох механізмів прийняття рішень;
- DDR (*DNA Damage Response*) — відповідь на пошкодження ДНК;
- DSB (*Double-Strand Break*) — дволанцюговий розрив ДНК;
- E1, E2, E3 — ферменти каскаду убіквітинування;
- ET (*Extra-Terminal domain*) — позатермінальний домен;
- HAT (*Histone Acetyltransferase*) — гістонова ацетилтрансфераза;
- IL (*Interleukin*) — інтерлейкін;
- KNN (*K-Nearest Neighbors*) — метод k-найближчих сусідів;
- MCC (*Matthews Correlation Coefficient*) — коефіцієнт кореляції Меттьюса;
- ML (*Machine Learning*) — машинне навчання;
- NF-κB — транскрипційний фактор, що регулює запальні процеси;
- PCA (*Principal Component Analysis*) — метод головних компонент;
- P-TEFb (*Positive Transcription Elongation Factor b*) — фактор елонгації транскрипції;
- POI (*Protein of Interest*) — білок-мішень;
- Pol II (*RNA Polymerase II*) — РНК-полімераза II;

PPI (*Protein–Protein Interaction*) — білок-білкова взаємодія;
PROTAC (*Proteolysis Targeting Chimeras*) — химерні молекули для таргетної деградації білків;
RBF (*Radial Basis Function*) — радіально-базисна функція ядра;
RF (*Random Forest*) — випадковий ліс;
RL (*Reinforcement Learning*) — навчання з підкріпленням;
SMILES (*Simplified Molecular Input Line Entry System*) — лінійне представлення хімічних структур;
SVM (*Support Vector Machine*) — метод опорних векторів;
TNF- α (*Tumor Necrosis Factor alpha*) — фактор некрозу пухлин α ;
TPE (*Tree-structured Parzen Estimator*) — деревоподібний оцінювач Парцена;
TSS (*Transcription Start Site*) — сайт початку транскрипції;
VHL (*Von Hippel–Lindau protein*) — E3-убіквітин-лігаза;
XGBoost (*eXtreme Gradient Boosting*) — алгоритм градієнтного бустингу дерев рішень.

ЗМІСТ

ВСТУП.....	11
РОЗДІЛ I.....	13
ОГЛЯД ЛІТЕРАТУРИ.....	13
1.1. Біологічна роль BRD4 у регуляції клітинних процесів	13
1.1.1. Загальна характеристика білків родини BET	13
1.1.2. Структурна організація та доменна будова BRD4.....	14
1.1.3. Функціональна роль BRD4 у регуляції транскрипції та епігенетичних механізмах.....	15
1.1.4. Участь BRD4 у молекулярних механізмах розвитку онкологічних і запальних захворювань	17
1.2. Сучасні підходи до інгібування BRD4.....	19
1.2.1. Механізми дії малих молекул-інгібіторів.....	19
1.2.2. Основні представники BET-інгібіторів	19
1.2.3. Обмеження традиційних інгібіторів: селективність, токсичність та розвиток резистентності	20
1.3. Концепція PROTAC у сучасній фармакології	21
1.3.1. Принципи технології таргетної деградації білків	21
1.3.2. Формування та стабілізація трикомпонентного комплексу POI–PROTAC–E3-лігаза	22
1.3.3. Порівняльна характеристика PROTAC та класичних інгібіторів ...	22
1.4. Структурна організація PROTAC-молекул	23
1.4.1. Ліганд, спрямований на білок-мішень.....	23
1.4.2. Ліганд для рекрутування E3-убіквітин-лігази	24
1.4.3. Лінкер як ключовий структурний елемент PROTAC та його вплив на ефективність деградації	25
1.5. Роль E3-убіквітин-лігаз у механізмі дії PROTAC.....	26
1.5.1. Убіквітин-протеасомна система як інструмент деградації.....	26
1.5.2. Основні типи E3-убіквітин-лігаз, що використовуються у PROTAC-технології.....	27

1.6. Обчислювальні підходи до проектування PROTAC: сучасний стан і прогалини	28
1.6.1. Структурно-орієнтоване моделювання трикомпонентних комплексів POI–PROTAC–E3-лігаза.....	28
1.6.2. Прогнозування біологічної активності PROTAC методами машинного навчання	29
1.6.3. Генеративні підходи до de novo дизайну лінкерів PROTAC	30
1.6.4. Позичування цієї роботи у контексті існуючих підходів	31
РОЗДІЛ II.....	32
МАТЕРІАЛИ І МЕТОДИ	32
2.1. Модель прогнозування біологічної активності сполук PROTAC	32
2.1.1. Формування набору даних та анотація біологічної активності	32
2.1.2. Представлення вхідних даних	33
2.1.3. Попередня обробка та зменшення розмірності ознак.....	34
2.1.4. Архітектура ансамблевої моделі	35
2.1.5. Оптимізація гіперпараметрів.....	36
2.1.6. Мета-класифікатор та схема навчання	37
2.1.7. Абляційний аналіз	38
2.1.8. Статистична валідація та оцінка невизначеності	38
2.2. Генерація та оптимізація PROTAC за допомогою REINVENT4 та Link-INVENT	39
2.2.1. Архітектура генеративної моделі та модуль Link-INVENT	39
2.2.2. Стратегія навчання з підкріпленням.....	39
2.2.3. Процедура генерації вибірки PROTAC	41
2.2.4. Ліганди BRD4 та CRBN, які були використані у якості лігандних фрагментів генерації PROTAC-сполук.....	42
2.3. Розрахунок фізико-хімічних дескрипторів та прогнозування фармакокінетичного профілю	43
2.4. Структурно-орієнтована оцінка PROTAC-кандидатів методом молекулярного докінгу.....	43
2.4.1. Відбір сполук для докінгу.....	44

	10
2.4.2. Підготовка структур білків та протокол докінгу	44
2.4.3. Критерії відбору та аналіз результатів	44
РОЗДІЛ III	46
РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ	46
3.1. Оцінка моделі передбачення біологічної активності PROTAC.....	46
3.1.1. Кореляційний аналіз передбачень базових класифікаторів.....	46
3.1.2. Порівняльний аналіз стратегій агрегації передбачень.....	47
3.1.3. Комбінаторний абляційний аналіз	48
3.1.4. Ваги мета-класифікатора та інтерпретація внеску експертів.....	50
3.1.5. Оцінка ефективності на тестовій вибірці.....	50
3.1.6. Епістемічна невизначеність ансамблю.....	53
3.1.7. Статистична валідація моделі методом Y-рандомізації	54
3.2. Генерація та оптимізація лінкерів PROTAC-сполук за допомогою Link- INVENT з підкріплювальним навчанням	56
3.2.1. Траєкторія навчання агента з підкріпленням.....	56
3.2.2. Новизна та внутрішня різноманітність згенерованої вибірки	57
3.2.3. Порівняльний аналіз фізико-хімічних властивостей апріорної та RL-оптимізованої вибірок PROTAC-сполук.....	59
3.3. Комплексний аналіз результатів молекулярного докінгу топ-100 PROTAC-сполук.....	64
3.3.1. Загальна характеристика результатів докінгу	64
3.3.2. Оцінка відтворюваності та збіжності результатів	65
3.3.3. Топ-10 PROTAC-сполук за середньою оцінкою докінгу	66
3.3.4. Аналіз взаємодій топ-3 PROTAC-сполук з трикомпонентним комплексом BRD4–CRBN.....	68
3.4. Обмеження запропонованої методології та практичне значення отриманих результатів.....	71
ВИСНОВКИ	73

ВСТУП

Таргетна деградація білків є сучасною терапевтичною стратегією, що дозволяє не лише пригнічувати активність патологічно значущих білків, а й спрямовано знижувати їх рівень у клітині. На відміну від класичних малих молекул-інгібіторів, цей підхід використовує внутрішньоклітинні механізми контролю якості білків, зокрема убіквітин-протеасомну систему (УПС), що відкриває можливості для впливу на мішені з обмеженою ефективністю традиційного інгібування.

Перспективним інструментом таргетної деградації є технологія PROTAC (Proteolysis Targeting Chimeras). PROTAC-молекули — це біфункціональні сполуки, що одночасно зв'язують білок-мішень та E3-убіквітин-лігазу, індукуючи утворення тернарного комплексу, убіквітинування мішені та її подальшу протеасомну деградацію. Такі молекули складаються з двох лігандів, з'єднаних лінкером, структура і довжина якого визначають ефективність деградації, селективність, здатність до утворення тернарного комплексу та фармакокінетичні властивості.

Актуальною мішенню для PROTAC-опосередкованої деградації є бромодомейн-вмісний білок 4 (Bromodomain-containing protein 4, BRD4) — епігенетичний регулятор транскрипції, залучений до контролю експресії генів, пов'язаних із проліферацією, виживанням клітин і запальними відповідями. Надмірна активність BRD4 пов'язана з онкологічними захворюваннями, зокрема лімфомами, лейкеміями й твердоклітинними пухлинами, а також із хронічним запаленням. Обмеження традиційних інгібіторів BRD4 — недостатня селективність, токсичність і резистентність — зумовлюють потребу в альтернативних підходах до цієї мішені.

Генеративні методи машинного навчання відкривають нові можливості для проєктування PROTAC-молекул, спрямованих на деградацію BRD4. Вони дозволяють створювати та оцінювати віртуальні структури з прогнозованими фізико-хімічними, фармакокінетичними та біологічними властивостями. *In silico* оптимізація лінкерів прискорює відбір перспективних кандидатів і підвищує ймовірність отримання молекул із бажаним профілем активності.

Таким чином, розробка генеративного підходу до дизайну PROTAC для селективного таргетування BRD4 є актуальною, оскільки поєднує біологічну значущість цієї мішені з можливостями сучасних обчислювальних методів і сприяє прискоренню пошуку перспективних терапевтичних агентів.

У зв'язку з цим **метою дипломної роботи** є розробка та валідація інтегрованої обчислювальної методології для *de novo* генерації та структурно-орієнтованого відбору PROTAC-молекул, спрямованих на деградацію білка BRD4, що поєднує ансамблеве машинне навчання для прогнозування біологічної активності, генеративну оптимізацію лінкерів на основі навчання з підкріпленням та молекулярний докінг для оцінки кандидатів.

Для досягнення поставленої мети у роботі передбачається розв'язання таких **завдань**:

1. Розробити гетерогенну ансамблеву модель машинного навчання для бінарної класифікації біологічної активності PROTAC-сполук.
2. Провести валідацію моделі на незалежній тестовій вибірці та застосувати процедуру Y-рандомізації для контролю наявності випадкових кореляцій і підтвердження узагальнювальної здатності моделі.
3. Реалізувати оптимізацію апіорної генеративної моделі Link-INVENT (платформа REINVENT4) із використанням навчання з підкріпленням, де ансамблева модель машинного навчання виступає функцією винагороди.
4. Здійснити *de novo* генерацію лінкерів PROTAC на основі RL-оптимізованої Link-INVENT моделі та сформувати вибірку сполук для подальшого аналізу.
5. Провести порівняльний аналіз фізико-хімічних і прогнозованих фармакокінетичних властивостей (MW, LogP, TPSA, Caco-2) апіорної та RL-оптимізованої вибірок PROTAC для оцінки відповідності згенерованих сполук простору bRo5-молекул.
6. Здійснити структурно-орієнтовану оцінку топ-100 RL-генерованих кандидатів методом молекулярного докінгу та ідентифікувати пріоритетні сполуки на основі комплексних критеріїв: оцінки докінгу, відтворюваності і збіжності результатів.

РОЗДІЛ І

ОГЛЯД ЛІТЕРАТУРИ

1.1. Біологічна роль BRD4 у регуляції клітинних процесів

1.1.1. Загальна характеристика білків родини BET

Родина білків із бромодоменом та позатермінальним доменом (Bromodomain and Extra-Terminal, BET) займає центральне місце в ієрархії епігенетичних регуляторів, виступаючи ключовими читачами гістонового коду [1]. До цієї родини належать чотири споріднені білки: BRD2, BRD3, BRD4, що характеризуються повсюдним (убіквітарним) рівнем експресії в тканинах ссавців, та BRDT, чия активність обмежена переважно гермінативними клітинами сім'яників (Рис. 1.1) [2]. Спільною структурною ознакою всіх представників BET є наявність двох висококонсервативних тандемних амінотермінальних бромодоменів (BD1 та BD2) та унікального позатермінального домену (ET), залученого до білок-білкових взаємодій [3].

Сучасні функціональні дослідження вказують на чіткий розподіл ролей між цими доменами: BD1 відіграє домінуючу роль у первинному розпізнаванні ацетильованих залишків лізину на хвостах гістонів хроматину, забезпечуючи специфічне зв'язування з промоторними ділянками, тоді як BD2 сприяє подальшому рекрутуванню білкових комплексів, необхідних для індукції та підтримки експресії генів [4]. Завдяки здатності інтегрувати сигнали ацетилювання гістонів у транскрипційні програми, білки родини BET регулюють ключові клітинні процеси, включаючи проліферацію, диференціацію, клітинний цикл та автофагію. Дисрегуляція цих механізмів, зокрема через гіперактивацію супер-енхансерів та онкогенних транскрипційних програм, тісно пов'язана з патогенезом широкого спектра захворювань, включаючи гострі лейкози, тверді пухлини (зокрема рак молочної та передміхурової залоз), а також системні запальні процеси [5]. Це робить родину BET, і особливо білок BRD4, критично важливою мішенню для розробки нових стратегій таргетної терапії.

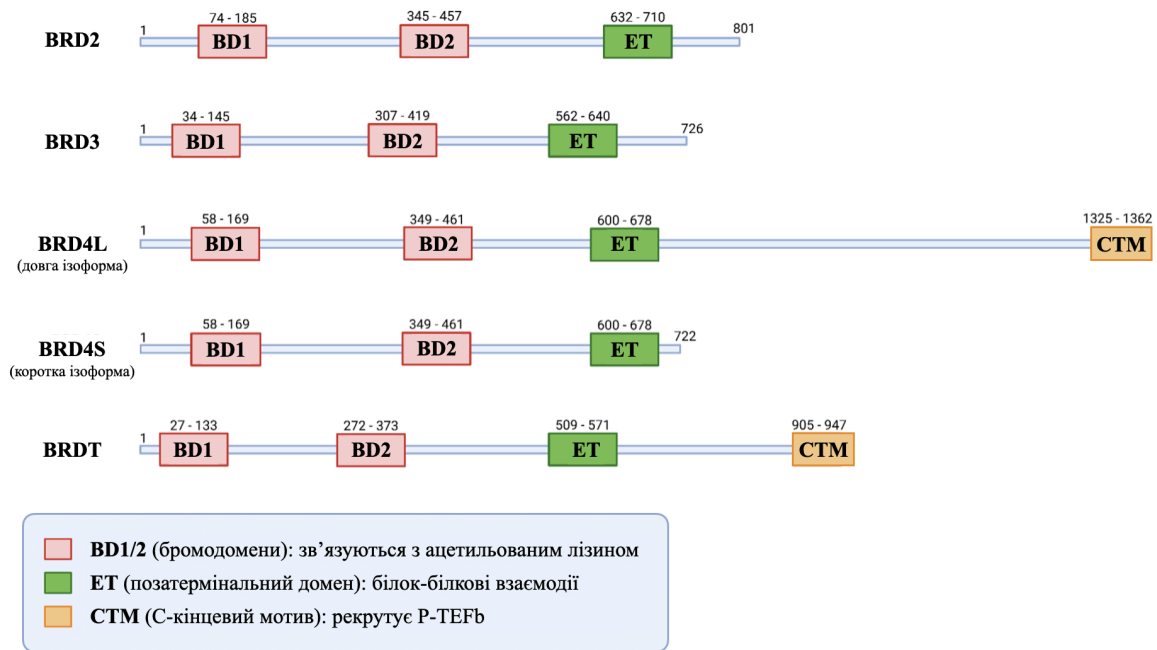


Рисунок 1.1. Структурна організація білків родини BET [2].

1.1.2. Структурна організація та доменна будова BRD4

BRD4 спочатку був ідентифікований як білок, що зв'язується з мітотичними хромосомами та залишається асоційованим з ацетильованим хроматином протягом усього клітинного циклу, забезпечуючи епігенетичну пам'ять для відновлення транскрипції генів у фазі G1 після мітозу [6, 7]. Білок BRD4 належить до еволюційно консервативної родини BET (Bromodomain and Extra-Terminal), і його структурна організація безпосередньо зумовлює його роль як ключового епігенетичного «читача» [8]. Білок BRD4 представлений трьома ізоформами — BRD4L, BRD4Sa і BRD4Sb, що характеризуються різними функціональними ролями у процесі транскрипції. Основу архітектури BRD4 становлять два тандемні амінотермінальні бромодомени, відомі як BD1 та BD2, які мають високий ступінь гомології, проте функціонально спеціалізовані. Кожен із цих доменів складається з лівозакрученого пучка чотирьох α -спіралей (αZ , αA , αB , αC), з'єднаних петлями ZA та BC, що формують гідрофобну кишеню для розпізнавання ацетильованих залишків лізину на хвостах гістонів H3 та H4 [9]. Сучасні структурні дослідження вказують на те, що BD1 демонструє

вищу спорідненість до мультиацетильованих пептидів гістонів і забезпечує первинне заякорювання білка на хроматині, тоді як BD2 частіше залучений до стабілізації транскрипційних комплексів.

Окрім бромодоменів, у центральній частині молекули розташований домен ET (Extra-Terminal), який є унікальною ознакою родини BET [10]. Домен ET не взаємодіє безпосередньо з ДНК або модифікованими гістонами, а натомість виступає у якості білок-білкова платформа для рекрутування різноманітних ефекторних молекул, таких як гістонова метилтрансфераза NSD3, деметилаза JMJD6 або білок ремоделювання хроматину CHD4 [3]. Важливою особливістю BRD4 є наявність довгої С-термінальної ділянки, відомої як C-terminal domain (CTD), що присутня лише у повнорозмірній ізоформі (BRD4-L). Цей домен містить велику кількість неструктурованих регіонів, які відіграють критичну роль у взаємодії з позитивним фактором елонгації транскрипції ψ (P-TEFb) та медіаторним комплексом [11].

Окрім структурних доменів BD1, BD2 та CTD, представники родини BET характеризуються наявністю додаткових функціональних мотивів, серед яких кластери сайтів фосфорилування для казеїн кінази 2 (NPS і CPS) та домен взаємодії, збагачений основними амінокислотними залишками, відомий як BID [12]. Ці мотиви сприяють динамічному переміщенню білка вздовж нуклеосом під час транскрипції. Така мультимодульна будова дозволяє BRD4 одночасно розпізнавати епігенетичні мітки та координувати збірку макромолекулярних транскрипційних ансамблів.

1.1.3. Функціональна роль BRD4 у регуляції транскрипції та епігенетичних механізмах

BRD4 відіграє фундаментальну роль у координації клітинного циклу та транскрипційної пластичності. Спочатку ідентифікований як мітотичне «закладення пам'яті», цей білок залишається асоційованим із хромосомами під час М-фази, коли більшість факторів дисоціюють [13]. Маркуючи сайти старту транскрипції (TSS) генів, запрограмованих на експресію одразу після

поділу, BRD4 забезпечує швидке відновлення транскрипції у фазі G1 та перехід клітини до S-фази.

Функціональна активність BRD4 базується на інтеграції сигналів гістонового коду з роботою РНК-полімерази II (Pol II) (Рис. 1.2). Повнорозмірна ізоформа (BRD4L) через свій PID-домен рекрутує позитивний фактор елонгації транскрипції b (P-TEFb), вивільняючи його з інгібіторного комплексу 7SK РНП [14]. Це стимулює фосфорилювання Pol II та зняття транскрипційної паузи. Важливо, що BRD4L володіє атиповою кіназною активністю, яка дозволяє фосфорилювати CTD-домен Pol II незалежно від P-TEFb [15], а також внутрішньою гістонацетилтрансферазною (НАТ) активністю. Остання забезпечує ацетилювання та евікцію (витіснення) нуклеосом, що призводить до декомпактизації хроматину в цільових локусах.

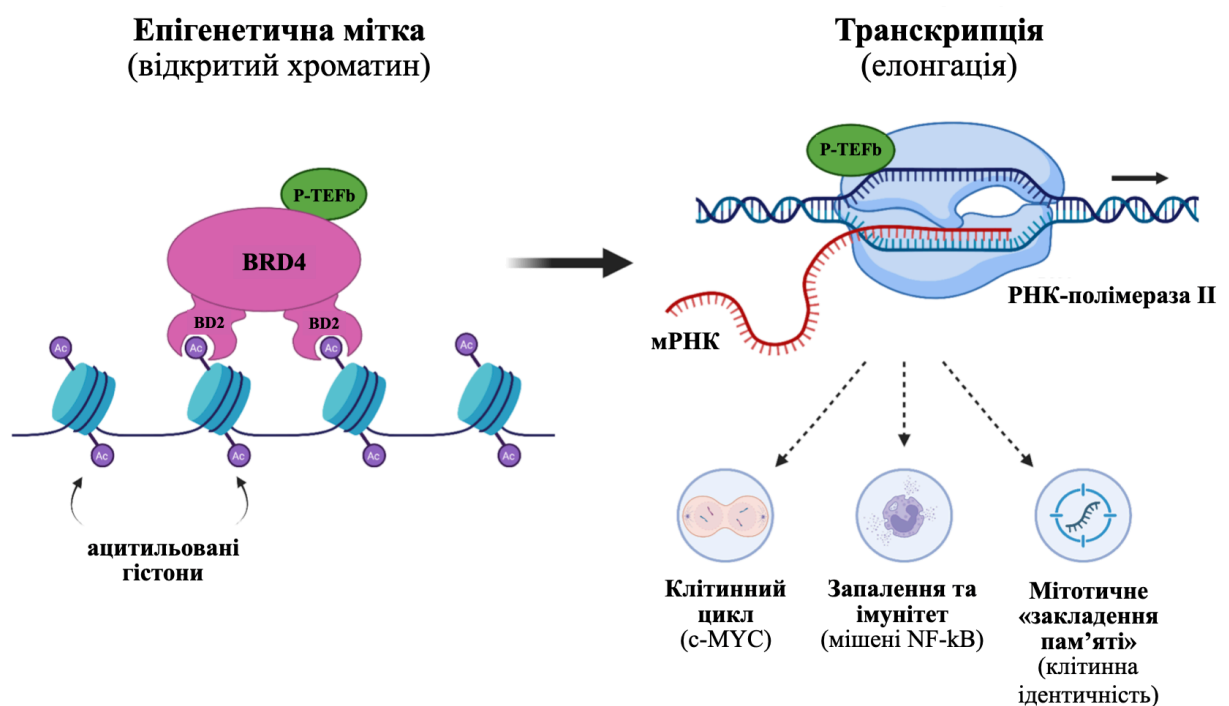


Рисунок 1.2. Функціональна роль BRD4 у регуляції епігенетичних механізмів і транскрипції: BRD4 розпізнає ацетилювані гістони, стабілізує відкритий стан хроматину та, рекрутуючи коактиваторні комплекси (зокрема P-TEFb), забезпечує фосфорилювання РНК-полімерази II і перехід до продуктивної елонгації транскрипції [2].

Окрім прямої участі в елонгації, BRD4 виступає молекулярним хабом для рекрутування транскрипційних факторів та комплексів ремоделювання хроматину (SWI/SNF, NuRD). У контексті суперпідсилювачів, згідно з моделлю фазового розділення, BRD4 формує високонасичені транскрипційні конденсати, що забезпечує синхронізовану відповідь клітини на зовнішні стимули [16].

На відміну від довгої форми, коротка ізоформа (BRD4S) позбавлена НАТ-активності та PID-домену, проте відіграє унікальну роль у підтриманні цілісності геному. BRD4S рекрутується до місць дволанцюгових розривів ДНК (DSB), де бере участь у формуванні репараційних комплексів [17]. Крім того, вона може діяти як хроматиновий інсулятор, залучаючи комплекс конденсину II для обмеження доступу репараційних білків до певних ділянок, регулюючи таким чином відповідь на пошкодження ДНК.

Динаміка та стабільність BRD4 жорстко контролюються посттрансляційними модифікаціями (PTMs). Фосфорилування кластера NPS кіназою CK2, метилювання доменом SETD6 та убіквітінування білком SPOP модулюють його здатність зв'язуватися з хроматином та визначають час напівжиття білка в клітині [12]. Таким чином, BRD4 є багатофункціональним архітектором епігеному, що поєднує функції «читача», кінази, НАТ та репараційного фактора.

1.1.4. Участь BRD4 у молекулярних механізмах розвитку онкологічних і запальних захворювань

Порушення регуляції активності BRD4 розглядається як один із важливих молекулярних механізмів розвитку багатьох патологічних станів. Це зумовлено його ключовою роллю у контролі транскрипції генів, пов'язаних із проліферацією клітин, запальною відповіддю та ремоделюванням тканин. Завдяки здатності розпізнавати ацетильовані залишки лізину на хвостах гістонів і рекрутувати транскрипційні комплекси, цей білок виступає центральним регулятором активності численних генетичних програм.

В онкологічних процесах особливе значення має участь BRD4 у регуляції так званих надпідсилювачів — великих кластерів регуляторних елементів хроматину, які контролюють експресію ключових генів клітинної ідентичності та проліферації. У багатьох пухлинах, зокрема при гострому мієлоїдному лейкозі [18], множинній мієломі [19] та аденокарциномі підшлункової залози [20], відбувається інтенсивне накопичення BRD4 у цих регуляторних ділянках, що призводить до гіперактивації транскрипції онкогенів, таких як *c-MYC*, *BCL2* та *CCND1*, що підтримують неконтрольовану проліферацію та стійкість до апоптозу [21].

BRD4 відіграє важливу роль у регуляції запальних реакцій. BRD4 є ключовим коактиватором транскрипційного фактора NF- κ B, взаємодіючи з ацетильованою субодиницею RelA та сприяючи активації генів прозапальних медіаторів [22, 23]. У результаті стимулюється синтез прозапальних цитокінів (IL-1 β , IL-6, TNF- α), що відіграють ключову роль у розвитку системного запалення, аутоімунних процесів та гіперзапальних станів. Гіперактивація BRD4 призводить до стійкої активації NF- κ B-залежних транскрипційних програм і лежить в основі патогенезу сепсису, «цитокінового шторму» та хронічних аутоімунних захворювань.

Останні дослідження також демонструють участь BRD4 у розвитку фіброзу різних органів. Зокрема, RNAi-скринінг виявив, що BRD4 є обов'язковим для активації гепатоцитарних зірчастих клітин у відповідь на TGF- β 1, а фармакологічне пригнічення BRD4 інгібіторами JQ1, I-BET та PFI-1 значно знижує експресію близько 30 генів, асоційованих із фіброгенезом, включаючи α -SMA, колаген I та фібронектин [24]. Надмірна активація відповідних генетичних програм призводить до посиленого синтезу компонентів позаклітинного матриксу (α -SMA, колаген I, фібронектин), що спричиняє прогресуюче рубцювання тканин і формування фіброзу легень, печінки та нирок [25].

Крім того, багато ДНК-вмісних вірусів використовують BRD4 як «якір» для прикріплення вірусного геному до хроматину господаря. Зокрема, вірус папіломи людини (HPV) через білок E2 тетерує свій геном до хроматину

клітини господаря у комплексі з BRD4, що є необхідним для реплікації та сегрегації вірусної ДНК [26]. Цей механізм є критично важливим для стабільної реплікації, сегрегації та підтримання вірусної ДНК у дочірніх клітинах.

1.2. Сучасні підходи до інгібування BRD4

1.2.1. Механізми дії малих молекул-інгібіторів

Стратегічну основу фармакологічного інгібування BRD4 становить конкурентне витіснення ацетильованих лізинових залишків з гідрофобної кишені бромодоменів BD1 та BD2. Ліганди BET-інгібіторів містять гетероциклічний фрагмент — міметик ацетил-лізину (Kac-mimic), — який імітує геометрію та характер водневих зв'язків нативного субстрату. Ключовим контактом є водневий зв'язок між карбонільною групою міметика та консервативним залишком аспарагіну (Asn140 у BD1 BRD4), а також гідрофобні взаємодії у WPF-шельфі (Trp81-Pro82-Phe83) [9]. Така архітектура забезпечує конкурентне витіснення ацетильованого хроматину, вивільнення BRD4 із суперпідсилювачів та транскрипційну репресію онкогенів на кшталт c-MYC [21].

Концепція «BD-селективного» інгібування відображає функціональну відмінність між BD1 і BD2: BD2-селективний інгібітор ABBV-744 у доклінічних моделях демонструє виражений протипухлинний ефект при суттєво кращому профілі токсичності порівняно з пан-BET-інгібіторами [4].

1.2.2. Основні представники BET-інгібіторів

Піонером серед BET-інгібіторів є JQ1 — тієно-тріазоло-1,4-діазепін, що зв'язується з BD1 та BD2 BRD4 зі значеннями K_d близько 50 нМ та 90 нМ відповідно і демонструє виражену антипроліферативну активність у моделях лейкозу, лімфоми та мієломи [9]. Серед клінічно досліджуваних BET-інгібіторів виділяють тріазолодіазепінові похідні (OTX015/МК-8628, ABBV-075/mivebresib), діазабензімідазольні сполуки (I-BET762/GSK525762)

та ізоксазолні класи [27]. У клінічних дослідженнях першої фази GSK525762 при гематологічних злоякісних пухлинах лише частина пацієнтів досягла часткової відповіді, а дозолімітуючою токсичністю була тромбоцитопенія. Паралельно вивчаються гетеробіфункціональні деградери на основі технології PROTAC, зокрема dBET6 та ARV-825, що показують більш глибоке та тривале пригнічення BRD4 [28].

1.2.3. Обмеження традиційних інгібіторів: селективність, токсичність та розвиток резистентності

Попри значний терапевтичний потенціал, клінічне застосування класичних BET-інгібіторів обмежене низкою фундаментальних проблем. По-перше, більшість пан-BET-інгібіторів не розрізняють BRD2, BRD3 та BRD4, що спричиняє позамішеневу токсичність [29]. По-друге, BET-інгібітори реалізують лише «окупаційно-керований» механізм дії: їхній ефект є зворотним і зникає після дисоціації молекули від мішені [30].

Ключовою перешкодою для клінічного впровадження є швидкий розвиток набутої резистентності. Серед описаних механізмів — компенсаторна надекспресія BRD2, яка заміщає функцію BRD4 на хроматині при тривалому лікуванні JQ1 та іншими BET-інгібіторами в широкому панелі пухлин [31]. Ще одним механізмом є адаптивне «перепрограмування кіному» (kinome reprogramming) — активація компенсаторних кіназних сигнальних мереж (RTK, PI3K, AKT, ERK), що долають пригнічення BRD4 [32]. Додатково в клітинних моделях потрійно-негативного раку молочної залози виявлено, що BET-резистентні клітини зберігають залежність від BRD4, який підтримує транскрипцію незалежно від бромодому — за рахунок гіперфосфорилювання та бромодоменнезалежного зв'язування з MED1 [33]. Сукупність цих обмежень зумовлює необхідність розробки принципово нових терапевтичних підходів — зокрема, технології PROTAC, яка забезпечує каталітичну деградацію BRD4.

1.3. Концепція PROTAC у сучасній фармакології

1.3.1. Принципи технології таргетної деградації білків

Технологія Proteolysis Targeting Chimeras (PROTAC) запропонована у 2001 році [34] як принципово новий підхід до терапії — не через блокування каталітичного сайту, а через примусову деградацію цільового білка засобами власної клітинної протеолітичної машинерії (Рис. 1.3).

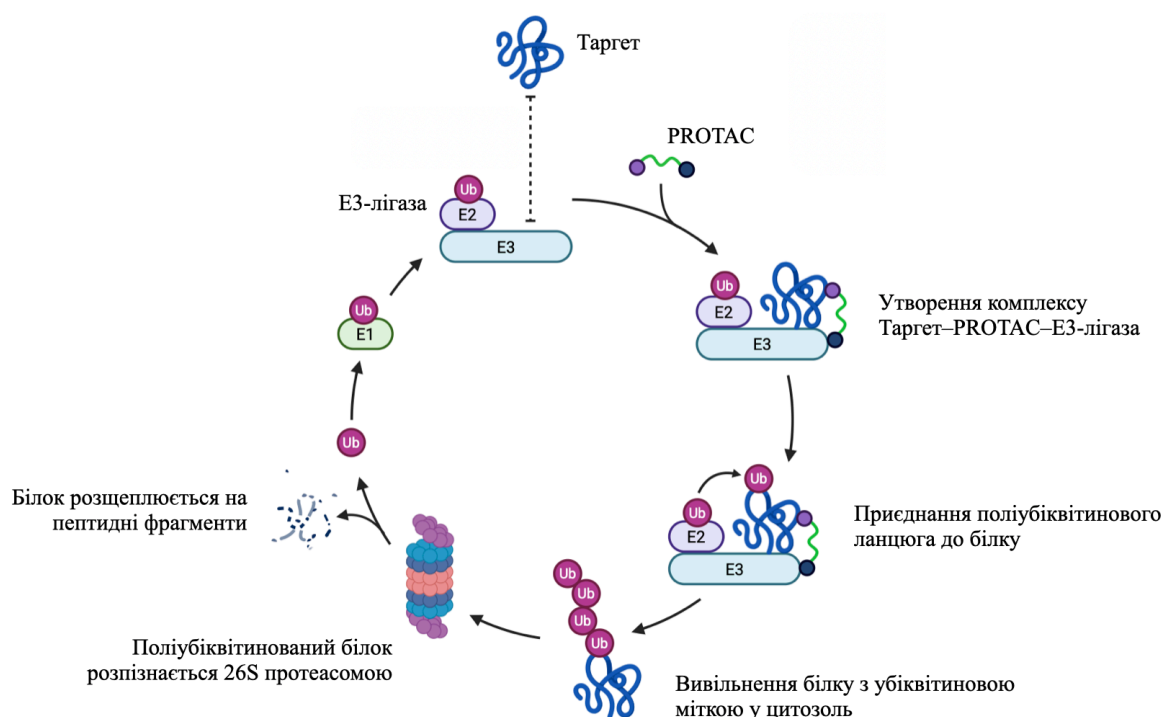


Рисунок 1.3. Механізм PROTAC-опосередкованої деградації білків: PROTAC індукує утворення комплексу між цільовим білком і E3-убіквітин-лігазою, що призводить до приєднання поліубіквітинового ланцюга, подальшого розпізнавання білка 26S протеасомою та його деградації до пептидних фрагментів. Схема створена за допомогою

BioRender (<https://www.biorender.com/>).

На відміну від класичних інгібіторів, що реалізують «окупаційно-керований» (occupancy-driven) механізм дії, PROTAC функціонують за «подійно-керованим» (event-driven) принципом: одна молекула здатна послідовно опосередковувати деградацію кількох молекул цільового білка,

оскільки після завершення убіквітинування PROTAC вивільняється і може повторно утворювати трикомпонентний комплекс [35]. Це відкриває можливість досягнення терапевтичного ефекту при субстехіометричних концентраціях, а також дозволяє таргетувати раніше важкодоступні для ліків (undruggable) мішені, що не мають зручного каталітичного сайту.

1.3.2. Формування та стабілізація трикомпонентного комплексу POI–PROTAC–E3-лігаза

Центральною молекулярною подією механізму дії PROTAC є формування трикомпонентного комплексу між цільовим білком (POI), молекулою PROTAC та E3-убіквітин-лігазою. Ефективність цього процесу визначається не лише афінністю кожного ліганду до відповідного білкового партнера, але й характером білок-білкових взаємодій (PPI) між POI та E3-лігазою в межах комплексу — так званим коефіцієнтом кооперативності (cooperativity factor, α), визначеним як $\alpha = K_d \text{ (бінарний)} / K_d \text{ (тернарний)}$ [36]. Кристалографічне дослідження трикомпонентного комплексу VHL–MZ1–BRD4BD2 показало, що молекула PROTAC складається «усередину» себе, утворюючи специфічні міжмолекулярні контакти між BRD4 і VHL та забезпечуючи виражену позитивну кооперативність [37].

Важливим феноменом є також «hook effect» — парадоксальне зниження ефективності деградації при надвисоких концентраціях PROTAC, коли молекули займають обидва білки окремо, не утворюючи продуктивного трикомпонентного комплексу. Математичний опис цього явища на основі рівноваг трьохкомпонентного зв'язування вперше деталізовано у роботі Douglass et al. [38].

1.3.3. Порівняльна характеристика PROTAC та класичних інгібіторів

Принципова перевага PROTAC над класичними інгібіторами полягає у кількох ключових аспектах, систематизованих у Таблиці 1.1.

Таблиця 1.1. Порівняльна характеристика механізмів дії, стехіометрії та біологічних ефектів класичних інгібіторів і PROTAC-деградерів

Характеристика	Класичні інгібітори	PROTAC-деградери
Механізм дії	Окупаційно-керований (occupancy-driven)	Подійно-керований (event-driven)
Стехіометрія	1:1 (стехіометрична)	Каталітична (субстехіометрична)
Вплив на білок	Тимчасове блокування функції	Повне усунення білка з клітини
Вимоги до сайту	Наявність активного/глибокого кишені	Будь-який сайт зв'язування
Ризик резистентності	Вищий (мутації сайту, надекспресія)	Нижчий (за рахунок деградації)

Каталітична природа механізму дії уможливорює повне усунення цільового білка, а відсутність вимоги до функціонального (каталітичного) сайту значно розширює простір потенційних мішеней — включно з транскрипційними факторами та адапторними білками [35]. Водночас висока молекулярна маса PROTAC (>700–900 Да, «поза правилом Ліпінського», bRo5) ускладнює пероральну біодоступність та проникність через клітинні мембрани. Попри це, накопичені клінічні дані демонструють принципову можливість подолання цих обмежень. Зокрема, кандидати ARV-110 (спрямований на андрогенний рецептор) та ARV-471 (спрямований на естрогенний рецептор) продемонстрували задовільні фармакокінетичні профілі, включаючи пероральну активність і клінічно значущу деградацію мішеней [39].

1.4. Структурна організація PROTAC-молекул

1.4.1. Ліганд, спрямований на білок-мішень

Ліганд, спрямований на білок-мішень (warhead, або POI-ліганд), визначає мішеневу специфічність PROTAC-молекули (Рис. 1.4). На відміну

від класичних інгібіторів, POI-ліганд не обов'язково повинен повністю інгібувати функцію цільового білка: достатнім є забезпечення афінності, необхідної для ініціювання та стабілізації трикомпонентного комплексу [40].

У PROTAC, спрямованих на BRD4, найпоширенішими warhead-лігандами є JQ1 та його похідні [29, 41, 42]. Важливим параметром дизайну є вибір точки приєднання лінкера (linker attachment point, LAP) до POI-ліганда, оскільки вона визначає просторову організацію трикомпонентного комплексу та істотно впливає на ефективність і селективність деградації [37].

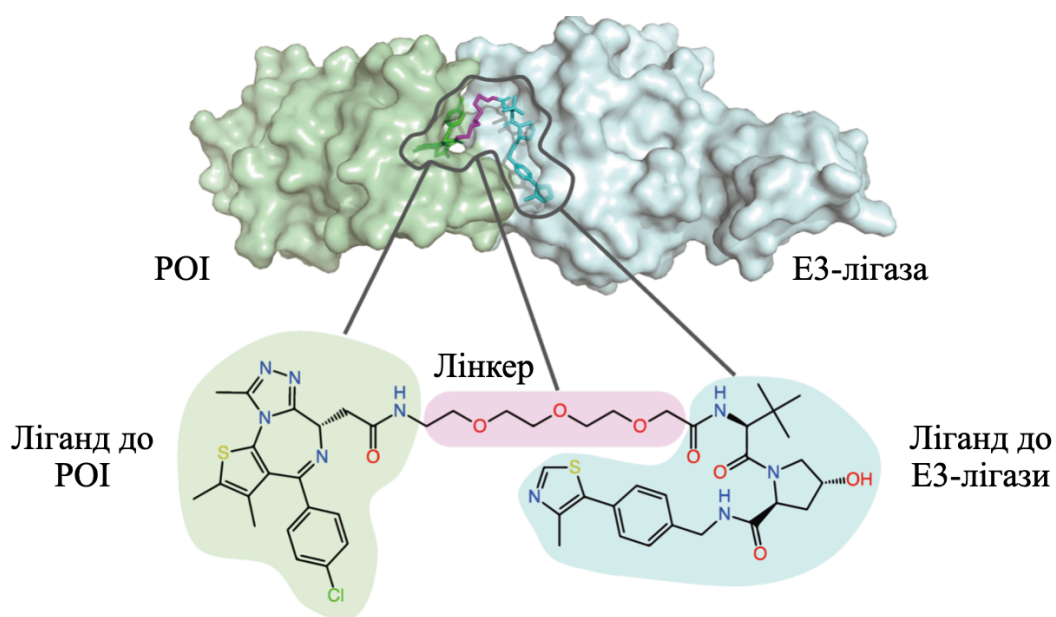


Рисунок 1.4. Структурна організація PROTAC-сполуки та утворення трикомпонентного комплексу: молекула PROTAC складається з ліганду до цільового білка (POI), ліганду до E3-убіквітин-лігази та лінкера, що з'єднує їх і забезпечує зближення білків для формування тернарного комплексу [43].

1.4.2. Ліганд для рекрутування E3-убіквітин-лігази

E3-ліганд забезпечує зв'язування PROTAC-молекули із субстрат-розпізнавальним компонентом E3-убіквітин-лігазного комплексу, тим самим визначаючи залучення конкретної лігази до процесу убіквітинування (Рис. 1.4). Серед найуживаніших E3-лігандів у BRD4-направлених PROTAC — похідні таліломіду (для CRBN, зокрема у ARV-825) та VHL-ліганд VHL-1 (для VHL, наприклад у MZ1) [29, 44]. Навіть незначні структурні модифікації

ЕЗ-ліганда можуть суттєво змінювати геометрію трикомпонентного комплексу та характер кооперативної взаємодії між його компонентами. У зв'язку з цим розробка нових лігандів до ширшого спектра ЕЗ-лігаз є одним із ключових напрямів сучасних досліджень [45, 46].

1.4.3. Лінкер як ключовий структурний елемент PROTAC та його вплив на ефективність деградації

Лінкер з'єднує РОІ-ліганд та ЕЗ-ліганд, визначаючи просторову організацію трикомпонентного комплексу, а також істотно впливаючи на фізико-хімічні властивості та фармакокінетичний профіль молекули (Рис. 1.4). На відміну від лігандних фрагментів, структура яких обмежена вимогами до специфічного зв'язування з мішенями, лінкер забезпечує значно ширший простір для раціональної оптимізації [47].

До ключових параметрів лінкера належать його довжина, хімічний склад та конформаційна жорсткість. Недостатня довжина лінкера може призводити до стеричних перешкод між РОІ та ЕЗ-лігазою, тоді як надмірна довжина зумовлює підвищену конформаційну рухомість і, як наслідок, зниження стабільності трикомпонентного комплексу. Типова ефективна довжина лінкера становить 3–20 атомів у ланцюзі [36].

Гнучкі алкільні та PEG-лінкери широко застосовувалися у ранніх поколіннях PROTAC, проте характеризуються підвищеною схильністю до окиснювального метаболізму та/або надмірною ліпофільністю. Натомість жорсткі лінкери, що містять гетероциклічні фрагменти (піперазин, піперидин, триазол, бензімідазол) або алкінові зв'язки, обмежують конформаційний простір молекули, потенційно сприяючи підвищенню кооперативності трикомпонентного комплексу та його метаболічної стабільності [47].

Перехід до використання більш жорстких лінкерів став одним із ключових чинників клінічної оптимізації PROTAC. Зокрема, у молекулі ARV-110 початковий гнучкий лінкер було замінено на піперидин-піперазинову конструкцію, що призвело до підвищення метаболічної стабільності та біологічної активності у клінічних дослідженнях [48].

Важливим аспектом є контроль топологічної полярної поверхні (topological polar surface area, TPSA), оскільки її високі значення ($>200 \text{ \AA}^2$) суттєво обмежують пасивну клітинну проникність PROTAC. У цьому контексті оптимізація лінкера — зокрема шляхом використання амідних ізостерів [49], метилювання [50] або циклізації [51] — є ефективною стратегією зниження TPSA без втрати біологічної активності.

Таким чином, лінкер слід розглядати не лише як структурний елемент, що поєднує функціональні фрагменти молекули, а як ключовий інструмент оптимізації, який визначає ефективність деградації, фармакокінетичні характеристики та загальний терапевтичний потенціал PROTAC [52].

1.5. Роль E3-убіквітин-лігаз у механізмі дії PROTAC

1.5.1. Убіквітин-протеасомна система як інструмент деградації

Убіквітин-протеасомна система (УПС) становить центральний механізм контрольованої внутрішньоклітинної деградації білків, що забезпечує підтримання протеостазу та регуляцію широкого спектра клітинних процесів. Деградація білків у межах УПС ініціюється ковалентним приєднанням убіквітину до ϵ -аміногрупи лізинових залишків субстратного білка через послідовний ферментативний каскад за участю E1-, E2- та E3-ферментів. При цьому саме E3-убіквітин-лігази відіграють вирішальну роль у забезпеченні субстратної специфічності, детермінуючи селективність усього процесу [53].

Формування поліубіквітинових ланцюгів, зокрема через Lys48-зв'язані кон'югати, слугує канонічним сигналом для протеасомної деградації. Протеасома 26S здійснює розпізнавання, розгортання та протеолітичне розщеплення мічених субстратів: 19S регуляторна частинка забезпечує селекцію та АТР-залежну транслокацію субстрату, тоді як 20S каталітичне ядро здійснює його гідроліз до олігопептидів [54].

У контексті PROTAC-технології зазначений ендогенний механізм експлуатується для індукції цілеспрямованої деградації білків. PROTAC-молекули функціонують як біфункціональні хімічні адаптери, що індукують

просторове зближення E3-лігази та білка-мішені (protein of interest, POI), тим самим спрямовуючи убіквітінування на нові, неендогенні субстрати (neo-substrates) [36]. Утворення трикомпонентного комплексу ініціює процес поліубіквітінування POI шляхом послідовного перенесення убіквітіну на доступні лізиніві залишки білка.

Досягнення ефективної деградації, як правило, потребує формування поліубіквітинового ланцюга критичної довжини (≥ 4 убіквітинових мономерів), після чого субстрат піддається транслокації до 20S ядра та протеолітичному розщепленню [54]. Принципово важливою характеристикою PROTAC є їх каталітичний механізм дії: після деградації POI молекула деградера дисоціює з комплексу та зберігає здатність до повторного залучення у наступні цикли убіквітінування, що зумовлює їх високу функціональну ефективність.

1.5.2. Основні типи E3-убіквітин-лігаз, що використовуються у PROTAC-технології

Попри те, що геном людини кодує понад 600 E3-убіквітин-лігаз, лише обмежений їх піднабір є придатним для використання у PROTAC-дизайні. Ключовим обмежувальним чинником є наявність високоафінних маломолекулярних лігандів, здатних забезпечити ефективне та селективне рекрутування відповідної лігази [55].

Найбільш валідованими E3-рекрутерами залишаються церебелон (CRBN) та білок фон Гіппеля–Ліндау (VHL). Вирішальним етапом у розвитку PROTAC-підходу стало відкриття лігандів до цих лігаз у 2012–2014 роках, включаючи IMiD-похідні (таліломід, леналідомід, помалідомід) для CRBN та синтетичні міметики HIF-1 α -пептиду для VHL, для яких було встановлено детальні структурні основи взаємодії на атомному рівні [56].

Паралельно інтенсивно досліджуються альтернативні E3-лігази, зокрема MDM2, cIAP, а також нещодавно ідентифіковані RNF4, RNF114, DCAF16 та DCAF11 [57, 58]. Частина з них активується за ковалентним механізмом

(наприклад, RNF114 у присутності німболіду), тоді як інші були ідентифіковані за допомогою хемопротеомних стратегій [59].

Вибір E3-лігази є визначальним фактором ефективності, селективності та контекст-залежної активності PROTAC. Зокрема, CRBN-рекрутуючі деградери демонструють підвищену ефективність у гематологічних злоякісних новоутвореннях, тоді як VHL-базовані системи характеризуються кращою продуктивністю у солідних пухлинах [60]. Таким чином, раціональний підбір пари «E3-лігаза — білок-мішень» є критичною передумовою оптимізації біологічної активності PROTAC.

Використання CRBN у поєднанні з BRD4 є обґрунтованим завдяки формуванню стабільних трикомпонентних комплексів з білками родини BET, що забезпечують ефективну поліубіквітинізацію та деградацію мішені. Крім того, CRBN характеризується сприятливим профілем експресії в клітинах гематологічного походження, де BRD4 відіграє ключову онкогенну роль, що додатково підсилює ефективність відповідних PROTAC-конструкцій.

1.6. Обчислювальні підходи до проєктування PROTAC: сучасний стан і прогалини

1.6.1. Структурно-орієнтоване моделювання трикомпонентних комплексів POI–PROTAC–E3-лігаза

Фундаментальною передумовою раціонального дизайну PROTAC є здатність передбачити тривимірну геометрію трикомпонентного комплексу POI–PROTAC–E3-лігаза. Піонерською у цьому напрямі стала платформа PRosettaC, розроблена Zaidman et al. [61]. Протокол чергує семплювання простору білок-білкових взаємодій і конформаційного простору лінкера: спочатку здійснюється глобальний докінг з PatchDock з геометричними обмеженнями на відстань між якірними атомами лігандів мішені та E3-лігази, потім — локальна оптимізація з RosettaDock та відбір лінкерних конформерів методом Rosetta Packer. На бенчмарку кристалічних структур PRosettaC продемонстрував атомну точність передбачення білкових ланцюгів і

лігандних поз, підтвердивши здатність структурно обґрунтовувати активність PROTAC для системи BTK–CRBN. Обмеженням методу є висока обчислювальна вартість і чутливість до якості вхідних структур.

Наступним кроком у точності структурного передбачення стала робота Ignatov et al. [62], які запропонували метод генерації ансамблів трикомпонентних комплексів на основі «мовчазного» конволюційного члена у функції докінгу. Ключовою інновацією є кластеризація моделей для кількох PROTAC, спрямованих на один POI, де найбільший консенсусний кластер відповідає нативно-подібній геометрії. Ансамбль моделей також дозволяє оцінити кінетику дисоціації та кооперативність тернарного комплексу, що корелюють із деградуючою активністю PROTAC.

1.6.2. Прогнозування біологічної активності PROTAC методами машинного навчання

Паралельно з методами структурного моделювання активно розвивається лігандно-орієнтований напрям — побудова предиктивних моделей класифікації деградуючої активності. Першою систематичною спробою стала робота Nori et al. [63], де для класифікації PROTAC використовували XGBoost із молекулярними відбитками Моргана. Хоча модель продемонструвала прийнятну базову точність, вона не враховувала білкового контексту системи POI–E3 і використовувала лише хімічні дескриптори молекули.

Принципово новий рівень архітектурної складності запровадив DeepPROTACs — глибока нейромережева платформа, яка інтегрує графові конволюційні мережі (GCN) для кодування структур зв'язуючих кишень POI та E3-лігази та двонаправлені LSTM-шари для кодування SMILES-рядка лінкера [64]. Навчаючись на 2832 анотованих PROTAC з бази PROTAC-DB, модель досягла точності 77.95% та AUROC = 0.847 на тестовій вибірці. Ключовою перевагою є явне врахування структурної інформації про обидва білки — на відміну від чисто хімічних підходів.

Ribes et al. [65] провели систематичне порівняння XGBoost та глибоких нейромереж на трьох схемах розбиття даних — випадковій, за мішенню та за хімічною подібністю. Центральний висновок роботи: при стандартному випадковому розбитті XGBoost досягає точності до 78.6%, порівнянної з нейромережами; проте при суворіших схемах розбиття (за мішенню) глибокі моделі виявляють вищу узагальнюючу здатність. Це свідчить про обмежену транслятивність простих ML-підходів за межі тренувального хімічного простору.

1.6.3. Генеративні підходи до *de novo* дизайну лінкерів PROTAC

Поряд із предиктивними моделями активно розробляються генеративні архітектури для *de novo* дизайну лінкерів. Zheng et al. [66] запропонували PROTAC-RL — трансформерну генеративну модель, попередньо навчену на 294 675 квазі-PROTAC молекулах із простору ZINC, з подальшим тонким налаштуванням на реальних PROTAC та RL-оптимізацією за фармакокінетичними компонентами функції винагороди. Застосована до мішені BRD4, система PROTAC-RL згенерувала 5000 кандидатів, з яких шість було синтезовано і три підтверджено в клітинних тестах на деградацію. Один провідний кандидат продемонстрував сприятливий фармакокінетичний профіль *in vivo* у мишей — що є рідкісним прикладом верифікації обчислювального PROTAC-дизайну в живому організмі. Фундаментальним обмеженням PROTAC-RL є генерація лише у просторі SMILES без урахування тривимірної конформації трикомпонентного комплексу.

Саме цей недолік долає PROTAC-INVENT — 3D-генеративна модель Li, Ran та Chen [67], яка генерує SMILES лінкера та одночасно будує тривимірну позу PROTAC у зв'язуючому сайті трикомпонентного комплексу. RL-оптимізація спрямовує генерацію до структур із кращими оцінками докінгу у системі POI–ЕЗ-лігаза, що дозволяє безпосередньо враховувати просторові вимоги трикомпонентного комплексу на етапі генерації — а не лише постфактум. Проте висока обчислювальна вартість 3D-семплювання обмежує практичну масштабованість методу.

1.6.4. Позиціонування цієї роботи у контексті існуючих підходів

Узагальнюючи огляд, можна виділити три ключові прогалини в існуючих обчислювальних підходах до PROTAC-дизайну, які адресує ця робота.

По-перше, більшість предиктивних моделей — як DeepPROTACs [62], так і підходи на основі XGBoost [63, 65] — використовують гомогенні простори ознак і один алгоритм машинного навчання. У даній роботі реалізовано гетерогенну ансамблевую архітектуру зі стекуванням, де кожен базовий класифікатор навчається на принципово різному поданні даних: фармакофорних дескрипторах, відбитках Моргана, фізико-хімічних параметрах RDKit, ембедингах ChemBERTa та білкових ембедингах ESM-2 з LoRA-адаптером. Це дозволяє зменшити кореляцію між помилками окремих алгоритмів і підвищити узагальнюючу здатність ансамблю.

По-друге, у роботі PROTAC-RL [66] власна модель активності використовується лише як постгенераційний фільтр. У даній роботі кастомна ансамблева модель є безпосереднім компонентом функції винагороди RL-агента у середовищі REINVENT4, що забезпечує цілеспрямовану оптимізацію хімічного простору лінкерів за критерієм прогнозованої деградуючої активності вже у процесі генерації.

По-третє, 3D-підходи на кшталт PROTAC-INVENT [67] хоча і враховують геометрію трикомпонентного комплексу в циклі генерації, мають обмежену масштабованість через ресурсоємну 3D-генерацію. У запропонованій методології структурно-орієнтована валідація (молекулярний докінг у MolSoft ICM-Pro) застосовується як фінальний етап відбору топ кандидатів — після попередньої фільтрації предиктивною моделлю, що є обчислювально обґрунтованим компромісом між точністю та продуктивністю.

РОЗДІЛ II

МАТЕРІАЛИ І МЕТОДИ

2.1. Модель прогнозування біологічної активності сполук PROTAC

Для прогнозування біологічної активності молекул PROTAC розроблено гетерогенну ансамблеву модель на основі методу стекування. Архітектура моделі ґрунтується на ансамблевому підході з використанням різних представлень даних, де кожен базовий алгоритм оперує власним піднабором хімічних або біологічних дескрипторів. Такий підхід сприяє зниженню кореляції між помилками окремих моделей та підвищує узагальнюючу здатність ансамблю. Повний код реалізації доступний у відкритому репозиторії: <https://github.com/VaryaMelnik/Ensemble-Model-for-PROTAC-Activity-Prediction.git>.

2.1.1. Формування набору даних та анотація біологічної активності

Первинний масив даних був сформований на основі спеціалізованої бази даних PROTAC-DB [68]. Початкова вибірка, що налічувала 9381 сполуку, пройшла етап стандартизації та попередньої обробки, включаючи видалення дублікатів і перевірку повноти інформації про біологічну активність. У результаті очищення фінальний набір даних склав 6087 унікальних об'єктів.

Для анотації біологічної активності було застосовано ієрархічну трирівневу систему класифікації. На I рівні (функціональна деградація) сполука вважалася активною за умови $DC_{50} \leq 1000$ нМ або $D_{max} \geq 70$ %. За відсутності таких даних використовувався II рівень (стабільність трикомпонентного комплексу), де критеріями активності були $K_d \leq 500$ нМ, $t_{1/2} \geq 30$ с або $\Delta G \leq -8$ ккал/моль. Якщо ці параметри також були відсутні, застосовувався III рівень (афінність до мішеней), за яким сполука вважалася активною у разі експериментально підтвердженої афінності (EC_{50} , K_d або $K_i \leq 1000$ нМ) одночасно до білка-мішені та E3-лігази. Сполуки, що не відповідали зазначеним критеріям або не мали достатніх експериментальних даних, класифікувалися як неактивні.

Для побудови та оцінювання моделей набір даних, що включає 6087 сполук (з них 1407 активних, 23.1 %), було поділено на тренувальну, валідаційну та тестову вибірки у співвідношенні 80 : 10 : 10 із застосуванням стратифікованого відбору за міткою активності, що забезпечує збереження початкового співвідношення активних і неактивних сполук у кожній підмножині. Зокрема, тренувальна вибірка містить 4869 сполук, з яких 1125 є активними (23.1 %), валідаційна — 609 сполук, з яких 141 активна (23.2 %), а тестова — також 609 сполук із 141 активною сполукою (23.2 %).

Тренувальна підмножина використовувалася для навчання базових моделей та побудови протоколу попередньої обробки даних, валідаційна — для налаштування мета-класифікатора та оптимізації порогу класифікації, тоді як тестова вибірка залишалася повністю незалежною і застосовувалася виключно для фінальної оцінки метрик та узагальнюючої здатності моделі.

2.1.2. Представлення вхідних даних

Вхідні дані представлені двома модальностями: структурними представленнями молекул у нотації SMILES (Simplified Molecular Input Line Entry System) та амінокислотними послідовностями білків — білка-мішені і ЕЗ-лігази.

Для кодування білкових послідовностей використано трансформерну модель ESM-2 (Evolutionary Scale Modeling) [69] конфігурації t33_650M_UR50D, доповнену адаптером QBind LoRA (Low-Rank Adaptation), спеціалізованим для моделювання взаємодій типу «білок–ліганд». Векторизація здійснювалася шляхом усереднення прихованих станів останнього шару моделі (mean pooling), у результаті чого для кожної послідовності формувався ембедінг розмірністю 1280 компонент.

Хімічні характеристики молекул описувалися кількома наборами дескрипторів:

- Фармакофорні відбитки — двовимірні фармакофорні ознаки, побудовані за методикою Гоббі [70];

- Молекулярні відбитки Моргана — циркулярні молекулярні відбитки з радіусом 3 та розмірністю 2048 біт [71];
- RDKit 2D дескриптори — набір приблизно з 200 фізико-хімічних параметрів (молекулярна маса, коефіцієнт розподілу LogP, полярна поверхня TPSA тощо) [72];
- ChemBERTa-2 — глибокі хімічні ембединги (chemical embeddings) розмірністю 768 компонент, отримані за допомогою трансформерної моделі, навченої в режимі багатозадачної регресії із застосуванням маскованого мовного моделювання [73].

2.1.3. Попередня обробка та зменшення розмірності ознак

Для кожного типу дескрипторів застосовано узгоджений протокол попередньої обробки. Числові (небінарні) ознаки піддавалися стандартизації за Z-критерієм, що забезпечувало їх приведення до нульового середнього та одиничної дисперсії. Зокрема, стандартизація застосовувалася до білкових ембедингів, RDKit 2D дескрипторів та хімічних ембедингів ChemBERTa.

Для фармакофорних дескрипторів додатково використовувався метод відсікання ознак за дисперсією з пороговим значенням $p = 0.005$, що дозволило усунути малоінформативні змінні з низькою варіабельністю.

З метою зменшення розмірності та зниження мультиколінеарності для білкових ембедингів (окремо для білка-мішені та ЕЗ-лігази) і хімічних ембедингів ChemBERTa застосовано метод головних компонент (Principal Component Analysis, PCA). Кількість компонент обиралася таким чином, щоб зберегти 99 % загальної дисперсії даних, після чого PCA-проекції білкових ембедингів конкатенували в єдине білкове представлення.

Важливо зазначити, що всі параметри попередньої обробки — включно зі статистиками стандартизації, порогом дисперсії та матрицями перетворення PCA — визначалися виключно на тренувальній вибірці. Надалі отримані параметри застосовувалися до валідаційних та тестових даних без повторного перенавчання, що забезпечує коректність оцінювання моделі та запобігає витоку інформації між підвибірками.

2.1.4. Архітектура ансамблевої моделі

Ансамбль першого рівня складається з чотирьох базових алгоритмів, кожен з яких працює з окремим представленням ознак, сформованим на основі комбінації хімічних дескрипторів та білкових ембедингів (Рис. 2.1). Такий підхід дозволяє інтегрувати різні представлення інформації та підвищити узагальнюючу здатність моделі.

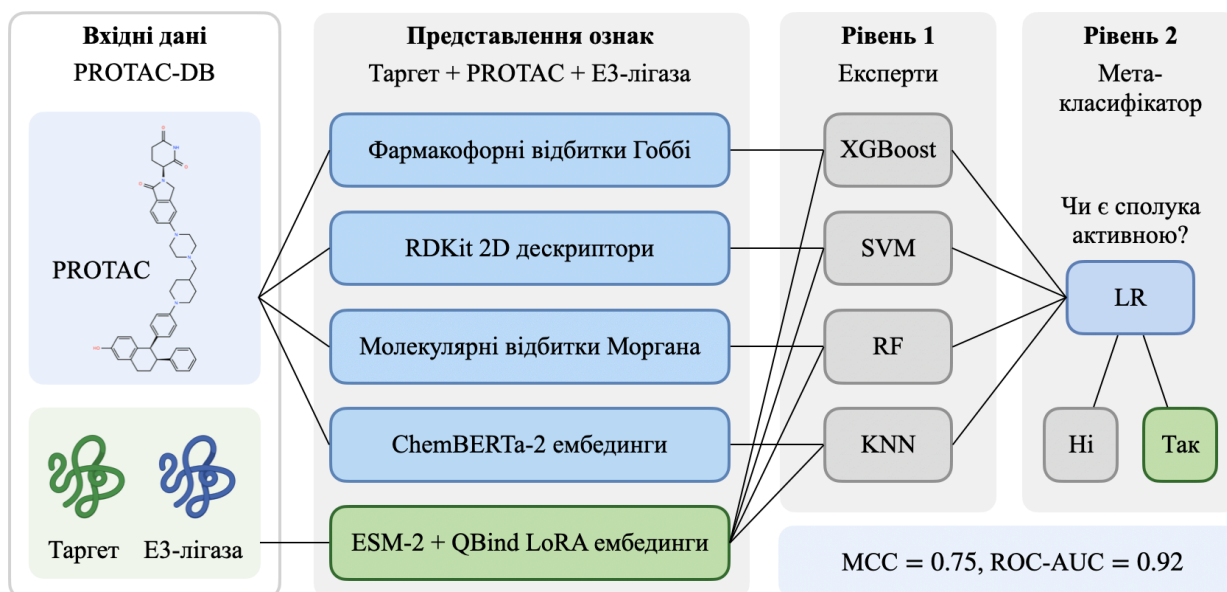


Рисунок 2.1. Архітектура ансамблевої моделі прогнозування біологічної активності PROTAC-молекул.

Усі базові моделі використовують спільне білкове представлення, сформоване шляхом конкатенації PCA-компонент ембедингів білка-мішені та E3-лігази. Попередньо кожен тип білкових ембедингів піддається стандартизації та незалежному зменшенню розмірності за допомогою методу головних компонент, після чого об'єднується в єдиний вектор ознак.

Експерт на основі екстремального градієнтного бустингу (Extreme Gradient Boosting, XGBoost) [74] навчається на поєднанні цього білкового представлення та фармакофорних дескрипторів, попередньо відібраних за дисперсією. Така конфігурація дозволяє моделі виявляти складні нелінійні взаємодії між фармакофорними характеристиками молекул і структурними особливостями білків.

Модель випадкового лісу (Random Forest) [75] використовує конкатенацію білкових ембедингів та стандартизованих відбитків Моргана, що забезпечує одночасне врахування локальних структурних фрагментів молекул і глобального білкового контексту.

Метод опорних векторів (Support Vector Machine, SVM) [76] працює у просторі, сформованому шляхом поєднання білкових ембедингів та стандартизованих фізико-хімічних дескрипторів RDKit 2D. Це дозволяє моделі ефективно враховувати як молекулярні властивості, так і білкові фактори, що впливають на біологічну активність.

Модель k-найближчих сусідів (k-Nearest Neighbors, KNN) [77] використовує об'єднаний простір ознак, що включає PCA-компоненти білкових ембедингів та зменшені за розмірністю (PCA) ембединги ChemBERTa. Пошук найближчих сусідів здійснюється у цьому просторі з використанням косинусної метрики подібності, що дозволяє ідентифікувати сполуки з подібними структурно-функціональними характеристиками.

2.1.5. Оптимізація гіперпараметрів

Оптимізація параметрів базових моделей здійснювалася із використанням фреймворку Optuna [78] на основі байєсівського підходу з алгоритмом деревоподібного оцінювача Парцена (Tree-structured Parzen Estimator, TPE). Цільовою метрикою оптимізації обрано коефіцієнт кореляції Меттьюза (Matthews Correlation Coefficient, MCC).

Модель XGBoost була налаштована на 951 ітерацію з швидкістю навчання 0.05 та максимальною глибиною дерев 4. Для зменшення ризику перенавчання застосовано стохастичну регуляризацію через випадковий піднабір рядків (`subsample = 0.76`) та ознак (`colsample_bytree = 0.33`).

Модель RF використовує 525 дерев із максимальною глибиною 17 та збалансовані ваги класів, що дозволяє коректно враховувати нерівномірний розподіл активних і неактивних сполук у наборі даних.

Модель KNN використовує 6 найближчих сусідів із косинусною метрикою подібності, причому ваги сусідів визначаються обернено

пропорційно до відстані, що підвищує чутливість моделі до локальних структурних аналогів.

Для SVM обрано радіально-базисну функцію ядра (Radial Basis Function, RBF) з параметрами ($\gamma \approx 0.006$) та ($C \approx 39.12$), що забезпечує моделювання нелінійної геометрії розділяючої гіперплощини.

2.1.6. Мета-класифікатор та схема навчання

Поєднання прогнозів базових моделей реалізовано за допомогою мета-класифікатора на основі логістичної регресії, який навчається на ймовірнісних виходах експертів першого рівня. Такий підхід відповідає класичній схемі стекування [79] і дозволяє інтегрувати інформацію з різних представлень ознак. Вибір логістичної регресії як мета-класифікатора зумовлений її здатністю ефективно інтерпретувати внесок кожного експерта через вагові коефіцієнти.

На етапі розробки архітектури було проведено порівняльне дослідження декількох стратегій агрегації:

- Жорстке голосування (hard voting): вибір класу за мажоритарним принципом;
- М'яке голосування (soft voting): пряме усереднення ймовірностей від усіх моделей;
- Мета-моделювання (meta-modeling): навчання алгоритму для зважування прогнозів.

Результати тестувань підтвердили перевагу мета-класифікатора, оскільки цей підхід краще оптимізує поріг класифікації та ефективніше зважує внески окремих алгоритмів порівняно з іншими підходами агрегації передбачень.

Навчання ансамблю здійснювалося із дотриманням протоколу розділення даних на тренувальну, валідаційну та тестову вибірки. Базові моделі навчалися на тренувальних даних, тоді як їхні прогнози на валідаційній вибірці використовувалися як вхідні ознаки для мета-

класифікатора. Така схема дозволяє уникнути витоку інформації між рівнями ансамблю та забезпечує коректне узагальнення моделі.

Оптимальний поріг прийняття рішення для фінальної бінарної класифікації визначався на основі валідаційної вибірки шляхом максимізації коефіцієнта кореляції Меттьюза (Matthews Correlation Coefficient, MCC). Це дозволяє враховувати баланс між класами та підвищує надійність моделі в умовах потенційного дисбалансу даних.

2.1.7. Абляційний аналіз

Для оцінювання внеску різних типів дескрипторів та їхніх комбінованих ефектів проведено розширений абляційний аналіз у форматі UpSet-дослідження. На відміну від традиційного підходу «пропустити один» (leave-one-out), цей метод передбачає тестування всіх можливих комбінацій моделей першого рівня ($2^n - 1$). Для кожної комбінації виконувалися повторне навчання мета-класифікатора та оптимізація порогу класифікації. Це дозволило оцінити відносний внесок різних модальностей даних та підтвердити доцільність використання гетерогенної ансамблевої архітектури.

2.1.8. Статистична валідація та оцінка невизначеності

Для перевірки відсутності перенавчання застосовано процедуру Y-рандомізації (Y-randomization, Y-scrambling) [80]. У межах цього підходу вектор цільових міток (Y_{train}) піддавався випадковому перемішуванню при незмінній матриці дескрипторів (X_{train}). Моделі, навчені на таких даних, не демонстрували прогностичної здатності, що перевищує рівень випадкового вгадування. Це підтверджує, що прогностична здатність моделі зумовлена реальними структурами зв'язку між дескрипторами та цільовою змінною.

Додатково для верифікації надійності прогнозів у методологію інтегровано аналіз епістемічної невизначеності. Вона оцінюється як стандартне відхилення прогнозів базових моделей ансамблю та відображає ступінь їх узгодженості.

2.2. Генерація та оптимізація PROTAC за допомогою REINVENT4 та Link-INVENT

2.2.1. Архітектура генеративної моделі та модуль Link-INVENT

Для дизайну нових молекул PROTAC у роботі використано фреймворк REINVENT4 — відкриту платформу для *de novo* дизайну, що базується на генеративних нейронних мережах та навчанні з підкріпленням (RL) [81].

Особливістю PROTAC є їхня біфункціональна природа: вони складаються з ліганду до білка-мішені та ліганду до E3-убіквітин-лігази, поєднаних хімічним містком — лінкером. Для вирішення задачі з'єднання двох фіксованих фрагментів було застосовано модуль Link-INVENT [82]. На відміну від генеративних підходів, що не передбачають використання початкових структурних шаблонів, Link-INVENT спеціалізується на фрагментній розробці сполук, де модель навчається пропонувати оптимальні лінкери, що забезпечують правильну просторову орієнтацію обох лігандів.

Генерація здійснюється авторегресивною мовною моделлю, яка базується на рекурентних нейронних мережах. Вона моделює ймовірність послідовності SMILES $x = (x_1, x_2, \dots, x_T)$ як добуток умовних імовірностей:

$$P_{\theta}(x) = \prod_{t=1}^T P_{\theta}(x_t | x_{<t}), \quad (2.1)$$

де θ — параметри моделі.

Початкова модель $P_{\theta}(x)$ попередньо натренована на великих масивах хімічних сполук, що гарантує генерацію синтаксично правильних та хімічно релевантних структур. Априорна модель для модуля Link-INVENT була попередньо натренована на наборі даних з бази ChEMBL.

2.2.2. Стратегія навчання з підкріпленням

Для спрямування генерації у бік сполук із заданими властивостями використовується навчання з підкріпленням, де модель діє як «агент», що максимізує очікувану винагороду (scoring function) $S(x)$:

$$\max_{\theta} \mathbb{E}_{x \sim P_{\theta}}[S(x)]. \quad (2.2)$$

У даній роботі застосовано алгоритм DAP (Difference between Augmented and Posterior). Його перевага — стабілізація навчання через обмеження відхилення від апіорного розподілу, що запобігає колапсу моделі в вузьку хімічну область. Функція втрат $\mathcal{L}(\theta)$ має вигляд:

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim P_{\theta}} \left[\left(\log P_{\theta}(x) - \log P_{\text{prior}}(x) - \sigma S(x) \right)^2 \right], \quad (2.3)$$

де $\sigma = 256$ — коефіцієнт інтенсивності навчання, що визначає ступінь впливу винагороди на оновлення ваг моделі.

Оптимізація проводилася в один етап, спрямований на підвищення прогнозованої біологічної активності генерованих PROTAC-молекул (Рис. 2.2). Функція винагороди $S_1(x)$ базувалася виключно на прогнозі гетерогенної ансамблевої моделі (див. Розділ 2.1):

$$S_1(x) = f_{\text{activity}}(x). \quad (2.4)$$

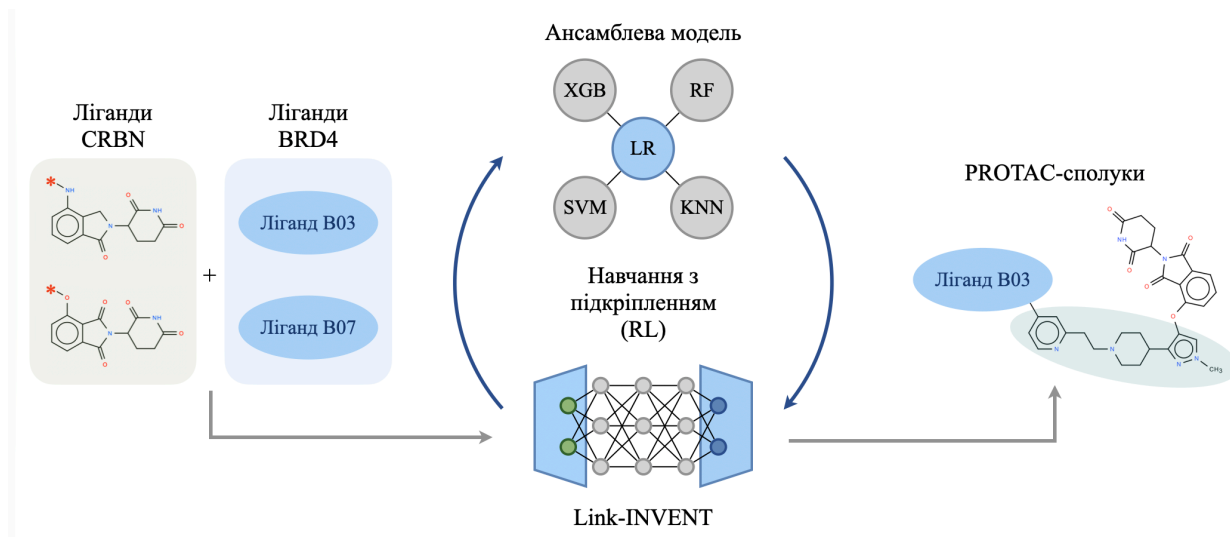


Рисунок 2.2. Схема методу *de novo* генерації лінкерів PROTAC із використанням навчання з підкріпленням.

Для забезпечення високої якості та структурного різноманіття згенерованих сполук було застосовано комплекс узгоджених технічних налаштувань. Зокрема, для підвищення стійкості та узагальнювальної

здатності моделі використано рандомізацію SMILES-послідовностей, що передбачало подачу різних еквівалентних неканонічних представлень одного й того самого хімічного запису під час навчання.

З метою уникнення генерації ідентичних структур та розширення простору пошуку було активовано обмеження на унікальність згенерованих послідовностей у межах кожного робочого батча розміром ($N = 128$).

Як параметри навчання було використано оптимізаційний алгоритм Adam із заданою швидкістю навчання ($\eta = 10^{-4}$), що забезпечувало стабільне та поступове оновлення параметрів моделі-агента.

Контроль збіжності здійснювався шляхом обмеження максимальної кількості ітерацій до 200. Таке обмеження є критично важливим для запобігання перенавчанню, тобто надмірному пристосуванню моделі до специфічних статистичних закономірностей функції винагороди. У результаті забезпечується збереження генеративної здатності моделі та адекватне покриття хімічного простору.

Усі використані гіперпараметри навчання, а також детальні налаштування генерації та оптимізації моделі наведено у конфігураційному файлі (Додаток 1). Оскільки функція винагороди базувалася на гетерогенній ансамблевій моделі власної розробки, для її використання в середовищі REINVENT4 було реалізовано спеціалізований програмний плагін, що забезпечує інтеграцію моделі в цикл навчання з підкріпленням та обчислення значення винагороди для згенерованих структур (Додаток 2).

2.2.3. Процедура генерації вибірки PROTAC

Після завершення RL-оптимізації генеративної моделі було згенеровано фінальну бібліотеку PROTAC-молекул з використанням оптимізованого агента та фіксованих лігандів (Рис. 2.2). Для забезпечення коректного статистичного порівняння аналогічно було проведено паралельну генерацію з апіорної моделі (Raw Prior), яка не зазнавала RL-оптимізації.

Генерація здійснювалося за допомогою фреймворку REINVENT4 ідентично для обох моделей. Для кожної з чотирьох пар лігандів генерувалося

по 200 молекул. Протокол генерації включав канонізацію SMILES, видалення дублікатів та генерацію без рандомізації атомного порядку. Конфігураційний файл наведено у Додатку 2.

Обидві вибірки використовувалися у подальшому статистичному порівнянні фізико-хімічних властивостей лінкерів (розділ 3.2.3) та аналізі різноманітності (розділ 3.2.2). Ідентичність лігандів є ключовою умовою коректного порівняння: усі відмінності у властивостях лінкерів і прогнозованій активності зумовлені виключно RL-оптимізацією.

2.2.4. Ліганди BRD4 та CRBN, які були використані у якості лігандних фрагментів генерації PROTAC-сполук

Для генерації PROTAC-сполук було обрано чотири пари лігандів, які представляють комбінації двох лігандів для BRD4 та двох лігандів для CRBN (Рис. 2.3). Вибір зазначених зв'язувальних фрагментів зумовлений їхньою здатністю до селективного зв'язування відповідних мішеней та підтвердженою ефективністю у складі PROTAC-систем. Леналідомід і талідомід є класичними лігандами E3-убіквітин-лігази CRBN і широко застосовуються для індукції деградації білків-мішеней. Водночас сполуки B03 та B07 були розроблені в ТОВ «Кемспейс» і представляють різні класи інгібіторів BRD4, що відрізняються за структурною організацією та фізико-хімічними властивостями.

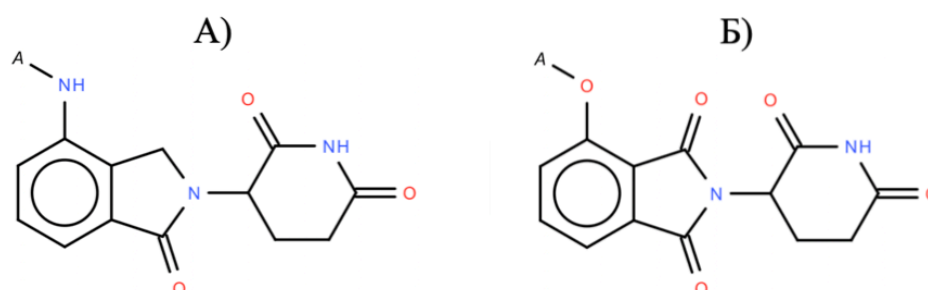


Рисунок 2.3. Ліганди CRBN, які були використані у якості лігандів для PROTAC: А) леналідомід; Б) талідомід. Літерою А позначено місця приєднання лінкера.

Комбінування цих лігандів у чотири пари дозволяє варіювати не лише геометрію потенційного трикомпонентного комплексу, але й просторові вимоги до лінкера, такі як довжина, гнучкість та топологія. Це є критично важливим, оскільки ефективність PROTAC значною мірою визначається здатністю лінкера забезпечувати оптимальну орієнтацію лігандів для індукції продуктивної білок-білкової взаємодії.

2.3. Розрахунок фізико-хімічних дескрипторів та прогнозування фармакокінетичного профілю

Для всебічної характеристики згенерованих бібліотек молекул PROTAC та їх подальшої фільтрації було проведено розрахунок набору молекулярних дескрипторів, що описують ADMET-профіль.

Розрахунок базових фізико-хімічних параметрів здійснювався за допомогою бібліотеки хімічної інформатики з відкритим вихідним кодом RDKit (версії 2023.09.5) для мови програмування Python. Ці параметри використовувалися для оцінки відповідності сполук правилам «Beyond Rule of Five» (bRo5).

Для детального прогнозування фармакокінетичних характеристик та профілю безпеки використовувався інструмент ADMET-AI [83]. Ця платформа базується на архітектурі графів нейронних мереж (ChemProp) та забезпечує високу точність прогнозування параметрів.

2.4. Структурно-орієнтована оцінка PROTAC-кандидатів методом молекулярного докінгу

Для структурно-орієнтованої валідації згенерованих PROTAC-молекул було застосовано молекулярний докінг з використанням програмного забезпечення ICM-Pro (Molsoft LLC). Докінг проводився для 100 найбільш перспективних кандидатів, відібраних з RL-оптимізованої вибірки за найвищим значенням прогнозованої ймовірності біологічної активності.

2.4.1. Відбір сполук для докінгу

Із фінальної RL-оптимізованої вибірки ($n = 766$) було відібрано 100 молекул із найвищою прогнозованою ймовірністю біологічної активності за результатами ансамблевого класифікатора. Пороговий рівень ймовірності для входження до топ-100 становив ≥ 0.75 . Обраний підхід зосереджує обчислювальні ресурси на найбільш перспективних структурах і зменшує навантаження при ресурсоємному докінгу PROTAC, який є значно складнішим за докінг малих молекул через моделювання двох білок-лігандних взаємодій.

2.4.2. Підготовка структур білків та протокол докінгу

Для CRBN використовувалися кристалографічні структури у комплексі з талідомідом (PDB ID: 4CI1) та леналідомідом (PDB ID: 4TZ4); для BRD4 — структури з лігандами B03 та B07. Перед докінгом структури BRD4 та CRBN були суміщені в єдину систему на основі відомої геометрії трикомпонентного комплексу, що дозволило моделювати одночасне зв'язування PROTAC з обома білками. Усі структури підготовлені в ICM-Pro та конвертовані в ICM-об'єкти; доданий водень при pH 7.4 та присвоєні заряди за силовим полем AMBER. Ко-кристалізовані ліганди слугували шаблонами для визначення боксу докінгу та стартових поз лігандів.

Кожну сполуку докували незалежно п'ять разів ($\text{scsRad} = 8.0$, $\text{effort} = 1.0$, $\text{nrun} = 5.0$). Лінкер та бічні ланцюги залишків у радіусі 4 Å від лігандів розглядалися як гнучкі. Середню оцінку топ-5 конформацій (за п'ятьма запусками) використовували для фінального ранжування.

2.4.3. Критерії відбору та аналіз результатів

Оцінювання результатів докінгу проводили за комплексом кількісних і структурних критеріїв, використовуючи середню оцінку п'яти найкращих конформацій, що враховує ван-дер-ваальсові, електростатичні та водневі взаємодії, а також ентропійні внески гнучкості лінкера.

Варіативність отриманих рішень оцінювалася за допомогою коефіцієнта варіації (CV), що визначається як відношення стандартного відхилення до середнього значення оцінки докінгу:

$$CV = \frac{\sigma}{\mu}, \quad (2.5)$$

де σ — стандартне відхилення, а μ — середнє значення оцінки докінгу. Менші значення CV свідчать про високу узгодженість та стабільність ансамблю згенерованих конформацій.

Збіжність докінг-оцінок оцінювалася як відхилення середнього значення від найкращої (мінімальної) оцінки серед розглянутих конформацій:

$$\Delta = \mu - E_{\min}, \quad (2.6)$$

де μ — середнє значення оцінки докінгу, а $E_{\min}$ — найнижча (найкраща) оцінка. Менші значення Δ вказують на кращу збіжність, тобто концентрацію рішень поблизу енергетичного мінімуму.

Додатково аналізували відтворюваність пози зв'язування між незалежними запусками та збереження ключових контактів із каталітично значущими залишками (CRBN: N140/N433; BRD4: W81/W374, Y97/Y390). Для найперспективніших комплексів також оцінювали просторове розташування білків і комплементарність інтерфейсу.

РОЗДІЛ III

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

3.1. Оцінка моделі передбачення біологічної активності PROTAC

3.1.1. Кореляційний аналіз передбачень базових класифікаторів

Ключовим фактором ефективності навчання ансамблю є забезпечення різноманітності базових моделей. У цьому дослідженні різноманітність забезпечувалась шляхом використання різнорідних представлень даних. Для оцінки ступеня незалежності прогнозів було проведено кореляційний аналіз передбачень, згенерованих чотирма базовими класифікаторами. Результати показано у вигляді матриць кореляції окремо для активних та неактивних сполук (Рис. 3.1).

Для неактивних сполук ($n = 468$) попарні кореляції знаходяться в межах $\rho \in [0.65, 0.81]$, що свідчить про помірно-сильну узгодженість (Рис. 3.1А). Найнижчий показник зафіксовано між XGBoost та SVM ($\rho = 0.65$), що пояснюється радикальною відмінністю у просторах ознак (фармакофорні дескриптори проти 2D-дескрипторів) та алгоритмічних парадигмах (градієнтний бустинг дерев рішень проти методу опорних векторів).

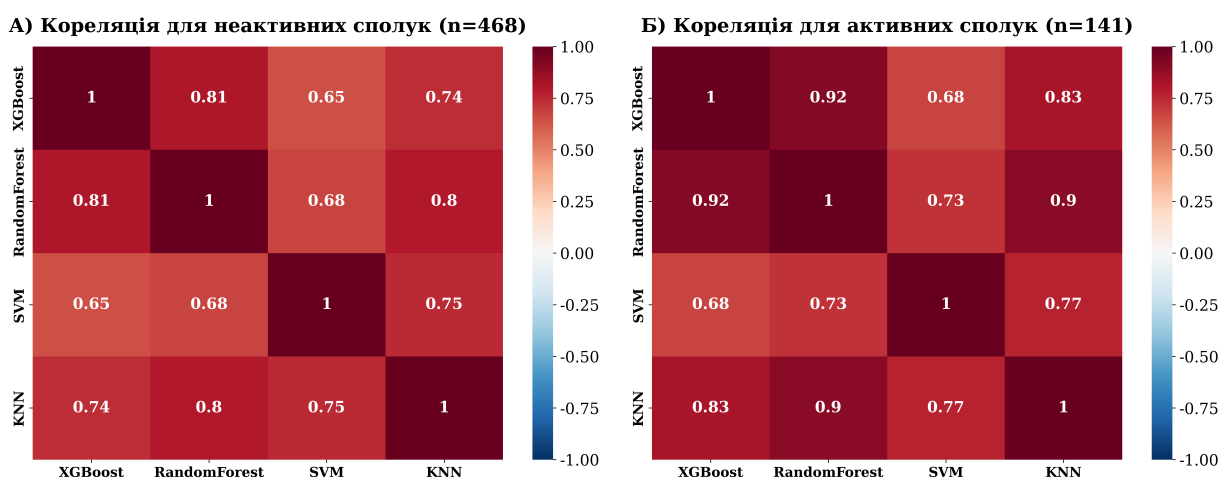


Рисунок 3.1. Матриці попарної кореляції ймовірностей прогнозів базових класифікаторів для неактивних (А) та активних (Б) сполук.

Натомість для активних сполук ($n = 141$) спостерігається зростання кореляції до $[0.68, 0.92]$ (Рис. 3.1Б). Максимальне значення ($\rho = 0.92$) між XGBoost та RandomForest вказує на певну інформаційну надмірність, зумовлену схожою природою алгоритмів та топологічних дескрипторів. Проте SVM стабільно демонструє найнижчу кореляцію з іншими експертами ($\rho \in [0.68, 0.77]$), виконуючи роль критичного диверсифікатора. Відсутність повної кореляції ($\rho < 1.0$) підтверджує, що кожен експерт вносить унікальний сигнал, що є необхідною умовою для подальшого стекування.

3.1.2. Порівняльний аналіз стратегій агрегації передбачень

Для вибору оптимального методу об'єднання результатів було порівняно три стратегії ансамблювання — жорстке голосування (Hard Voting), м'яке голосування (Soft Voting) та мета-моделювання на основі логістичної регресії (Meta-modelling) — з результатами найкращого окремого класифікатора (KNN). Порівняння проводилося за метриками ROC-AUC та MCC (Рис. 3.2).

Жорстке голосування показало найнижчу ефективність за обома метриками ($AUC = 0.89$; $MCC = 0.67$), поступившись навіть одиночній моделі KNN ($AUC = 0.91$; $MCC = 0.66$). Це свідчить про те, що агрегація дискретних передбачень без урахування апостеріорних ймовірностей призводить до втрати інформації про впевненість класифікаторів. М'яке голосування ($AUC = 0.92$; $MCC = 0.69$) демонструє кращу дискримінативну здатність, однак максимальна ефективність ($AUC = 0.92$; $MCC = 0.75$) досягається лише із застосуванням мета-моделі. При цьому показово, що значення ROC-AUC для м'якого голосування та мета-моделі збіглися (0.92), тоді як MCC мета-моделі перевершує м'яке голосування на 9.3 %. Ця різниця обумовлена природою метрик: ROC-AUC є незалежною від порогу мірою дискримінативної здатності, тоді як MCC відображає якість бінарної класифікації при конкретному порозі. Отже, мета-модель перевершує м'яке голосування завдяки кращій оптимізації порогу класифікації та внесків окремих алгоритмів.

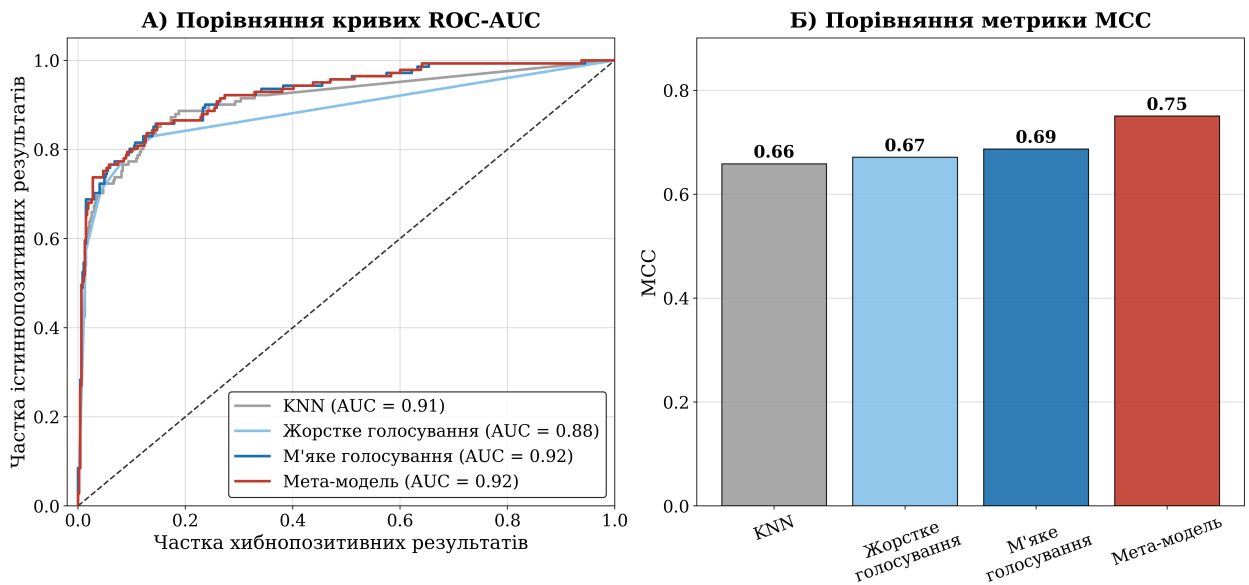


Рисунок 3.2. Порівняльний аналіз стратегій агрегації прогнозів: жорсткого голосування, м'якого голосування та мета-моделювання у зіставленні з найкращою базовою моделлю (KNN): А — метрика ROC-AUC; Б — коефіцієнт кореляції Меттьюза (MCC).

3.1.3. Комбінаторний абляційний аналіз

З метою кількісного оцінювання внеску кожного базового класифікатора та визначення оптимального складу ансамблю проведено вичерпне комбінаторне абляційне дослідження (exhaustive combinatorial ablation study). Для всіх $2^4 - 1 = 15$ підмножин чотирьох базових алгоритмів (від одиничних класифікаторів до повного ансамблю) незалежно навчались та оцінювались мета-моделі. Результати систематизовано на UpSet-діаграмі (Рис. 3.3).

Аналіз одиничних класифікаторів виявив, що найвищий індивідуальний MCC демонструє KNN (0.660), тоді як RF (0.655), XGBoost (0.649) та SVM (0.649) показують схожі результати, що підтверджує відсутність домінантного алгоритму та рівноправне взаємодоповнення. Варто зазначити, що жоден з одиничних класифікаторів не досягає $MCC > 0.660$ (пунктирна лінія «найкращого окремого експерта»), що ілюструє неспроможність будь-якого окремого алгоритму конкурувати з ансамблем.

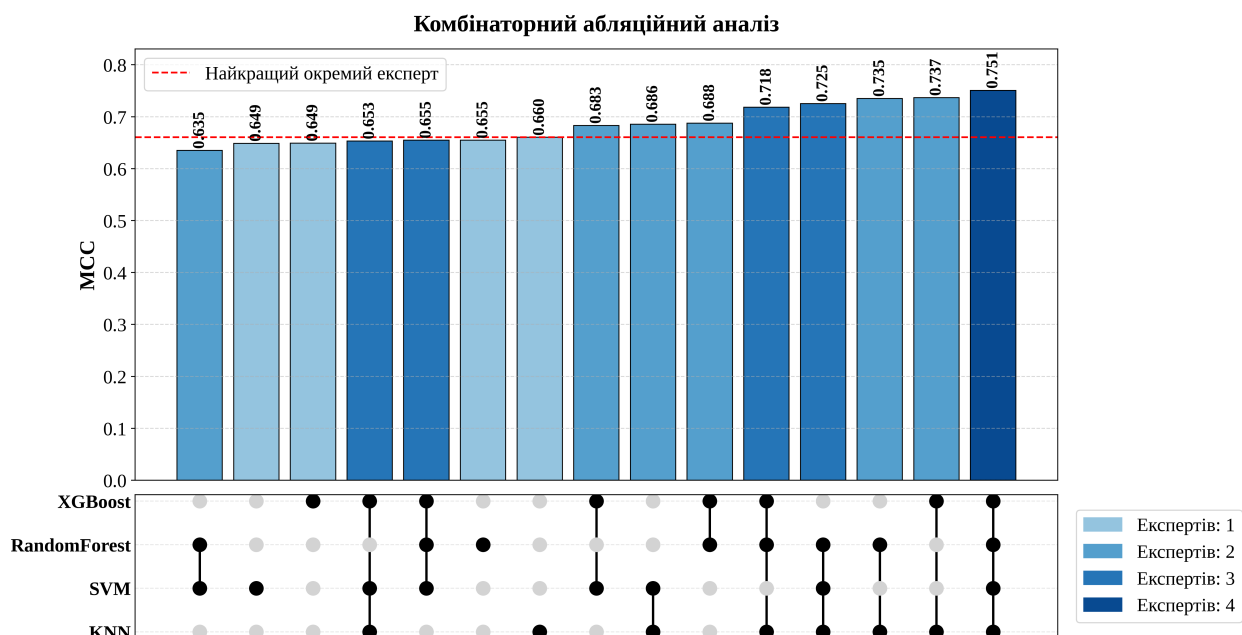


Рисунок 3.3. Результати комбінаторного абляційного аналізу складу ансамблю. Діаграма перетинів (UpSet plot) демонструє зміну метрики MCC залежно від обраної підмножини базових класифікаторів.

Найвище значення MCC досягається при об'єднанні всіх чотирьох моделей (темно-синій стовпчик крайній праворуч). Показник MCC становить 0.75, що на 13,8 % перевищує результат найкращого одиночного класифікатора. Це підтверджує синергетичний ефект повного ансамблю та обґрунтовує архітектурне рішення включати всі чотири спеціалізовані представлення молекул.

Можна помітити, що майже всі комбінації з високими показниками ($MCC > 0.71$) включають модель KNN. Наприклад, дуети XGBoost і KNN (0.74) та RF і KNN (0.74) продемонстрували кращі метрики в порівнянні з більшістю моделей з трьома експертами.

Цікавим спостереженням є те, що не кожне додавання моделі покращує результат. Ансамбль з трьох моделей (XGBoost + SVM + KNN) показав результат $MCC = 0.65$, що навіть нижче за показник одиночного KNN (0.66). Це вказує на те, що певні комбінації (зокрема за участю SVM у деяких зв'язках) можуть вносити шум або конфліктуючі прогнози, що знижує загальну точність.

3.1.4. Ваги мета-класифікатора та інтерпретація внеску експертів

Оцінка параметрів логістичної регресії дозволила визначити питому вагу кожного базового класифікатора в ансамблі: 1.09 (XGBoost), 1.64 (RF), 1.69 (SVM) та 2.28 (KNN). Коефіцієнти для всіх моделей є додатними, що демонструє конструктивний внесок кожного базового класифікатора. Найвищий ваговий коефіцієнт KNN (2.28) обумовлений тим, що KNN-представлення (PCA-зменшені ембединги ChemBERTa та білкові ембединги) надає унікальну інформацію про молекулярну подібність у семантичному просторі — інформацію, яку жоден інший класифікатор не може відтворити зі своїх дескрипторів. SVM займає друге місце (1.69), підтверджуючи цінність 2D-дескрипторів. Близькість ваг XGBoost (1.09) та RF (1.64) при їх взаємно найвищій кореляції ($\rho = 0.92$ на активних сполуках) вказує на часткову інформаційну надмірність, однак це не перешкоджає загальному позитивному внеску.

3.1.5. Оцінка ефективності на тестовій вибірці

Фінальна оцінка ефективності ансамблю проводилась на тестовій вибірці, яка жодного разу не використовувалась у процесі тренування чи оптимізації. Результати зведено у Таблиці 3.1.

Таблиця 3.1. Результати кількісного оцінювання якості класифікації розробленої моделі за основними статистичними метриками.

Метрика	Значення
Точність (Accuracy)	0.915
Влучність (Precision)	0.874
Повнота (Recall)	0.738
F1-метрика	0.800
MCC	0.751
ROC-AUC	0.924
AUPRC	0.832

Отримані значення свідчать про стабільно високу ефективність ансамблю у задачі бінарної класифікації активності PROTAC. Особливо важливим є показник $MCC = 0.75$, оскільки він є найбільш інформативним для незбалансованих наборів даних. Значення, близьке до 1.0, означає практично ідеальну класифікацію, тоді як значення 0 відповідає випадковому вгадуванню. Висока влучність (0.87) при помірній повноті (0.74) підтверджує, що обраний поріг $\tau = 0.80$ зміщує баланс у бік мінімізації хибнопозитивних результатів. Така стратегія була цілеспрямовано обрана з огляду на подальше використання моделі як функції оцінювання (scoring function) у середовищі підкріплювального навчання. У такому сценарії хибнопозитивні передбачення можуть призводити до накопичення помилкових сигналів винагороди, що, у свою чергу, спричиняє зсув генеративного процесу у бік хімічно неефективних або біологічно неактивних структур.

Матриця помилок (Рис. 3.4) наочно демонструє розподіл помилок класифікації: з 468 неактивних сполук лише 15 було хибно класифіковано як активні ($FP = 15$), тоді як з 141 активної сполуки 37 було пропущено ($FN = 37$), а 104 — коректно ідентифіковано ($TP = 104$).

Для кількісної характеристики помилок першого та другого роду обчислено специфічність і чутливість — дві взаємодоповнюючі міри, що розкривають поведінку класифікатора окремо для кожного класу.

Специфічність (Specificity, або True Negative Rate) визначає частку коректно ідентифікованих неактивних сполук:

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{453}{453 + 15} = \frac{453}{468} \approx 0.97. \quad (3.1)$$

Значення 0.97 підтверджує те, що модель хибно маркує як активні лише 3.2 % справді неактивних сполук — вкрай низький рівень.

Чутливість (Sensitivity, або True Positive Rate, Recall) характеризує частку коректно виявлених активних сполук:

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{104}{104 + 37} = \frac{104}{141} \approx 0.74. \quad (3.2)$$

Значення 0.738 означає, що модель пропускає близько 26.2 % справді активних PROTAC-кандидатів.

Асиметрія між специфічністю (0.97) та чутливістю (0.74) є прямим наслідком обраного порогу $\tau = 0.80$, проте її прикладне обґрунтування визначається цільовим призначенням моделі. Розроблений ансамбль мав використовуватись як оціночна функція у системі генеративного молекулярного дизайну на основі навчання з підкріпленням у Link-INVENT. У цьому контексті, помилкове віднесення сполук до активного класу є особливо шкідливим: генератор отримує хибний сигнал про активність і починає систематично генерувати молекули у неперспективній ділянці хімічного простору, що призводить до деградації траєкторії навчання. Натомість, помилкова класифікація сполук як неактивних є менш критичною: генератор просто не отримує заохочення за цю конкретну молекулу, не порушуючи загальну логіку генерації. Отже, висока специфічність є ключовою вимогою до оціночної функції в RL-керованому молекулярному дизайні, оскільки забезпечує надійність сигналу винагороди.

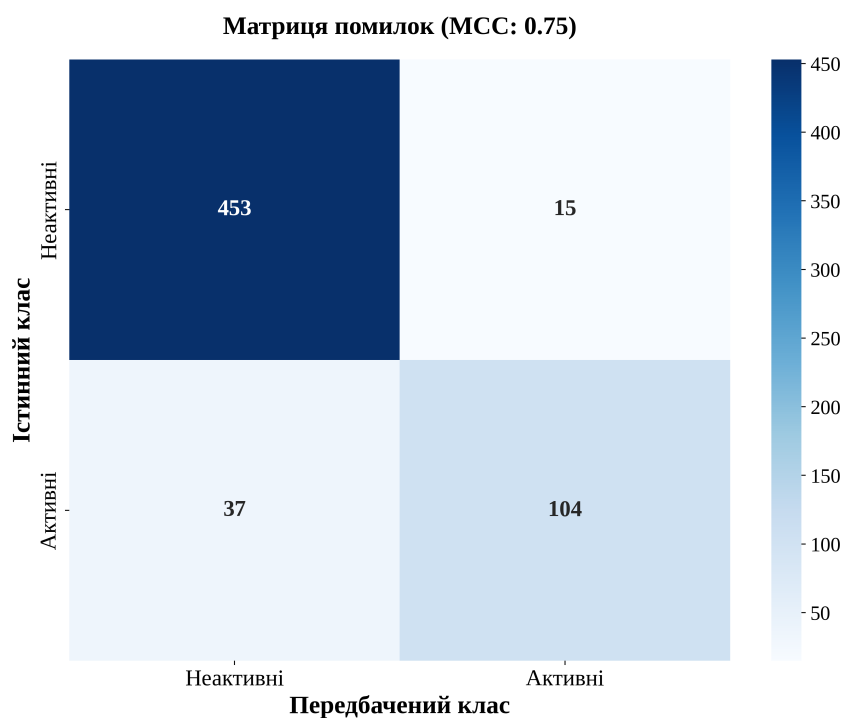


Рисунок 3.4. Матриця помилок фінальної ансамблевої моделі, оціненої на незалежній тестовій вибірці.

3.1.6. Епістемічна невизначеність ансамблю

Важливою характеристикою ансамблю є здатність кількісно оцінювати власну впевненість у прогнозах. Епістемічна невизначеність визначається як стандартне відхилення ймовірностей, генерованих чотирма базовими класифікаторами для кожного окремого зразка.

Гістограма розподілу епістемічної невизначеності (Рис. 3.5) розкриває якісну відмінність між коректними та хибними передбаченнями. Правильні передбачення (блакитні стовпці) демонструють різко виражений пік при $\sigma \approx 0.02 - 0.03$: переважна більшість коректно класифікованих зразків обробляється ансамблем із високим ступенем консенсусу між базовими класифікаторами. Розподіл хибних передбачень (червоні стовпці) є значно більш рівномірним і пологим у широкому діапазоні $\sigma \in [0.02, 0.30]$, що відображає те, що більшість помилок пов'язана з підвищеною розбіжністю передбачень між алгоритмами.

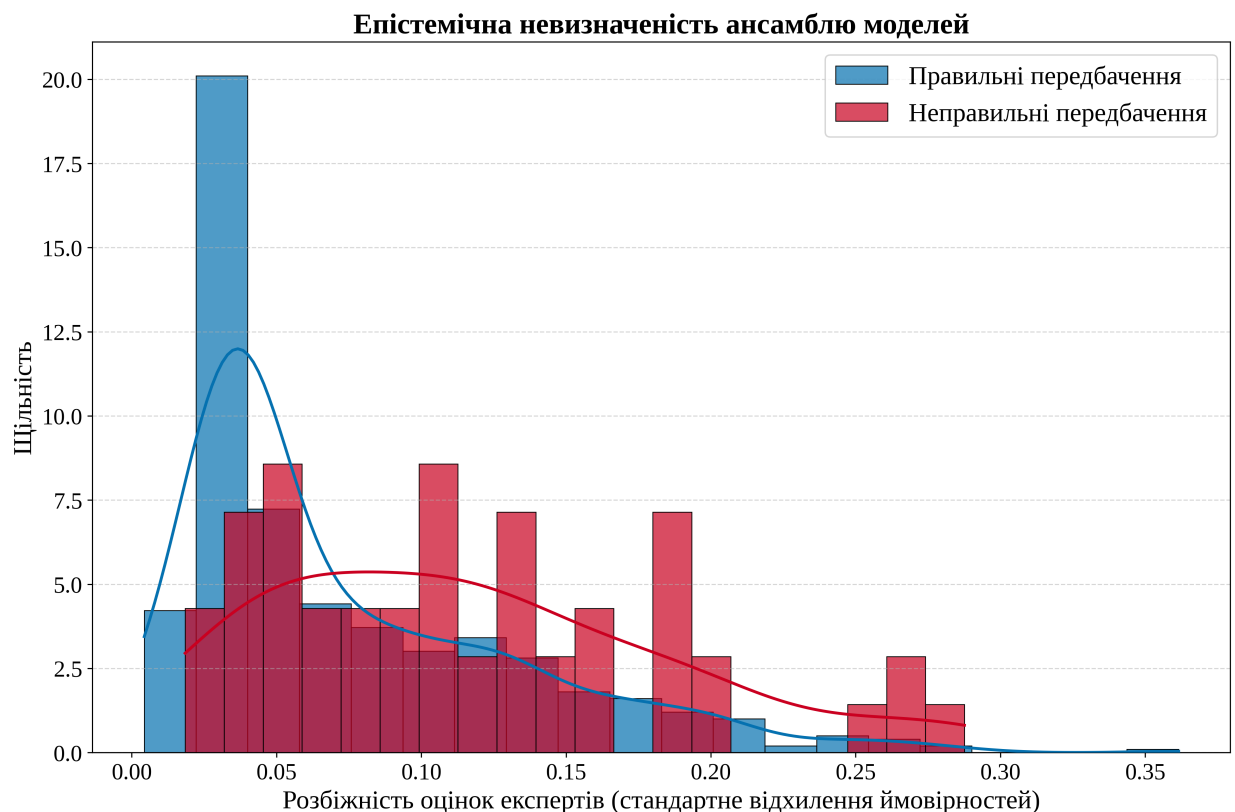


Рисунок 3.5. Розподіл епістемічної невизначеності (стандартного відхилення прогнозів базових моделей) для тестової вибірки.

При порозі невизначеності $\sigma < 0.20$ до категорії «впевнених» прогнозів потрапляє 95.07 % тестового набору. Обраний поріг є обґрунтованим емпірично: він відповідає точці, де щільність розподілу хибних передбачень суттєво перевищує щільність правильних ($\sigma \approx 0.15 - 0.20$), тобто саме з цього значення невизначеності хибні прогнози починають кількісно переважати. Для підмножини «впевнених» зразків $MCC = 0.76$, що на 0.01 одиниць (або $\approx 1.1\%$ відносно базового значення) перевищує загальний $MCC = 0.75$ по всьому тестовому набору. Незначний, але стабільний приріст підтверджує, що фільтрація за порогом невизначеності дійсно відсіює найбільш сумнівні передбачення, хоча й ціною виключення 4.93 % зразків із автоматичної класифікації — вони можуть бути передані на додаткову експертну оцінку або повторний скринінг.

3.1.7. Статистична валідація моделі методом Y-рандомізації

Необхідною умовою довіри до будь-якої прогностичної моделі машинного навчання є підтвердження того, що її передбачувальна здатність ґрунтується на змістовних закономірностях у структурі даних, а не на випадкових статистичних кореляціях або артефактах процедури навчання. З цією метою застосовано процедуру Y-рандомізації (Y-randomization, або Y-scrambling) — стандартний метод статистичної валідації у хемінформатиці. Вектор цільових міток (Y_{train}) було випадково перемішано при незмінній матриці дескрипторів (X_{train}), що руйнує зв'язок між хімічними ознаками молекул і класовою належністю. Результати рандомізованої моделі на тестовій вибірці зведено у Таблиці 3.2 та проілюстровано матрицею помилок (Додаток 3).

Деградація всіх метрик при Y-рандомізації є різкою та однозначною. MCC впав з 0.75 до 0.11 — значення наближене до нуля, що відповідає відсутності будь-якої справжньої прогностичної здатності. ROC-AUC знизився з 0.92 до 0.56, фактично наблизившись до 0.5, що характерно для випадкового класифікатора. Це падіння на 0.36 одиниці (39.4 % відносно

базового значення) є особливо показовим, оскільки ROC-AUC є мірою незалежною від порогу і не залежить від вибору τ . Невелике відхилення від очікуваних значень ($MCC \approx 0$, $ROC-AUC \approx 0.5$) після Y-рандомізації може пояснюватися стохастичними флуктуаціями та незбалансованістю даних.

Таблиця 3.2. Показники ефективності класифікації на випадково перемішаних даних (Y-рандомізації) у порівнянні з вихідною моделлю.

Метрика	Основна модель	Y-рандомізована модель
MCC	0.75	0.11
ROC-AUC	0.92	0.56
Точність (Accuracy)	0.92	0.54
Влучність (Precision)	0.87	0.28
Повнота (Recall)	0.74	0.61
F1-метрика	0.80	0.38
AUPRC	0.83	0.28

Характер помилок рандомізованої моделі також є діагностично значущим. Матриця помилок демонструє, що модель, навчена на перемішаних мітках, класифікує майже половину неактивних сполук як активні ($FP = 224$ з 468), генеруючи практично симетричний розподіл помилок — поведінку, типову для моделі, що не розрізняє класи. Натомість порівняно висока повнота рандомізованої моделі ($Recall = 0.61$) при вкрай низькій влучності ($Precision = 0.28$) вказує на тривіальну стратегію: модель схильна класифікувати більшість зразків як активні, що при незбалансованому наборі даних механічно підвищує повноту без жодної реальної дискримінативної здатності.

Показово також, що ваги мета-класифікатора рандомізованої моделі набули частково від'ємних значень: -0.29 (XGBoost), -0.37 (RF), $+0.02$ (SVM), $+0.22$ (KNN), тоді як всі ваги основної моделі є позитивними та значно більшими за абсолютною величиною $[1.09, 1.64, 1.69, 2.28]$. Це

свідчить про те, що при рандомізованих мітках мета-класифікатор не здатен узгоджено інтегрувати сигнали від базових алгоритмів.

Епістемічна невизначеність рандомізованої моделі (Додаток 4) також якісно відрізняється від основної: обидва розподіли — для правильних та хибних передбачень — є практично ідентичними та зосереджені у діапазоні $\sigma \in [0.05, 0.20]$, без виражених піків консенсусу. Відсутність чіткої диференціації між розподілами підтверджує, що базові класифікатори, навчені на шумових мітках, не формують узгоджених сигналів навіть для зразків, які рандомізована модель випадково класифікує правильно.

Отримані результати однозначно підтверджують, що висока ефективність основної ансамблевої моделі ($AUC = 0.92$; $MCC = 0.75$) ґрунтується на реальних структурно-активнісних залежностях (SAR) у даних, а не є артефактом процедури навчання.

3.2. Генерація та оптимізація лінкерів PROTAC-сполук за допомогою Link-INVENT з підкріплювальним навчанням

3.2.1. Траєкторія навчання агента з підкріпленням

З метою спрямованої генерації лінкерів PROTAC-сполук із підвищеною прогнозованою біологічною активністю було проведено оптимізацію генеративної моделі Link-INVENT методом навчання з підкріпленням (RL) протягом 200 кроків. Як цільова функція винагороди використовувався класифікатор біологічної активності PROTAC-сполук — гетерогенна ансамблева модель, навчена на даних бази PROTAC-DB (детально описана у розділі 2.1). Для стабілізації навчання застосовувався алгоритм DAP із коефіцієнтом інтенсивності $\sigma = 256$.

Аналіз траєкторії навчання (Додаток 5) демонструє стійке зростання середнього балу валідних сполук: від 0.62 для апріорної моделі до 0.70 – 0.71 на фінальних кроках. На етапах 75 – 80 та близько 120 спостерігаються локальні максимуми (0.68 та 0.69 відповідно), що

супроводжуються короткочасними зниженнями та відображають дослідження нових ділянок хімічного простору. Сумарний приріст середнього значення становить 0.09 – 0.10 ($\approx 14 - 15\%$ від початкового рівня). Відсутність ознак перенавчання та колапсу різноманітності (Рис. 3.7) підтверджує ефективність обраної стратегії RL-оптимізації з використанням DAP.

3.2.2. Новизна та внутрішня різноманітність згенерованої вибірки

Критичним аспектом оцінки якості генеративної моделі є здатність продукувати не лише активні, але й хімічно нові та різноманітні структури. Тому було проаналізовано дві характеристики фінальної вибірки сполук: новизну та внутрішню різноманітність (Рис. 3.6 та 3.7, відповідно).

Новизна визначалась як одиниця мінус максимальна подібність Танімото між згенерованою сполукою та сполуками з бази даних PROTAC-DB: $\text{Novelty}(x) = 1 - \max_{y \in Y} T(x, y)$, де $T(x, y)$ — коефіцієнт подібності Танімото, Y — база даних PROTAC-DB.

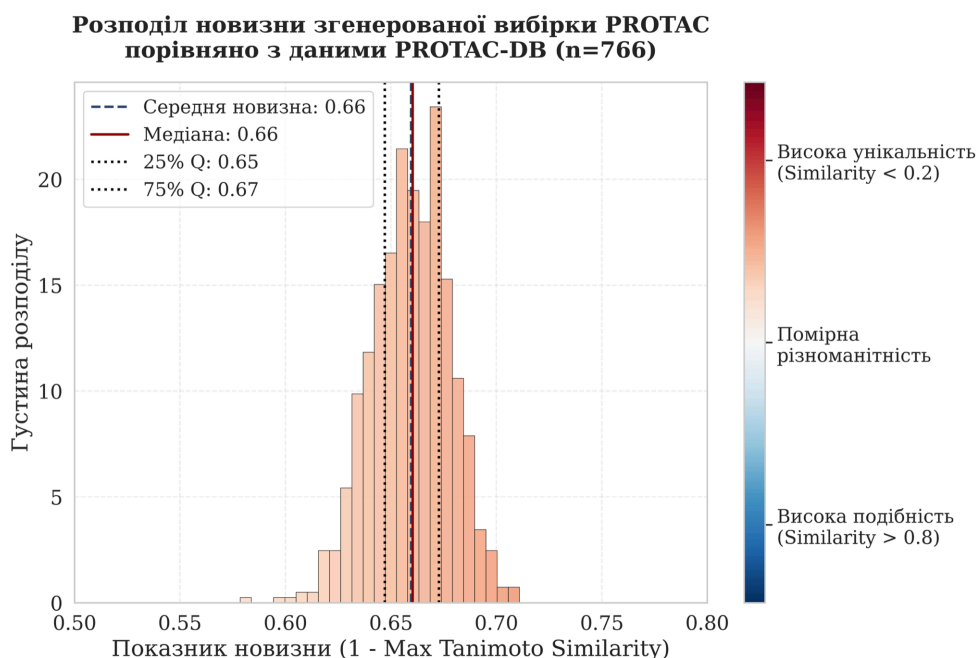


Рисунок 3.6. Розподіл новизни згенерованої вибірки PROTAC-сполук, що характеризує ступінь структурної відмінності від відомих сполук.

Середнє значення новизни по вибірці становить 0.66 (медіана 0.66; міжквартильний діапазон (IQR) 0.65 – 0.67). Вузкий та симетричний розподіл із незначним правим хвостом вказує на те, що переважна більшість згенерованих структур є структурно новими відносно бази даних при одночасній відсутності надмірної структурної розбіжності, яка могла б свідчити про хімічну нереалістичність. Отриманий рівень новизни (~ 0.66) є прийнятним для задач молекулярного дизайну: він означає, що згенеровані структури не є простими аналогами відомих PROTAC, а представляють оригінальний хімічний простір.

Внутрішню різноманітність оцінено через попарні відстані Танімото між усіма 292 995 парами згенерованих сполук: $D_{\text{int}} = 1 - T(x_i, x_j)$.

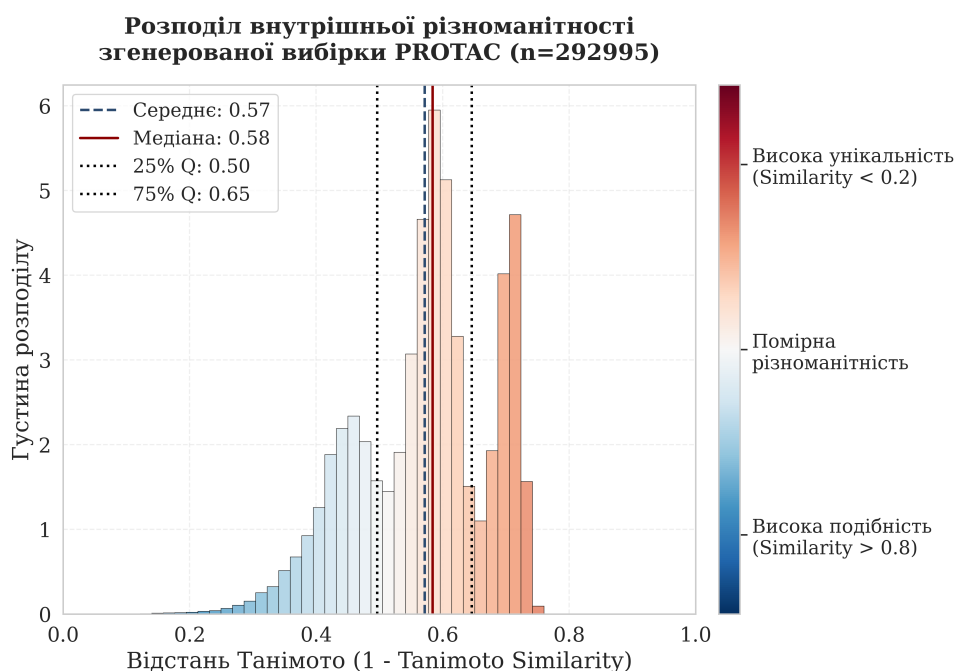


Рисунок 3.7. Розподіл попарних відстаней Танімото (1 – подібність) між PROTAC-сполуками, що відображає внутрішню різноманітність згенерованої вибірки.

Середня внутрішня відстань становить 0.57 (медіана 0.58; IQR 0.50 – 0.65). Розподіл є лівоасиметричним з модою в діапазоні 0.60 – 0.75, що відповідає помірному рівню різноманітності. Кількість пар зі значеннями

відстані менше 0.2 є незначною, що вказує на відсутність суттєвої кластеризації або режиму колапсу генерації — патологічної поведінки, при якій модель продукує численні копії обмеженого набору структур із максимальним значенням функції винагороди.

3.2.3. Порівняльний аналіз фізико-хімічних властивостей апіорної та RL-оптимізованої вибірок PROTAC-сполук

Для з'ясування того, які структурні зміни в лінкерах зумовлюють підвищення прогнозованої біологічної активності, було проведено порівняльний аналіз чотирьох ключових фізико-хімічних дескрипторів між апіорною вибіркою ($n = 551$) та RL-оптимізованою вибіркою ($n = 766$) з використанням непараметричного критерію Манна–Уїтні (Рис. 3.8).

Ймовірність біологічної активності (Рис. 3.8А). Найбільш прямим підтвердженням ефективності RL-оптимізації є статистично значуще підвищення прогнозованої ймовірності активності ($p \ll 0.001$). Медіанне значення ймовірності активності зросло з 0.61 для апіорної вибірки до 0.70 для RL-оптимізованої, що відповідає відносному приросту $\approx 15\%$. Розмір ефекту Cliff's $\Delta = -0.74$ свідчить про великий ефект: у 73.9% всіх попарних порівнянь молекула з RL-вибірки характеризується вищою прогнозованою активністю, ніж молекула з апіорної. Це підтверджує, що зрушення розподілів є не статистичним артефактом великих вибірок, а відображає справжнє фокусування генеративної моделі на хімічному підпросторі з високою прогнозованою активністю.

Довжина лінкера (Рис. 3.8Б). Статистично значуще збільшення довжини лінкера зафіксовано в RL-оптимізованій вибірці ($p \ll 0.001$): медіана зросла з 16.0 до 17.0 атомів. Однак Cliff's $\Delta = -0.26$ вказує на малий розмір ефекту — незважаючи на значущість p -значення, перекриття між розподілами є суттєвим. Це узгоджується з відомими структурними вимогами до PROTAC: лінкер має бути достатньо довгим для просторового зближення лігандів, проте RL-агент уникає надмірного подовження, яке погіршує

фармакокінетичні характеристики. Незначний приріст медіанної довжини свідчить про те, що модель не схильна до надмірного подовження лінкера, яке могло б призвести до погіршення фармакокінетичних характеристик.

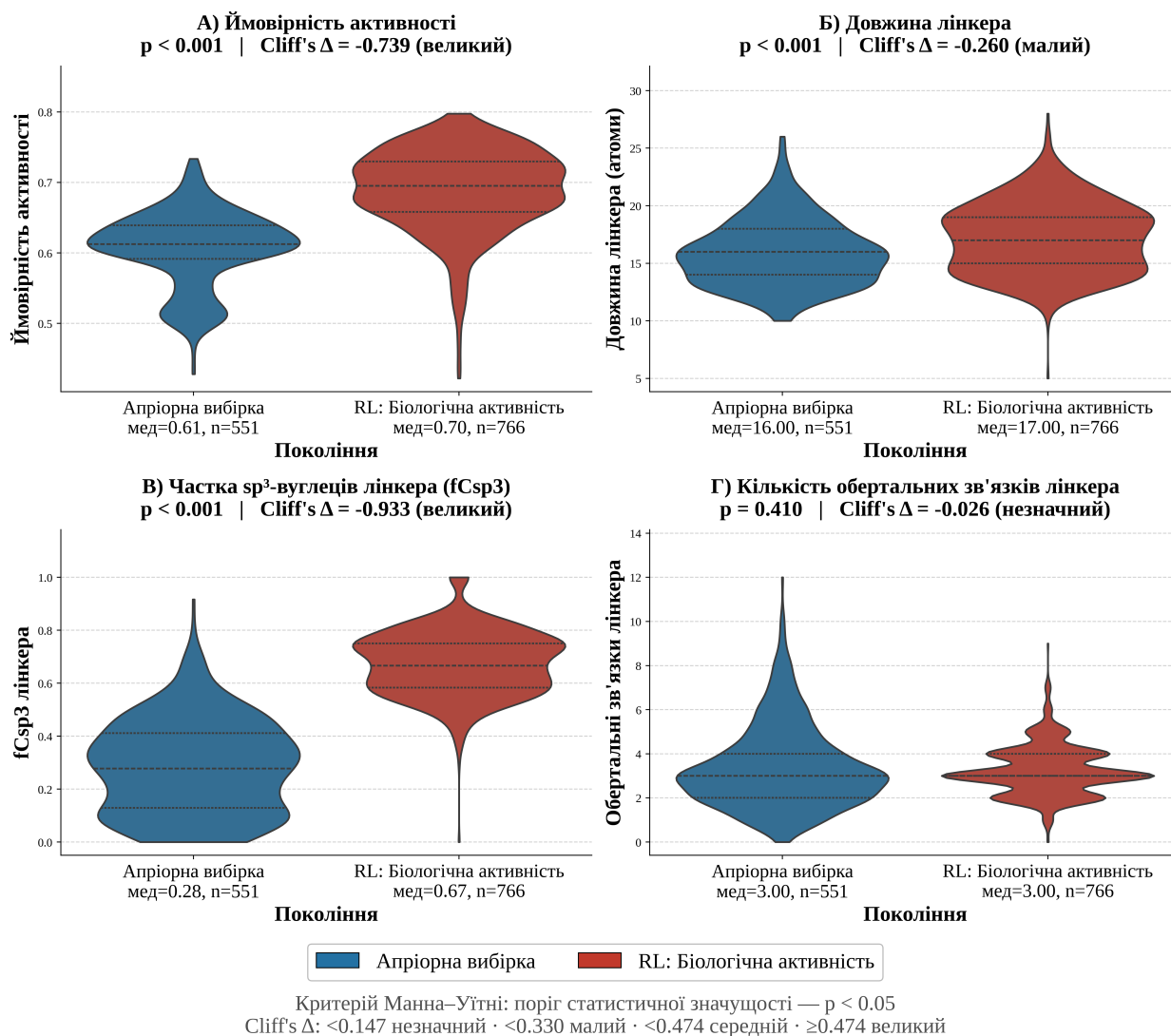


Рисунок 3.8. Порівняльний аналіз властивостей лінкерів у апріорній та RL-оптимізованій вибірках PROTAC-сполук: А) розподіл прогнозованої ймовірності біологічної активності; Б) довжина лінкера (у атомах); В) частка sp^3 -гібридизованих атомів вуглецю (fCsp³); Г) кількість обертальних зв'язків.

Частка sp^3 -вуглеців лінкера — fCsp³ (Рис. 3.8В). Найбільш виражені та статистично значущі відмінності між вибірками спостерігаються за показником fCsp³ ($p \ll 0.001$). Медіанне значення fCsp³ зросло більш ніж удвічі — з 0.28 в апріорній вибірці до 0.67 в RL-оптимізованій.

$\text{Cliff's } \Delta = -0.93$ є найбільшим серед усіх аналізованих дескрипторів і відповідає великому ефекту, що межує з повним розділенням розподілів: у 93.3 % попарних порівнянь fCsp^3 RL-молекули вище за апіорну. Така виражена зміна свідчить про принципову зміну хімічної природи лінкерів: від планарних, ароматичних або слабконасичених структур до тривимірних, аліфатичних фрагментів (піперидин, піперазин, проліноподібні мотиви), що асоціюються з кращою розчинністю та зниженням токсичності.

Кількість обертальних зв'язків лінкера (Рис. 3.8Г). На відміну від описаних вище дескрипторів, кількість обертальних зв'язків не зазнала статистично значущих змін у процесі RL-оптимізації ($p = 0.41$). $\text{Cliff's } \Delta = -0.03$ підтверджує незначний розмір ефекту: розподіли двох вибірок практично ідентичні за медіаною (3.0 для обох), тоді як RL-вибірка характеризується дещо меншою дисперсією. Такий результат є непрямим свідченням того, що ансамблева модель є нечутливою до гнучкості лінкера в аналізованому діапазоні, або що оптимальна гнучкість вже була представлена в апіорній вибірці.

З метою оцінки впливу RL-оптимізації на фармакокінетично релевантні властивості згенерованих PROTAC-сполук було проведено порівняльний аналіз чотирьох ключових дескрипторів — молекулярної маси (MW), ліпофільності (LogP), полярної площі поверхні (TPSA) та клітинної проникності (Caco-2) — між апіорною та RL-оптимізованою вибірками із застосуванням непараметричного критерію Манна–Уїтні (Рис. 3.9).

Молекулярна маса (MW) (Рис. 3.9А). RL-оптимізована вибірка демонструє статистично значуще збільшення MW ($p \ll 0.001$): медіанне значення зросло з 816.26 Да до 843.36 Да. $\text{Cliff's } \Delta = -0.34$ відповідає середньому розміру ефекту, що свідчить про реальне, але помірне зміщення розподілу. Обидва значення залишаються значно нижчими за умовну верхню межу ≈ 1200 Да, прийнятну для PROTAC. Це узгоджується з природою PROTAC-молекул, де збільшення маси зумовлене оптимізацією лінкера для ефективного зближення білка-мішені та Е3-лігази.

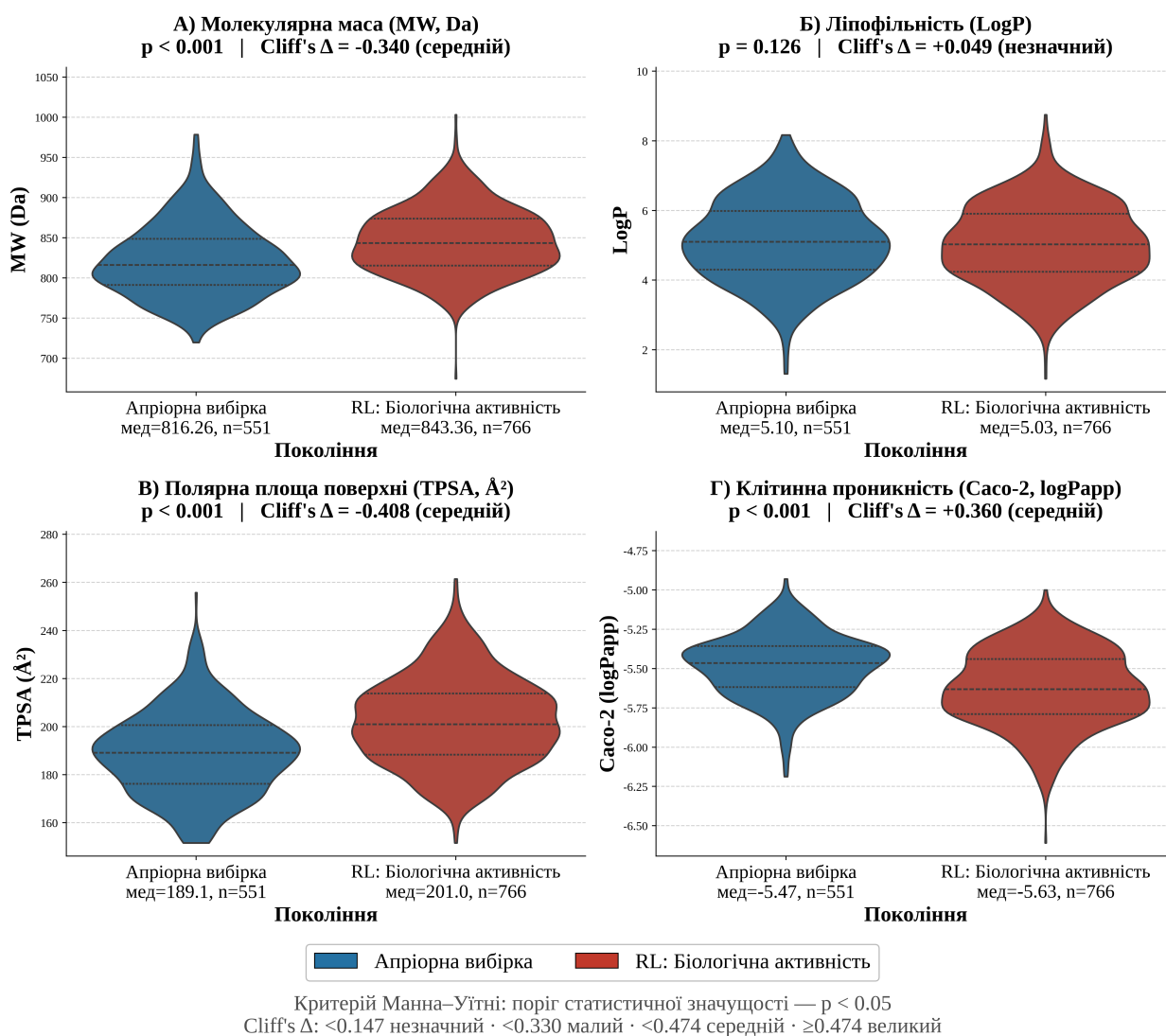


Рисунок 3.9. Порівняльний аналіз фізико-хімічних властивостей апріорної та RL-оптимізованої вибірок PROTAC-сполук: А) молекулярна маса (MW, Da); Б) ліпофільність (LogP); В) полярна площа поверхні (TPSA, Å²); Г) клітинна проникність (Caco-2).

Ліпофільність (LogP) (Рис. 3.9Б). На відміну від інших дескрипторів, зміни ліпофільності між вибірками не є статистично значущими ($p = 0.126$), що підтверджується Cliff's $\Delta = +0.05$ — незначний ефект. Медіанні значення залишаються практично незмінними (5.10 для апріорної та 5.03 для RL-вибірки). Форма розподілів також є подібною, що свідчить про збереження загального балансу між гідрофобними та гідрофільними властивостями молекул.

Полярна площа поверхні (TPSA) (Рис. 3.9В). Для TPSA спостерігається статистично значуще зростання ($p \ll 0.001$): медіанне значення збільшилось з $189.1 \text{ \AA}^2$ до $201.0 \text{ \AA}^2$. Cliff's $\Delta = -0.41$ відповідає середньому розміру ефекту. Підвищення TPSA може відображати збагачення молекул полярними функціональними групами, що потенційно сприяють формуванню додаткових міжмолекулярних взаємодій у складі трикомпонентного комплексу. Водночас збільшення TPSA традиційно асоціюється зі зниженням мембранної проникності, що вказує на можливий компроміс між активністю та фармакокінетичними властивостями.

Клітинна проникність (Caco-2) (Рис. 3.9Г). RL-оптимізована вибірка демонструє статистично значуще зниження проникності ($p \ll 0.001$): медіанне значення LogPapp зменшилось з -5.47 до -5.63 . Cliff's $\Delta = +0.36$ (знак позитивний, оскільки менш негативні значення апріорної вибірки переважають у парних порівняннях) відповідає середньому розміру ефекту. Такий результат є очікуваним з огляду на одночасне зростання молекулярної маси та TPSA, які є ключовими факторами, що обмежують проникність для великих бінарних молекул, зокрема PROTAC.

Важливо підкреслити, що попри виявлені зсуви розподілів, усі чотири проаналізовані дескриптори — MW, LogP, TPSA та Caco-2 — для обох вибірок залишаються в межах фізико-хімічного простору, характерного для PROTAC та інших bRo5 сполук. Зокрема, медіанні значення молекулярної маси становлять 816.26 Да для апріорної та 843.36 Да для RL-оптимізованої вибірки, що є значно нижчими за умовну верхню межу $\approx 1200 \text{ Да}$, прийнятну для PROTAC. Ліпофільність практично не змінюється (медіана LogP: 5.10 проти 5.03), залишаючись у межах $\approx 2 - 6$, характерних для фармакокінетично реалістичних bRo5-сполук. Полярна площа поверхні залишається в межах $200 - 250 \text{ \AA}^2$, які описані як допустимі для проникних великих молекул. Водночас передбачена клітинна проникність (Caco-2) знижується з -5.47 до -5.63 , що відповідає діапазону обмеженої, але реалістичної проникності, типової для PROTAC.

3.3. Комплексний аналіз результатів молекулярного докінгу топ-100 PROTAC-сполук

3.3.1. Загальна характеристика результатів докінгу

До молекулярного докінгу було відібрано 100 сполук, згенерованих алгоритмом навчання з підкріпленням та ранжованих за ймовірністю біологічної активності. Для кожної сполуки виконано 5 незалежних запусків докінгу для оцінки зв'язування, відтворюваності та збіжності результатів докінгу. У Таблиці 3.3. наведено статистичні показники глобального розподілу оцінок докінгу усіх запусків.

Таблиця 3.3. Глобальна статистика оцінок докінгу ($n = 100$ сполук, 5 незалежних запусків на сполуку)

Показник	Значення
Мінімальна оцінка	-38.99
Максимальна оцінка	-14.53
Середнє значення	-26.76
Медіана	-26.28
Стандартне відхилення	4.36

Медіана оцінок становить -26.28 , що свідчить про загалом високу спорідненість відібраних RL-сполук до цільових білків. Стандартне відхилення у 4.36 відображає природну варіативність хімічного простору, охопленого генеративною моделлю.

На Рис. 3.10 представлено розподіл оцінок докінгу для кожної з 100 сполук, відсортованих у порядку зростання середньої оцінки: найкраща оцінка (синя крапка), середня оцінка (червона крапка) та варіабельність між запусками (червоні відрізки — ± 1 стандартне відхилення). Спостережувана закономірність є типовою для стохастичного докінгу: сполуки з найнижчими середніми оцінками демонструють найбільшу різницю між найкращим і

середнім значеннями (блакитна тінь на Рис. 3.10). Це зумовлено розширенням конформаційного простору для великих гнучких молекул, що ускладнює пошук глобального мінімуму енергії.

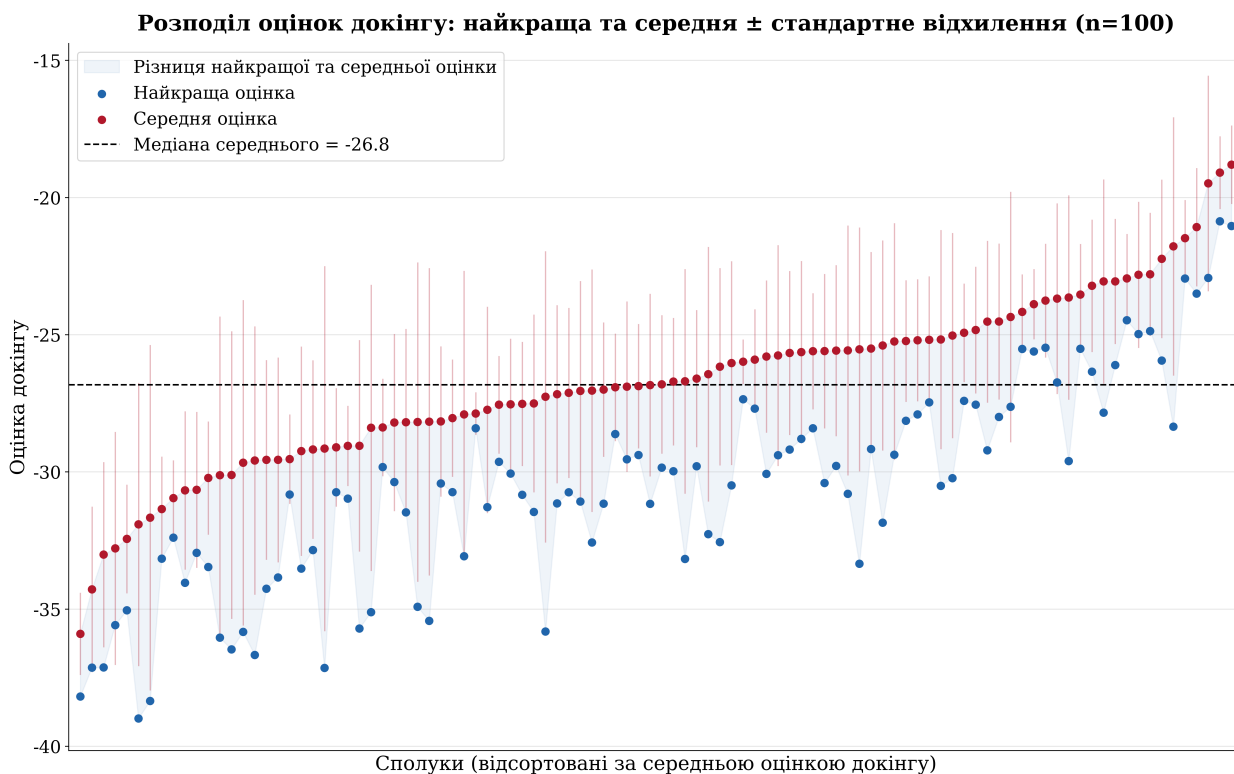


Рисунок 3.10. Розподіл оцінок докінгу: найкраща та середня $\pm$ стандартне відхилення для 100 сполук (відсортовані за середньою оцінкою).

3.3.2. Оцінка відтворюваності та збіжності результатів

Критичним аспектом надійності результатів комп'ютерного докінгу є відтворюваність. Для її кількісної оцінки використовувався коефіцієнт варіації (CV, %), що відображає відносне стандартне відхилення між незалежними запусками. Результати представлено на Рис. 3.11.

Панель А (Рис. 3.11) ілюструє зв'язок між середньою оцінкою докінгу та CV. Поріг прийнятної прецизійності CV = 10 % (синя пунктирна лінія) та поріг відбору за оцінкою докінгу –32 (червона пунктирна лінія) утворюють чотири квадранти. Сполуки у лівому нижньому квадранті — тобто ті, що мають і сильне зв'язування (низька оцінка докінгу, < -32), і низький CV

($< 10\%$) — є пріоритетними кандидатами, оскільки поєднують потенційну ефективність із аналітичною надійністю (виділено чорним контуром, $n = 3$).

Панель Б (Рис. 3.11) демонструє збіжність результатів через співвідношення між середньою та найкращою оцінками. Ідеальна збіжність відповідає розміщенню точок на лінії «Найкращий = Середній». Фактичні дані розташовані нижче цієї лінії, що відповідає стохастичній природі докінгу: найкраща докінг-оцінка за визначенням не перевищує середню. Обрані топ-3 кандидати розміщені поблизу діагоналі та мають низькі значення CV, що свідчить про їхню високу збіжність і надійність результатів.

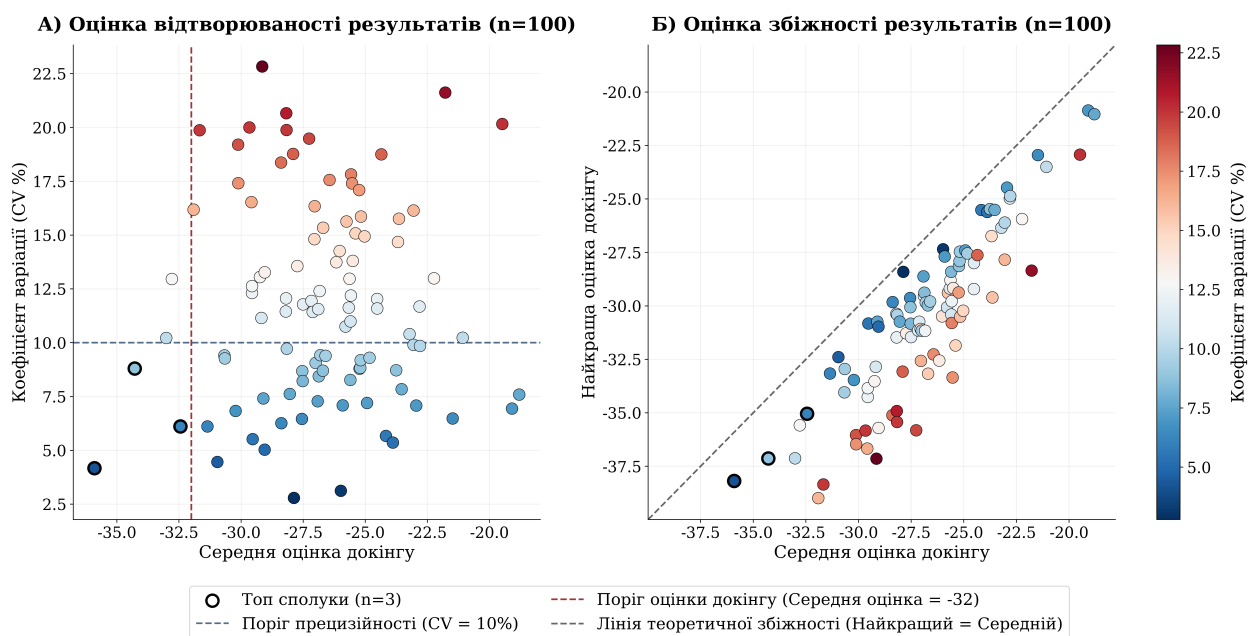


Рисунок 3.11. А) Оцінка відтворюваності результатів: коефіцієнт варіації (CV%) та середня оцінка докінгу. Б) Оцінка збіжності результатів: найкраща та середня оцінка докінгу. Кольорова шкала відображає CV%.

3.3.3. Топ-10 PROTAC-сполук за середньою оцінкою докінгу

У Таблиці 3.4 представлено 10 сполук з найнижчими значеннями середньої оцінки докінгу. Ранжування за середньою оцінкою є більш репрезентативним критерієм відбору, ніж найкращий разовий результат, оскільки відображає стабільну здатність молекули зв'язуватися з цілью

незалежно від конкретного запуску. Всі представлені сполуки мають середню оцінку нижче -30.67 , а три перші — нижче -33 .

Таблиця 3.4. Топ-10 сполук за середньою оцінкою докінгу. Топ-3 сполуки виділено сірим кольором.

Сполука	Середня оцінка	Найкраща оцінка	Std	CV (%)	Збіжність
P34	-35.90	-38.19	1.50	4.18	2.28
P87	-34.28	-37.13	3.02	8.81	2.85
P50	-33.01	-37.13	3.37	10.22	4.11
P92	-32.79	-35.58	4.25	12.97	2.80
P96	-32.44	-35.04	1.98	6.12	2.60
P57	-31.91	-38.99	5.16	16.18	7.08
P6	-31.67	-38.35	6.29	19.87	6.68
P3	-31.35	-33.16	1.92	6.12	1.81
P17	-30.95	-32.40	1.38	4.46	1.44
P54	-30.67	-34.04	2.89	9.41	3.37

Вибір топ-3 ґрунтується на виконанні наступних вимог: $CV < 10\%$, середня оцінка < -32 , збіжність < 3 (Таблиця 3.4, виділено сірим кольором). Сполука P34 посідає перше місце завдяки поєднанню найсильнішого середнього зв'язування (-35.90) та найнижчого CV серед усього набору (4.18%) при збіжності 2.28 — молекула стабільно і відтворювано займає вигідну позу в активному центрі. Сполука P87 посідає друге місце зі середньою оцінкою -34.28 при CV 8.81% та збіжності 2.85. Сполука P96 замикає трійку лідерів із середньою оцінкою -32.44 та з найнижчим CV у вибірці (6.12%) при збіжності 2.60. Ці три сполуки формують пріоритетний список для подальшої розробки.

Сполука P57 є показовим прикладом того, чому ранжування виключно за найкращою оцінкою може бути оманливим: попри те, що для цієї сполуки

було зафіксовано найнижче значення оцінки докінгу серед усіх протестованих сполук (-38.99), її CV 16.18% та збіжність 7.08 виключають її з ідеальної вибірки (критерії CV $< 10\%$ та збіжність < 3). Висока варіативність між запусками може вказувати на декілька конкурентних зон зв'язування або нестабільність оптимальної конформації, що знижує надійність прогнозу.

3.3.4. Аналіз взаємодій топ-3 PROTAC-сполук з трикомпонентним комплексом BRD4–CRBN

Усі три сполуки відтворюють канонічну просторову організацію трикомпонентного комплексу (Рис. 3.12). Фрагмент, спрямований на бромодомен BD1 білка BRD4, у кожному випадку займає ацетил-лізин-зв'язуючу кишеню в канонічній орієнтації [9], формуючи ідентичний набір гідрофобних контактів із WPF-шельфом (Trp81, Pro82, Phe83) та неполярним ядром Val87, Leu92 та Leu94. На протилежному полюсі глутаримідний фрагмент CRBN-ліганда у всіх трьох сполуках фіксується у класичному триптофановому кластері (Trp386/Trp400/Phe402 у сполук P34 і P96; Trp388/Trp402/Phe404 у сполуки P87 — зсув нумерації +2 відображає альтернативну конформацію ліганд-зв'язуючої петлі у відповідній моделі) щонайменше одним водневим зв'язком довжиною $2.9 \text{ \AA}$ з δ -азотом імідазольного кільця гістидину (His378/380) [84]. Отже, усі три сполуки відтворюють базову орієнтацію у кишенях бромодомену BD1 та CRBN, що свідчить про оптимальний підбір довжини лінкерів у кожному з випадків.

Аналіз отриманих моделей свідчить, що сконструйовані лінкерні фрагменти виконують не лише роль інертних просторових спейсерів, але й функціонують як активні елементи інтерфейсу зв'язування, формуючи додаткові стабілізуючі взаємодії в межах трикомпонентного комплексу. Лінкерні ділянки всіх кандидатів проходять уздовж міжбілкового інтерфейсу без стеричних конфліктів, зберігаючи ван-дер-ваальсові контакти з петлею Asn351/353–Pro352/354–His353.

Цікаво, що Сполука 96 формує унікальний водневий зв'язок із карбоксилатним ϵ -киснем залишку Glu377 (2.8 Å), який не спостерігається для інших кандидатів і забезпечує додаткову, четверту, точку полярної фіксації на CRBN-полюсі. Водночас для Сполуки 96 характерне незначне подовження канонічних IMiD-взаємодій із His378 і Trp380 (обидві 2.9 Å) порівняно зі сполукою P34 (2.8 – 2.9 Å). Така тенденція узгоджується з концепцією компенсаторного геометричного балансу: формування додаткового контакту з Glu377 супроводжується локальним перерозподілом напруження у глутаримідному фрагменті, що проявляється у незначному послабленні базових взаємодій. Натомість сполука P34, позбавлена розширеної мережі полярних контактів, реалізує канонічні IMiD-взаємодії у більш оптимальних, геометрично вигідних конфігураціях.

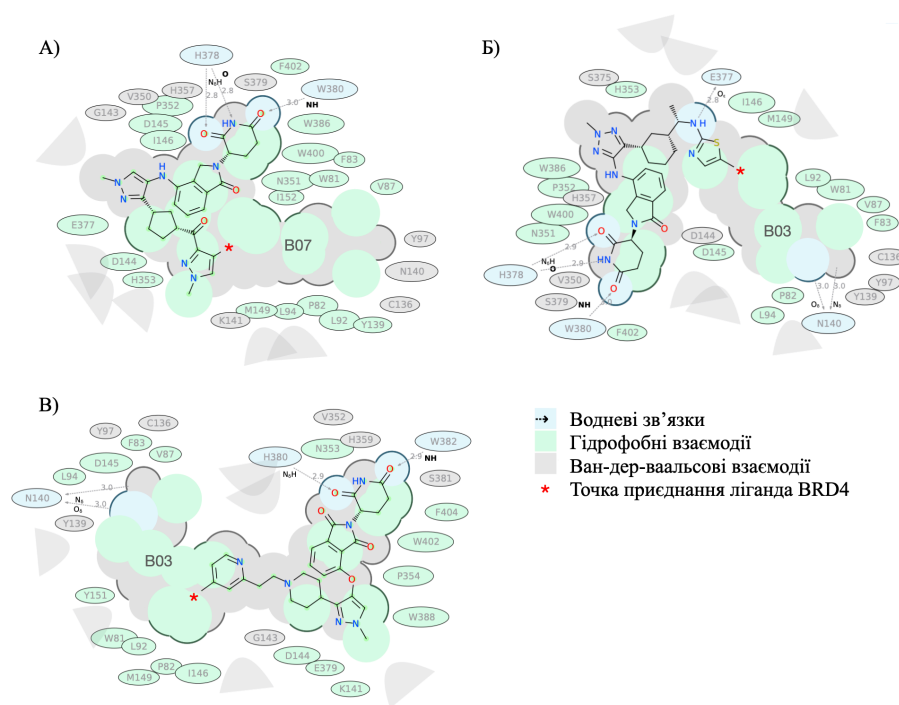


Рисунок 3.12. Порівняльний аналіз міжмолекулярних взаємодій у трикомпонентному комплексі BRD4–PROTAC–CRBN для топ-3 кандидатів:

А) P34; Б) P87; В) P96. Показано схеми контактів із підписаними амінокислотними залишками білків BRD4 і CRBN, що беруть участь у зв'язуванні. Позначено водневі зв'язки, гідрофобні взаємодії, ван-дер-ваальсові контакти та точки приєднання BRD4-ліганду до лінкера.

Узагальнюючи результати структурного аналізу, сполука P34 демонструє найбільш оптимальний профіль взаємодій у межах трикомпонентного комплексу BRD4–CRBN. Її перевага зумовлена реалізацією коротких і геометрично сприятливих канонічних контактів без структурних компромісів. Така збалансована архітектура зв'язування вказує на найвищий потенціал стабільності комплексу серед досліджених кандидатів. Взаємодії сполуки P34 також наочно проілюстровані на Рис. 3.13.

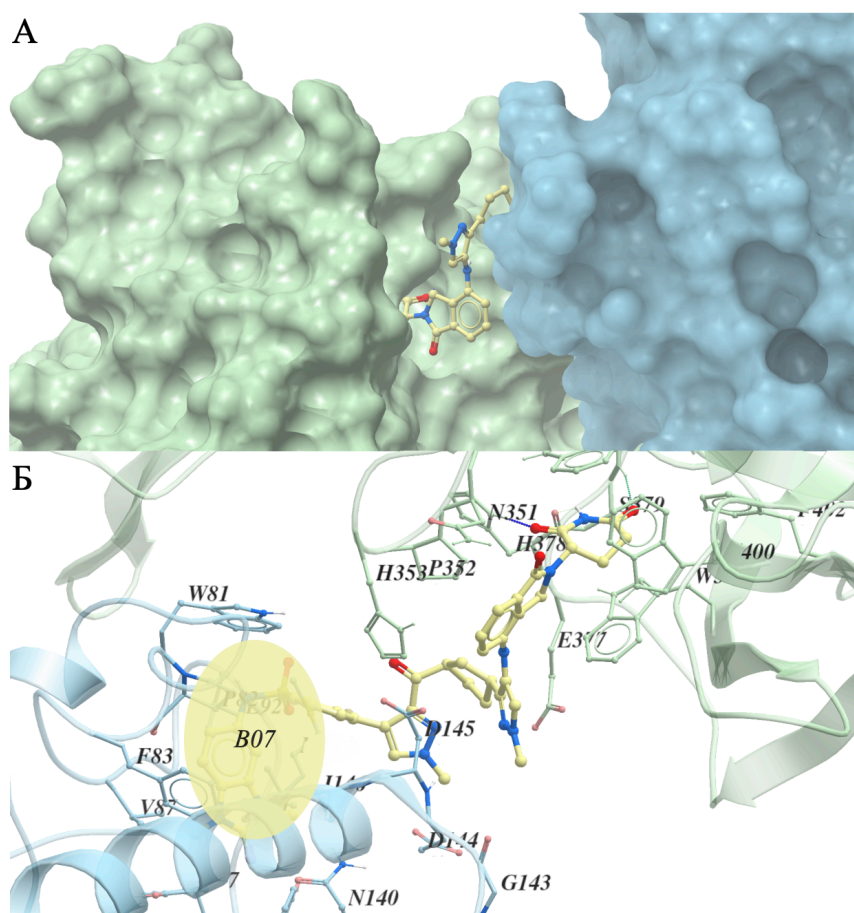


Рисунок 3.13. Комплекс BRD4–P34–CRBN: А) Поверхнева модель трикомпонентного комплексу BRD4–P34–CRBN; Б) Взаємодії сполуки P34 з BRD4 і CRBN із підписаними амінокислотними залишками. BRD4 — блакитний, CRBN — зелений, сполука P34 — жовта. Ліганд до BRD4 замасковано жовтим овалом.

3.4. Обмеження запропонованої методології та практичне значення отриманих результатів

Обмеження ансамблевої моделі зумовлені природою навчальних даних PROTAC-DB, які поєднують результати різнорідних експериментів. Використання бінарної анотації «активний/неактивний» спрощує складну залежність структура–активність, унаслідок чого модель більш ефективно ідентифікує загальні патерни активних PROTAC, ніж прогнозує точні кількісні параметри. Водночас досягнуті метрики демонструють її здатність до надійної пріоритизації сполук, що є ключовим для ранніх етапів дизайну.

Структурно-орієнтована валідація на основі молекулярного докінгу забезпечує оцінку геометричної сумісності з білковими мішенями, однак не враховує повною мірою складні біофізичні процеси, що визначають ефективність деградації *in cellulo*. Це визначає доцільність подальшого застосування методів молекулярної динаміки та вільної енергії зв'язування для уточнення отриманих результатів.

Каскадний характер методології передбачає накопичення невизначеності на кожному етапі моделювання. Водночас така багаторівнева фільтрація дозволяє суттєво звузити простір кандидатів і сфокусувати експериментальні ресурси на найбільш перспективних сполуках.

Незважаючи на перелічені методологічні обмеження, запропонована методологія має конкретну і вимірювану практичну цінність.

По-перше, сформована методологія дозволяє за лічені дні пройти шлях від визначення білкової мішені та пари лігандів до структурно-обґрунтованого короткого списку кандидатів, тоді як аналогічна задача у класичному режимі раціонального дизайну та послідовного синтезу серій займає місяці. Це принципово змінює економіку ранніх етапів проєктування PROTAC.

По-друге, генеративний етап забезпечує систематичне покриття хімічного простору лінкерів, недоступного для мануальної оцінки. Важливо, що згенеровані сполуки не є простими аналогами відомих PROTAC, а

охоплюють оригінальні структурні мотиви, формуючи продуктивну стартову точку для медичної хімії.

По-третє, методологія є архітектурно-агностичною щодо конкретної мішені та E3-ліганду: заміна лігандів та повторне навчання з підкріпленням агенту Link-INVENT дозволяє адаптувати підхід для будь-якої пари POI/E3, що робить його трансляційно цінним за межами модельного завдання деградації BRD4. Структура запропонованої методології — гетерогенний ансамбль як оціночну функцію, Link-INVENT з RL, докінговий фільтр з оцінкою відтворюваності — формує універсальну архітектуру, яку можна адаптувати до нових терапевтичних задач без необхідності у суттєвих структурних змінах.

Нарешті, ідентифіковані пріоритетні сполуки (P34, P87, P96) слід розглядати не як готові кандидати на доклінічні випробування, а як обґрунтовані обчислювально гіпотези, що першими заслуговують на експериментальну перевірку. Така позиція повністю відповідає сучасній парадигмі обчислювально-керованого дизайну, у якій *in silico* інструменти виконують роль ефективного фільтра гіпотез, що раціоналізує, але не замінює експериментальну верифікацію.

ВИСНОВКИ

У дипломній роботі розроблено та валідовано обчислювальну методологію для *de novo* генерації та структурно-орієнтованого відбору PROTAC-молекул, спрямованих на деградацію білка BRD4. Основні результати дослідження:

1. Розроблено гетерогенну ансамблевую модель машинного навчання для бінарної класифікації біологічної активності PROTAC-сполук. Модель інтегрує декілька базових алгоритмів, що працюють із різними типами молекулярних та білкових представлень, а фінальна агрегація передбачень здійснюється за допомогою мета-класифікатора.

2. Проведено валідацію моделі на незалежній тестовій вибірці та підтверджено її узагальнювальну здатність. Оцінювання якості класифікації розробленої моделі підтвердило її ефективність та надійність ($MCC = 0.75$, $ROC-AUC = 0.92$, $\text{точність} = 0.92$, $\text{специфічність} = 0.97$). Додатково застосовано процедуру Y-рандомізації, яка підтвердила відсутність випадкових кореляцій, а аналіз епістемічної невизначеності дозволив оцінити надійність прогнозів ансамблю.

3. Реалізовано оптимізацію апіорної генеративної моделі Link-INVENT на платформі REINVENT4 із використанням навчання з підкріпленням. Розроблена ансамблева модель була інтегрована в генеративний підхід як функція винагороди, що спрямовувала агент на генерацію структур із підвищеною прогнозованою біологічною активністю. Оптимізацію проведено для чотирьох пар лігандів BRD4 та CRBN.

4. Здійснено *de novo* генерацію лінкерів PROTAC на основі RL-оптимізованої Link-INVENT моделі та сформовано вибірку кандидатних сполук для подальшого аналізу. RL-оптимізована вибірка продемонструвала статистично значуще підвищення прогнозованої активності порівняно з апіорною вибіркою. Водночас було збережено структурну різноманітність згенерованих молекул, що свідчить про відсутність колапсу генерації в межах вузького хімічного класу.

5. Проведено порівняльний аналіз фізико-хімічних і прогнозованих фармакокінетичних властивостей апіорної та RL-оптимізованої вибірок PROTAC. Встановлено, що RL-оптимізовані сполуки зберігають параметри MW, logP і TPSA у межах, характерних для bRo5-молекул. Прогнозована проникність Caco-2 дещо знижується після оптимізації, однак залишається в допустимому діапазоні, що вказує на збереження фармакокінетично прийняттого профілю згенерованих кандидатів.

6. Здійснено структурно-орієнтовану оцінку топ-100 RL-генерованих кандидатів методом молекулярного докінгу та ідентифіковано пріоритетні сполуки. Відбір проводився з урахуванням середньої оцінки докінгу, відтворюваності та збіжності результатів між незалежними запусками. За результатами аналізу відібрано три пріоритетні сполуки: P34 (–35.90, CV 4.18 %), P87 (–34.28, CV 8.81 %) та P96 (–32.44, CV 6.12 %). Якісний аналіз взаємодій підтвердив формування канонічного трикомпонентного комплексу BRD4–CRBN і активну роль лінкерів у стабілізації міжбілкового інтерфейсу для топ-3 кандидатів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Patel K, Solomon PD, Walshe JL, Ford DJ, Wilkinson-White L, Payne RJ, Low JKK, Mackay JP. BET-Family Bromodomains Can Recognize Diacetylated Sequences from Transcription Factors Using a Conserved Mechanism. *Biochemistry*. 2021;60(9):648–662. doi: 10.1021/acs.biochem.0c00816.
2. Lindner M, Lisińska D, Kędzierzyńska A, Majchrzak-Celińska A. Targeting BRD4—A Promising Therapeutic Option for Glioblastoma? *International Journal of Molecular Sciences*. 2026; 27(5):2268. <https://doi.org/10.3390/ijms27052268>.
3. Rahman S, Sowa ME, Ottinger M, Smith JA, Shi Y, Harper JW, Howley PM. The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3. *Mol Cell Biol*. 2011;31(13):2641–2652. doi: 10.1128/MCB.01341-10.
4. Divakaran A, Harki DA, Pomerantz WCK. Recent progress and structural analyses of domain-selective BET inhibitors. *Med Res Rev*. 2023;43(4):972–1018. doi: 10.1002/med.21942.
5. Stathis A, Bertoni F. BET Proteins as Targets for Anticancer Treatment. *Cancer Discov*. 2018;8(1):24–36. doi: 10.1158/2159-8290.CD-17-0605.
6. Dey A, Chitsaz F, Abbasi A, Misteli T, Ozato K. The double bromodomain protein Brd4 binds to acetylated chromatin during interphase and mitosis. *Proc Natl Acad Sci U S A*. 2003;100(15):8758–8763. doi: 10.1073/pnas.1433065100.
7. Zhao R, Nakamura T, Fu Y, Lazar Z, Spector DL. Gene bookmarking accelerates the kinetics of post-mitotic transcriptional re-activation. *Nat Cell Biol*. 2011;13(11):1295–1304. doi: 10.1038/ncb2341.
8. Wu SY, Chiang CM. The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation. *J Biol Chem*. 2007;282(18):13141–13145. doi: 10.1074/jbc.R700001200.
9. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O, et al. Selective inhibition of BET bromodomains. *Nature*. 2010;468(7327):1067–1073. doi: 10.1038/nature09504.

10. Pasioneck I, Machowska A, Ridenour JB, Donczew M, Donczew R. The extra-terminal domain drives the role of BET proteins in transcription. *bioRxiv* [Preprint]. 2025. doi: 10.1101/2025.05.12.653540.
11. Jang MK, Mochizuki K, Zhou M, Jeong HS, Brady JN, Ozato K. The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell*. 2005;19(4):523–534. doi: 10.1016/j.molcel.2005.06.027.
12. Chiang CM. Phospho-BRD4: transcription plasticity and drug targeting. *Drug Discov Today Technol*. 2016;19:17–22. doi: 10.1016/j.ddtec.2016.05.003.
13. Donati B, Lorenzini E, Ciarrocchi A. BRD4 and Cancer: going beyond transcriptional regulation. *Mol Cancer*. 2018;17(1):164. doi: 10.1186/s12943-018-0915-9.
14. Yang Z, Yik JH, Chen R, He N, Jang MK, Ozato K, Zhou Q. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell*. 2005;19(4):535–545. doi: 10.1016/j.molcel.2005.06.029.
15. Devaiah BN, Lewis BA, Cherman N, Hewitt MC, Albrecht BK, Robey PG, et al. BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain. *Proc Natl Acad Sci U S A*. 2012;109(18):6927–6932. doi: 10.1073/pnas.1120422109.
16. Sabari BR, Dall’Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*. 2018;361(6400):eaar3958. doi: 10.1126/science.aar3958.
17. Floyd SR, Pacold ME, Huang Q, Clarke SM, Lam FC, Cannell IG, et al. The bromodomain protein Brd4 insulates chromatin from DNA damage signalling. *Nature*. 2013;498(7453):246–250. doi: 10.1038/nature12147.
18. Zuber J, Shi J, Wang E, Rappaport AR, Herrmann H, Sison EA, et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature*. 2011;478(7370):524–528. doi: 10.1038/nature10334.
19. Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi J, Jacobs HM, et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell*. 2011;146(6):904–917. doi: 10.1016/j.cell.2011.08.017.

20. Wang YH, Sui YN, Yan K, Wang LS, Wang F, Zhou JH. BRD4 promotes pancreatic ductal adenocarcinoma cell proliferation and enhances gemcitabine resistance. *Oncol Rep.* 2015;33(4):1699–1706. doi: 10.3892/or.2015.3774.
21. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell.* 2013;153(2):320–334. doi: 10.1016/j.cell.2013.03.036.
22. Huang B, Yang XD, Zhou MM, Ozato K, Chen LF. Brd4 coactivates transcriptional activation of NF-kappaB via specific binding to acetylated RelA. *Mol Cell Biol.* 2009;29(5):1375–1387. doi: 10.1128/MCB.01365-08.
23. Zou Z, Huang B, Wu X, Zhang H, Qi J, Bradner J, et al. Brd4 maintains constitutively active NF-κB in cancer cells by binding to acetylated RelA. *Oncogene.* 2014;33(18):2395–2404. doi: 10.1038/onc.2013.179.
24. Ding N, Hah N, Yu RT, Sherman MH, Benner C, Leblanc M, et al. BRD4 is a novel therapeutic target for liver fibrosis. *Proc Natl Acad Sci U S A.* 2015;112(51):15713–15718. doi: 10.1073/pnas.1522163112.
25. Wei Q, Gan C, Sun M, Xie Y, Liu H, Xue T, et al. BRD4: an effective target for organ fibrosis. *Biomark Res.* 2024;12(1):92. doi: 10.1186/s40364-024-00641-6.
26. Jang MK, Shen K, McBride AA. Papillomavirus genomes associate with BRD4 to replicate at fragile sites in the host genome. *PLoS Pathog.* 2014;10(5):e1004117. doi: 10.1371/journal.ppat.1004117.
27. Kulikowski E, Rakai BD, Wong NCW. Inhibitors of bromodomain and extraterminal proteins for treating multiple human diseases. *Med Res Rev.* 2021;41(1):223–245. doi: 10.1002/med.21730.
28. Winter GE, Mayer A, Buckley DL, Erb MA, Roderick JE, Vittori S, et al. BET Bromodomain Proteins Function as Master Transcription Elongation Factors Independent of CDK9 Recruitment. *Mol Cell.* 2017;67(1):5–18. doi: 10.1016/j.molcel.2017.06.004.
29. Zengerle M, Chan KH, Ciulli A. Selective Small Molecule Induced Degradation of the BET Bromodomain Protein BRD4. *ACS Chem Biol.* 2015;10(8):1770–1777. doi: 10.1021/acscchembio.5b00216.

30. Lai AC, Crews CM. Induced protein degradation: an emerging drug discovery paradigm. *Nat Rev Drug Discov.* 2017;16(2):101–114. doi: 10.1038/nrd.2016.211.
31. Archasappawat S, Jacques J, Lee E, Hwang CI. BRD2 Upregulation as a Pan-Cancer Adaptive Resistance Mechanism to BET Inhibition. *bioRxiv [Preprint]*. 2025. doi: 10.1101/2025.08.30.673282.
32. Kurimchak AM, Shelton C, Duncan KE, Johnson KJ, Brown J, O'Brien S, et al. Resistance to BET Bromodomain Inhibitors Is Mediated by Kinome Reprogramming in Ovarian Cancer. *Cell Rep.* 2016;16(5):1273–1286. doi: 10.1016/j.celrep.2016.06.091.
33. Shu S, Lin CY, He HH, Witwicki RM, Tabassum DP, Roberts JM, et al. Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer. *Nature.* 2016;529(7586):413–417. doi: 10.1038/nature16508.
34. Sakamoto KM, Kim KB, Kumagai A, Mercurio F, Crews CM, Deshaies RJ. Protacs: chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proc Natl Acad Sci U S A.* 2001;98(15):8554–8559. doi: 10.1073/pnas.141230798.
35. Békés M, Langley DR, Crews CM. PROTAC targeted protein degraders: the past is prologue. *Nat Rev Drug Discov.* 2022;21(3):181–200. doi: 10.1038/s41573-021-00371-6.
36. Hu Z, Crews CM. Recent Developments in PROTAC-Mediated Protein Degradation: From Bench to Clinic. *Chembiochem.* 2022;23(2):e202100270. doi: 10.1002/cbic.202100270.
37. Gadd MS, Testa A, Lucas X, Chan KH, Chen W, Lamont DJ, et al. Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nat Chem Biol.* 2017;13(5):514–521. doi: 10.1038/nchembio.2329.
38. Douglass EF Jr, Miller CJ, Sparer G, Shapiro H, Spiegel DA. A comprehensive mathematical model for three-body binding equilibria. *J Am Chem Soc.* 2013;135(16):6092–6099. doi: 10.1021/ja311795d.

39. Chirnomas D, Hornberger KR, Crews CM. Protein degraders enter the clinic — a new approach to cancer therapy. *Nat Rev Clin Oncol*. 2023;20(4):265–278. doi: 10.1038/s41571-023-00736-3.
40. Bondeson DP, Smith BE, Burslem GM, Buhimschi AD, Hines J, Jaime-Figueroa S, et al. Lessons in PROTAC Design from Selective Degradation with a Promiscuous Warhead. *Cell Chem Biol*. 2018;25(1):78–87. doi: 10.1016/j.chembiol.2017.09.010.
41. Lu J, Qian Y, Altieri M, Dong H, Wang J, Raina K, et al. Hijacking the E3 Ubiquitin Ligase Cereblon to Efficiently Target BRD4. *Chem Biol*. 2015;22(6):755–763. doi: 10.1016/j.chembiol.2015.05.009.
42. Winter GE, Buckley DL, Paulk J, Roberts JM, Souza A, Dhe-Paganon S, et al. Phthalimide conjugation as a strategy for in vivo target protein degradation. *Science*. 2015;348(6241):1376–1381. doi: 10.1126/science.aab1433.
43. Qiang B, Shi W, Song Y, Wu M. PROflow: An iterative refinement model for PROTAC-induced structure prediction. *arXiv [Internet]*. 2024;2405.06654. Available from: <https://arxiv.org/abs/2405.06654>.
44. Raina K, Lu J, Qian Y, Altieri M, Gordon D, Rossi AM, et al. PROTAC-induced BET protein degradation as a therapy for castration-resistant prostate cancer. *Proc Natl Acad Sci U S A*. 2016;113(26):7124–7129. doi: 10.1073/pnas.1521738113.
45. Rodríguez-Gimeno A, Galdeano C. Drug Discovery Approaches to Target E3 Ligases. *Chembiochem*. 2025;26(1):e202400656. doi: 10.1002/cbic.202400656.
46. Cui J, Qu Q, Pan M. Targeted Protein Degradation by Repurposing Transmembrane E3 Ubiquitin Ligases. *Chembiochem*. 2026;27(5):e202500870. doi: 10.1002/cbic.202500870.
47. Dong Y, Ma T, Xu T, Feng Z, Li Y, Song L, et al. Characteristic roadmap of linker governs the rational design of PROTACs. *Acta Pharm Sin B*. 2024;14(10):4266–4295. doi: 10.1016/j.apsb.2024.04.007.
48. Cecchini C, Tardy S, Scapozza L. Linkers as Game-changers in PROTAC Technology: Emphasizing General Trends in PROTAC Pharmacokinetics for their

Rational Design. *Chimia (Aarau)*. 2022;76(4):341–345. doi: 10.2533/chimia.2022.341.

49. Lebraud H, Wright DJ, Johnson CN, Heightman TD. Protein Degradation by In-Cell Self-Assembly of Proteolysis Targeting Chimeras. *ACS Cent Sci*. 2016;2(12):927–934. doi: 10.1021/acscentsci.6b00280.

50. Garcia Jimenez D, Ermondi G, Jandova Z, Vallaro M, Caron G, Arnhof H. Linker Methylation as a Strategy to Enhance PROTAC Oral Bioavailability. *J Med Chem*. 2025;68(15):16666–16677. doi: 10.1021/acs.jmedchem.5c01497.

51. Wurz RP, Dellamaggiore K, Dou H, Javier N, Lo MC, McCarter JD, et al. A 'Click Chemistry Platform' for the Rapid Synthesis of Bispecific Molecules for Inducing Protein Degradation. *J Med Chem*. 2018;61(2):453–461. doi: 10.1021/acs.jmedchem.6b01781.

52. Churcher I. Protac-Induced Protein Degradation in Drug Discovery: Breaking the Rules or Just Making New Ones? *J Med Chem*. 2018;61(2):444–452. doi: 10.1021/acs.jmedchem.7b01272.

53. Ciechanover A. Proteolysis: from the lysosome to ubiquitin and the proteasome. *Nat Rev Mol Cell Biol*. 2005;6(1):79–87. doi: 10.1038/nrm1552.

54. Finley D. Recognition and processing of ubiquitin-protein conjugates by the proteasome. *Annu Rev Biochem*. 2009;78:477–513. doi: 10.1146/annurev.biochem.78.081507.101607.

55. Jiang H, Xiong H, Gu SX, Wang M. E3 ligase ligand optimization of clinical PROTACs. *Front Chem*. 2023;11:1098331. doi: 10.3389/fchem.2023.1098331.

56. Ishida T, Ciulli A. E3 Ligase Ligands for PROTACs: How They Were Found and How to Discover New Ones. *SLAS Discov*. 2021;26(4):484–502. doi: 10.1177/2472555220965528.

57. Nafie MS, Diab MK, Yassen ASA, Elshamy AM, El Tohamy MR, Tawfik HO, et al. Next-Generation Proteolysis-Targeting Chimeras in Precision Oncology: Multifunctional Designs, Emerging Modalities, and Translational Prospects. *Drug Dev Res*. 2025;86(8):e70192. doi: 10.1002/ddr.70192.

58. Liu Y, Yang J, Wang T, Luo M, Chen Y, Chen C, et al. Expanding PROTACtable genome universe of E3 ligases. *Nat Commun.* 2023;14(1):6509. doi: 10.1038/s41467-023-42233-2.
59. Diehl CJ, Ciulli A. Discovery of small molecule ligands for the von Hippel-Lindau (VHL) E3 ligase and their use as inhibitors and PROTAC degraders. *Chem Soc Rev.* 2022;51(19):8216–8257. doi: 10.1039/d2cs00387b.
60. Yin L, Shu P, Peng X. Context Matters: E3 Ligase-Ligand Pairing Strategies for Optimized PROTAC Performance. *Protein Cell.* 2025. doi: 10.1093/procel/pwaf107.
61. Zaidman D, Prilusky J, London NJM. PROsettaC: Rosetta Based Modeling of PROTAC Mediated Ternary Complexes. *J Chem Inf Model.* 2020;60(10):4894–4903. doi: 10.1021/acs.jcim.0c00589.
62. Ignatov M, Jindal A, Kotelnikov S, et al. High Accuracy Prediction of PROTAC Complex Structures. *J Am Chem Soc.* 2023;145(13):7123–7135. doi: 10.1021/jacs.2c09387.
63. Nori D, Coley CW, Mercado R. De Novo PROTAC Design Using Graph-Based Deep Generative Models. *arXiv preprint arXiv:2211.02660.* 2022.
64. Li F, Hu Q, Zhang X, Sun R, Liu Z, Wu S, et al. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. *Nat Commun.* 2022;13(1):7133. doi: 10.1038/s41467-022-34807-3.
65. Ribes S, Nittinger E, Tyrchan C, Mercado R. Modeling PROTAC degradation activity with machine learning. *Artif Intell Life Sci.* 2024;6:100104. doi: 10.1016/j.ailsci.2024.100104.
66. Zheng S, Tan Y, Wang Z, Li C, Zhang Z, Sang X, et al. Accelerated rational PROTAC design via deep learning and molecular simulations. *Nat Mach Intell.* 2022;4(9):739–748. doi: 10.1038/s42256-022-00527-y.
67. Li B, Ran T, Chen H. 3D based generative PROTAC linker design with reinforcement learning. *Brief Bioinform.* 2023;24(5):bbad323. doi: 10.1093/bib/bbad323.

68. Weng G, Gao H, Li C, Liang X, Hu D, Chen H, et al. PROTAC-DB: an integrated database for proteolysis targeting chimeras. *Nucleic Acids Res.* 2021;49(D1):D1381–D1387. doi: 10.1093/nar/gkaa846.
69. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Nayfach S, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 2023;379(6637):1123–1130. doi: 10.1126/science.ade2574.
70. Gobbi A, Poppinger D. Genetic optimization of combinatorial libraries. *Biotechnol Bioeng.* 1998;61(1):47–54. doi: 10.1002/(sici)1097-0290(199824)61:1<47::aid-bit9>3.0.co;2-z.
71. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50(5):742–754. doi: 10.1021/ci100050t.
72. Landrum G. RDKit: Open-source cheminformatics software. 2016. Available from: <http://www.rdkit.org>.
73. Ahmad W, Simon E, Iyer S, Schneider J. ChemBERTa-2: Towards Safe and Effective Molecular Self-Supervised Learning. arXiv preprint arXiv:2209.01712. 2022.
74. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785–794. doi: 10.1145/2939672.2939785.
75. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. doi: 10.1023/A:1010933404324.
76. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273–297. doi: 10.1007/BF00994018.
77. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory.* 1967;13(1):21–27. doi: 10.1109/TIT.1967.1053964.
78. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2019. p. 2623–2631. doi: 10.1145/3292500.3330701.
79. Wolpert DH. Stacked generalization. *Neural Netw.* 1992;5(2):241–259. doi: 10.1016/S0893-6080(05)80023-1.

80. Golbraikh A, Tropsha A. Beware of q^2 ! *J Mol Graph Model*. 2002;20(4):269–276. doi: 10.1016/S1093-3263(01)00123-1.
81. Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, et al. Reinvent 4: modern AI-driven generative molecule design. *J Cheminform*. 2024;16:20. doi: 10.1186/s13321-024-00812-z.
82. Guo J, Knuth F, Margreitter C, Janet JP, Papadopoulos K, Engkvist O, et al. Link-INVENT: generative linker design with reinforcement learning. *Digital Discovery*. 2023;2(2):392–408. doi: 10.1039/D2DD00115B.
83. Swanson K, Walther P, Leitz J, Mukherjee S, Wu JC, Shivnaraine RV, et al. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics*. 2024;40(7):btae416. doi: 10.1093/bioinformatics/btae416.
84. Fischer E, Böhm K, Lydeard J, et al. Structure of the DDB1–CRBN E3 ubiquitin ligase in complex with thalidomide. *Nature*. 2014;512:49–53. doi: 10.1038/nature13527.

Додаток 1.

Конфігураційний файл для оптимізації Link-INVENT за допомогою навчання з підкріпленням

```
# =====
# REINVENT 4: Link-INVENT PROTAC Optimization
# =====

run_type = "staged_learning"
device = "cpu"
tb_logdir = "BRD4_project/tb_logs_protac_run_full_3"
json_out_config = "_staged_learning.json"

[parameters]
summary_csv_prefix = "protac_linkinvent_full_3"

# LinkInvent setup
prior_file = "priors/linkinvent.prior"
agent_file = "priors/linkinvent.prior"
smiles_file = "BRD4_project/warheads_final_2.smi"
batch_size = 128

[learning_strategy]
type = "dap"
sigma = 256
rate = 0.0001

# =====
# STAGE 1: PROTAC Active Probability
# =====

[[stage]]
chkpt_file = 'BRD4_project/protac_active_full_3.chkpt'
termination = "simple"
max_score = 0.75
min_steps = 50
max_steps = 200

[stage.scoring]
type = "geometric_mean"

[[stage.scoring.component]]
[stage.scoring.component.PROTACPredictor]
[[stage.scoring.component.PROTACPredictor.endpoint]]
name = "PROTAC_Active_Prob"
weight = 1.0 # 100% focus on the ensemble meta-learner output

params.model_dir = "/home/va.melnik@ad.loc/chemai_storage/personal_data/varia_mel/PROTAC_model/
PROTAC_final/models/final_model_FULL"
params.target = "BRD4"
params.e3 = "CRBN"
params.emb_base_dir = "/home/va.melnik@ad.loc/chemai_storage/personal_data/varia_mel/
PROTAC_model/PROTAC_final/data/processed"
```

Додаток 2.

Плагін інтеграції ансамблевої моделі в цикл навчання з підкріпленням

```

from __future__ import annotations

__all__ = ["PROTACPredictor"]
from typing import List
import logging
import os
import joblib
import numpy as np
from pydantic.dataclasses import dataclass

from .component_results import ComponentResults
from .add_tag import add_tag
from reinvent.scoring.utils import suppress_output
from ..normalize import normalize_smiles

# Import your custom modules
# Ensure that the directory containing 'src' is in your PYTHONPATH
from src.chem import (
    batch_process_2D_descriptors,
    batch_process_morgan,
    batch_process_pharmacophore,
    batch_process_chemberta_mtr,
)
from src.trainer import PROTACTrainer

logger = logging.getLogger("reinvent")

@add_tag("__parameters")
@dataclass
class Parameters:
    """Parameters for the PROTAC Predictor scoring component"""
    model_dir: List[str]
    target: List[str]
    e3: List[str]
    emb_base_dir: List[str]

@add_tag("__component")
class PROTACPredictor:
    def __init__(self, params: Parameters):
        self.smiles_type = "rdkit_smiles"

        self.model_dir = params.model_dir[0]
        self.target = params.target[0]
        self.e3 = params.e3[0]
        self.emb_base_dir = params.emb_base_dir[0]

        logger.info(f"Initializing PROTAC Predictor with Target: {self.target}, E3: {self.e3}")

        # 1. Load Processors & Model
        logger.info(f>Loading processors and meta-ensemble from {self.model_dir}...")

```

```

        self.processors = joblib.load(os.path.join(self.model_dir,
"multi_view_processors_FULL.joblib"))
        self.trainer = PROTACTrainer.load_models(self.model_dir)

# 2. Load Pre-computed Protein Embeddings
target_emb_path = os.path.join(self.emb_base_dir, f"{self.target}_embedding.npy")
e3_emb_path = os.path.join(self.emb_base_dir, f"{self.e3}_embedding.npy")

if not os.path.exists(target_emb_path):
    raise FileNotFoundError(f"Target embedding not found at {target_emb_path}")
if not os.path.exists(e3_emb_path):
    raise FileNotFoundError(f"E3 embedding not found at {e3_emb_path}")

self.t_emb_single = np.load(target_emb_path)
self.e_emb_single = np.load(e3_emb_path)

@normalize_smiles
def __call__(self, smiles: List[str]) -> ComponentResults:
    num_mols = len(smiles)

    # Handle empty batches
    if num_mols == 0:
        return ComponentResults([np.array([])])

    with suppress_output():
        # 3. Extract Raw Chemical Features
        # use_cache=False to prevent I/O bottlenecks during generation of novel SMILES
        X_phar = np.array(batch_process_pharmacophore(smiles, use_cache=False))
        X_morg = np.array(batch_process_morgan(smiles, radius=3, n_bits=2048))
        X_svm_d = np.array(batch_process_2D_descriptors(smiles))
        X_mtr = np.array(batch_process_chemberta_mtr(smiles, use_cache=False))

# 4. Tile embeddings to match the batch size
t_tr = np.tile(self.t_emb_single, (num_mols, 1))
e_tr = np.tile(self.e_emb_single, (num_mols, 1))

# 5. Apply Transformers
t_proc = self.processors['pca_t'].transform(self.processors['sc_t'].transform(t_tr))
e_proc = self.processors['pca_e'].transform(self.processors['sc_e'].transform(e_tr))
prot_stack = np.hstack([t_proc, e_proc])

views_infer = {
    "XGBoost": np.hstack([prot_stack, self.processors['vt_phar'].transform(X_phar)]),
    "RandomForest": np.hstack([prot_stack, self.processors['sc_m3'].transform(X_morg)]),
    "SVM": np.hstack([prot_stack, self.processors['sc_svm'].transform(X_svm_d)]),
    "KNN": np.hstack([t_proc, e_proc,
self.processors['pca_mtr'].transform(self.processors['sc_mtr'].transform(X_mtr))])
}

# 6. Predict Active Probability
_, meta_probs, _ = self.trainer.ensemble_predict(views_infer)
scores = np.array(meta_probs, dtype=float)
return ComponentResults([scores])

```

Додаток 3.

Результати Y-рандомізації (матриця помилок)

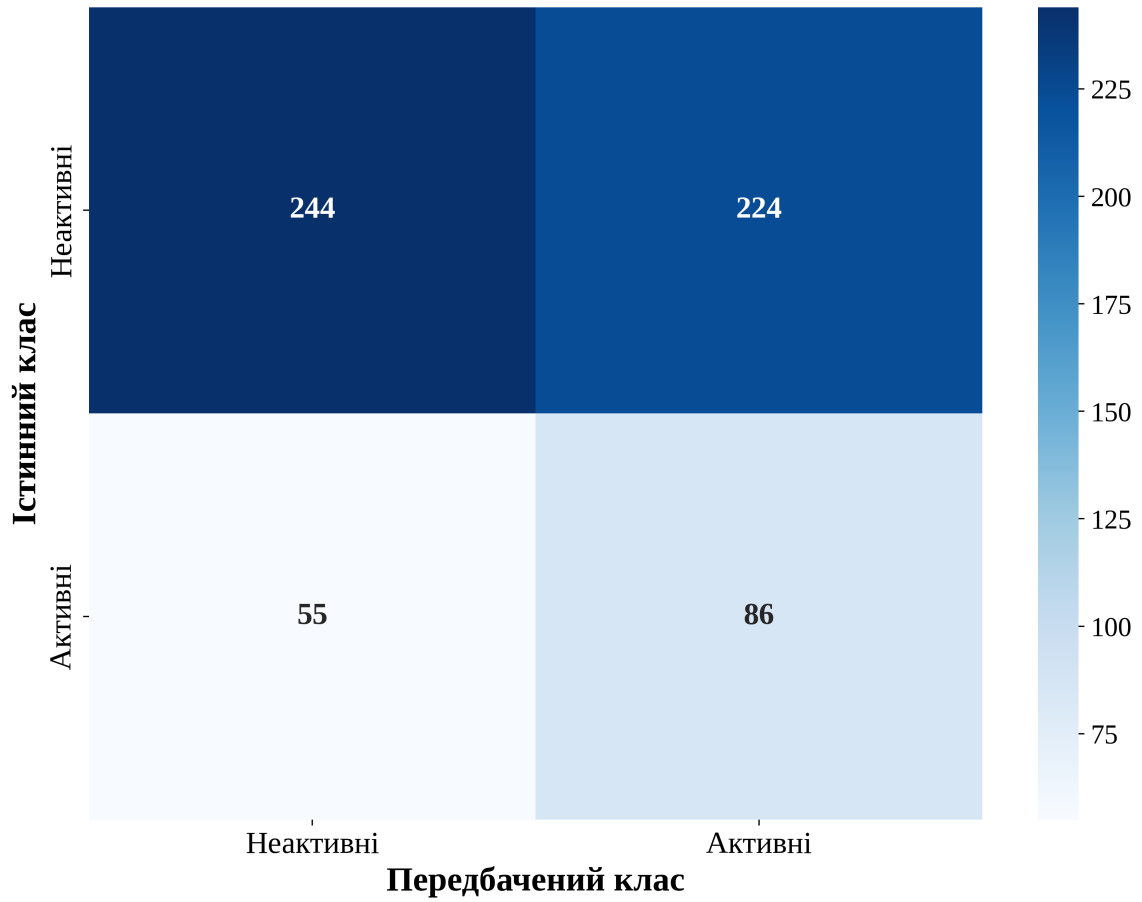


Рисунок. Матриця помилок фінальної ансамблевої моделі, що була натренована на випадково перемішаних даних (Y-рандомізації) незалежної тестової вибірки.

Додаток 4.

Результати Y-рандомізації (епістемічна невизначеність)

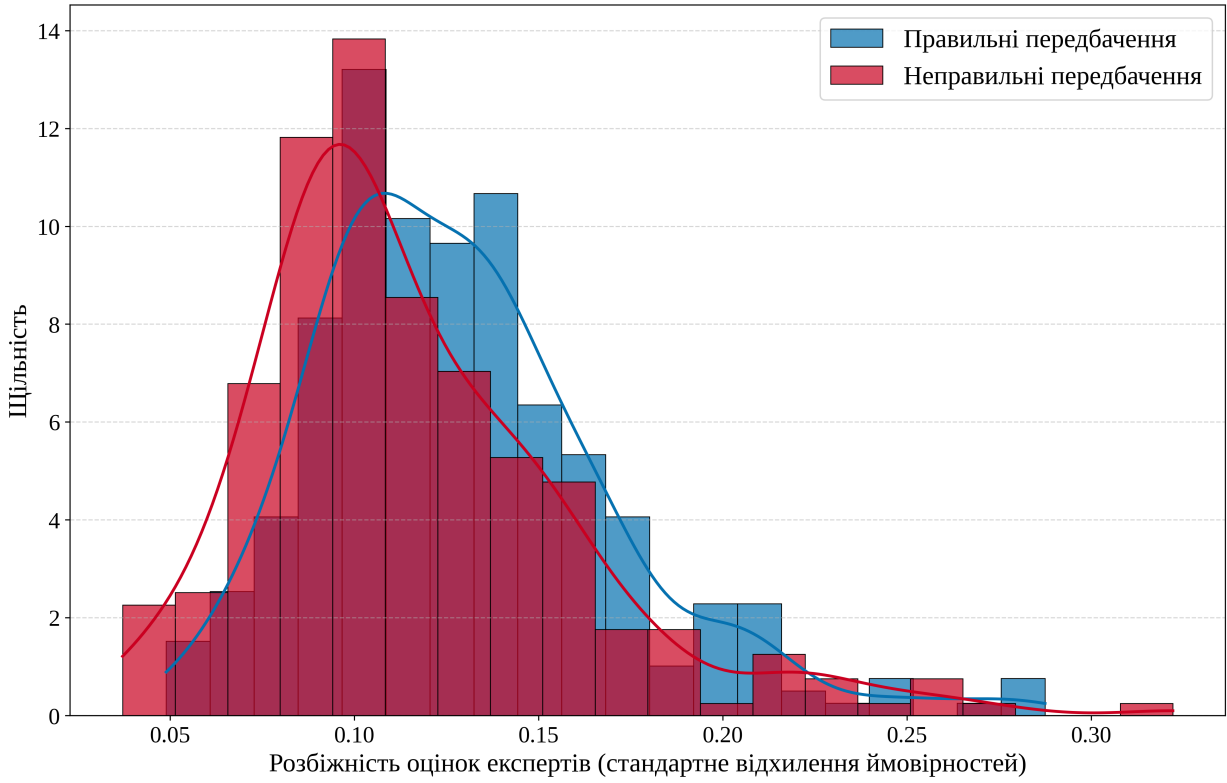


Рисунок. Розподіл епістемічної невизначеності (стандартного відхилення прогнозів базових моделей) для випадково перемішаних даних (Y-рандомізації) незалежної тестової вибірки.

Додаток 5.

Траєкторія RL-оптимізації біологічної активності PROTAC-сполук

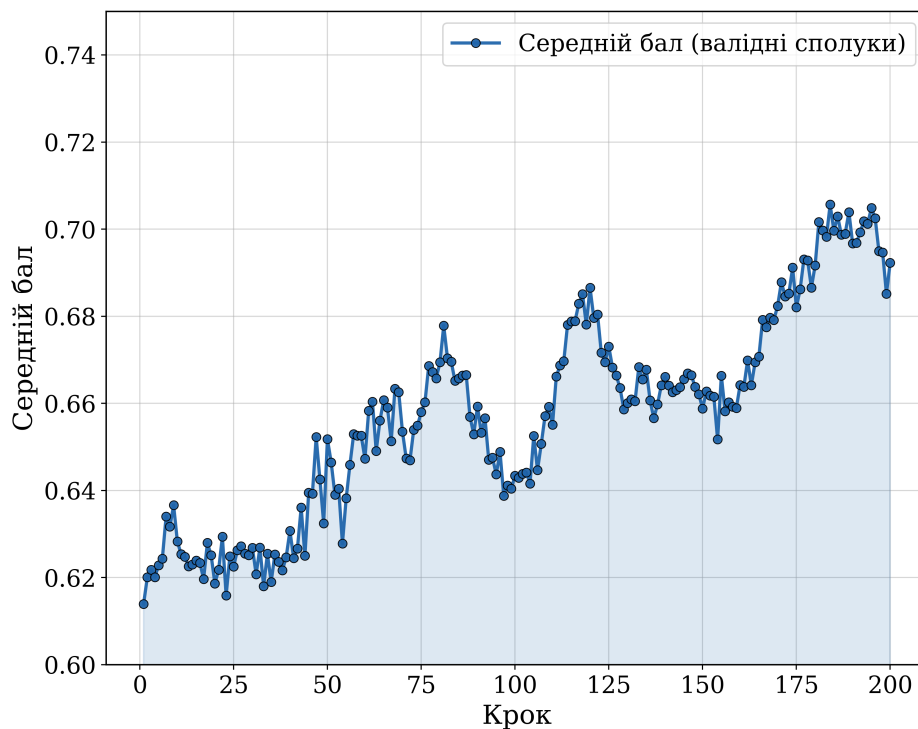


Рисунок. Динаміка середнього значення функції винагороди валідних PROTAC-сполук у процесі RL-оптимізації.