Інститут фізики Національної академії наук України
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

ВОЙЦІЦЬКИЙ ТАРАС ВАСИЛЬОВИЧ

УДК 577.32, 539.196.3, 544.165, 004.85

ДИСЕРТАЦІЯ

ВИЗНАЧЕННЯ ПАРАМЕТРІВ НЕКОВАЛЕНТНИХ ВЗАЄМОДІЙ МАЛИХ ОРГАНІЧНИХ МОЛЕКУЛ ІЗ ПРОТЕЇНАМИ НА ОСНОВІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

104 – ФІЗИКА ТА АСТРОНОМІЯ
10 – ПРИРОДНИЧІ НАУКИ

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело _____Т.В. Войціцький

Науковий керівник Єсилевський Семен Олександрович, доктор фізико-математичних наук, провідний науковий співробітник

Київ - 2026

АНОТАЦІЯ

Войціцький Т.В. Визначення параметрів нековалентних взаємодій малих органічних молекул із протеїнами на основі моделей машинного навчання. - Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 104 - Фізика та астрономія (10 - Природничі науки). - Інститут фізики НАН України, Київ, 2026.

Дисертаційна робота присвячена розробці та оптимізації методів машинного навчання для передбачення кількісних характеристик біологічної активності та структурних параметрів молекулярних комплексів протеїн-мала органічна молекула. Дослідження спрямоване на розв’язання актуальної науково-прикладної проблеми підвищення точності та швидкості обчислювального моделювання нековалентних взаємодій у біологічних системах, що має фундаментальне значення для фізики біологічних систем та практичне застосування у процесі розробки нових лікарських засобів.

Взаємодія між протеїнами та малими органічними молекулами (лігандами) є фундаментальним фізичним процесом, що лежить в основі функціонування живих систем. Розуміння природи цієї взаємодії має критичне практичне значення для раціонального пошуку ліків, який передбачає дослідження мільйонів малих молекул з метою ідентифікації таких, що можуть змінювати біологічну активність молекулярних мішеней (найчастіше протеїнів) внаслідок зв’язування з ними. Віртуальний скринінг (обчислювальна фільтрація великих бібліотек сполук) є ключовим інструментом на початкових етапах розробки лікарських засобів. Однак класичні обчислювальні підходи, зокрема молекулярний докінг, значною мірою досягли плато своєї практичної точності, що зумовлює актуальність розробки нових методів на основі машинного навчання.

Метою роботи є розробка та оптимізація методів машинного навчання для передбачення кількісних характеристик біологічної активності та структурних характеристик молекулярних комплексів протеїн-мала органічна молекула на основі аналізу закономірностей нековалентних взаємодій та фізично обґрунтованого вибору вхідних репрезентацій, навчальних даних і архітектур штучних нейронних мереж. **Об'єктом дослідження** є взаємодія між протеїнами та малими органічними молекулами та її зв'язок із кількісними характеристиками їх біологічної активності та просторовою структурою утворених молекулярних комплексів. **Предметом дослідження** є моделі машинного навчання для передбачення кількісних характеристик біологічної активності та просторової структури молекулярних комплексів протеїн-мала органічна молекула.

У фізичному контексті задача передбачення біологічної активності та структури молекулярних комплексів зводиться до апроксимації багатовимірної залежності між координатами та типами атомів системи та її кількісними характеристиками зв'язування, якість якої принципово залежить від повноти представлення статистик різних порядків в тренувальних даних моделей машинного навчання - за прямою аналогією до ієрархії функцій розподілу в статистичній фізиці багаточастинкових систем.

У роботі використано методи обчислювальної біофізики для встановлення фізичної природи процесів молекулярного розпізнавання, аналізу нековалентних взаємодій та обґрунтування вибору репрезентацій вхідних даних, а також методи чисельного комп'ютерного моделювання для реалізації розроблених підходів у вигляді алгоритмів та моделей машинного навчання.

Дисертація складається з чотирьох розділів. У **першому розділі** проведено огляд сучасного стану досліджень взаємодій між протеїнами та малими органічними молекулами, який охоплює фізико-хімічні механізми зв'язування (кінетику, термодинаміку та моделі зв'язування), експериментальні методи визначення структури комплексів та афінності (сили) зв'язування, класичні обчислювальні

підходи моделювання взаємодій, а також роль сучасних методів машинного навчання, включаючи графові нейронні мережі, трансформери та дифузійні моделі.

У **другому розділі** представлено розробку графової нейронної мережі 3DProtDTA для передбачення афінності зв'язування лігандів з протеїнами. Запропоновано новий підхід до кодування структури протеїну на рівні амінокислотних залишків у вигляді графів, що інтегрує просторову інформацію, отриману за допомогою AlphaFold, разом із фізико-хімічними властивостями амінокислот та еволюційними характеристиками. Проведено систематичний підбір гіперпараметрів та порівняння архітектур графових нейронних мереж, методів графового пулінгу та типів вхідних репрезентацій. Модель 3DProtDTA перевершує наявні аналоги на 1-20% за загальноприйнятими критеріями (середньоквадратичне відхилення, коефіцієнт конкордації та коефіцієнт кореляції) на еталонних наборах даних Davis та KIBA. Також досліджено здатність моделі узагальнювати результати на "холодних" білкових мішенях, що не зустрічалися під час навчання.

У **третьому розділі** проведено детальний статистичний аналіз геометричних параметрів основних типів нековалентних взаємодій у системі протеїн-мала молекула, включаючи гідрофобні контакти, взаємодії з π системами, водневі та галогенні зв'язки та катіон-аніонні взаємодії. На основі виявлених статистичних закономірностей розроблено алгоритм генерації штучних комплексів сайт зв'язування протеїну-ліганд, які імітують фізико-хімічні закономірності реальних (експериментально визначених) комплексів. Розроблений підхід розв'язує проблему обмеженої кількості наявних експериментальних даних для тренування моделей машинного навчання. Ефективність підходу підтверджена на вдосконаленій генеративній дифузійній моделі PocketCFDM: застосування штучних комплексів для аугментації тренувальних даних дозволило підвищити частку правильних передбачень положення лігандів до 10% за критерієм середньоквадратичного відхилення координат атомів у тривимірному просторі (RMSD) та знизити частку передбачень зі стеричними зіткненнями на 3-6%.

У **четвертому розділі** представлено створення ArtiDock - методу молекулярного докінгу на основі глибинного навчання, оптимізованого для високопродуктивного віртуального скринінгу. ArtiDock поєднує обчислювально ефективну архітектуру штучних нейронних мереж із фізично обґрунтованою післяобробкою передбачень, що дозволяє досягти пропускної здатності на рівні або кращої за найшвидші традиційні інструменти. На тестовому наборі даних PLINDER ArtiDock передбачає положення лігандів із медіаною RMSD 2.0 Å та медіаною частки правильно відтворених міжатомних контактів (lDDT-PLI) 0.71, порівняно з 2.8 Å та 0.64 для найефективнішого класичного методу Glide. ArtiDock стабільно перевершує класичні методи в різноманітних реалістичних сценаріях, включаючи білкові сайти з кофакторами, іонами та молекулами води. На розділеному за часом публікації даних тестовому наборі PoseX ArtiDock демонструє конкурентоспроможну точність порівняно з найефективнішими інструментами машинного навчання, досягаючи 73.5% успішних передбачень з RMSD < 2 Å та найкращих результатів серед усіх порівняних підходів (63.9%) з урахуванням хімічної валідності передбачень. У сценарії, де вхідні конформації протеїнів були отримані з експериментальних структур незв'язаних з лігандом, ArtiDock зберігає перевагу, досягаючи 62.4% точних та валідних передбачень.

Наукова новизна отриманих результатів. Розроблено графову штучну нейронну мережу 3DProtDTA для передбачення афінності зв’язування лігандів, яка кодує структури протеїнів на рівні амінокислотних залишків та навчається на структурах, передбачених методом AlphaFold; модель є точнішою за наявні підходи машинного навчання на 1-20%. Запропоновано новий підхід до збільшення кількості тренувальних даних для моделей молекулярного докінгу шляхом генерації штучних комплексів протеїн-ліганд, які імітують статистичні закономірності нековалентних взаємодій реальних комплексів, що розв’язує проблему обмеженої кількості наявних експериментальних даних. Підтверджено, що тренування моделей із використанням штучних комплексів дозволяє підвищити частку правильних передбачень до 10%. Створено новий метод молекулярного докінгу ArtiDock, який поєднує точність

сучасних штучних нейронних мереж зі швидкістю, необхідною для віртуального скринінгу великих хімічних баз даних; метод передбачає положення та конформації лігандів на 9-10% краще за класичні методи за критерієм RMSD та на 2-3% за критеріями RMSD та хімічної коректності порівняно з іншими підходами машинного навчання.

Практичне значення отриманих результатів. Розроблені підходи дозволяють прискорити процес *in silico* пошуку нових лікарських засобів шляхом швидкого та точного передбачення афінності та геометрії зв'язування між терапевтичними цілями (протеїнами) та потенційними ліками (малими органічними молекулами). Методологія генерації штучних комплексів протеїн-ліганд дозволяє підвищити точність моделей машинного навчання для роботи з новими класами протеїнів та лігандів, для яких відсутні експериментально визначені дані. Результати досліджень стали основою програмного забезпечення, що застосовується у проєктах віртуального скринінгу.

Ключові слова: нековалентні взаємодії, електростатичні взаємодії, міжфазні границі, ефекти діелектричного середовища, малі органічні молекули, біоактивні молекули, біомолекули, біополімери, обчислювальна хімія, обчислювальна біофізика, молекулярне моделювання, віртуальний скринінг, машинне навчання, глибинне навчання, графові нейронні мережі

ABSTRACT

T. Voitsitskyi. Determination of noncovalent interaction parameters for small organic molecule-protein systems using machine learning models. - Qualifying scientific work on rights of a manuscript.

Thesis for the degree of Doctor of Philosophy on specialty 104 - Physics and Astronomy (10 - Natural Sciences). - Institute of Physics of National Academy of Sciences of Ukraine, Kyiv, 2026.

The dissertation is devoted to the development and optimization of machine learning methods for predicting quantitative characteristics of biological activity and structural parameters of protein-small organic molecule complexes. The study aims to address the relevant scientific and applied problem of improving the accuracy and speed of computational modeling of non-covalent interactions in biological systems, which is of fundamental importance for the physics of biological systems and has practical applications in the drug discovery process.

The interaction between proteins and small organic molecules (ligands) is a fundamental physical process underlying the functioning of living systems. Understanding the nature of this interaction is of critical practical importance for rational drug discovery, which involves screening millions of small molecules to identify those capable of modulating the biological activity of molecular targets (most commonly proteins) through binding. Virtual screening (computational filtering of large compound libraries) is a key tool in the early stages of drug development. However, classical computational approaches, in particular molecular docking, have largely reached a plateau in their practical accuracy, making the development of new machine learning-based methods highly relevant.

The goal of the work is the development and optimization of machine learning methods for predicting quantitative characteristics of biological activity and structural characteristics of protein-small organic molecule complexes, based on the analysis of non-covalent interaction patterns and physically motivated selection of input features,

training data and artificial neural network architectures. **The object of the study** is the interaction between proteins and small organic molecules and its relationship with the quantitative characteristics of their biological activity and the spatial structure of the resulting molecular complexes. **The subject of the study** is machine learning models for predicting quantitative characteristics of biological activity and the spatial structure of protein-small organic molecule complexes.

In the physical context, the task of predicting biological activity and the structure of molecular complexes reduces to approximating a multidimensional dependence between the coordinates and atom types of the system and its quantitative binding characteristics, the quality of which critically depends on how completely statistics of different orders are represented in the training data of machine learning models - in direct analogy to the hierarchy of distribution functions in the statistical physics of many-particle systems.

The work employs methods of computational biophysics to establish the physical nature of molecular recognition processes, analyze non-covalent interactions, and justify the choice of input data features, as well as numerical computer simulation methods to implement the developed approaches in the form of algorithms and machine learning models.

The dissertation consists of four chapters. The **first chapter** provides a review of the current state of research on protein-small organic molecule interactions, covering the physicochemical mechanisms of binding (kinetics, thermodynamics, and binding models), experimental methods for determining complex structure and binding affinity, classical computational approaches to modeling interactions, and the role of modern machine learning methods, including graph neural networks, transformers, and diffusion models.

The **second chapter** presents the development of the 3DProtDTA graph neural network for predicting ligand-protein binding affinity. A novel approach to encoding protein structure at the amino acid residue level as graphs is proposed, integrating spatial information obtained via AlphaFold together with physicochemical properties of amino acids and evolutionary features. Systematic hyperparameter tuning and comparison of graph neural network architectures, graph pooling methods, and input representation types

are conducted. The 3DProtDTA model outperforms existing counterparts by 1-20% according to established criteria (root mean square deviation, concordance coefficient, and correlation coefficient) on the benchmark Davis and KIBA datasets. The model's ability to generalize to "cold" protein targets not encountered during training is also investigated.

The **third chapter** presents a detailed statistical analysis of the geometric parameters of the main types of non-covalent interactions in protein-small molecule systems, including hydrophobic contacts, interactions with π-systems, hydrogen and halogen bonds, and cation-anion interactions. Based on the identified statistical patterns, an algorithm for generating artificial protein-ligand binding site complexes is developed, mimicking the physicochemical regularities of real (experimentally determined) complexes. The developed approach addresses the problem of the limited number of available experimental data for training machine learning models. The effectiveness of the approach is validated on an improved generative diffusion model, PocketCFDM: using artificial complexes for training data augmentation improved the fraction of correct ligand pose predictions by up to 10% according to the RMSD criterion for atomic coordinates in three-dimensional space, and reduced the fraction of predictions with steric clashes by 3-6%.

The **fourth chapter** presents ArtiDock - a deep learning-based molecular docking method optimized for high-throughput virtual screening. ArtiDock combines a computationally efficient neural network architecture with physically motivated post-processing of predictions, achieving throughput at or better than the fastest classical tools. On the PLINDER test dataset, ArtiDock predicts ligand poses with a median RMSD of 2.0 Å and a median fraction of correctly reproduced interatomic contacts (lDDT-PLI) of 0.71, compared to 2.8 Å and 0.64 for the best-performing classical method, Glide. ArtiDock consistently outperforms classical methods across a variety of realistic scenarios, including protein binding sites containing cofactors, ions, and water molecules. On the time-split PoseX test set, ArtiDock demonstrates competitive accuracy compared to the best-performing machine learning tools, achieving 73.5% successful predictions with RMSD < 2 Å and the best results among all compared approaches (63.9%) when

accounting for the chemical validity of predictions. In a scenario where input protein conformations were taken from experimental structures of ligand-unbound proteins, ArtiDock retains its advantage, achieving 62.4% accurate and valid predictions.

Scientific novelty of the results. The 3DProtDTA graph neural network for predicting ligand binding affinity has been developed; it encodes protein structures at the amino acid residue level and is trained on structures predicted by AlphaFold, outperforming existing machine learning approaches by 1-20%. A novel approach to augmenting training data for molecular docking models has been proposed, based on the generation of artificial protein-ligand complexes that mimic the statistical patterns of non-covalent interactions in real complexes, thereby addressing the problem of limited experimental data availability. It has been confirmed that training models with artificial complexes improves the fraction of correct predictions by up to 10%. A new molecular docking method, ArtiDock, has been created, combining the accuracy of modern neural networks with the speed required for virtual screening of large chemical databases; the method predicts ligand poses and conformations 9-10% better than classical methods by the RMSD criterion, and 2-3% better by combined RMSD and chemical correctness criteria compared to other machine learning approaches.

Practical significance of the results. The developed approaches enable acceleration of the *in silico* drug discovery process through fast and accurate prediction of binding affinity and binding geometry between therapeutic targets (proteins) and potential drug candidates (small organic molecules). The methodology for generating artificial protein-ligand complexes improves the accuracy of machine learning models when applied to new classes of proteins and ligands for which no experimentally determined data are available. The research findings form the basis of software used in virtual screening projects.

Keywords: non-covalent interactions, electrostatic interactions, interfaces, dielectric environment effects, small organic molecules, bioactive molecules, biomolecules, biopolymers, computational chemistry, computational biophysics, molecular modeling, virtual screening, machine learning, deep learning, graph neural networks

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Наукові праці, в яких опубліковані основні наукові результати дисертації

1. **T. Voitsitskyi**, R. Stratiichuk, I. Koleiev, L. Popryho, Z. Ostrovsky, P. Henitsoi, I. Khropachov, V. Vozniak, R. Zhytar, D. Nechepurenko, S. Yesylevskyy, A. Nafiiev, and S. Starosyla. **3DProtDTA: a deep learning model for drug-target affinity prediction based on residue-level protein graphs**. *RSC Adv.*, vol. 13, no. 15, pp. 10261–10272, Mar. 2023, DOI: 10.1039/D3RA00281K.

2. **T. Voitsitskyi**, V. Bdzhola, R. Stratiichuk, I. Koleiev, Z. Ostrovsky, V. Vozniak, I. Khropachov, P. Henitsoi, L. Popryho, R. Zhytar, S. Yesylevskyy, A. Nafiiev, and S. Starosyla. **Augmenting a training dataset of the generative diffusion model for molecular docking with artificial binding pockets**. *RSC Adv.*, vol. 14, no. 2, pp. 1341–1353, Jan. 2024, DOI: 10.1039/D3RA08147H.

3. **T. Voitsitskyi**, I. Koleiev, R. Stratiichuk, O. Kot, R. Kyrylenko, I. Savchenko, V. Husak, S. Yesylevskyy, S. Starosyla, and A. Nafiiev. **ArtiDock: Accurate Machine Learning Approach to Protein–Ligand Docking Optimized for High-Throughput Virtual Screening**. *J. Chem. Inf. Model.*, vol. 66, no. 4, pp. 2117–2128, Feb. 2026, DOI: 10.1021/acs.jcim.5c02777.

Апробація матеріалів дисертації

1. **T. Voitsitskyi**, O. Gnatyuk, F. Naeimaeirouhani, O. Skorokhod, G. Dovbeshko, G. Gilardi. **Universal force field minimization for predicting secondary structure changes in proteins due to post-translational modifications**. XXVI Galyna Puchkovska International School-Seminar “Spectroscopy of Molecules and Crystals” (ISSSMC-2024). Постерна доповідь. Вересень 2024, Воянув, Польща.

2. **Т. Войціцький**, В. Бджола, І. Колеєв, С. Єсилевський, С. Старосила. **Аналіз нековалентних взаємодій в системі мала молекула-білок для оптимізації процесу тренування моделей машинного навчання**. Підсумкова наукова конференція ІФ НАН України ПНК-2024. Усна доповідь. Лютий 2025, Київ, Україна.

3. **T. Voitsitskyi**, I. Koleiev, S. Starosyla, S. Yesylevskyy. **Accurate machine learning approach to protein-ligand docking**. «NANOBIOPHYSICS: Fundamental and Applied Aspects» (NBP-2025). Усна доповідь. Жовтень 2025, Харків, Україна.

4. **T. Voitsitskyi**. **Protein-Ligand Docking with Machine Learning**. APS Satellite Workshop «Applied Problems of Theoretical and Computational Biophysics». Усна доповідь. Березень 2026, Київ, Україна.

5. **T. Voitsitskyi**, I. Koleiev, S. Starosyla, S. Yesylevskyy. **Molecular docking of ligands in heteroatom-containing protein binding sites using machine learning**. Biophysics Week 2026 «Biophysics in Ukraine: Human Energy and Willpower in Science and Education During the Wartime». Усна доповідь. Березень 2026, Київ, Україна.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

NMR - Nuclear magnetic resonance. Ядерний магнітний резонанс.

Cryo-EM - Cryogenic electron microscopy. Кріоелектронна мікроскопія.

ML - Machine Learning. Машинне навчання.

DL - Deep Learning. Глибинне навчання.

GNN - Graph Neural Networks. Графові нейронні мережі.

SMILES - Simplified Molecular Input Line Entry System. Специфікація спрощеного представлення молекул у вигляді текстового рядка.

RMSD - Root mean square deviation of atomic positions. Корінь середньоквадратичного відхилення координат атомів у тривимірному просторі.

ВСТУП

Обґрунтування вибору теми дослідження

Взаємодія між біологічними макромолекулами, зокрема протеїнами (білками), та малими органічними молекулами (лігандами) є фундаментальним фізичним процесом, що лежить в основі функціонування живих систем. Утворення комплексу ліганду з протеїном у сайті зв’язування визначається законами кінетики та термодинаміки. Ключовими параметрами кінетики зв'язування є константи асоціації (зв'язування) та дисоціації, які визначаються співвідношенням рівноважних концентрацій комплексу та його вільних компонентів. Саме ці константи традиційно слугують для кількісної оцінки афінності (сили) зв'язування протеїну з лігандом. З погляду термодинаміки, зв’язування стає можливим лише за умови зменшення вільної енергії системи, що досягається балансом між утворенням енергетично вигідних нековалентних взаємодій (ентальпійна компонента вільної енергії) та змінами ступенів свободи молекулярної системи (ентропійна компонента вільної енергії).

Розуміння природи зв’язування ліганду з протеїном має критичне практичне значення, оскільки на ньому базується тривалий і коштовний процес розробки лікарських засобів. На початкових етапах потрібно дослідити мільйони малих органічних молекул, щоб ідентифікувати найбільш перспективні, які мають високу афінність зв’язування. Після утворення комплексу з молекулярними мішенями (зазвичай це протеїни), відібрані молекули можуть впливати на їхню активність і забезпечувати терапевтичний ефект, тобто проявляти біологічну активність. Особливо важливу роль відіграє визначення структури комплексів протеїн-мала молекула, оскільки наявність точної геометрії комплексу, дозволяє оцінити вклад кожної міжмолекулярної взаємодії та планувати модифікації малих молекул для покращення їх властивостей.

Експериментальні методи визначення біологічної активності та структурних характеристик молекулярних комплексів є високоточними, але занадто

трудомісткими для масового відбору малих молекул. Саме тому, на початкових етапах розробки лікарських засобів ключову роль відіграє віртуальний скринінг - використання обчислювальних методів для швидкої фільтрації великих бібліотек сполук. Традиційно скринінг виконується у два етапи: первинна фільтрація найшвидшими методами оцінки афінності зв’язування та подальше передбачення структури комплексу та уточнення афінності методами молекулярного докінгу. Віртуальний скринінг дозволяє залишити найперспективніші малі молекули в тій кількості, що робить можливим застосування повільних, але високоточних обчислювальних методів, чи одразу експериментальних методів.

Сьогодні вважається, що класичні обчислювальні підходи скринінгу, зокрема молекулярний докінг, значною мірою досяг плато своєї практичної точності. У цьому контексті методи машинного навчання (ML), зокрема глибинне навчання (DL) штучних нейронних мереж, відкривають нові можливості для апроксимації складних фізичних функцій, що описують міжатомні взаємодії. На сьогодні описано багато підходів, що тою чи іншою мірою застосовують машинне навчання як для етапу первинної фільтрації, так і для молекулярного докінгу та часто перевершують традиційні підходи за точністю і швидкістю. Попри значний прогрес ML підходів для скринінгу лікарських засобів, досі існують суттєві обмеження, пов’язані з дефіцитом якісних експериментальних даних для навчання ML моделей та компромісом між точністю та обчислювальною ефективністю.

У фізичному контексті, задача передбачення афінності та просторового положення ліганду у комплексі протеїн-ліганд зводиться до апроксимації складної багатовимірної залежності між координатами та типами атомів системи, з одного боку, та її кількісними характеристиками зв'язування - з іншого. Якість такої апроксимації методами машинного навчання принципово залежить від того, наскільки повно в тренувальних даних представлена статистика різних порядків: одновимірні (маргінальні) розподіли окремих геометричних та топологічних параметрів, парні розподіли (кореляції другого порядку), а також розподіли вищих порядків. Ця ієрархія є прямою аналогією до ієрархії функцій розподілу в

статистичній фізиці багаточастинкових систем, де повна багаточастинкова функція розподілу послідовно редукується до функцій нижчих порядків, кожна з яких містить лише частину інформації про систему. Зменшення дефіциту тренувальних даних та підвищення точності ML-моделей можна розглядати як пошук мінімального порядку статистик, якого достатньо для адекватного відтворення характеристик реальних комплексів.

У дисертаційній роботі такий статистичний підхід застосовується у трьох самодостатніх, але концептуально споріднених задачах, кожна з яких розв'язується шляхом використання інформації з багатовимірного розподілу певного типу. У задачі передбачення афінності зв'язування тривимірні структури протеїнів, отримані з генеративної нейронної мережі, трактуються як вибірка з багатовимірного розподілу конформацій, вивченого відповідною генеративною моделлю; це дозволяє забезпечити структурне представлення для широкого спектра протеїнів, значна частка яких не має експериментально визначених структур. У задачі розширення тренувальних даних молекулярного докінгу перевіряється гіпотеза про достатність лише одновимірних статистик геометричних параметрів нековалентних взаємодій для генерації реалістичних штучних сайтів зв'язування. У задачі високопродуктивного молекулярного докінгу проміжною репрезентацією, яку передбачає нейронна мережа, обрано матрицю міжмолекулярних міжатомних відстаней; ця матриця є компактною статистичною характеристикою контактів між усіма атомами сайту зв'язування протеїну та ліганду, а відновлення з неї тривимірного положення ліганду виконується окремим двоетапним фізично обґрунтованим алгоритмом післяобробки.

З огляду на це, розробка нових методів, які поєднують фізичне розуміння природи міжмолекулярних взаємодій із сучасними алгоритмами глибинного навчання, є актуальною науковою проблемою фізики біологічних систем.

Мета і завдання дослідження

Метою роботи є розробка та оптимізація методів машинного навчання для передбачення кількісних характеристик біологічної активності та структурних характеристик молекулярних комплексів протеїн-мала органічна молекула на основі аналізу закономірностей нековалентних взаємодій та фізично обґрунтованого вибору вхідних репрезентацій, навчальних даних і архітектур штучних нейронних мереж.

Для досягнення мети були поставлені такі *завдання*:

1. Розробка методу передбачення афінності зв'язування малих молекул із протеїнами на основі штучних нейронних мереж, використовуючи структурну інформацію про протеїн та ліганд, який би перевершував наявні аналоги за точністю.
2. Проведення статистичного аналізу фізико-хімічних параметрів основних типів нековалентних взаємодій у відомих експериментальних комплексах протеїн-ліганд.
3. Розробка алгоритму генерації модельних систем, які імітують комплекси протеїн-ліганд на рівні сайту зв'язування та відтворюють фізико-хімічні закономірності у комплексах з експериментально визначеною структурою, для збільшення кількості тренувальних даних моделей машинного навчання для молекулярного докінгу.
4. Розробка методу високопродуктивного молекулярного докінгу на основі штучних нейронних мереж, який би перевершував класичні підходи та інші методи машинного навчання за точністю та швидкістю.

Об'єкт дослідження - взаємодія між протеїнами та малими органічними молекулами та її зв'язок із кількісними характеристиками їх біологічної активності та просторовою структурою утворених молекулярних комплексів.

Предмет дослідження - моделі машинного навчання для передбачення кількісних характеристик біологічної активності та просторової структури молекулярних комплексів протеїн-мала органічна молекула.

Методи дослідження

1. Методи обчислювальної біофізики для встановлення фізичної природи процесів молекулярного розпізнавання, аналізу нековалентних взаємодій та обґрунтування вибору репрезентацій вхідних даних, архітектур моделей машинного навчання та їхніх параметрів.

2. Методи чисельного комп'ютерного моделювання для реалізації розроблених теоретичних підходів у вигляді алгоритмів та моделей машинного навчання для прогнозування кількісних характеристик біологічної активності, генерації аугментованих навчальних даних, та прогнозування структур молекулярних систем.

Наукова новизна отриманих результатів

1. Запропоновано підхід до передбачення афінності зв'язування малих молекул із протеїнами на основі графового представлення структури протеїну на рівні амінокислотних залишків із використанням тривимірних структур передбачених методом AlphaFold як вибірки з багатовимірного розподілу конформацій протеїнів. Це дозволяє забезпечити структурне представлення для широкого спектра протеїнів, включно з тими, які не мають експериментально визначених структур. Реалізована на цій основі графова штучна нейронна мережа 3DProtDTA перевершує наявні ML-підходи за точністю на 1-20%.

2. Запропоновано новий підхід до збільшення кількості тренувальних даних для ML моделей молекулярного докінгу, що ґрунтується на одновимірних статистичних розподілах геометричних параметрів нековалентних взаємодій. Підхід реалізовано у вигляді алгоритму генерації штучних комплексів сайт

зв'язування протеїну-ліганд, які імітують статистичні закономірності нековалентних взаємодій реальних (експериментально визначених) комплексів, що розв’язує проблему обмеженої кількості наявних експериментальних даних.

3. Встановлено, що використання лише одновимірних статистик нековалентних взаємодій є достатнім для генерації штучних комплексів сайт зв'язування протеїну-ліганд. Тренування моделей із використанням штучних комплексів дозволяє підвищити частку правильних передбачень положення та конформації лігандів до 10%.

4. Створено метод молекулярного докінгу ArtiDock, який поєднує точність сучасних штучних нейронних мереж зі швидкістю, необхідною для віртуального скринінгу великих хімічних баз даних на початкових етапах розробки ліків. Модель навчено на наборі даних зі структурно обґрунтованим розбиттям на основі кластеризації за схожістю протеїнових послідовностей та профілів протеїн-лігандних взаємодій, що мінімізує витік даних між навчальними та тестовими вибірками та забезпечує об'єктивну оцінку узагальнювальної здатності моделі. Запропоновано новий підхід до післяобробки, що перетворює передбачену матрицю міжмолекулярних відстаней на хімічно валідне положення ліганду. Метод передбачає положення та конформації лігандів на 9-10% краще за критерієм RMSD порівняно з класичними методами та на 2-3% за критеріями RMSD та хімічної коректності порівняно з іншими ML підходами.

Особистий внесок здобувача

Кожен із розділів 2, 3 та 4 дисертації відповідає науковій праці, опублікованій зі співавторами. Особистий внесок здобувача полягає у наступному:

- У науковій публікації [1], результати якої описані в розділі 2, розроблено графову штучну нейронну мережу 3DProtDTA для передбачення афінності зв'язування малих молекул із протеїнами на основі графового представлення структури протеїну на рівні амінокислотних залишків з використанням AlphaFold-передбачених структур: реалізовано програмні модулі генерації ознак, підбору гіперпараметрів, тренування й тестування моделі; досліджено наявні методи передбачення афінності та виконано систематичне порівняння з ними; написано текст публікації та підготовлено таблиці й рисунки.

- У науковій публікації [2], результати якої описані в розділі 3, розроблено алгоритм генерації штучних комплексів сайт зв'язування протеїну-ліганд, що відтворюють одновимірні статистичні розподіли геометричних параметрів нековалентних взаємодій реальних експериментально визначених комплексів: реалізовано програмні модулі попередньої обробки даних, генерації сайтів зв'язування та ознак, тренування й тестування ML-моделі молекулярного докінгу; досліджено наявні ML-підходи до молекулярного докінгу; написано текст публікації та підготовлено таблиці й рисунки.

- У науковій публікації [3], результати якої описані в розділі 4, розроблено метод молекулярного докінгу ArtiDock: реалізовано модель ArtiDock та алгоритм післяобробки, що перетворює передбачену матрицю міжмолекулярних відстаней на хімічно валідне положення ліганду; підготовлено навчальні дані зі структурно обґрунтованим розбиттям, виконано підбір гіперпараметрів, тренування й оцінювання моделі; проведено порівняння розробленої моделі з іншими методами молекулярного докінгу; написано текст публікації та підготовлено таблиці й рисунки.

Апробація матеріалів дисертації

Матеріали дисертаційної роботи представлені у вигляді доповідей на міжнародних наукових конференціях:

- T. Voitsitskyi, O. Gnatyuk, F. Naeimaeirouhani, O. Skorokhod, G. Dovbeshko, G. Gilardi. Universal force field minimization for predicting secondary structure changes in proteins due to post-translational modifications. XXVI Galyna Puchkovska International School-Seminar “Spectroscopy of Molecules and Crystals” (ISSSMC-2024). Постерна доповідь. Вересень 2024, Воянув, Польща.
- Т. Войціцький, В. Бджола, І. Колеєв, С. Єсилевський, С. Старосила. Аналіз нековалентних взаємодій в системі мала молекула-білок для оптимізації процесу тренування моделей машинного навчання. Підсумкова наукова конференція ІФ НАН України ПНК-2024. Усна доповідь. Лютий 2025, Київ, Україна.
- T. Voitsitskyi, I. Koleiev, S. Starosyla, S. Yesylevskyy. Accurate machine learning approach to protein-ligand docking. «NANOBIOPHYSICS: Fundamental and Applied Aspects» (NBP-2025). Усна доповідь. Жовтень 2025, Харків, Україна.
- T. Voitsitskyi. Protein-Ligand Docking with Machine Learning. APS Satellite Workshop «Applied Problems of Theoretical and Computational Biophysics». Усна доповідь. Березень 2026, Київ, Україна.
- T. Voitsitskyi, I. Koleiev, S. Starosyla, S. Yesylevskyy. Molecular docking of ligands in heteroatom-containing protein binding sites using machine learning. Biophysics Week 2026 «Biophysics in Ukraine: Human Energy and Willpower in Science and Education During the Wartime». Усна доповідь. Березень 2026, Київ, Україна.

Структура та обсяг дисертації

Дисертація містить вступ, чотири розділи, висновки та додатки. Повний обсяг дисертації складає 197 сторінок та містить 28 рисунків та 23 таблиці. Список використаних джерел - 185 найменувань на 13 сторінках. Додатки - 19 сторінок, що містять 3 рисунки та 7 таблиць.

Зв’язок роботи з науковими програмами

Робота виконувалась у відділі фізики біологічних систем, Інституту фізики Національної академії наук України в рамках тем:

- 1.4. В/218, Державний реєстраційний номер: 0123U100990, Фізичні ефекти та молекулярні механізми взаємодії біологічних молекул та біологічних систем різного рівня структурної організації з малими молекулами, нанорозмірними та супрамолекулярними комплексами (2023-2027 рр.).
- 2020-21, 2023 - НФДУ № 2020.02/0027, «Низькорозмірні графеноподібні дихалькогеніди перехідних металів з контрольованими полярними та електронними властивостями для перспективних застосувань у наноелектроніці та біомедицині».

Практичне значення отриманих результатів

Розроблені підходи дозволяють прискорити процес *in silico* пошуку нових лікарських засобів (віртуальний скринінг) шляхом швидкого та точного передбачення афінності та геометрії зв’язування між терапевтичними цілями (протеїнами) та потенційними ліками (малими органічними молекулами). Методологія генерації штучних комплексів сайт зв'язування протеїну-ліганд дозволяє підвищити точність моделей машинного навчання для роботи з новими класами протеїнів та лігандів, для яких відсутні експериментально визначені дані. Результати досліджень стали основою програмного забезпечення, що застосовується у проєктах віртуального скринінгу компанії Receptor.AI (https://www.receptor.ai/).

1. СУЧАСНИЙ СТАН ДОСЛІДЖЕНЬ ВЗАЄМОДІЙ МІЖ ПРОТЕЇНАМИ ТА МАЛИМИ ОРГАНІЧНИМИ МОЛЕКУЛАМИ

1.1. Фізико-хімічні механізми взаємодій

1.1.1. Кінетика зв'язування протеїнів з лігандами

Кінетичні характеристики взаємодії між протеїном та лігандом є ключовими для розуміння того, з якою швидкістю формується комплекс і наскільки він є стабільним у часі. Цей процес описується двома протилежними подіями: асоціацією (утворенням комплексу) та дисоціацією (його розпадом). Обидва етапи характеризуються відповідними константами швидкості, що відображають ефективність і динаміку зв'язування. У найпростішій моделі взаємодії білка P з лігандом L, процес формування комплексу PL залежить від константи швидкості асоціації k_{on} ($M^{-1}\cdot c^{-1}$), що відображає ймовірність утворення комплексу за одиницю часу при заданій концентрації реагентів та k_{off} (c^{-1}), що вказує на ймовірність розпаду комплексу.

У стані рівноваги швидкість асоціації дорівнює швидкості дисоціації [4]:

$$k_{on}[P][L] = k_{off}[PL], \tag{1.1}$$

де квадратні дужки являють собою рівноважну концентрацію відповідних реагентів чи комплексу.

Звідси визначається константа зв'язування K_b (M^{-1}) та константа дисоціації K_d (M):

$$K_b = \frac{1}{K_d} = \frac{k_{on}}{k_{off}} = \frac{[PL]}{[P][L]}. \tag{1.2}$$

Для простоти викладу у наведених формулах концентрації використовуються замість термодинамічних активностей, що є стандартним наближенням для систем протеїн-ліганд за фізіологічних умов [4].

Значення K_d є одним з основних показників афінності зв'язування: чим нижче значення K_d, тим сильніше ліганд зв'язується з білком. Таким чином, висока швидкість асоціації при низькій швидкості дисоціації свідчить про утворення стабільного комплексу з високою афінністю зв'язування.

Попри те, що K_b та K_d не є прямими кінетичними параметрами, їх можна отримати з кінетичних даних. Це підкреслює важливий аспект: афінність зв'язування визначається співвідношенням між швидкостями утворення та розпаду комплексу. Такий підхід особливо цінний у розробці ліків, де знання кінетичних параметрів дозволяє оцінити, наскільки швидко лікарський засіб взаємодіє з молекулярними мішенями (найчастіше білками) та скільки часу зберігається ця взаємодія. Таким чином, вивчення кінетики взаємодії у системі білок-ліганд має чітке практичне спрямування, зокрема в розробці нових ліків [4].

1.1.2. Термодинамічна природа взаємодій

Взаємодія протеїну з лігандом відбуваються в складній термодинамічній системі, що включає розчинені речовини, власне молекули протеїну та ліганду, та розчинник, який переважно складається з води та буферних речовин. Динаміка взаємодії та обмін енергією в цій системі регулюються принципами термодинаміки. Серед термодинамічних параметрів вільна енергія Гіббса є особливо важливою, оскільки вона вимірює здатність системи виконувати максимальну оборотну роботу в умовах постійної температури та тиску [5].

Зв'язування білок-ліганд є можливим лише тоді, коли зміна вільної енергії Гіббса (ΔG) є від'ємною за постійної температури та тиску при досягненні термодинамічної рівноваги. Величина цього від'ємного ΔG безпосередньо впливає на стабільність отриманого комплексу і, відповідно, на афінність зв'язування ліганду. Важливо, що ΔG залежить від стану, тобто на нього впливають лише початковий і кінцевий термодинамічні стани системи та не залежить від конкретного шляху між цими станами.

Стандартна вільна енергія зв'язування ($\Delta G°$), виміряна за стандартних умов тиску 1 атм., температури 298 К та ефективних концентрацій білка та ліганду при 1 М, безпосередньо пов'язана з константою зв'язування K_b через співвідношення Гіббса:

$$\Delta G° = -RT \ln K_b, \quad (1.3)$$

де R - універсальна газова стала (1.987 кал·К$^{-1}$·моль$^{-1}$), а T - абсолютна температура. Згідно з цим співвідношенням, вища K_b відповідає більш негативній стандартній вільній енергії, що означає більшу стабільність комплексу та сильнішу афінність зв'язування між протеїнами та лігандами. Таким чином, кінетичні параметри, зокрема константи швидкості k_{on} і k_{off}, прямо пов'язані з термодинамічними.

У будь-який момент часу ΔG може бути представлена таким рівнянням:

$$\Delta G = \Delta G° + RT \ln Q, \quad (1.4)$$

де Q - коефіцієнт реакції, визначений як відношення концентрації комплексу до добутку концентрацій вільного білка та ліганду в будь-який момент часу. У стані рівноваги, де Q дорівнює K_b (рівняння 1.2), ΔG дорівнює нулю.

ΔG можна розділити на ентальпійний (ΔH) та ентропійний (ΔS) внески:

$$\Delta G = \Delta H - T\Delta S. \quad (1.5)$$

Ентальпія визначає загальну енергію термодинамічної системи, включаючи внутрішню енергію молекул розчиненої речовини та розчинника та енергію, необхідну для просторового розташування системи [6]. Від'ємна ΔH відповідає екзотермічним реакціям, при яких утворюються енергетично вигідні нековалентні взаємодії між атомами, тоді як додатна ΔH вказує на ендотермічні процеси, пов'язані з порушенням вигідних нековалентних взаємодій. Хоча ентальпія зв'язування в цілому відображає утворення чи порушення нековалентних взаємодій на інтерфейсі (міжфазній границі) зв'язування, вона також включає енергетичні внески від реорганізації розчинника та порушення взаємодій протеїн-розчинник і

ліганд-розчинник, кожен з яких може впливати на загальну зміну ентальпії як позитивно, так і негативно [7].

Ентропія кількісно визначає розподіл теплової енергії в системі. Її також можна розглядати як міру невпорядкованості атомів та молекул в системі. Додатні значення ΔS свідчать про загальне підвищення невпорядкованості (збільшення ступенів свободи) в термодинамічній системі. Загальна зміна ентропії при зв'язуванні охоплює внески ентропії розчинника (ΔS_{solv}), конформаційної ентропії (ΔS_{conf}), трансляційної та ротаційної ентропії ($\Delta S_{r/t}$):

$$\Delta S = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t}. \quad (1.6)$$

ΔS_{solv} зазвичай робить сприятливий внесок у зв'язування через вивільнення структурованих молекул розчинника на межі зв'язування. ΔS_{conf}, пов'язана зі змінами ступенів конформаційної свободи для білка та ліганду, може підвищувати або зменшувати загальну ентропію залежно від того, знижує чи збільшує утворення комплексу ці ступені свободи [8]. $\Delta S_{r/t}$ зазвичай має несприятливий внесок через втрату незалежного руху вільного ліганду та білка при утворенні комплексу.

Отже, щоб зв'язування відбулося, система повинна подолати ентропійні "штрафи" (зазвичай, негативну $\Delta S_{r/t}$ і потенційно негативну ΔS_{conf}) або через значний позитивний приріст ΔS_{solv}, або через сприятливі ентальпійні внески (від'ємний ΔH) [4], [9].

1.1.3. Воронка зв'язування

Утворення комплексу протеїну та ліганду може відбуватися спонтанно, лише якщо ΔG системи є негативною. Величина цієї негативної зміни вільної енергії визначає загальну стабільність отриманого комплексу. Зв'язування ліганду має схожість зі згортанням (фолдингом) білка в тому, що обидва процеси регулюються зменшенням вільної енергії Гіббса. Попри ці подібності, фундаментальна відмінність між згортанням та зв'язуванням ліганду полягає в тому, що перше характеризується

внутрішньомолекулярними взаємодіями в межах одного поліпептидного ланцюга, тоді як друге визначається міжмолекулярними взаємодіями між окремими молекулярними одиницями [10]. Таким чином, для опису цих процесів були запропоновані воронкоподібні ландшафти вільної енергії, які називаються воронкою фолдингу та воронкою зв'язування [11], [12].

Оскільки ΔH і ΔS можна вважати основними рушійними силами, які впливають на взаємодію білка з лігандом, важливо підкреслити складну взаємодію між ентальпійним і ентропійним внеском у процеси зв'язування. Наприклад, зв'язування, що характеризується численними сильними нековалентними взаємодіями, зазвичай призводить до значного від'ємного ΔH. Однак цей сприятливий ентальпійний внесок часто збігається зі значним зменшенням ентропії, що пояснюється зменшенням молекулярної свободи при утворенні комплексу, тим самим зменшуючи загальну величину негативного ΔG. І навпаки, значне збільшення ентропії, яке зазвичай виникає внаслідок вивільнення розчинника, часто супроводжується ентальпійними "штрафами" (позитивним ΔH) через порушення нековалентних взаємодій. Цей взаємний баланс між ентальпією та ентропією відомий як явище ентальпійно-ентропійної компенсації [13].

Ентальпійно-ентропійна компенсація була темою постійних дискусій і досліджень у науковій літературі. Критика в першу чергу зосереджується на інтерпретації цього явища компенсації як можливого оманливого через обмежені експериментальні температурні діапазони, обмежену варіацію спостережуваних значень вільної енергії, потенційні випадкові або систематичні експериментальні помилки або упередження, що виникають через вибірку даних. Проте, багато досліджень термодинаміки зв'язування в біологічних системах, калориметричного аналізу взаємодій білок-ліганд та теоретичних досліджень підтверджують уявлення про те, що ентальпійно-ентропійна компенсація є фізичним явищем, що часто спостерігається [4].

Основа цієї компенсації, ймовірно, пов’язана з утворенням і руйнуванням слабких нековалентних взаємодій. Різні фактори модулюють це явище, включаючи термодинаміку розчинника, що охоплює гідрофобний ефект, події сольватації/десольватації та структурні зміни води; гнучкість у ділянці зв’язування ліганду; структурні характеристики ліганду; та зміни міжмолекулярних сил під час процесу зв’язування [13], [14], [15].

Розуміння та кількісне визначення внесків ентальпії та ентропії у міжмолекулярне зв’язування є критично важливими в медичній хімії та розробці ліків. В ідеалі стратегії оптимізації ліків повинні бути спрямовані на максимізацію сприятливих ентальпійних або ентропійних внесків, одночасно мінімізуючи ентальпійні або ентропійні “штрафи”. Цей підхід спрямований на суттєве зниження вільної енергії зв’язування, тим самим підвищуючи ефективність і специфічність потенційних ліків [7].

1.2. Моделі зв’язування протеїнів з лігандами

Для опису та інтерпретації механізмів, що лежать в основі зв’язування протеїн-ліганд використовуються 3 моделі: модель замка та ключа, модель індукованої відповідності та модель конформаційного відбору. Ці моделі поглиблюють розуміння того, як протеїни взаємодіють з лігандами, зокрема з погляду їх структурної динаміки та гнучкості.

Найбільш рання модель замка і ключа [16], запропонована Емілем Фішером, передбачає, що і білок, і ліганд мають жорсткі структури, які точно доповнюють одна одну, аналогічно ключу, який точно вставляється в замок. Згідно з цією моделлю, специфічність виникає цілком зі структурної комплементарності; отже, лише ліганд відповідної форми може успішно взаємодіяти з сайтом зв’язування білка. Однак цей статичний погляд не може пояснити експериментальні спостереження, де початкові структурні невідповідності між білками та лігандами все ж призводять до зв’язування.

Пізніше була запропонована гіпотеза індукованої відповідності [17], яка вводить концепцію структурної адаптивності в місці зв’язування білка. Ця модель передбачає, що місце зв’язування є гнучким і контакт з лігандом може викликати конформаційні зміни всередині білка. Ці зміни призводять до більш стабільного комплексу білок-ліганд через утворення сильніших взаємодій. Ця концепція краще узгоджується з експериментальними результатами, де білки демонструють незначні структурні зміни після зв’язування ліганду.

Проте як модель замка і ключа, так і модель індукованої відповідності все ще передбачають, що білки переважно статичні або демонструють гнучкість, локалізовану лише в місці зв’язування. Жодна з них адекватно не враховує притаманну білкам конформаційну динаміку. Усуваючи цей недолік, модель конформаційного відбору [18] виникла з теорії ландшафту вільної енергії, підкреслюючи, що білки природним чином існують як динамічні ансамблі багатьох конформацій. Згідно з цією моделлю, ліганди переважно зв’язуються з вже наявними конформаціями білка у розчині, які найкраще відповідають структурам лігандів, таким чином зміщуючи рівноважний розподіл конформаційних станів білка в бік зв’язаної з лігандом конформації.

Перш ніж може відбутися зв’язування білка з лігандом, спочатку повинні відбутися початкові контакти або зіткнення між молекулами білка та ліганду. Цей початковий етап переважно регулюється молекулярною дифузією, процесом, керованим ентропією, який виникає внаслідок кінетичної енергії молекул [19]. Молекулярна дифузія в системі білок-ліганд-розчинник охоплює випадкові броунівські рухи молекул. Ці рухи є результатом як власної кінетичної енергії самих молекул розчиненої речовини, так і частих зіткнень між розчиненими речовинами та меншими молекулами розчинника.

При постійній температурі та тиску молекули розчинника демонструють випадковий рух завдяки своїй тепловій енергії, максимізуючи ентропію. Ці молекули розчинника часто стикаються з більшими молекулами розчиненої речовини, що полегшує їхні

поступальні та обертальні рухи. Отже, такі рухи значно збільшують ймовірність випадкових зіткнень між молекулами білка та ліганду, утворюючи таким чином тимчасові контакти, необхідні для формування комплексу [20]. Електростатичне притягання між протилежно зарядженими молекулами білка та ліганду може значно підвищити швидкість асоціації, дозволяючи молекулам подолати типові обмеження дифузії та зв'язуватися швидше, ніж це було б можливо за допомогою однієї лише дифузії [21]. Кількісне моделювання таких процесів зіткнення, керованих дифузією, можливе за допомогою обчислювальних методів, таких як симуляція броунівської динаміки, яка може забезпечити детальні прогнози швидкості асоціації та розуміння молекулярних шляхів, що призводять до формування комплексу білок-ліганд [22].

1.2.1. Модель замка і ключа

Дослідження з використанням геометрично-комплементарних колоїдних частинок широко досліджували модель замка та ключа як за допомогою експериментальних, так і теоретичних методологій [4]. Ці дослідження в сукупності вказують на те, що максимізація ступеня невпорядкованості системи визначає взаємодією в колоїдних системах типу "замок і ключ". Таким чином, взаємодія типу "замок і ключ" між колоїдними частинками є переважно явищем, керованим ентропією, що відбувається навіть у сценаріях відсутності явних сил притягання або за наявності електростатичного відштовхування. В першу чергу, взаємодія залежить від геометричних факторів, таких як розмір та форма поверхні молекул, що взаємодіють [23].

Початкові зіткнення між комплементарними поверхнями протеїну та ліганду можуть ініціювати зміщення молекул розчинника, що оточують розчинену речовину, ефективно збільшуючи доступний об'єм для молекул розчинника. Отже, це зміщення різко збільшує ентропію розчинника. Перед формуванням комплексу молекули води зазвичай утворюють структуровану гідратну оболонку навколо поверхонь розчиненої речовини [24]. Хоча ця сольватація спочатку призводить до негативної ΔH через сприятливі водневі зв'язки та ван-дер-ваальсові взаємодії між

розчинником і розчиненою речовиною, вона одночасно зменшує ΔS_{solv}, оскільки обмежує ступені свободи для гідратованих молекул води. Після початкового зіткнення білок-ліганд деякі взаємодії всередині структурованої водної мережі порушуються, вивільняючи раніше обмежені молекули води. Це спричиняє позитивну ΔH через руйнування нековалентних взаємодій, що компенсується молекулярною кінетичною енергією, тоді як вивільнення води значно збільшує ΔS_{solv} в системі [23].

Таким чином, модель замка та ключа для формування комплексу протеїну та ліганду полягає в оптимальному витісненні структурованої води завдяки високій комплементарності поверхонь взаємодії. Крім того, за припущення щодо жорсткості поверхонь, властивого моделі замка та ключа, під час процесу зв'язування не відбувається істотної зміни ΔS_{conf}. Отже, збільшення ентропії розчинника має бути достатньо великим, щоб подолати не лише позитивну ентальпію, пов'язану з початковим процесом десольватації, але й втрату ентропії внаслідок зменшення свободи трансляції та обертання ліганду ($\Delta S_{r/t}$). Хоча сприятливі внески ентальпії від нових нековалентних взаємодій, що утворюються на поверхні зв'язування, також відіграють певну роль у зниженні вільної енергії системи, збільшення ΔS_{solv} є основним фактором зв'язування. Отже, модель замка та ключа є фундаментально керованою ентропією, що підкреслює роль динаміки розчинника та геометричної сумісності у розпізнаванні білка лігандом.

1.2.2. Модель індукованої відповідності

Модель індукованої відповідності була продемонстрована зокрема за допомогою методів динамічної комбінаторної хімії, коли ідентифікували що протеїн під назвою діастереомер 4 зв'язує свій ліганд, NMe4I демонструючи індуковану зміну конформації під час контакту [25]. Дослідження методами ядерно-магнітно-резонансної (NMR) спектроскопії надали докази цього індукованого механізму, тоді як ізотермічна титраційна калориметрія (ІТК) дозволила виявити, що процес зв'язування сильно залежить від ентальпії ($\Delta G° = -9.1$

ккал·моль$^{-1}$, $T\Delta S°$ = -0,2 ккал·моль$^{-1}$, $\Delta H°$ = -9,3 ккал·моль$^{-1}$). Структурні дані рентгенівської кристалографії також свідчать про те, що зв'язування сахаридів з білком пермеазою відбувається через індуковану відповідність [26]. Крім того, дослідження ІТК показують, що зв'язування нітрофеніл-α-галактозиду з пермеазою відбувається виключно внаслідок негативної зміни ентальпії (-10,3 ккал·моль$^{-1}$), компенсуючи несприятливу негативну зміну ентропії (-3,8 ккал·моль$^{-1}$) [27].

На відміну від зв'язування за моделлю замка і ключа, індукована відповідність передбачає необхідність попередніх контактів для досягнення відповідності форми сайту зв'язування через неповну початкову сумісність поверхні. Початкові міжмолекулярні взаємодії, що встановлюються після зіткнення, повинні бути достатньо міцними, щоб підтримувати тимчасовий комплекс протягом часу, необхідного для наступних конформаційних коригувань. Початкове зміщення молекули розчинника на межах зв'язування, хоча й потенційно менш масштабне, ніж у сценарії "замок і ключ", забезпечує сприятливий внесок від ΔS_{solv}, необхідний для стабілізації проміжного комплексу. Подальші структурні коригування оптимізують поверхню молекул, полегшуючи додаткове витіснення розчинника, послідовно підвищуючи ентропію [6].

Поза ΔS_{solv}, загальна зміна ентропії також регулюється ΔS_{conf} і $\Delta S_{r/t}$ (рівняння 1.6). Як правило, сценарії індукованої відповідності передбачають негативні значення ΔS_{conf} і $\Delta S_{r/t}$ через обмеження конформаційної, ротаційної та трансляційної свободи молекул у комплексі. Таким чином, індукована відповідність зазвичай характеризується незначними загальними змінами ентропії. Отже, в моделі індукованої відповідності, як правило, домінує ентальпія. Негативна ΔH в результаті нововстановлених взаємодій повинна бути достатньо великою, щоб компенсувати не тільки позитивну ΔH спричинену порушенням початкових взаємодій з розчинником, але й можливу негативну ΔS. Величина негативної зміни ентальпії може бути пов'язана або з кількістю нековалентних взаємодій між протеїном і лігандом, або з їх силою. Це підкреслює вирішальну роль міжмолекулярних нековалентних взаємодій в стабілізації комплексу [6].

1.2.3. Модель конформаційного відбору

Модель конформаційного відбору заснована на теорії ландшафтів вільної енергії. Вона інтерпретує зв'язування білка з лігандом через рівноважний розподіл та перерозподіл конформаційних станів протеїну у системі [28].

Спочатку протеїни перебувають у стані численних конформацій, розподілених в нижній частині воронкоподібного ландшафту вільної енергії, де формування кожного стану представлене різними ймовірностями. Це дозволяє ліганду вибірково взаємодіяти з конформацією білка, геометрія сайту зв'язування якого оптимально підходить для розміщення ліганду, без необхідності істотних початкових конформаційних коригувань [29]. Цей етап, схожий на модель замка і ключа, головним чином зумовлений збільшенням ентропії розчинника в результаті витіснення та вивільнення молекул води, спочатку структурованих навколо поверхонь розчиненої речовини. Після цього, внутрішньомолекулярна гнучкість білка дозволяє конформаційні коригування, що призводять до сильніших міжмолекулярних нековалентних взаємодій з лігандом, подібно до моделі індукованої відповідності [30]. На цьому етапі ентальпія системи значно зменшується через утворення сприятливих нековалентних взаємодій.

Утворення стабільного комплексу білок-ліганд згодом змінює висоту енергетичних бар'єрів між зв'язаним станом протеїну і його суміжними конформаційними станами. Ця модифікація призводить до зміщення рівноважної популяції конформацій в бік стану зв'язаного з лігандом, що спричиняє перерозподіл конформаційного ансамблю білка [31].

У моделі конформаційного відбору немає чіткого визначення домінівного фактора (ентропія чи ентальпія) зниження вільної енергії системи. Значне збільшення ΔS_{solv}, досягнуте під час початкового селективного зв'язування, потенційно може бути компенсовано одночасними втратами ΔS_{conf} та $\Delta S_{r/t}$ під час наступних конформаційних коригувань. І навпаки, зменшення ΔH від утворення сприятливих взаємодій білок-ліганд може бути врівноважено енергетичними "штрафами", що

виникають внаслідок десольватації та руйнування вже наявних нековалентних взаємодій поблизу сайтів зв’язування. Однак, у рамках парадигми конформаційного відбору процеси початкового селективного зв’язування та подальших конформаційних коригувань чітко відображають послідовне переважання ентропії та ентальпії у зниженні вільної енергії Гіббса системи [4].

1.2.4. Взаємозв’язок моделей зв’язування

Моделі замка і ключа, індукованої відповідності та конформаційного відбору представляють взаємопов’язані концепції в описі взаємодій білок-ліганд у рамках воронкоподібного ландшафту вільної енергії.

Механізм замка і ключа можна розглядати як граничний випадок конформаційного відбору. У такому сценарії білки демонструють мінімальну конформаційну гнучкість, де нативний стан переважно займає єдиний глобальний мінімум енергії без значних варіацій популяції в кількох конформаційних підстанах [31]. І навпаки, високогнучкі білки мають кілька дискретних мінімумів енергії, що відповідають ансамблям конформацій, що дозволяє ліганду вибірково зв’язуватися з найбільш комплементарною конформацією серед цих підстанів, що характерно для механізму конформаційного відбору.

Відмінність між моделями полягає в перерозподілі конформаційних популяцій. Конформаційний відбір за своєю суттю передбачає переміщення популяцій між різними підстанами після зв’язування ліганду, тоді як модель замка і ключа не має такого динамічного перерозподілу, оскільки постулює жорсткі взаємодії без значних конформаційних коригувань. Також, у механізмі конформаційного відбору ліганд вибирає найбільш комплементарний конформер з ансамблю в межах одного білка, а модель замка і ключа розглядає сценарій, у якому ліганд вибирає найбільш підхожий протеїн з колекції унікальних за первинною структурою макромолекул, що доступні у клітинному середовищі [32].

Індукована відповідність є одним із ключових кроків механізму конформаційного відбору. Встановлення додаткових сприятливих взаємодій між лігандом та білком через індуковані конформаційні зміни підвищує спорідненість зв'язування та стабільність, тим самим сприяючи більш значному зсуву популяції до зв'язаної конформації. Таким чином, механізм індукованої відповідності може оптимізувати та розширити процес відбору, посилюючи взаємодії між партнерами зв'язування, тим самим прискорюючи зсуви рівноваги в конформаційних популяціях [32].

З іншого боку, для ефективного зв'язування за індукованою відповідністю, початкові селективні взаємодії повинні бути достатньо стійкими, щоб підтримувати проміжний комплекс протягом певного часу. Ця передумова означає, що навіть у рамках механізму індукованою відповідності повинна існувати початкова стадія конформаційного відбору, щоб дозволити подальші індуковані корекції [6].

Таким чином, описані моделі зв'язування можуть існувати одночасно або послідовно, охоплюючи широкий спектр подій зв'язування білок-ліганд. Їхня взаємодія підкреслює складність та адаптивність процесів розпізнавання та взаємодії з макромолекулами в живих системах.

1.3. Експериментальні методи визначення взаємодій протеїнів з лігандами

Протягом десятиліть було розроблено цілу низку лабораторних методів для кількісного вимірювання параметрів зв'язування протеїну з лігандом. Кожен метод базується на власному фізичному принципі, має свої, переваги й обмеження.

Ядерний магнітний резонанс (NMR), рентгенівська кристалографія та кріоелектронна мікроскопія (cryo-EM) дозволяють отримати тривимірні структури білок-лігандних комплексів з достатньою роздільною здатністю для завдань розробки ліків. Експериментально отримані високороздільні структури дозволяють безпосередньо побачити, як ліганд розташовується у сайті зв'язування білка, які нековалентні взаємодії він утворює з амінокислотами, і яку конформацію приймає

комплекс. Це є фундаментом для структурного дизайну ліків - процесу оптимізації молекул лігандів з метою підвищення афінності та специфічності зв’язування.

Класичні підходи, такі як рівноважний діаліз, спектрофотометричні титрування та радіолігандні зв’язувальні аналізи, історично заклали основу досліджень взаємодій білок-ліганд. Сучасні біофізичні методи, зокрема ізотермічна титраційна калориметрія (ІТК), поверхневий плазмонний резонанс, мас-спектрометрія, та зсув температури плавлення (TSA, Thermal Shift Assay) значно розширили можливості точного визначення афінності та механізмів зв’язування [33].

1.3.1. Визначення структури комплексу

Рентгеноструктурний аналіз базується на дифракції рентгенівських променів на впорядкованих кристалах білка. Кристал діє як “підсилювач” розсіювання, створюючи дифракційну картину, з якої після відновлення фаз (методом молекулярного заміщення або експериментально) будується карта електронної густини та оптимізується атомна модель. Результатом є координати атомів з високою точністю; для білкових комплексів типова роздільна здатність становить 1.5-3.5 Å [34]. Для аналізу необхідний чистий, концентрований (~10 мг/мл) розчин білка, здатний до утворення стабільних кристалів. Метод не має строгих обмежень по розміру молекули, проте великі гнучкі комплекси та мембранні білки кристалізуються важко . Отримана структура є представленням однієї конформації у твердому стані, що може не повністю відображати динаміку білка в розчині.

Комплекси методом рентгеноструктурного аналізу отримують шляхом співкристалізації або дифузії ліганду в готові кристали. Роздільна здатність 2,0-2,5 Å дозволяє чітко визначити положення ліганду та його контакти, тоді як при 3,0-3,5 Å деталі стають менш визначеними [35]. Рентгенографія залишається основним методом у розробці ліків (близько 84% структур у Protein Data Bank станом на 2024 р.) завдяки здатності досягати атомарного рівня деталізації [36].

NMR використовує магнітні властивості ядер для визначення атомної структури та динаміки молекул у розчині. У сильному магнітному полі атомні ядра зазнають резонансного збудження під дією радіочастотних імпульсів і в процесі релаксації випромінюють сигнали, специфічні для їхнього хімічного оточення. Багатовимірні спектри дозволяють розрахувати ансамбль структур (зазвичай 10-20 моделей), що відображає можливі конформації та гнучкість білка [37]. Метод потребує високої концентрації білка (>0.2-0.3 мМ), стабільності зразка та його монодисперсності. Верхня межа розміру білка для детального аналізу становить ~50 кДа; для молекул 5-25 кДа зазвичай необхідне ізотопне мічення. Точність отриманих моделей для добре визначених ділянок сягає 1-2 Å, що порівнювано з рентгенівськими структурами середньої якості.

NMR дозволяє встановити взаєморозміщення ліганду та амінокислот, визначаючи контакти та відстані між групами [38]. Ключові переваги методу - проведення аналізу в умовах, близьких до фізіологічних (без потреби в кристалізації) та можливість спостерігати кінетику зв’язування в реальному часі, виявляючи навіть слабкі взаємодії. Основний недолік - значна тривалість збору та аналізу даних (від днів до тижнів), що робить метод менш високопродуктивним порівняно з рентгенівською дифракцією.

Cryo-EM (зокрема, метод одиничних частинок) дозволяє візуалізувати біомолекули в нативному стані шляхом їх миттєвого заморожування в аморфний лід без утворення кристалів . Електронний мікроскоп фіксує тисячі 2D-проєкцій окремих частинок, які комп’ютерними алгоритмами реконструюються у 3D-карту електронної густини . Це дозволяє будувати атомні моделі та визначати координати лігандів [39]. Метод вимагає потужних обчислювальних ресурсів. Найкращі результати отримують для великих комплексів (>150 кДа); хоча робота з меншими білками складніша. Важливою перевагою є мала витрата зразка (3-4 мкг) та відсутність потреби у кристалізації.

Сучасна cryo-EM досягає роздільної здатності 3-4 Å, а в ідеальних випадках наближається до 1.2 Å [40]. Карти часто мають варіабельну роздільну здатність, що дозволяє оцінити гнучкість доменів протеїну. Cryo-EM стала проривним методом для вивчення мембранних білків та великих макромолекул, які важко кристалізувати. Вона дозволяє візуалізувати конформаційні зміни (наприклад, цикл активації рецептора), ставши потужним доповненням до рентгеноструктурного аналізу та NMR [34].

1.3.2. Визначення афінності та механізмів зв'язування

Рівноважний діаліз - один із найстаріших методів (використовується з 1930-х років), де досліджуваний білок та ліганд розділені напівпроникною мембраною, крізь яку вільно дифундує лише малий ліганд. Після досягнення рівноваги концентрація вільного ліганду вирівнюється по обидва боки, що дозволяє розрахувати константу дисоціації (K_d) на основі розподілу речовини між камерами [41]. Перевагами підходу є відсутність потреби у міченні чи іммобілізації компонентів, що наближає умови до фізіологічних, а також простота обладнання. Головними недоліками є тривалий час встановлення рівноваги, необхідність великих об'ємів зразка та потреба у чутливому методі детекції ліганду, що обмежує пропускну здатність. Метод досі широко застосовується у фармакокінетиці для оцінки зв'язування ліків з білками плазми крові.

Спектрофотометричні методи базуються на вимірюванні змін оптичного поглинання або флуоресценції (зокрема, залишків триптофану чи тирозину) під час утворення комплексу білок-ліганд. Аналіз змін спектра при титруванні дозволяє побудувати криву насичення та розрахувати K_d. Головними перевагами є швидкість виконання, відсутність потреби в радіоактивних ізотопах та доступність обладнання. Однак метод працює лише за умови зміни оптичних властивостей при зв'язуванні; іноді необхідне введення флуоресцентних міток, що може вплинути на афінність, а результати можуть спотворюватися через розсіювання світла чи неспецифічне

поглинання. У деяких випадках, як-от при зв’язуванні гему, візуальна зміна кольору одразу свідчить про утворення комплексу [42].

Радіоліганднe зв’язування традиційно використовується для вивчення рецепторів і базується на застосуванні радіоактивно мічених лігандів, що дозволяє виявляти пікомолярні концентрації речовини. Розрізняють три основні типи аналізу: сатураційний (визначення K_d та кількості сайтів зв’язування), конкурентний (визначення концентрації напівмаксимального інгібування IC_{50} та K_i) та кінетичний (розрахунок констант швидкості k_{on} та k_{off}) [43]. Метод дозволяє працювати з мембранними препаратами та зрізами тканин. Основними недоліками є необхідність роботи з радіоактивними матеріалами, складність синтезу специфічних мічених лігандів та можливий вплив неспецифічної сорбції на результати.

Ізотермічна титраційна калориметрія вважається “золотим стандартом” для визначення параметрів взаємодії. Метод прямо вимірює тепловий ефект (виділення або поглинання тепла) при титруванні білка лігандом, що дозволяє за один експеримент отримати повний термодинамічний профіль: константу зв’язування (K_b), ентальпію, ентропію та стехіометрію (число лігандів на білок) [4]. Ключовими перевагами є відсутність потреби у міченні чи іммобілізації молекул та проведення аналізу в нативних умовах розчину. До обмежень належать необхідність значних кількостей зразка високої чистоти, низька пропускна здатність (одне титрування триває 30-90 хв) та неможливість детекції процесів із нульовою зміною ентальпії.

Поверхневий плазмонний резонанс - це оптичний метод, що в реальному часі відстежує кінетику взаємодії молекул за зміною показника заломлення на поверхні сенсорного чипа. Зазвичай білок-мішень іммобілізують на золотій підкладці, а аналіт подають у потоці, що дозволяє детектувати зв’язування за збільшенням маси на поверхні без використання міток. Метод генерує сенсограми, аналіз яких дає константи швидкості асоціації k_{on} та k_{off}, що є критичним для оцінки тривалості дії ліків [44]. До недоліків належать ризик зміни конформації білка при іммобілізації,

вплив ефектів масоперенесення (дифузії) та висока чутливість до відмінностей у складі буферів.

Мас-спектрометрія дозволяє досліджувати білок-лігандні комплекси за допомогою іонізації у м'яких умовах (native MS), розрізняючи сигнали вільного білка та комплексу. Це дає змогу визначити частку зв'язаного білка для розрахунку K_d, а також встановити точну стехіометрію взаємодії. Додатковий підхід - воднево-дейтерієвий обмін - дозволяє локалізувати сайт зв'язування білка, визначаючи захищені від обміну ділянки. Метод вирізняється надзвичайною чутливістю (до фемтомолярних кількостей) та швидкістю. Проте результати можуть спотворюватися через розпад слабких комплексів у газовій фазі або утворення неспецифічних агрегатів під час іонізації, тому кількісні параметри часто потребують підтвердження іншими методами [45].

Зсув температури плавлення виявляє взаємодію лігандів через зміну термостабільності білка: специфічне зв'язування зазвичай стабілізує структуру, підвищуючи температуру плавлення (T_m) [46]. Метод використовує флуоресцентний барвник, який зв'язується з гідрофобними ділянками, що експонуються при денатурації, дозволяючи спостерігати процес на стандартному апараті для полімеразної ланцюгової реакції з можливістю флуоресцентної детекції. Завдяки простоті та високій продуктивності (сотні зразків за експеримент), метод став популярним методом первинного скринінгу. Однак це переважно якісний підхід: величина зсуву ΔT_m не завжди корелює з афінністю, а результати можуть бути спотворені неспецифічною взаємодією, агрегацією або інтерференцією ліганду з барвником. Тому дані TSA потребують верифікації кількісними методами [47].

1.4. Класичні обчислювальні методи моделювання взаємодій протеїнів з лігандами

Сучасні обчислювальні методи відіграють ключову роль у вивченні тривимірної структури білково-лігандних комплексів та оцінці афінності зв'язування. Це

особливо важливо на ранніх стадіях розробки лікарських засобів. Вони дозволяють моделювати взаємодію білків із малими органічними лігандами *in silico*, що значно прискорює і здешевлює пошук нових кандидатів у ліки порівняно з суто експериментальними підходами [48]. Далі розглянуто найбільш поширені класичні обчислювальні методи: підходи до передбачення структури та динаміки комплексу (гомологічне моделювання білка, молекулярний докінг, молекулярна динаміка) та оцінки афінності зв'язування і вільної енергії (MM/PBSA, MM/GBSA, FEP).

1.4.1. Передбачення структури комплексу

Гомологічне моделювання дозволяє передбачити просторову структуру білка на основі відомої структури схожого білка-шаблону, спираючись на принцип, що подібні послідовності формують схожі 3D-структури. Процес включає пошук шаблону, вирівнювання послідовностей та побудову моделі координат. Якість моделі залежить від відсотка ідентичності послідовностей: при >30-50% відхилення від реальної структури зазвичай складає лише 1-2 Å [49]. Це робить метод ефективним, коли експериментальна структура відсутня, дозволяючи використовувати моделі для молекулярного докінгу та дизайну ліків. Головним обмеженням є необхідність наявності близького гомолога; крім того, неконсервативні ділянки (наприклад, петлі) часто моделюються з похибками. Тому отримані моделі зазвичай потребують енергетичної мінімізації або молекулярної динаміки перед використанням для усунення стеричних конфліктів.

Молекулярний докінг - це метод комп'ютерного моделювання, призначений для передбачення оптимальної орієнтації та конформації (положення) малого органічного ліганду при його зв'язуванні з білком (вимагає відомої вхідної структури білка), а також для оцінки приблизної енергії їхньої взаємодії (скорингу). Докінг базується на двох головних компонентах: пошуковому алгоритмі та функції оцінки (скорингу). Пошуковий алгоритм генерує велику кількість можливих положень і конформацій ліганду в сайті зв'язування білка (з урахуванням гнучкості ліганду, а інколи й окремих амінокислот у рецепторі), тоді як функція скорингу

швидко оцінює енергетичну доцільність кожного варіанта і пріоритезує їх. Завдяки спрощеним енергетичним розрахункам докінг може дуже швидко (за секунди-хвилини) передбачати положення ліганду, що робить можливим перевірку відносно великих бібліотек сполук на предмет зв’язування до даної цілі - так званий віртуальний скринінг. У типовому випадку програма докінгу перебирає тисячі або мільйони молекул і повертає список найбільш перспективних кандидатів (хітів) для подальшого дослідження. Цей підхід значно підвищує ефективність пошуку ліків: наприклад, в успішних кампаніях структурного віртуального скринінгу частка активних сполук серед відібраних докінгом кандидатів сягала 10-40%, що набагато вище випадкового відбору [50].

Традиційний докінг також має низку істотних обмежень. По-перше, більшість докінг методів розглядають білок як жорстку структуру (або допускають лише обмежену гнучкість декількох заданих бічних ланцюгів в активному центрі). Насправді ж білки є динамічними молекулами, і конформація сайту зв’язування може змінюватися при зв’язуванні ліганду. Якщо оптимальна для ліганду конформація білка відрізняється від його статичної структури, докінг може не знайти правильне положення. Це особливо критично для так званих криптичних кишень - тимчасових сайтів зв’язування, які відсутні у незв’язаній структурі білка і відкриваються лише при певних змінних умовах або при зв’язуванні ліганду. По-друге, спрощені функції оцінки не завжди адекватно визначають реальну енергію зв’язування. Вони можуть давати багато хибнопозитивних результатів. Точність прогнозу абсолютних значень афінності за допомогою докінгу, як правило, невисока. На практиці докінг скоріше слугує для пріоритезації та відбору кандидатів, а не для передбачення точних значень ΔG [50].

Докінг широко застосовується на етапі пошуку хітів при віртуальному скринінгу великих бібліотек у пошуках нових хімічних класів лігандів до даного білка-мішені. Також докінг використовують для передбачення положення відомого ліганду в активному центрі білка (наприклад, щоб зрозуміти, які взаємодії забезпечують афінність інгібітора). На стадії оптимізації ліків докінг може допомогти оцінити, як

модифікація структури ліганду вплине на його зв’язування (хоча для кількісної оцінки афінності краще залучати більш точні методи). Часто докінг використовується у комбінації з іншими методами: наприклад, найкращі пози лігандів, знайдені докінгом, далі перевіряють шляхом молекулярної динаміки та/або перерахунку енергії зв’язування більш точними методами (MM/PBSA, FEP).

Молекулярна динаміка (МД) імітує фізичний рух атомів у часі, вирішуючи рівняння класичної механіки для багаточастинкової системи. Метод дозволяє моделювати динамічну поведінку комплексу (індуковану відповідність, флуктуації, нековалентні зв’язки) у наближених до фізіологічних умовах, використовуючи силові поля (найчастіше AMBER чи CHARMM) [51].

На відміну від статичного докінгу, МД-симуляції генерують ансамблі конформацій, що дозволяє оцінювати стабільність комплексу та враховувати гнучкість білка. МД часто використовують після докінгу для релаксації структури та усунення нереалістичних стеричних контактів, що підвищує точність моделей. Метод також дозволяє вивчати кінетику зв’язування, алостеричні ефекти та вплив мутацій. Попри зростання доступності завдяки GPU-прискоренню [52], основними обмеженнями залишаються значна ресурсомісткість (дні розрахунків) та залежність результатів від якості параметризації силового поля і тривалості моделювання.

1.4.2. Оцінка афінності зв’язування

MM/PBSA (Molecular Mechanics/Poisson-Boltzmann Surface Area) та MM/GBSA (Molecular Mechanics/Generalized Born Surface Area) - це методи післяобробки даних молекулярної динаміки для оцінки вільних енергій зв’язування [53]. Вони називаються методами кінцевого стану, оскільки потребують інформації лише про вільні та зв’язані стани, забезпечуючи вигідний баланс: вони точніші за функції скорингу докінгу, але значно швидші за строгі “алхімічні” розрахунки [54].

Основою є термодинамічний цикл, який пов’язує зв’язування у розчині з комбінацією газофазних (вакуумних) енергій і сольватаційних вкладів. Для реакції P

+ L → PL у воді вільна енергія зв'язування ΔG_{bind} отримується як різниця між комплексом і розділеними компонентами з урахуванням молекулярно-механічної енергії, сольватації та ентропії:

$$\Delta G_{bind} \approx \Delta E_{MM} + \Delta G_{solv} - T\Delta S \ , \quad (1.7)$$

де ΔE_{MM} - зміна молекулярно-механічної енергії при зв'язуванні; ΔG_{solv} - зміна вільної енергії сольватації при зв'язуванні; $T\Delta S$ - ентропійний внесок. Полярний компонент сольватації обчислюється через рівняння Пуассона-Больцмана (у MM/PBSA) або узагальнене наближення Борна (у MM/GBSA), а неполярний оцінюється за площею поверхні.

Попри ряд спрощень при обрахунку, методи кінцевого стану успішно відтворюють відносні тенденції афінності та є популярними для пріоритезації лігандів. Абсолютні значення енергій часто зміщені й потребують калібрування. Методи широко застосовують для післяобробки результатів докінгу та декомпозиції енергії взаємодії по окремих амінокислотах для структурної оптимізації [54].

"Алхімічні" методи, такі як збурення вільної енергії (Free Energy Perturbation, FEP) та термодинамічне інтегрування, вважаються найточнішими для розрахунку афінності, оскільки прямо оцінюють різницю вільної енергії на основі формалізму статистичної механіки. Метод використовує параметр зв'язку λ ($0 \leq \lambda \leq 1$) для моделювання поступової трансформації одного ліганду в інший одночасно у сайті зв'язування білка та у розчині. Різниця в роботі цих перетворень дає відносну вільну енергію зв'язування $\Delta\Delta G$ [55].

Завдяки застосуванню графічних процесорів (GPU) та новим алгоритмам "алхімічні" методи стали доступними для практичного використання, демонструючи точність у межах 1-2 ккал/моль [56], що дозволяє надійно пріоритезувати навіть структурно близькі сполуки. FEP широко застосовується на етапі оптимізації ліків для побудови фізично обґрунтованих моделей структура-активність (QSAR), спрямовуючи синтез і скорочуючи кількість необхідних експериментів. Основними обмеженнями

залишаються висока обчислювальна вартість та складність налаштування: необхідна точна параметризація силових полів і врахування змін сумарного заряду. Також методи можуть давати похибки, якщо зв’язування лігандів супроводжується значними конформаційними змінами білка, які важко відтворити за доступний час симуляції.

1.5. Роль сучасних методів машинного навчання у моделюванні взаємодій протеїнів з лігандами

В останні роки машинне навчання (ML), особливо глибинне навчання (DL), стало потужною альтернативою класичним обчислювальним методам. Завдяки навчанню на великих наборах даних прикладів протеїн-ліганд, моделі ML можуть робити висновки про силу зв’язування та передбачати комплекси з кращою швидкістю та часто точністю. У цьому розділі наведено загальний огляд основних архітектур ML моделей, що застосовуються для прогнозування сили зв’язування та структури комплексу.

Графові нейронні мережі враховують локальне хімічне оточення та топологію взаємодій білок-ліганд, підвищуючи точність прогнозування сили зв’язування та конформацій. Трансформери застосовують глобальні механізми уваги, які кодують далекосяжні залежності та мультимодальну (послідовність та структуру) інформацію. Дифузійні моделі пропонують генеративний підхід до дослідження зв’язування, тісно узгоджуючись з фізичною інтуїцією зв’язування як стохастичного процесу. Ці методи не є взаємозаперечними; насправді, передові моделі часто інтегрують ідеї з усіх трьох. Результатом є новий інструментарій для *in silico* розробки ліків. Завдяки навчанню на експериментальних даних та використанню сучасних обчислювальних потужностей, методи DL можуть забезпечувати швидкі та точні передбачення. Очікується, що подальший розвиток архітектур моделей та стратегій навчання (включаючи включення фізичних знань та вдосконалення методів узагальнення) ще більше скоротить розрив між результатами обчислювальних

методів та експериментальною реальністю, що зрештою прискорить процес розробки ліків у найближчі роки.

1.5.1. Графові нейронні мережі

Графові нейронні мережі (GNN) обробляють молекули як графи з вершинами (наприклад, атомами або амінокислотними залишками) та ребрами (просторовими контактами), що дозволяє отримати внутрішнє представлення молекулярної структури.

Ранні DL підходи до оцінки афінності використовували одновимірні послідовності протеїну (первинна структура) або SMILES (рядок символів за правилами однозначного опису складу та структури молекули хімічної речовини) ліганду, оброблені CNN (convolutional neural network, згорткові нейронні мережі) або RNN (recurrent neural networks, рекурентні нейронні мережі), але такі лінійні представлення втрачають важливу просторову інформацію. Моделі GNN зберігають тривимірну інформацію, дозволяючи мережі вивчати, які молекулярні підструктури та взаємодії керують зв’язуванням. Наприклад, модель GraphDTA використовує GNN для кодування структур лігандів та CNN для послідовностей білків, покращуючи прогнозування спорідненості порівняно з методами, що базуються виключно на послідовностях [57]. Новіші підходи, такі як 3DProtDTA, розширюють цю ідею, будуючи як графи білків, так і лігандів, використовуючи вершини для амінокислот [1]. Ключовим фактором стала доступність передбачених структур білків за рахунок AlphaFold [58], який надає тривимірні координати для побудови графів білків для довільної послідовності. Маючи 3D-структури, моделі на основі GNN можуть явно моделювати контакти зв’язування та геометричну комплементарність. Загалом, графові нейронні мережі пропонують природну основу для передбачення афінності зв’язування: операції передачі інформації між вершинами агрегують характеристики хімічного середовища, що дозволяє моделі зробити висновок про те, як функціональні групи лігандів взаємодіють з ділянками білка, що корелює з силою зв’язування [59].

Графові нейронні мережі широко використовуються і для прогнозування структури комплексу білок-ліганд. У докінгу можна представити сайт зв’язування білка та ліганд як два графи, що взаємодіють. Моделі на основі GNN можуть ітеративно оновлювати положення ліганду, обмінюючись інформацією між вершинами графа білка та ліганду. Одним із перших прикладів таких підходів є EquiBind, еквіваріантна GNN модель, яка безпосередньо прогнозує 3D-положення та орієнтацію ліганду в сайті зв’язування білка [60]. У більш загальному випадку, GNN, що включають просторові характеристики (відстані, кути), можуть ефективно кодувати геометрію взаємодій білок-ліганд та передбачати ймовірні конформації зв’язування. Ці моделі часто навчаються розміщувати комплементарні атоми лігандів поблизу функціональних груп білка (наприклад, донорів водневих зв’язків біля акцепторів) завдяки передачі повідомлень через графи. Методи докінгу на основі GNN швидко розвиваються; нещодавні тести показують, що новіші GNN здатні перевершувати класичні інструменти докінгу [61]. Загалом, графові нейронні мережі забезпечують гнучкий та потужний засіб для прогнозування складних структур.

1.5.2. Глибинне навчання на основі трансформерів

Трансформери - це специфічні DL архітектури, спочатку розроблені для аналізу послідовностей, які спираються на механізми багатоголової самоуваги (multi-head self-attention) для захоплення далекосяжних залежностей між елементами послідовності [62]. Сьогодні трансформери є архітектурною основою великих мовних моделей (LLM).

У завданнях аналізу взаємодій білок-ліганд трансформери застосовувалися для вивчення інформативних представлень як білкових послідовностей, так і хімічних структур. З боку білків, великі мовні моделі, такі як ESM [63], обробляють амінокислотні послідовності, вивчаючи біохімічні патерни та зв’язки віддалених залишків. Такі моделі можуть створювати представлення білків, які кодують їх структурні та функціональні особливості, що є цінним, зокрема, для прогнозування афінності лігандів. З боку лігандів, трансформери можна застосовувати для

представлення SMILES. Нещодавнім прикладом є Interformer, уніфікована модель, яка поєднує графові представлення з механізмом уваги трансформера, щоб зосередитися на конкретних нековалентних взаємодіях [64]. Моделі на основі трансформера можуть фіксувати як локалізовані взаємодії, так і ширший контекст (наприклад, алостеричні ефекти або конформаційну гнучкість білка) при прогнозуванні зв'язування.

Трансформери також революціонізували проблеми прогнозування структури, що підтверджується успіхом AlphaFold [58]. Нейронна мережа AlphaFold2 - це, по суті, спеціалізований трансформер, який ітеративно аналізує білкові послідовності та попарні карти відстаней для прогнозування 3D-координат. Його модуль Evoformer, що базується на multi-head self-attention, обробляє наявну еволюційну інформацію для послідовності, фіксуючи взаємодії на великі відстані та схеми контактів, які визначають згортання білка.

Наступним кроком стала адаптація архітектури трансформерів для передбачення комплексів білок-ліганд (ко-фолдинг моделі) [65]. Моделі обробляють білок та ліганд як спільну систему: наприклад, можна об'єднати представлення ліганду (як псевдозалишки або набір атомів) з білковою послідовністю та подати це в мережу трансформерів, яка передбачає об'єднану структуру. Ко-фолдинг підходи продемонстрували високу точність у прогнозуванні структури комплексів, часто конкуруючи зі спеціалізованими методами докінгу. Серед основних недоліків таких підходів порівняно з класичним докінгом є висока обчислювальна вартість, та те, що моделі не враховують хімічні обмеження належним чином. Загалом, підходи до ко-фолдингу на основі трансформерів добре справляються з одночасною оптимізацією конформації білка та розміщення ліганду. Вони ефективно розглядають докінг як передбачення структури: враховуючи наявність малої органічної молекули в системі, шари уваги трансформера ітеративно позиціюють ліганд там, де він максимізує вивчені моделлю закономірності контакту білок-ліганд. Результатом часто є дуже точні положення, особливо якщо керуватися попередніми знаннями про схожі білки та їх комплекси, закодованими в навчених моделях. У міру

розвитку цих підходів вони поєднуються зі стратегіями швидкої вибірки для подальшого підвищення точності докінгу.

1.5.3. Дифузійні моделі

Дифузійні моделі - це клас генеративних моделей, які навчаються шляхом ітеративного корегування випадкового шуму до структурованих результатів. У контексті моделювання комплексу протеїн-ліганд, дифузійні підходи нещодавно продемонстрували значний успіх, особливо для генерації конформацій зв’язування.

Безпосереднє прогнозування значення афінності (завдання регресії) не є типовим застосуванням дифузійних моделей, які є переважно генеративними. Проте, дифузійні моделі швидко набули популярності для докінгу та генерації складних структур. Популярним прикладом є підхід DiffDock [66], який переосмислив докінг як генеративну траєкторію на просторі конформацій ліганду. У DiffDock положення, орієнтація та торсійні кути ліганду розглядаються як неперервні змінні в геометричному многовиді, а модель дифузії вчиться коригувати випадкові розміщення в реалістичні конформації зв’язування. DiffDock визначає процес дифузії за ступенями свободи ліганду (3D-трансляція, обертання, внутрішні кути) таким чином, що ранні кроки додають шум (зміщення ліганду), а потім модель навчається звертати цей процес, поступово спрямовуючи ліганд у сайт зв’язування білка. Шляхом вибірки з цієї моделі можна генерувати кілька правдоподібних конформацій зв’язування, ефективно виконуючи ймовірнісний докінг. Моделі дифузії природним чином враховують невизначеність та кілька режимів зв’язування: результатом є розподіл поз, а не одне передбачення. Це вигідно у випадках, коли ліганд може зв’язуватися в більш ніж одній орієнтації або коли гнучкість білка створює мультимодальний ландшафт зв’язування.

Спираючись на DiffDock, з’явилося кілька вдосконалених версій, наприклад, DiffDock-PP генерує конформації на заздалегідь визначеному сайті зв’язування для підвищення точності [67], а SurfDock поєднує дифузію з графовими представленнями на основі поверхні, щоб краще враховувати комплементарність

форми білка [68]. Ще однією інновацією є DynamicBind, який використовує дифузійні еквіваріантні геометричні мережі, щоб дозволити частковий рух білка під час докінгу [69]. DynamicBind здатен адаптувати конформацію дозволяє ідентифікувати конформації білків, зв'язаних з лігандом, починаючи з незв'язаної структури.

Підсумовуючи, дифузійні моделі вносять генеративний аспект до прогнозування складних структур. Вони здатні вивчати конформаційний простір починаючи з випадкового або незв'язаного стану. Процес дифузії імітує знаходження лігандом сайту зв'язування білка та перехід у низькоенергетичну конформацію. Результатом часто є високоякісні передбачення, які можна вдосконалити класичними методами. Дійсно, у таких еталонних наборах даних, як PoseX, моделі докінгу на основі дифузії входять до числа найкращих, перевершуючи багато класичних програм докінгу [61]. Важливим спостереженням є те, що гібридні стратегії можуть дати найкращі результати: пози, згенеровані дифузійними моделями, можуть бути оброблені з мінімізацією енергії на основі фізики, поєднуючи сильні сторони кожної стратегії. У міру розвитку досліджень з'являються моделі ко-фолдингу на основі дифузії, які генерують повноцінні складні структури шляхом дифузії як білка, так і ліганду разом у 3D-просторі [65], [70]. Дифузійні моделі запровадили нову парадигму для прогнозування структури білок-ліганд, пропонуючи різноманітність у передбачених конформаціях, врахування невизначеності та покращене представлення геометричної складності через призму генеративної ймовірності.

Попри значний прогрес у розумінні фізико-хімічних основ взаємодій протеїн-ліганд та у розробці методів їх моделювання, наявні підходи (як класичні, так і засновані на машинному навчанні) мають суттєві обмеження в точності та обчислювальній ефективності. Відкритими залишаються питання оптимального представлення молекулярної інформації для задач передбачення афінності, подолання дефіциту тренувальних даних для підходів молекулярного докінгу на основі машинного навчання, а також досягнення практично прийнятного балансу між точністю та швидкістю при великомасштабному віртуальному скринінгу. Саме ці проблеми

визначають спрямованість подальших розділів дисертації, в яких послідовно розроблено та валідовано нові алгоритми та моделі машинного навчання для вирішення кожної з них.

2. ВИЗНАЧЕННЯ АФІННОСТІ ЗВ'ЯЗУВАННЯ МІЖ МАЛИМИ МОЛЕКУЛАМИ ТА ПРОТЕЇНАМИ

Сьогодні відкриття лікарських засобів залишається повільним і дорогим процесом, попри всі нещодавні наукові та технологічні досягнення. Зазвичай розробка нового препарату триває кілька років, а його вартість може перевищувати мільярд доларів США [71]. Понад 30 % усіх препаратів, що переходять в II фазу клінічних випробувань, і понад 58 % тих, що доходять до III фази, зазнають невдачі [72]. Було повідомлено, що серед 108 нових і перепрофільованих препаратів, які провалилися на II фазі у 2008-2010 роках, 51 % не мали достатньої ефективності [73]. Це спостереження підкреслило необхідність створення нових *in silico* методів, здатних зменшити рівень невдач шляхом фільтрації сполук із низькою передбачуваною ефективністю на ранніх етапах процесу відкриття ліків.

У цьому контексті особливий інтерес становлять обчислювальні методи оцінки афінності зв'язування між лігандом та білком-мішенню [74], оскільки показник афінності зазвичай вважається одним із найкращих предикторів кінцевої ефективності препарату. Точне прогнозування сили зв'язування має вирішальне значення для виключення неефективних молекул і запобігання їхньому потраплянню на етап клінічних випробувань, тому за останні роки було розроблено велику кількість відповідних обчислювальних підходів.

Найточнішу оцінку афінності можна отримати з атомістичних МД симуляцій (класичних, квантових або гібридних) у поєднанні з одним із сучасних методів розрахунку вільної енергії зв'язування ліганду. Проте така точність досягається ціною надзвичайно високих обчислювальних витрат, що робить ці методи практично непридатними для масштабного віртуального скринінгу. Саме тому найпоширенішим підходом для прогнозування афінності зв'язування у сучасній розробці ліків є молекулярний докінг, який забезпечує прийнятний компроміс між точністю та обчислювальною ефективністю [75]. Однак вважається, що емпіричні функції оцінювання, які застосовуються у молекулярному докінгу, уже досягли

практичної межі своєї точності, яку навряд чи можна підвищити без суттєвого збільшення обчислювального навантаження.

Щоб обійти ці недоліки, були розроблені класичні ML методи для визначення афінності. Ці методи не залежать від прямого розрахунку фізичних взаємодій між білковою мішенню та лігандом. Вони повністю базуються на знаннях про експериментальні дані та спираються на припущення, що схожі ліганди зазвичай взаємодіють зі схожими білковими мішенями, що кодується у попередньо натренованих нейронних мережах. Застосування таких моделей до будь-якого заданого ліганду відбувається надзвичайно швидко, що дає змогу використовувати їх для скринінгу кількості сполук, недосяжних для методів молекулярного докінгу чи молекулярної динаміки. Проте навіть найуспішніші методи класичного ML, такі як KronRLS [76] і SimBoost [77], нещодавно були перевершені сучаснішими DL методами.

Перші DL підходи для оцінки афінності використовували амінокислотну послідовність протеїну та спрощену лінійну репрезентацію лігандів за допомогою системи SMILES. Обидві послідовності потім кодувалися за допомогою одновимірних згорткових нейронних мереж (1D CNN) та/або блоків довго-короткотермінової пам’яті (LSTM) [78], [79]. Такі лінійні представлення також виявилися ефективними для прогнозування сили зв’язування у поєднанні з генеративними змагальними мережами (GAN) [80]. Проте очевидно, що лінійні представлення призводять до великої втрати інформації; тому збереження відомостей про тривимірну організацію як білків, так і лігандів може покращити результати. Справді, поява графових нейронних мереж (GNN), які зберігають інформацію про зв’язки та 3D-розташування атомів, значно підвищила точність моделей на більшості наборів даних для оцінки афінності [57].

На відміну від методів, що базуються на послідовності, ефективність GNN моделей сильно залежить від доступності й точності тривимірних структур білкових мішеней. Дефіцит таких структур обмежував розмір навчальних вибірок і стримував

прогрес GNN підходів для передбачення афінності. Успіх AlphaFold [58] у передбаченні структур білків відкрив нові можливості для створення більш точних моделей. Тривимірні структури білкових доменів, для яких раніше не було експериментально визначених структур, тепер забезпечують безпрецедентний обсяг даних для навчання ML моделей і підвищення їхньої точності.

У цьому розділі, запропоновано новий метод побудови графів білкових структур на рівні амінокислотних залишків, використовуючи тривимірну структуру, передбачену AlphaFold, і підібрано найкращі архітектури GNN для такого типу даних. Створено нову DL модель для прогнозування афінності - 3DProtDTA [1]. Під час тестування на загальновідомих наборах даних для оцінки афінності, створена модель перевершила інші відомі підходи за всіма порівняльними метриками. Розглянуто перспективи застосування та подальшого вдосконалення цієї моделі.

2.1. Інструменти для порівняння методів машинного навчання для передбачення афінності

Для об'єктивної оцінки точності розробленої моделі 3DProtDTA необхідно визначити інструментарій порівняння, що включає три компоненти. По-перше, набір методів-конкурентів - класичних ML підходів та методів глибинного навчання. По-друге, стандартизовані набори даних, на яких здійснюється порівняння і які широко використовуються в літературі для забезпечення відтворюваності результатів. По-третє, кількісні критерії оцінювання, що дозволяють однозначно зіставити точність різних підходів. Нижче кожен з цих компонентів описано детально.

Для оцінки 3DProtDTA, її було порівняно з різними класичними ML методами та методами DL, які на момент порівняння вважалися найсучаснішими та найточнішими.

Підходи, засновані на подібності, KronRLS [76] та SimBoost [77], використовують матрицю подібності, обчислену за допомогою сервера кластеризації структур

PubChem Structure Clustering Server (PubChem Sim, https://pubchem.ncbi.nlm.nih.gov) для представлення лігандів, а також матрицю подібності білків, побудовану на основі алгоритму Smith-Waterman [81] для представлення білків. Модель KronRLS базується на регуляризованому методі найменших квадратів, тоді як SimBoost є методом, заснованим на алгоритмі градієнтного бустингу.

Метод DeepDTA [79] використовував 1D CNN для обробки амінокислотних послідовностей білків та SMILES-представлень лігандів. Модель GANsDTA [80] запропонувала напівконтрольований підхід на основі генеративно-змагальних мереж (GANs) для прогнозування афінності зв'язування, також використовуючи білкові послідовності та SMILES лігандів. Ті самі початкові представлення білків і лігандів застосовувалися у методі DeepCDA [78], де автори використали комбінацію блоків CNN і LSTM для кодування. Автори методу GraphDTA [57] вперше запропонували використовувати GNN для обробки графів лігандів, тоді як білки, як і раніше, кодувалися за допомогою CNN, застосованих до послідовностей.

2.1.1. Набори даних для оцінки методів

Було використано 2 добре відомі еталонні набори даних для порівняння методів передбачення афінності: Davis [82] і KIBA [83].

Набір Davis містить пари білків кіназ і їхніх відповідних малих органічних молекул-інгібіторів з експериментально визначеними значеннями константи дисоціації K_d. Значення K_d, виражені в молях на літр (M), було перетворено за рівнянням

$$pK_d = -\log_{10}\left(\frac{K_d}{10^9}\right), \tag{2.1}$$

для звуження діапазону можливих значень (оптимізація процесу тренування), і отримані перетворені значення використовувалися як мітки для тренування та тестування моделі, аналогічно до інших підходів у порівнянні. У цьому наборі даних міститься 442 протеїни та 68 лігандів.

Набір KIBA включає значення афінності, що походять з однойменного підходу, у якому біоактивності інгібіторів із різних джерел (таких як K_i, K_d та концентрація напівмаксимального інгібування IC_{50}) комбінуються у єдиний інтегрований показник. Ці значення були попередньо оброблені за допомогою алгоритму SimBoost [77], а отримані значення використовувалися як мітки для навчання моделі. Початково набір KIBA містив 467 білків і 52498 лігандів. Для створення високоякісної вибірки, ті ж автори [77] відфільтрували дані, залишивши лише ті ліганди та мішені, для яких було принаймні 10 експериментальних значень, що звузило кількість до 229 унікальних білків і 2111 унікальних лігандів.

Кількість значень афінності та унікальних білків та малих органічних молекул у цих наборах даних узагальнено в Таблиці 2.1.

Таблиця 2.1. Набори даних для порівняння методів передбачення афінності.

Набір даних	**Білки**	**Ліганди**	**Значення афінності**
Davis	442	68	30056
KIBA	229	2111	118254

Для обох наборів даних було використано ізоморфні SMILES як початкове представлення лігандів, а також UniProt [84] ідентифікатори білків для набору KIBA, отримані з підходу DeepDTA [79]. У наборі Davis білки з однаковими UniProt ідентифікаторами могли відповідати різним записам (через мутації та фосфорилювання). Тому кожній білковій мішені було присвоєно унікальний ідентифікатор, який враховував код UniProt, тип мутації, статус фосфорилювання, належність до білкового комплексу та домен білка. Рисунок 2.1 демонструє розподіл значень афінності для наборів Davis і KIBA.

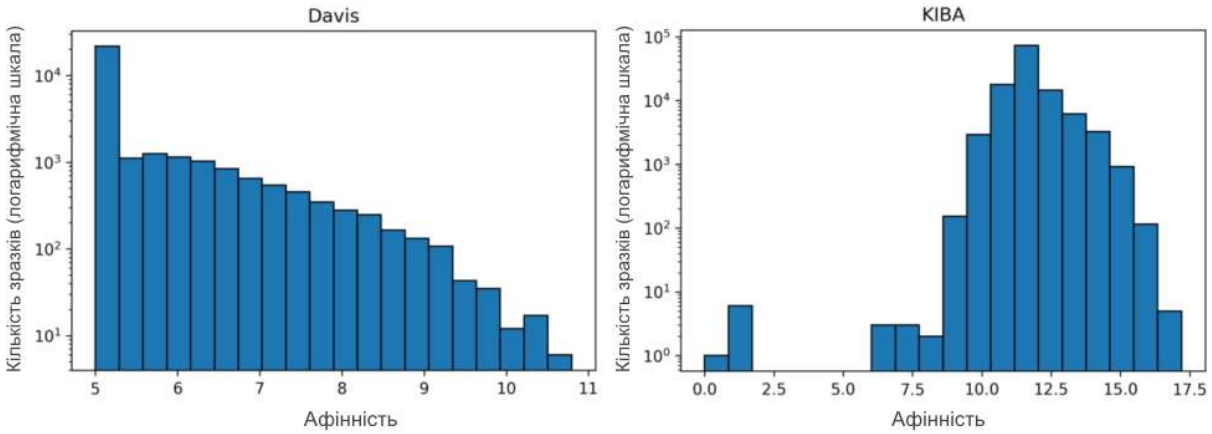

Рисунок 2.1. Розподіл модифікованих значень афінності для наборів даних Davis (ліворуч) і KIBA (праворуч).

2.1.2. Критерії оцінювання

Було обрано 3 поширені критерії оцінювання методів передбачення афінності зв'язування, що використовувались в інших підходах.

Середньоквадратична похибка (mean squared error, MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2, \tag{2.2}$$

де n - кількість зразків, y_i - експериментально визначена афінність (мітка), p_i - передбачена афінність.

Коефіцієнт конкордації [85] (CI):

$$CI = \frac{1}{Z}\sum_{\delta_i > \delta_j} h(b_i - b_j), \tag{2.3}$$

де b_i - передбачене значення для більшої афінності δ_i, b_j - передбачене значення для меншої афінності δ_j, Z - нормувальна константа, h(x) - це ступінчаста функція (h(x)=1 при x>0, h(x)=0.5 при x=0, h(x)=0 при x<0).

Індекс r_m^2 (оптимізований коефіцієнт кореляції) [86]:

$$r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2}), \quad (2.4)$$

де r^2 та r_0^2 це квадрати коефіцієнтів кореляції з урахуванням та без урахування вільного члена відповідно.

2.2. Розробка методу машинного навчання для передбачення афінності

2.2.1. Представлення малих органічних молекул

Для представлення лігандів у передбаченні афінності зв'язування з білком, було використано модифіковані молекулярні графи, спочатку запропоновані в методі Chemi-Net [87], разом зі стандартними фінгерпринтами Моргана (Morgan fingerprints) [88].

Для побудови обох типів представлень було застосовано Python API хемоінформатичного пакета RDKit v. 2021.03, використовуючи ізоморфні SMILES. Було обчислено 1024-бітні вектори класичних фінгерпринтів Моргана з радіусом 2, а також 1024-бітні альтернативні вектори, заснованих на атомних ознаках [89], з тим самим радіусом. Обидва вектори було об'єднано в єдиний 2048-бітний вектор. Графи лігандів створювалися на атомному рівні - кожна вершина графа відповідала одному атому молекули не включаючи гідрогени, а кожне ребро - ковалентному зв'язку між атомами.

Рисунок 2.2 показує розподіл кількості атомів у використаних малих молекулах, а Таблиця 2.2 містить повний перелік ознак, застосованих для графового представлення молекул.

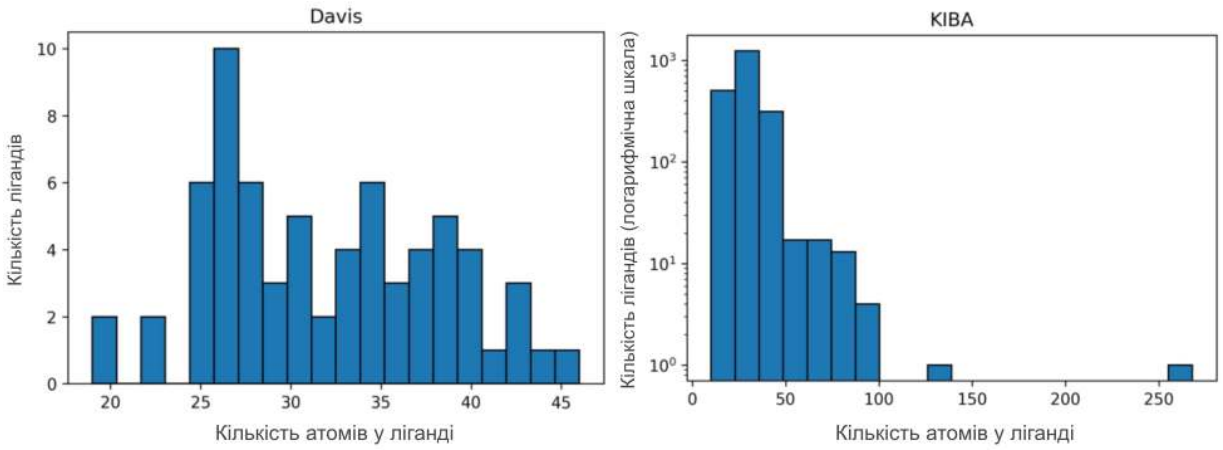

Рисунок 2.2. Розподіл кількості атомів лігандів (не включаючи гідрогенів) для наборів даних Davis (ліворуч) і KIBA (праворуч).

Таблиця 2.2. Характеристики графа для представлення лігандів.

Назва	**Розмірність**
Характеристики вершин (атомів)	
Унітарний код за типом атома (C, N, O, F, S, Cl, Br, P, I)	9
Маса атома	1
Кількість ковалентних зв’язків	1
Кількість зв’язаних гідрогенів	1
Унітарний код за гібридизацією атома (sp^2, sp^3)	2
Чи є атом частиною циклу (1 - так, 0 - ні)	1
Чи є атом частиною ароматичного кільця (1 - так, 0 - ні)	1
Чи є атом гідрофобним (1 - так, 0 - ні)	1
Чи є атом металом (1 - так, 0 - ні)	1
Чи є атом галогеном (1 - так, 0 - ні)	1
Чи є атом донором водневого зв’язку (1 - так, 0 - ні)	1
Чи є атом акцептором водневого зв’язку (1 - так, 0 - ні)	1
Чи є атом позитивно зарядженим (1 - так, 0 - ні)	1
Чи є атом негативно зарядженим (1 - так, 0 - ні)	1
Характеристики ребер (ковалентних зв’язків)	
Унітарний код типу зв’язку (одинарний, подвійний, потрійний, ароматичний)	4
Чи знаходиться зв’язок у кон’югованій системі (1 - так, 0 - ні)	1
Чи знаходиться зв’язок у циклі (1 - так, 0 - ні)	1

2.2.2. Представлення протеїнів

Для тренування моделі на інформативній репрезентації протеїну, було розроблено граф на рівні залишків амінокислот, побудований на основі тривимірних структур, згенерованих AlphaFold [58]. Близько 50% білків в обох наборах даних мають експериментально визначені 3D-структури, доступні в Protein Data Bank [36]. Проте, було свідомо використано передбачення AlphaFold для всіх білків, щоб уніфікувати підхід і уникнути додаткової складної попередньої обробки експериментальних структур, які часто є неповними або містять артефакти.

Для набору KIBA усі структури були отримані з бази даних білкових структур AlphaFold (https://alphafold.ebi.ac.uk/). Для набору Davis були завантажені лише білки без мутацій, наявні у базі AlphaFold; решту структур було згенеровано вручну за допомогою прискореної реалізації алгоритму AlphaFold у середовищі Google Colaboratory - ColabFold [90].

Щоб уникнути небажаного шуму з ділянок білків, які слабо або взагалі не пов’язані з місцями зв’язування лігандів, було використано анотації доменів UniProt [84] для визначення ділянок зв’язування. Оскільки обидва набори даних містять тільки кіназні ферменти та ліганди з інгібіторною активністю щодо кіназ, у білкових структурах було залишено лише ті домени, які мають кіназну активність, або пов’язані з нею (позначені в UniProt як protein kinase, histidine kinase, PI3K/PI4K, PIPK, AGC-kinase або CBS).

Цей етап попередньої обробки не лише зменшив шум у даних, але й усунув більшість залишків з низьким показником достовірності pLDDT (predicted Local Distance Difference Test). У AlphaFold pLDDT визначається на безперервній шкалі від 0 до 100 (чим більше - тим краще) і відображає якість передбачення структури. Значення pLDDT нижче 70, як правило, трапляються у передбачених структурах, якщо вони є неструктурованими у фізіологічних умовах або мають мало еволюційної інформації у базах даних [58]. Оскільки кіназні домени добре вивчені та

широко представлені в базах даних експериментальних структур, їх збереження і видалення решти ділянок дозволило підвищити середнє значення pLDDT.

Рисунок 2.3 показує розподіл кількості амінокислот в оброблених 3D-структурах (А) та розподіл pLDDT (Б) для наборів Davis і KIBA.

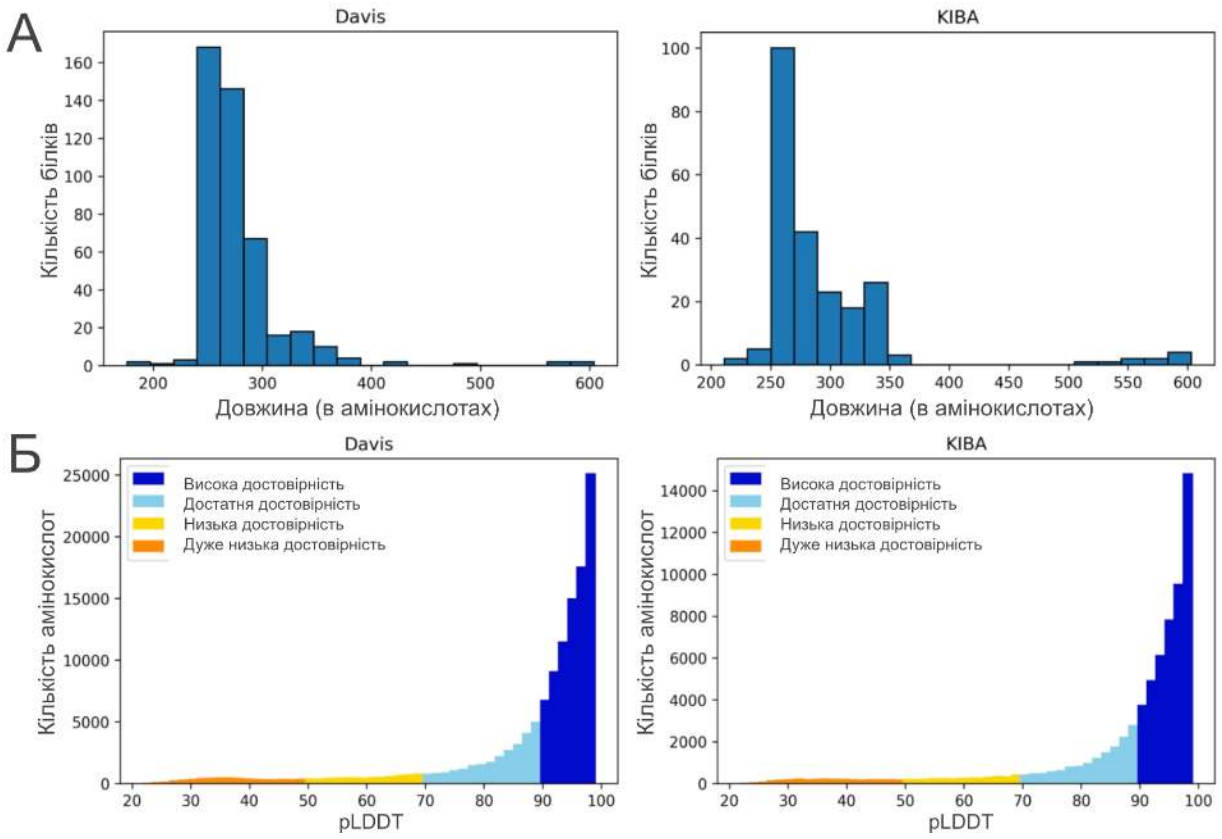

Рисунок 2.3. Розподіл кількості амінокислот (А) та значень pLDDT (Б) в оброблених тривимірних структурах білків (після видалення частин, не пов'язаних із кіназною активністю) для наборів даних Davis (ліворуч) та KIBA (праворуч).

Відфільтровані тривимірні структури було перетворено у графи на рівні амінокислотних залишків за допомогою Biopython v. 1.79 [91] та бібліотеки Pteros [92], [93]. Надихнувшись ідеями Open Drug Discovery Toolkit [94], було розроблено підхід до кодування властивостей білка в ознаках ребер графа. Ребро утворювалось, якщо дві амінокислоти формували ковалентний або нековалентний контакт на відстані, що не перевищує певного порогу. Тип цієї взаємодії визначався згідно з Таблицею 2.3. Цей підхід дозволив зменшити кількість ребер у графі білка,

порівняно з традиційними методами, які застосовують однаковий поріг відстані для всіх типів контактів між залишками.

Таблиця 2.3. Типи зв'язків і відповідні порогові відстані, використані для визначення ребер графів і їх ознак.

Тип зв'язку	**Порогова відстань (Å)**
Ковалентний	-
Гідрофобний контакт	4
Водневий зв'язок	3.5
Катіон - аніон	4
Катіон - π (ароматичне кільце)	5
Перпендикулярний π-π стекінг	5
Паралельний π-π стекінг	5

Побудований граф є спрямованим, оскільки водневі зв'язки та електростатичні взаємодії несиметричні та вимагають визначення ролей кожного із залучених залишків. Наприклад, для ребра, створеного водневим зв'язком між вершинами a та b, важливо ідентифікувати, яка вершина є донором, а яка - акцептором. Отже, ребро a → b отримує інші характеристики, ніж b → a. Аналогічно, для катіон - аніон взаємодій слід розрізняти катіонні та аніонні залишки, а для катіон-π - катіонні та ароматичні залишки. З іншого боку, ковалентні зв'язки, π-стекінг і гідрофобні контакти є симетричними.

Для кожного з 20 стандартних типів амінокислот було додано 7 фізико-хімічних властивостей AAPHY7 [95] і 23 значення BLOSUM62, що базуються на вирівнюваннях гомологічних послідовностей білків [96], отриманих із підходу GraphSol [97]. Крім того, до ознак вершин і ребер графів було включено характеристики структур та послідовностей (Таблиця 2.4).

Таблиця 2.4. Характеристики графа для представлення білків.

Назва	**Розмірність**
Характеристики вершин (амінокислотних залишків)	
Площа поверхні, доступна для розчинника	1
Кут Phi	1
Кут Psi	1
Унітарний код вторинної структури, до якої входить вершина (alpha helix, isolated beta-bridge residue, strand, 3-10 helix, turn, bend)	6
AAPHY7	7
BLOSUM62	23
Чи фосфорильований (1 - так, 0 - ні, однаковий для всіх вершин)	1
Чи є мутацією (1 - так, 0 - ні, однаковий для всіх вершин)	1
Характеристики ребер (ковалентних та нековалентних зв'язків)	
Ковалентний зв'язок (1 - так, 0 - ні)	1
Гідрофобний контакт (1 - так, 0 - ні)	1
Водневий зв'язок (донор → акцептор; 1 - так, 0 - ні)	1
Водневий зв'язок (акцептор → донор; 1 - так, 0 - ні)	1
Катіон - аніон (катіон → аніон; 1 - так, 0 - ні)	1
Катіон - аніон (аніон → катіон; 1 - так, 0 - ні)	1
Катіон - π (π → катіон; 1 - так, 0 - ні)	1
Катіон - π (катіон → π; 1 - так, 0 - ні)	1
Паралельний π-π стекінг (1 - так, 0 - ні)	1
Перпендикулярний π-π стекінг (1 - так, 0 - ні)	1

Як дві останні характеристики вершин графів, було додано статус фосфорилювання та статус мутації. Кожна з цих ознак може набувати значення 1 (фосфорильований/мутований) або 0 (не фосфорильований/не мутований) і є спільною для всіх вершин конкретного графа білка. Це має критичне значення для набору Davis, оскільки він містить білки з однаковими послідовностями, але різними станами - фосфорильованими/нефосфорильованими або білки з мутаціями з внутрішньою тандемною дуплікацією, яку неможливо однозначно представити на рівні вершини. Ігнорування цієї інформації призвело б до ситуації, коли ідентичні графи білків мали б різну афінність зв'язування з одним і тим самим лігандом.

2.2.3. Архітектура моделі

Для навчання на характеристиках графів лігандів і білків було використано GNN, після яких застосовувалися шари щільних нейронних мереж (fully connected neural network, FCNN), як зображено на Рисунку 2.4.

Рисунок 2.4. Загальна схема архітектури моделі 3DProtDTA.

Конкретна архітектура (гіперпараметри) моделі 3DProtDTA підбиралась таким чином, щоб досягти найкращих результатів крос-валідації на наборах даних. В

підбір було включено використання або одного типу GNN, або комбінації кількох типів GNN. Розглядалися такі типи графових нейронних мереж: Graph Attention Network (GAT) [98]; Graph Convolutional Network (GCN) [99]; Graph Isomorphism Network (GIN) [100]; Graph Isomorphism Network з урахуванням характеристик ребер (GINE) [101]; GCN для вивчення молекулярних фінгерпринтів (GMF) [102].

Окрім вибору типу GNN, параметри налаштування включали: конфігурацію шарів FCNN (2-3 шари з 256-4096 елементами); значення дропаут-коефіцієнтів (0.1-0.7); вибір функцій активації для GNN-шарів (ReLU, LeakyReLU, Sigmoid); використання або невикористання нормалізації; тип графового пулінгу (метод конвертації графів у одновимірну репрезентацію) або їх комбінації. Детальні межі та можливі значення налаштувань визначені у вихідному коді в репозиторії GitHub: https://github.com/vtarasv/3d-prot-dta.git.

Модель 3DProtDTA була розроблена за допомогою бібліотеки для машинного навчання PyTorch [103], а графові нейронні мережі - за допомогою бібліотеки PyTorch Geometric [104].

2.2.4. Планування експерименту

Процес розробки та порівняння моделі складався з таких етапів:

1. Підбір найкращої архітектури моделі для використання у порівняльному тестуванні;
2. Безпосереднє порівняльне тестування з іншими підходами;
3. Крос-валідаційні (cross-validation) експерименти для оцінки впливу окремих рішень при виборі характеристик вхідних даних і архітектури моделі;
4. Оцінка ефективності моделі на “холодних” білках (тобто тих, які не зустрічалися під час навчання) і перевірка тестування моделі на різних типах кіназ.

Для коректної оцінки нашого підходу було використано той самий розподіл даних на навчальні та тестові множини, що й в інших методах в порівнянні. Зокрема, набори KIBA і Davis було поділено на шість рівних частин, одна з яких використовувалася як незалежна тестова множина. Решта частин застосовувалася для підбору гіперпараметрів шляхом 5-разового перехресного затверджування (5-fold cross-validation). Усі навчальні підмножини, отримані під час перехресного затверджування, використовувалися для навчання моделі. Після цього п’ять навчених моделей застосовувалися для прогнозування спорідненості у тестовій множині. Для кожної метрики обчислювалося середнє значення, яке потім порівнювалося з іншими підходами. Навчальні та тестові розділення наборів даних були взяті з репозиторію DeepDTA на GitHub [79].

Під час підбору гіперпараметрів моделі 3DProtDTA оптимізували такі параметри:

1. Тип GNN або комбінацію типів;
2. Кількість параметрів GNN;
3. Використання та тип функції активації після шару GNN (ReLU, Leaky ReLU, sigmoid);
4. Конфігурацію FCNN;
5. Рівні dropout-регуляризації та використання нормалізації;
6. Тип або комбінацію типів графового пулінгу.

Оптимізація гіперпараметрів проводилася за допомогою програмного забезпечення Optuna v.3.0.3 [105], що базується на алгоритмі деревоподібного оцінювача Парзена (Tree-structured Parzen Estimator).

Модель навчали 700 епох з розміром пакета (batch size, кількість зразків на одну ітерацію тренування) 32, використовуючи оптимізатор Adam із темпом навчання (learning rate) 0.0001 і функцією втрати MSE (середньоквадратична похибка). Метою

оптимізації було мінімізувати MSE. Інші метрики/критерії використовувалися для тестування точності моделі, але не оптимізувалися під час навчання.

Після автоматичного підбору за допомогою Tree-structured Parzen Estimator було проведено серію ручних експериментів перехресного затверджування, у яких усі компоненти найкращої архітектури залишалися незмінними, за винятком:

1. Типу архітектури GNN для білка;
2. Типу архітектури GNN для ліганду;
3. Типу графового пулінгу для білка та ліганду;
4. Використання лише графу ліганду або лише фінгерпринтів Моргана як єдиного представлення ліганду.

Ці експерименти були спрямовані на глибоке вивчення впливу архітектури GNN, механізмів графового пулінгу та типу вхідних ознак ліганду.

Остання серія експериментів була проведена для оцінки здатності моделі 3DProtDTA узагальнювати результати на “холодних” білкових мішенях і на різних типах кіназ. Для цього білки в обох наборах даних було анотовано за допомогою записів бази даних InterPro [106]. Основними групами кіназ в обох наборах були тирозинові протеїнкінази (InterPro ідентифікатор IPR008266) та серин/треонінові протеїнкінази (InterPro ідентифікатор IPR008271). Після цього було випадково обрано по 10 тирозинових та 10 серин/треонінових кіназ із кожного набору даних й усі зразки, що містили ці білки, вилучено у незалежні тестові множини “холодних” білків. З решти даних було створено три навчальні підмножини (окремо для Davis і KIBA): (1) усі решта зразків без фільтрації; (2) зразки без тирозинових кіназ; (3) зразки без серин/треонінових кіназ. Таким чином, для кожного еталонного набору було побудовано три моделі: (1) натреновану на всіх даних; (2) натреновану без тирозинових кіназ; (3) натреновану без серин/треонінових кіназ. Для всіх “холодних” тестових підмножин було зібрано й порівняно значення за критеріями оцінювання.

2.3. Результати розробки та порівняння

2.3.1. Порівняння з іншими методами машинного навчання

Найкращу архітектуру моделі з оптимальними гіперпараметрами, що була використана для порівняння з іншими підходами, наведено на Рисунку 2.5.

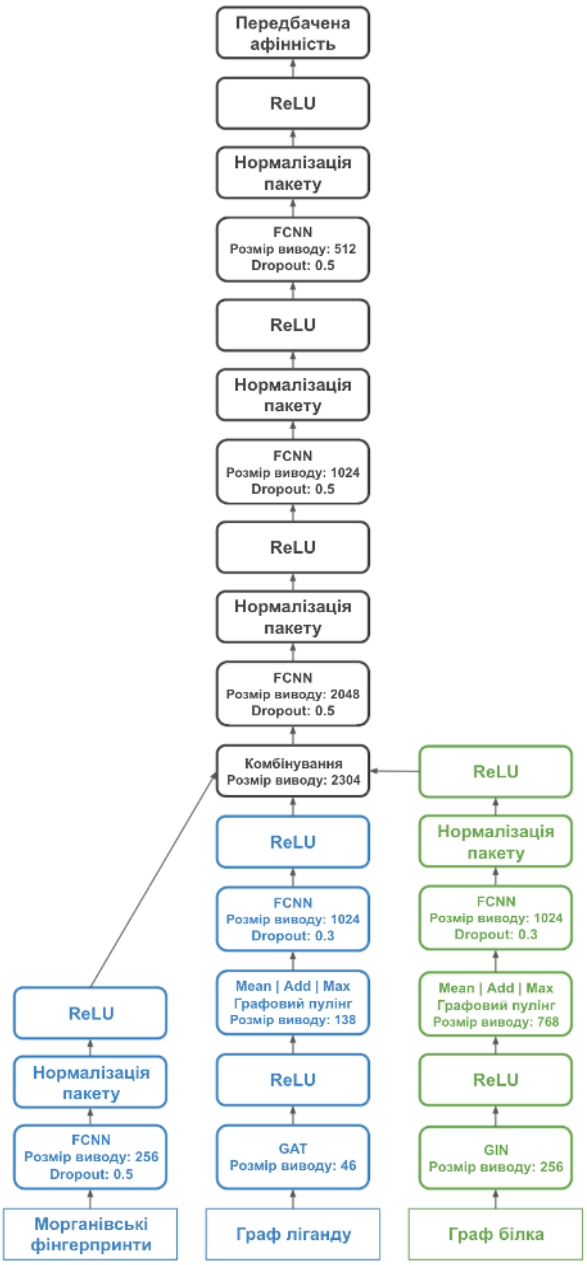

Рисунок 2.5. Детальна схема оптимізованої архітектури моделі 3DProtDTA.

Таблиці 2.5 та 2.6 порівнюють підходи за критеріями оцінювання, отриманими для наборів Davis та KIBA.

Таблиця 2.5. Середньоквадратичне відхилення (MSE), коефіцієнт конкордації (CI), та коефіцієнт кореляції (r_m^2) між реальними та передбаченими значеннями афінності малих молекул до білків з тестових даних Davis. Найкращі значення виділено жирним шрифтом.

Підхід	**MSE**	**CI**	**r_m^2**
KronRLS	0.379	0.871	0.407
SimBoost	0.282	0.872	0.644
DeepDTA	0.261	0.878	0.63
GANsDTA	0.276	0.881	0.653
DeepCDA	0.248	0.891	0.649
GraphDTA	0.229	0.893	0.63
3DProtDTA	**0.184**	**0.917**	**0.722**

Таблиця 2.6. Середньоквадратичне відхилення (MSE), коефіцієнт конкордації (CI), та коефіцієнт кореляції (r_m^2) між реальними та передбаченими значеннями афінності малих молекул до білків з тестових даних KIBA. Найкращі значення виділено жирним шрифтом.

Підхід	**MSE**	**CI**	**r_m^2**
KronRLS	0.411	0.782	0.342
SimBoost	0.222	0.836	0.629
DeepDTA	0.194	0.863	0.673
GANsDTA	0.224	0.866	0.675
DeepCDA	0.176	0.889	0.682
GraphDTA	0.139	0.891	0.673
3DProtDTA	**0.138**	**0.893**	**0.784**

Відповідно до отриманих результатів, модель 3DProtDTA суттєво перевершує інші підходи за всіма критеріями на наборі Davis. Для набору KIBA єдиним показником, який продемонстрував істотне покращення порівняно з конкурентами, був r_m^2, тоді як інші дві метрики були подібними (але все ж трохи вищими) до найкращих результатів альтернативних моделей.

2.3.2. Вплив вхідних даних та гіперпараметрів на точність моделі

Атомарні молекулярні графи та фінгерпринти Моргана є широко застосовуваними представленнями при створенні моделей, які використовують малі молекули як вхідні дані. Результати навчання моделі на кожному представленні окремо, а також на їхній комбінації наведено на Рисунку 2.6.

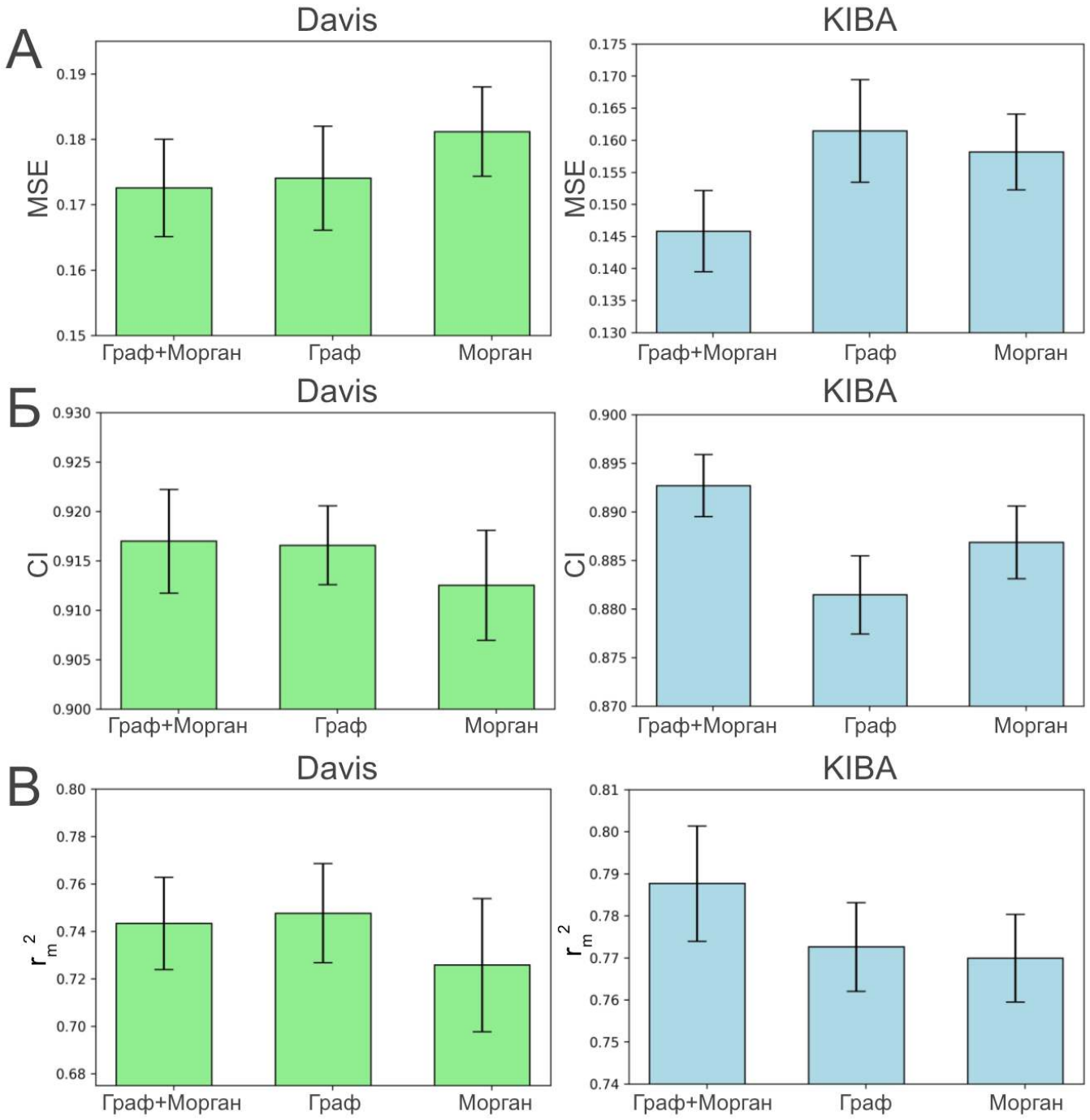

Рисунок 2.6. Середнє значення MSE (А), CI (Б) та r_m^2 (В), після 5-разового перехресного затверджування для наборів даних Davis (ліворуч) та KIBA (праворуч). Порівняно моделі, що використовували молекулярні графи та фінгерпринти Моргана (Граф+Морган), лише графи (Граф) та лише фінгерпринти (Морган) як вхідні репрезентації ліганду. Вус відображає стандартне відхилення.

Результати демонструють майже однакову ефективність моделей, що використовують лише графове представлення молекул та комбіноване представлення лігандів для набору даних Davis. Втім, для набору KIBA комбіноване представлення виявилося значно кращим.

Було оцінено молекулярну різноманітність у кожному наборі даних як 1-коефіцієнт подібності Танімото, розрахований на основі фінгерпринтів Моргана з радіусом 3 та усереднений для всіх пар лігандів у наборі. Отримані значення становили 0.882 для Davis і 0.892 для KIBA. Попри схожу молекулярну різноманітність, точність моделі для двох наборів даних суттєво відрізняється порівнюючи графове та комбіноване представлення. Це можна пояснити двома чинниками: (1) різною кількістю лігандів - 68 у Davis проти 2111 у KIBA (Таблиця 2.1); (2) значно ширшим діапазоном кількості атомів у молекулах набору KIBA (Рисунок 2.2). Ймовірно, лише графове представлення є недостатньо інформативним для навчання на всій множині лігандів KIBA, особливо для невеликої групи сполук, що містять понад 50 атомів (не включаючи гідрогени).

В середньому, модель, навчена на поєднанні графового представлення та фінгерпринтів Моргана, продемонструвала найкращі показники MSE, CI та r_m^2. Таким чином, ці два представлення містять корисну для прогнозування афінності інформацію, що не перекривається повністю.

Графовий пулінг (graph pooling) - це критично важливий етап, який використовується для створення одновимірного латентного представлення фіксованої довжини зі змінних за розміром вихідних даних з графових нейронних мереж, яке потім передається до шарів щільних нейронних мереж. Існують три найпоширеніші типи методів пулінгу: mean pooling (усереднення), max pooling (вибір максимуму), add pooling (сумування). Операція пулінгу відбувається по кожній з характеристик всіх вершин графу і в результаті дозволяє отримати одновимірне представлення рівне за розміром кількості характеристик вершини. Порівняння результатів для різних методів пулінгу наведено на Рисунку 2.7.

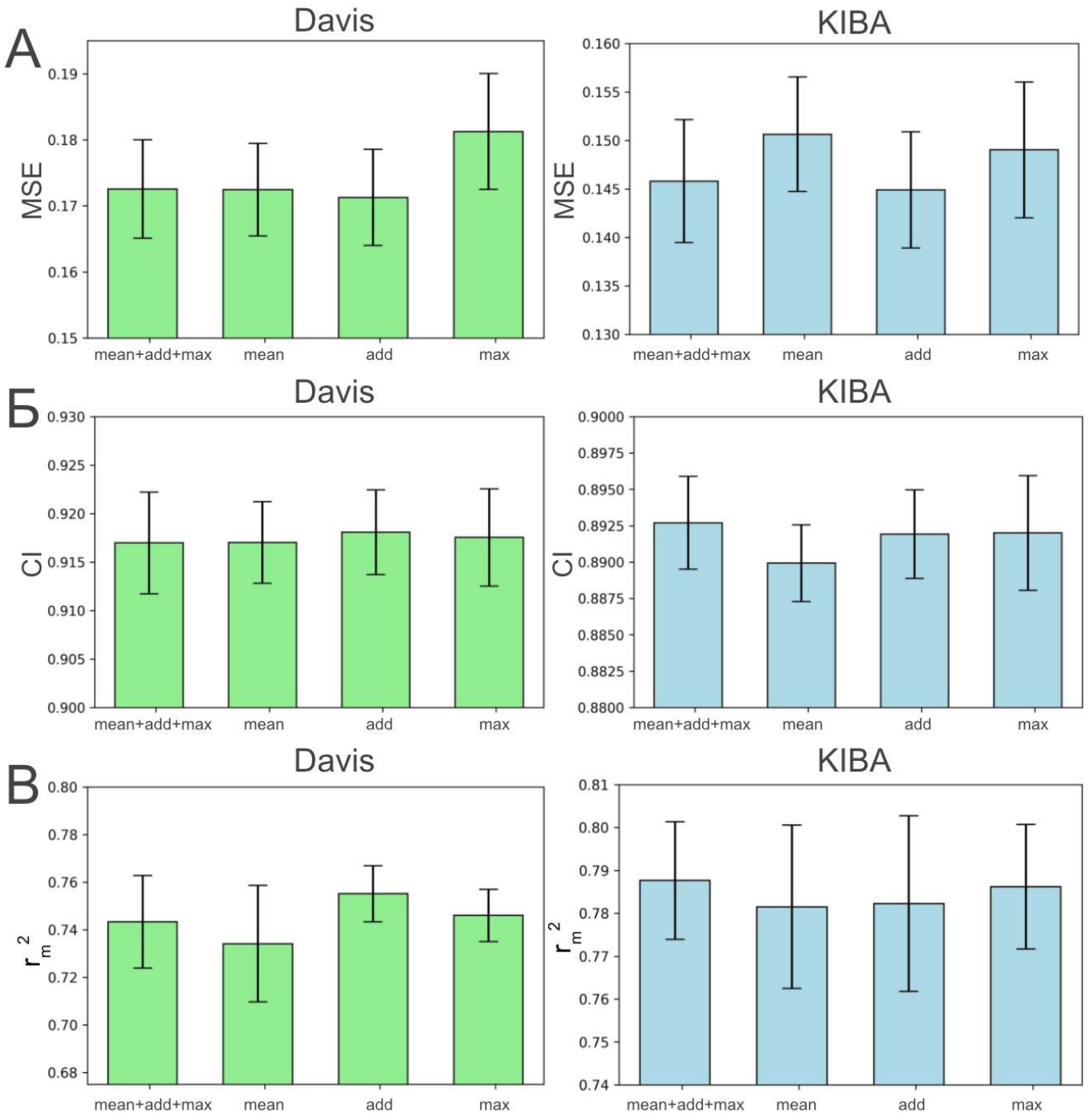

Рисунок 2.7. Середнє значення MSE (A), CI (Б) та r_m^2 (B), після 5-разового перехресного затверджування для наборів даних Davis (ліворуч) та KIBA (праворуч). Порівняно моделі, що використовували mean пулінг, max пулінг, add пулінг, або комбінування всіх типів пулінгу (mean+add+max) для конвертації графового представлення даних у одновимірне. Вус відображає стандартне відхилення.

Як видно з результатів, add пулінг є найкращим вибором для набору даних Davis, тоді як для KIBA він дає подібну точність до комбінованого підходу, у якому три типи пулінгу об’єднуються шляхом конкатенації. Add пулінг забезпечив найвищі значення MSE, CI та r_m^2 в середньому для двох наборів даних.

Рисунок 2.8 показує, що жоден тип GNN не демонструє значної переваги над іншими при обробці графів лігандів. Середні результати для двох наборів даних за всіма трьома критеріями оцінювання є дуже близькими для всіх розглянутих архітектур. Важливо підкреслити, що єдиною GNN архітектурою, яка враховувала характеристики ребер графа, була GINE. Отже, специфічні характеристики зв’язків ліганду (наприклад, тип ковалентного зв’язку) не відіграють вирішальної ролі у прогнозуванні афінності.

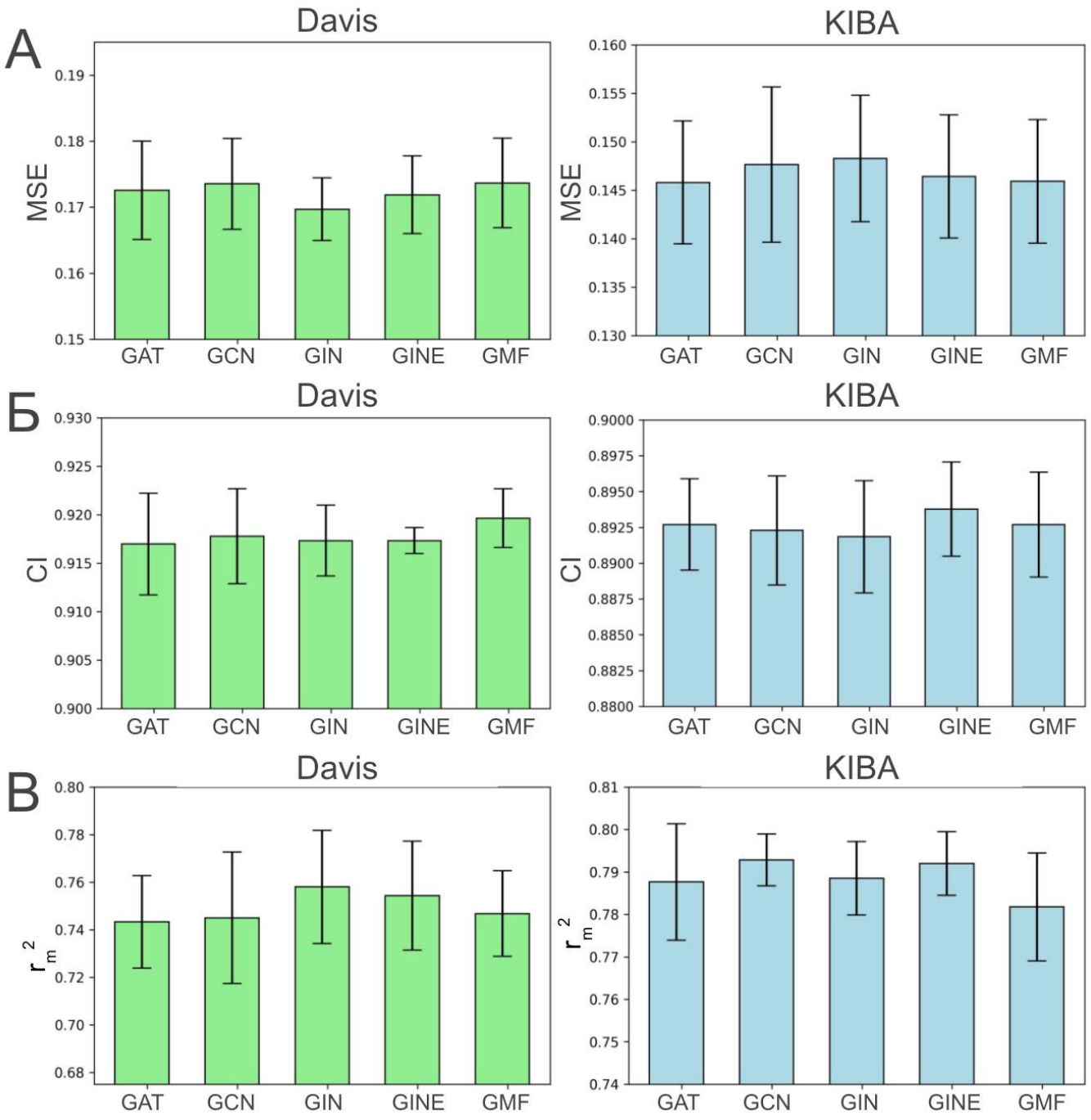

Рисунок 2.8. Середнє значення MSE (A), CI (Б) та r_m^2 (В), після 5-разового перехресного затверджування для наборів даних Davis (ліворуч) та KIBA (праворуч). Порівняно моделі, що використовували GAT, GCN, GIN, GINE або GMF архітектури графових нейронних мереж для навчання на графах ліганду. Вус відображає стандартне відхилення.

Відповідно до Рисунка 2.9, важко виділити конкретний тип GNN для графа білка, який стабільно демонструє найкращі результати. GAT, GCN, GIN та GINE показали дуже схожі середні результати, тоді як GMF архітектура продемонструвала значно нижчу точність. Аналогічно до результатів, отриманих для графів лігандів, GINE, яка враховує характеристики ребер графа білка, не показала помітного покращення метрик порівняно з іншими типами GNN. Це свідчить, що інформації про факт ковалентної чи нековалентної взаємодії амінокислотних залишків достатньо для передбачення афінності без необхідності врахування конкретного типу взаємодії.

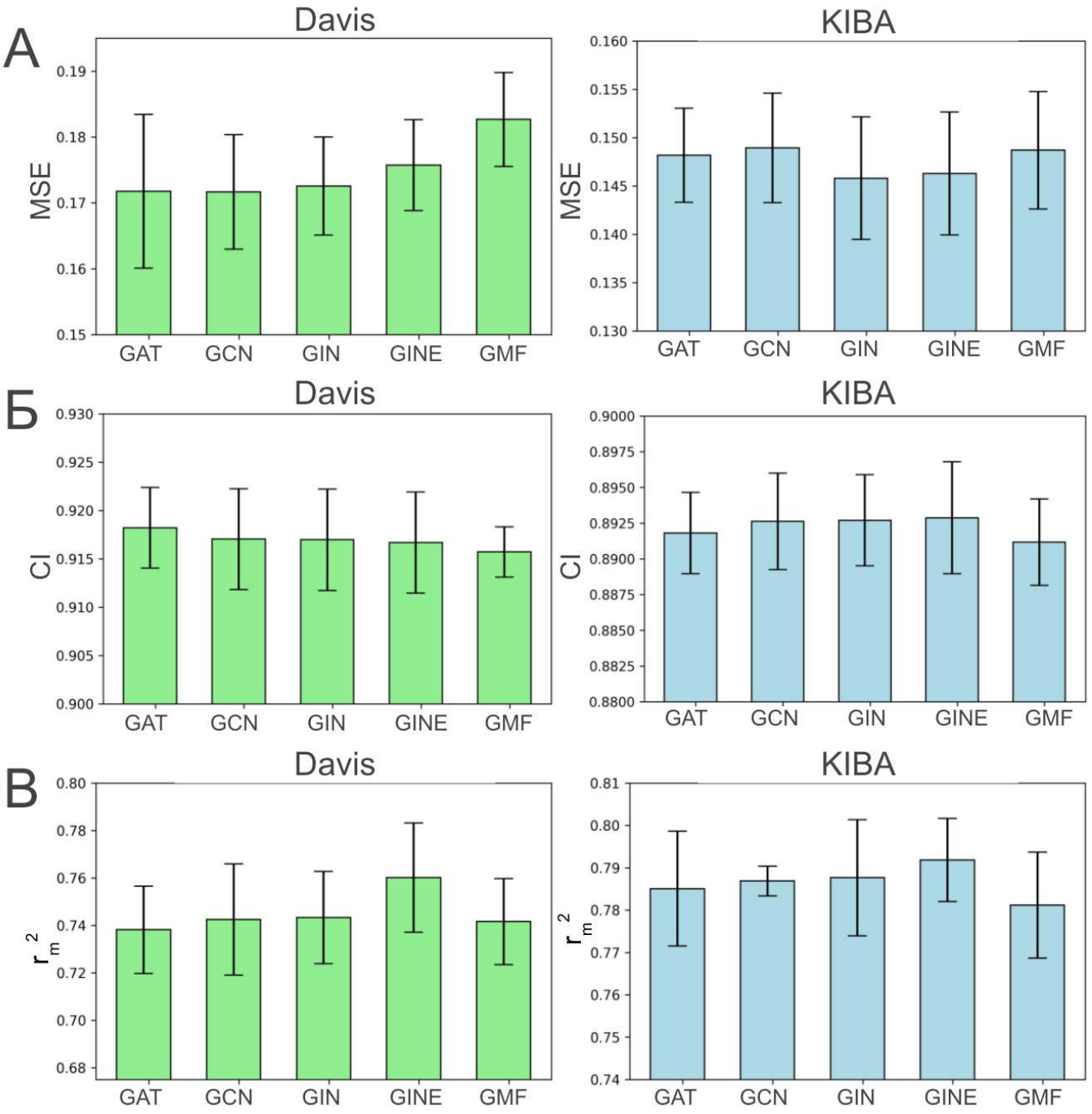

Рисунок 2.9. Середнє значення MSE (A), CI (Б) та r_m^2 (B), після 5-разового перехресного затверджування для наборів даних Davis (ліворуч) та KIBA (праворуч). Порівняно моделі, що використовували GAT, GCN, GIN, GINE або GMF архітектури графових нейронних мереж для навчання на графах білка. Вус відображає стандартне відхилення.

2.3.3. Точність моделі для протеїнів, відсутніх у навчальних даних

Точність моделі 3DProtDTA, натренованої на всіх типах кіназ, а також на вибірках без тирозинових кіназ або без серин/треонінових кіназ, була оцінена як для “холодних” білкових мішеней (тобто таких, які не зустрічалися під час навчання), так і для “холодних” типів кіназ. Отримані результати наведено у Таблицях 2.7 і 2.8.

Таблиця 2.7. Значення MSE, CI та r_m^2 для наборів даних, що включають усі зразки з відомою афінністю до однієї з 10 тестових тирозинових або серин/треонінових протеїнкіназ. Тестові протеїнкінази що були вилучені з навчальних даних. Результати представлено для набору даних Davis. Найкращі значення виділено жирним шрифтом.

Навчальні дані	**Тест, холодні тирозинові кінази**			**Тест, холодні серин/треонінові кінази**		
	MSE	**CI**	**r_m^2**	**MSE**	**CI**	**r_m^2**
Всі білки	0.305	0.88	0.536	0.311	0.889	0.509
Без тирозинових кіназ	0.719	0.746	0.254	**0.284**	**0.909**	**0.572**
Без серин/треонінових кіназ	**0.239**	**0.909**	**0.629**	0.683	0.682	0.136

Таблиця 2.8. Значення MSE, CI та r_m^2 для наборів даних, що включають усі зразки з відомою афінністю до однієї з 10 тестових тирозинових або серин/треонінових протеїнкіназ. Тестові протеїнкінази що були вилучені з навчальних даних. Результати представлено для набору даних KIBA. Найкращі значення виділено жирним шрифтом.

Навчальні дані	**Тест, холодні тирозинові кінази**			**Тест, холодні серин/треонінові кінази**		
	MSE	**CI**	**r_m^2**	**MSE**	**CI**	**r_m^2**
Всі білки	0.383	0.668	0.599	**0.236**	0.857	**0.7**
Без тирозинових кіназ	0.904	0.634	0.155	0.238	**0.858**	0.693
Без серин/треонінових кіназ	**0.359**	**0.729**	**0.625**	0.589	0.713	0.277

Результати показують, що модель, натренована на всіх типах кіназ, демонструє практично однакову точність для тестових підмножин, що містять тирозинові та серин/треонінові кінази в наборі Davis. Натомість у наборі KIBA модель, навчена на всіх типах кіназ, показує значно кращі результати для серин/треонінових кіназ.

Моделі, натреновані без тирозинових кіназ, продемонстрували найкращу або майже найкращу точність на тестових підмножинах із серин/треоніновими кіназами - і навпаки: моделі, навчені без серин/треонінових кіназ, - на тестах із тирозиновими. Їхня точність навіть вища, ніж у моделі, навченої на всіх типах кіназ. Це свідчить, що додавання нових типів кіназ до навчального набору даних може мати негативний вплив на точність передбачення для інших типів кіназ.

Можливим розв’язанням цієї проблеми є розробка багатозадачної моделі (multi-task learning), яка міститиме спільні шари нейронних мереж для всіх типів кіназ на початкових рівнях і окремі вихідні шари для кожного типу кіназ. Розробка й

перевірка такої архітектури виходить за межі цього дослідження, однак є перспективним напрямом подальших робіт.

Крім того, моделі, натреновані без певного типу кіназ, очікувано показують значно гірші результати на тестах, що містять виключені типи. Це цілком природна поведінка, яка підкреслює, що модель працює найкраще саме для тих типів білків, які були представлені у вибірці для навчання.

2.4. Обмеження та перспективи розробленого методу

Хоча використання білкових структур із бази даних AlphaFold забезпечує узгодженість в алгоритмі підготовки даних, цей підхід має певні обмеження. Зокрема, він може вносити невизначеність, якщо для одного білка існує кілька експериментально визначених структур, які використовуються як шаблони для передбачення структури. У таких випадках AlphaFold потенційно може створити змішану або усереднену конформацію, що поєднує кілька альтернативних станів білка, які не є функціонально релевантними. Можливим шляхом покращення було б використання всіх доступних експериментально визначених структур білка разом із прогнозами AlphaFold. Однак наш аналіз показав, що графи білків на рівні амінокислотних залишків, побудовані для кількох випадково вибраних експериментальних структур кіназ, є майже ідентичними графам, згенерованим на основі передбачених структур AlphaFold. Ймовірними причинами цього є: (1) велика кількість експериментально визначених структур доменів кіназ, доступних як шаблонів для передбачення AlphaFold, що позитивно впливає на точність прогнозів; (2) помірний рівень деталізації, використаний у графах на рівні залишків, який не враховує атомних деталей. Використання структур AlphaFold має також низку додаткових переваг: дозволяє включити до аналізу білки, для яких немає експериментально відомих структур, і водночас уникнути проблем, пов’язаних із неповними або неоднозначними структурами в експериментальних даних.

Іншим напрямом покращення є перехід до білкових графів атомного рівня замість графів на рівні залишків, а також використання ширшого набору характеристик вершин і ребер. Зокрема, можуть бути залучені B-фактори (показники рухливості) та інші метрики гнучкості білків, такі як матриці кореляції рухів.

2.5. Висновки розділу

Розділ описує розробку нової DL моделі для передбачення афінності зв’язування між протеїном та лігандом (малою органічною молекулою), яка отримала назву 3DProtDTA.

Відмінною особливістю цієї моделі є графове представлення як білків, що отримане з передбаченої тривимірної структури білка, так і лігандів, яке зберігає ключову інформацію про зв’язки та просторову організацію молекул, водночас не створюючи надмірного обчислювального навантаження. Характеристики вхідних даних білків для навчання моделі були згенеровані з використанням бази даних передбачених білкових структур AlphaFold, що дозволило охопити всі білки у двох наборах даних для тестування методів передбачення афінності.

Було перевірено широкий спектр архітектур, побудованих на основі графових нейронних мереж, а також різноманітні їхні комбінації, щоб досягти найкращих показників точності. Модель 3DProtDTA перевершує наявні на момент розробки аналоги на 1-20% за загальноприйнятими критеріями, такими як середньоквадратичне відхилення (MSE), коефіцієнт конкордації (CI), та коефіцієнт кореляції (r_m^2) між реальними та передбаченими значеннями афінності на загальноприйнятих наборах даних і має значний потенціал для подальшого вдосконалення.

Усі алгоритми та дані, використані для навчання та оптимізації моделі, знаходяться у відкритому доступі в репозиторії GitHub (https://github.com/vtarasv/3d-prot-dta.git).

3. ГЕНЕРАЦІЯ ШТУЧНИХ КОМПЛЕКСІВ ПРОТЕЇН-МАЛА МОЛЕКУЛА НА ОСНОВІ АНАЛІЗУ НЕКОВАЛЕНТНИХ ВЗАЄМОДІЙ ДЛЯ РОЗШИРЕННЯ ТРЕНУВАЛЬНИХ ДАНИХ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Експериментально визначені або точно передбачені структури білково-лігандних комплексів є надзвичайно цінними для пояснення і передбачення сили зв'язування (афінності) між білком і лігандом. Афінність відображає баланс всіх нековалентних взаємодій, що стабілізують комплекс, у порівнянні з вільними компонентами. Структура однозначно показує, де саме розмістився ліганд у білку, які групи ліганду звернуті до білка, а які - до розчинника.

Висока афінність часто спостерігається, коли ліганд глибоко сидить у кишені білка, максимально використовуючи доступний об'єм і взаємодіючи з багатьма сусідніми групами. Це, зокрема, максимізує кількість ван-дер-ваальсових контактів. Навпаки, якщо структура показує, що значна частина поверхні ліганду залишається експонованою назовні або ж ліганд не заповнює значну частину сайту зв'язування, то, ймовірно, такі взаємодії є слабкими. Якщо ж у структурі спостерігаються стеричні зіткнення (атом ліганду знаходиться занадто близько підходить до атома білка) - це ознака напруги, яка буде енергетично невигідною і послаблюватиме афінність.

Багато малих органічних лігандів містять ароматичні кільця та алкільні групи. Ці неполярні фрагменти вступають у гідрофобні взаємодії з неполярними сайтами зв'язування білка. Структура показує, наскільки ліганд захований в гідрофобному оточенні: кількісно це оцінюють через площу поверхні, що схована від розчинника при комплексоутворенні. Чим більша площа неполярної поверхні, тим сильніший гідрофобний ефект - молекули води витісняються з контактної поверхні, що дає виграш в ентропії [107]. Дослідження показали, що видалення навіть однієї молекули води з гідрофобного активного центру може підвищити афінність ліганду у десятки разів [108]. Якщо в структурі видно, що ліганд не заповнив певну

гідрофобну зону сайту зв’язування, це означає, що можна оптимізувати афінність додаванням до ліганду замісника, який займе цю кишеню, збільшивши площу гідрофобного контакту.

Структурні дані дозволяють ідентифікувати всі донорно-акцепторні пари між лігандом і протеїном. Водневі зв’язки - один з ключових факторів афінності. Кожен водневий зв’язок дає виграш в ентальпії, що може на порядок зменшити K_d. Особливо важливі зв’язки, що утворюють оптимальні пари: тобто якщо ліганд дозволяє утворити контакт донор-акцептор, який інакше б не утворився на поверхні білка [109]. Структурний аналіз показує, чи є у комплексі такі контакти. Наприклад, якщо в активному центрі ферменту є Asp/Glu, що у відсутності ліганду зв’язаний лише з водою, а ліганд має аміногрупу, яка витісняє воду та утворює прямий водневий зв’язок з цією амінокислотою - це часто обумовлює сильне зв’язування. Навпаки, якщо ліганд має полярну групу, спрямовану в гідрофобну зону сайту зв’язування, це знижує афінність. Окрім водневих зв’язків, катіон-аніонні взаємодії (сольові містки) та інші електростатичні контакти теж суттєво зміцнюють комплекс. Структура дозволяє визначити пари заряджених груп на близькій відстані (наприклад, основна аміногрупа ліганду та карбоксильна група амінокислотного залишку білка). Такі контакти також можуть суттєво зменшувати K_d.

Загалом, експериментально визначена чи передбачена структура системи білок-ліганд дає якісну і кількісну інформацію: знаючи точну геометрію комплексу, можна оцінити вклад кожної нековалентної взаємодії в енергію зв’язування. Структура допомагає пояснити різниці в активності аналогів лігандів (розуміючи, які контакти відрізняються) і планувати хімічні модифікації для покращення наявних малих молекул. Втім, варто зазначити, що розрахунок афінності лише зі структури є приблизним. Іноді навіть наявність багатьох сприятливих контактів не гарантує високої афінності, через додаткові фактори ентропії, змін у конфігурації білка, вплив розчинника тощо.

Хоча детальний аналіз структури комплексу протеїн-ліганд дає важливу інформацію про афінність та її оптимізацію, отримання таких структур експериментально (наприклад, методами NMR, рентгеноструктурного аналізу чи cryo-EM) є вкрай трудомістким та дорогим процесом. Це унеможливлює використання експериментальних підходів для швидкої оцінки великої кількості потенційних сполук, що необхідно, наприклад, у проєктах віртуального скринінгу з високою пропускною здатністю. Саме тому виникає гостра потреба у швидких та ефективних обчислювальних (*in silico*) методах, здатних прогнозувати положення малих молекул у сайтах зв'язування білків.

Молекулярний докінг відіграє центральну роль у сучасній *in silico* розробці ліків. Донедавна класичний (фізичний) докінг був єдиним доступним методом прогнозування положень малих молекул у сайтах зв'язування білків-мішеней, достатньо швидким, щоб бути корисним у проєктах віртуального скринінгу з високою пропускною здатністю. Попри його надзвичайну важливість, спостерігається помітна стагнація у покращенні універсальності, точності, обчислювальних витрат та прогностичних можливостей докінгу [110]. Базуючись на спрощених емпіричних потенціалах міжмолекулярних взаємодій та відсутності явного розчинника, функції оцінки докінгу, ймовірно, досягли плато практичної точності. Попри низку останніх розробок у цій галузі, усі вони є радше поступовими вдосконаленнями та налаштуваннями для конкретних галузей, аніж технологічними проривами.

Однак, останні досягнення в DL методологіях для прогнозування положення лігандів у сайтах зв'язування білків пропонують альтернативу алгоритмам докінгу. Ці методи можна розділити на дві основні групи. Моделі першої групи використовують підхід на основі регресії. Вони розроблені з метою дуже швидкого передбачення, що перевершує традиційний докінг. Геометричні векторні перцептрони (Geometric Vector Perceptrons, GVP) [111] та еквіваріантні графові нейронні мережі (Equivariant Graph Neural Networks, EGNN) [112] є найпопулярнішими архітектурами для таких типів моделей. Такі методи, як EquiBind

[60] та TANKBind [113] продемонстрували ефективність у прогнозуванні структури зв’язування білок-ліганд без попередніх знань про сайт зв’язування (також відоме як сліпий докінг), бувши при цьому на кілька порядків швидшими за традиційні методи докінгу. Проте прогнозовані такими підходами комплекси нерідко містять нереалістичні конформації лігандів та численні стеричні зіткнення, що за якістю й надійністю структури відводить ці методи на другий план порівняно з класичним докінгом.

Друга, і на момент написання найсучасніша група підходів, використовує генеративні ML моделі, які мають потенціал бути на рівні або кращими за класичні методи докінгу з погляду точності та якості структури. DiffDock [66], перший та один із найпопулярніших інструментів у цій галузі, демонструє вищу точність порівняно з деякими традиційними методами докінгу та попередніми ML моделями у сценаріях сліпого докінгу. Він створює набагато менше стеричних зіткнень, ніж його конкуренти, і генерує реалістичні конформації лігандів. Однак підвищена точність DiffDock досягається ціною значного обчислювального навантаження, яке можна порівняти з навантаженням традиційних методів докінгу. Оскільки існує значний потенціал для покращення швидкості передбачення, генеративні моделі є перспективним напрямком в цій галузі.

Основним обмеженням, що перешкоджає подальшому вдосконаленню генеративних моделей прогнозування положення ліганду, є недостатня кількість тренувальних даних. Загальна кількість якісних експериментально визначених комплексів білок-ліганд, отриманих за допомогою методів рентгеноструктурного аналізу, cryo-EM або NMR, становить менше ніж 20000. База даних PDBbind [114], яка є класичним набором даних для машинного навчання у прогнозуванні сайтів зв’язування білків [115], [116] та положення лігандів у цих сайтах [60], [66], [113], містить 19443 білок-лігандних комплекси з анотаціями афінності зв’язування (версія 2020 року). З них 2709 комплексів включають пептидні ліганди або декілька молекул в одному сайті зв’язування; 3827 мають низьку роздільну здатність (2.5 Å і більше); 3523 містять ліганди з низькою афінністю (K_d/K_i/IC_{50} > 10 μM) і 216 не мають

достовірного значення афінності. Таким чином, існує лише 10270 високоякісних комплексів, які можна використовувати для тренування моделей.

Ще однією проблемою експериментальних комплексів обмеженість даних про ліганди. У PDBbind є 12815 окремих малих молекул-лігандів, з яких лише 1655 зустрічаються у двох або більше комплексах (Рисунок 3.1). Це вказує на те, що при тренуванні моделі здебільшого стикаються з конкретним лігандом, зв'язаним з одним білком, без будь-якої інформації про можливу варіативність його способів зв'язування. Це може призвести до перенавчання (overfitting) на єдиній позі ліганду. Крім того, білки представлені дуже нерівномірно. PDB [36] ідентифікатори 19443 протеїн-лігандних комплексів пов'язані з 4749 унікальними ідентифікаторами UniProt [84], але 1500 найпоширеніших ідентифікаторів припадає на ~80% від загальної кількості комплексів (Рисунок 3.1). Те ж саме стосується і сімейств білків: загалом присутні 1646 окремих сімейств та надсімейств білків, але на 100 найпоширеніших сімейств припадає майже 60% комплексів (Рисунок 3.1). Таким чином, попри значне загальне різноманіття білків, існує очевидна надмірна представленість деяких білків та їх сімейств, що може призвести до перенавчання та дисбалансу точності моделі в сторону найпоширеніших білків.

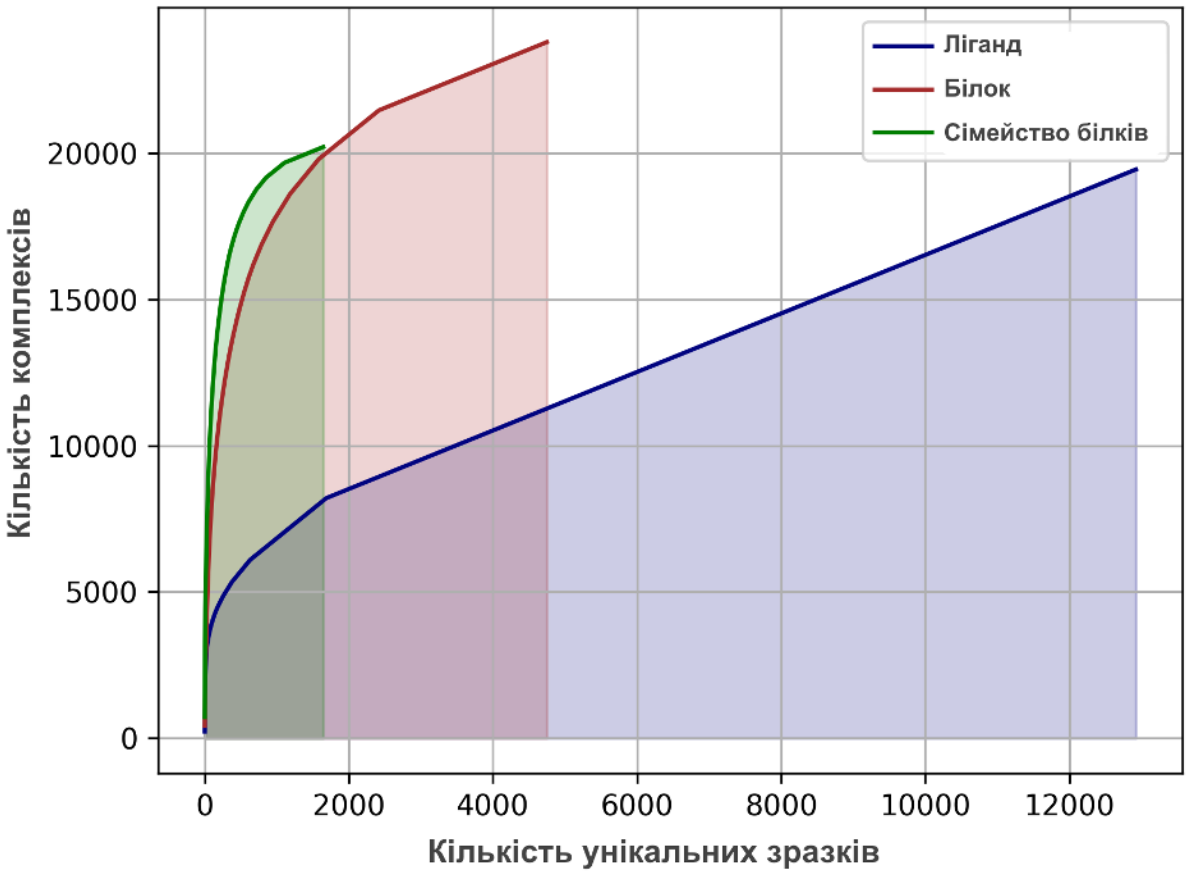

Рисунок 3.1. Кумулятивні суми зразків (ідентифікаторів) у базі даних PDBbind, віднесених до унікального ліганду, унікального білка або унікального сімейства білків. Зверніть увагу, що кількість білків та сімейств може перевищувати загальну кількість комплексів у PDBbind, оскільки одна структура PDB може бути віднесена до кількох записів.

Недостатня загальна кількість зразків, обмежене різноманіття лігандів та дисбаланс у представленні білків у доступних даних для тренування моделі погіршують точність ML підходів до прогнозування положення ліганду. Ці обмеження особливо помітні порівнюючи кількість та якість тренувальних даних з потребою працювати на довільних білкових мішенях та довільних лігандах з величезного хімічного простору [117]. Очевидно, що набір даних експериментально визначених структур не буде зростати достатньо швидко, щоб задовольнити потреби галузі ML прогнозування зв’язування ліганд-білок, що стрімко розвивається, тому потрібні інші підходи для подолання нестачі тренувальних даних.

У цьому розділі описано розробку підходу розширення (аугментації) тренувального набору білок-лігандних комплексів штучними даними, які імітують реальні сайти зв’язування білків за низкою структурних характеристик [2]. Статистичні розподіли параметрів нековалентних взаємодій штучних сайтів підлаштовуються під відповідні розподіли в реальних комплексах білок-ліганд. Ідея підходу базується на припущенні, що кількість сприятливих геометрій нековалентних взаємодій між амінокислотами білка та хімічними групами лігандів є скінченною і достатньо добре представлена у доступних експериментальних даних. Чого, ймовірно, не вистачає, так це вибірки всіх можливих комбінацій таких взаємодій у межах сайту зв’язування. Іншими словами, припущення полягає в тому, що доступні експериментальні структури надають достатню статистику щодо бажаної хімічної природи взаємодіючих пар атомів, відстаней між атомами та орієнтацій відповідних хімічних груп. Однак у реальних білках спостерігається лише дуже мала частка всіх можливих комбінацій таких пар. Якщо згенерувати велику кількість штучних комплексів, які слідують тому ж статистичному розподілу взаємодіючих пар атомів та груп, що й реальні, але охоплюють набагато ширше різноманіття їх комбінацій, можливо, вдасться розширити тренувальну вибірку інформативними даними та тренувати модель на більш повному наборі даних.

Дане припущення було перевірено експериментально. Було розроблено алгоритм для генерації штучних сайтів зв’язування, створено набір даних, що складається зі штучних та реальних комплексів, та оцінено дифузійну генеративну DL модель, яка натхненна DiffDock і навчена на таких доповнених даних - PocketCFDM (Pocket Conformation Fitting Diffusion Model). Порівняно результати навчання моделі, отримані з використанням лише штучних комплексів, лише експериментальних комплексів та їх комбінації. Було показано, що PocketCFDM перевершує DiffDock, з погляду коректності згенерованих поз лігандів (менше стеричних зіткнень та більше сприятливих нековалентних взаємодій). Обговорено майбутні перспективи методології щодо покращення її точності та швидкості.

3.1. Основні типи нековалентних взаємодій в біологічних системах

3.1.1. Гідрофобний ефект

Гідрофобний ефект є однією з найпотужніших рушійних сил у біології, що відповідає за згортання білків, формування мембран та зв'язування багатьох лігандів. Його природа є контрінтуїтивною: це не пряме притягання між неполярними молекулами, а скоріше ефект, керований властивостями розчинника - води. Молекули води мають високу когезійну енергію завдяки розгалуженій мережі водневих зв'язків. Внесення неполярної (гідрофобної) молекули у воду "розриває" цю мережу, змушуючи молекули води на межі з гідрофобом утворювати високовпорядковану, схожу на кристалічну, структуру. Цей процес супроводжується значним зменшенням ентропії системи. Коли два або більше гідрофоби асоціюють, вони мінімізують свою контактну поверхню з водою, що призводить до вивільнення частини впорядкованих молекул води в менш впорядкований розчинник. Це викликає велике та сприятливе збільшення ентропії розчинника ($\Delta S_{solv} > 0$), що є рушійною силою класичного гідрофобного ефекту. У цій моделі ентальпійний внесок (ΔH) є незначним або навіть несприятливим [107].

Однак, термодинамічний профіль гідрофобного ефекту не є універсальним і сильно залежить від контексту. Детальні калориметричні дослідження зв'язування лігандів з білками показали існування некласичного гідрофобного ефекту, який є ентальпійно-керованим ($\Delta H < 0$) [118]. Це спостерігається, коли зв'язування відбувається у відносно жорсткій, білковій кишені, де ключову роль відіграє реорганізація невеликої кількості специфічно розташованих молекул води, а не їх вивільнення. У таких випадках сприятливий ентальпійний внесок може походити від утворення більш оптимальних водневих зв'язків між молекулами води, що залишилися в активному центрі, або від сприятливих ван-дер-ваальсових контактів між лігандом та білком, які стають можливими після витіснення води. Важливо розрізняти гідрофобний ефект, який є термодинамічним явищем, що змушує неполярні групи зближуватися, та ван-дер-ваальсові сили (зокрема, лондонські

дисперсійні сили), які є електростатичними взаємодіями, що діють між цими групами вже після їх зближення.

3.1.2. Ван-дер-ваальсові контакти та стеричні зіткнення

Ван-дер-ваальсові сили є універсальними взаємодіями, що діють між усіма атомами. Вони виникають через тимчасові флуктуації в розподілі електронної густини, які створюють миттєві диполі. Ці диполі індукують відповідні диполі в сусідніх атомах, що призводить до слабкого, короткотривалого притягання, відомого як лондонські дисперсійні сили. Хоча окремі контакти є дуже слабкими (0.1-1 ккал/моль), їх сумарний внесок у білок-лігандному комплексі є значним через велику кількість атомів, що знаходяться в тісному контакті на комплементарних поверхнях [119].

Проте, коли атоми підходять занадто близько, їхні електронні орбіталі починають перекриватися. Згідно з принципом заборони Паулі, два електрони не можуть займати один і той самий квантовий стан. Це призводить до надзвичайно сильного відштовхування, яке стрімко зростає зі зменшенням відстані. Це і є стеричне зіткнення.

Для моделювання цих взаємодій широко використовується потенціал Леннарда-Джонса, який описує потенційну енергію $V(r)$ взаємодії між двома незв'язаними атомами як функцію відстані r між ними:

$$V(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \quad (3.1)$$

- Член притягання ($-\left(\frac{\sigma}{r}\right)^{6}$): описує лондонські дисперсійні сили, які домінують на оптимальних відстанях і сприяють асоціації молекул.
- Член відштовхування ($\left(\frac{\sigma}{r}\right)^{12}$): описує сильне відштовхування Паулі, що виникає на дуже коротких відстанях, коли електронні оболонки атомів починають перекриватися. Цей член моделює стеричні зіткнення.

Параметр ε (глибина потенційної ями) визначає силу притягання, а σ - це відстань, на якій потенційна енергія дорівнює нулю.

Зв’язування ліганду, який має високу комплементарність форми з протеїном, максимізує площу контакту у сайті зв'язування. Це дозволяє підсумовувати тисячі слабких, але сприятливих дисперсійних взаємодій, що дає значний негативний внесок в ΔH. Водночас будь-яке стеричне зіткнення робить величезний позитивний внесок в ΔH, що робить зв’язування неможливим. Таким чином, ван-дер-ваальсові взаємодії є головним рушієм специфічності форми: ліганд повинен ідеально заповнити сайт зв’язування, щоб максимізувати сприятливі дисперсійні сили та уникнути стеричних зіткнень.

3.1.3. Водневий зв’язок

Водневий зв’язок є найважливішою спрямованою нековалентною взаємодією в біології. Він утворюється між атомом водню, ковалентно зв’язаним з електронегативним атомом (донором, D-H, таким як O-H або N-H), та іншим електронегативним атомом, що має вільну електронну пару (акцептором, A, таким як кисень або азот). Ця взаємодія має переважно електростатичну природу, але також містить елементи поляризації та часткового ковалентного зв’язку. Енергія водневих зв’язків може варіюватися в широкому діапазоні (1-10 ккал/моль) залежно від типу донора та акцептора, їх оточення та геометрії. Ключовою особливістю водневого зв’язку є його висока спрямованість: він є найміцнішим, коли кут D-H···A наближається до 180°, а відстань між D та A є оптимальною. Саме ця геометрична вибагливість робить водневі зв’язки головними архітекторами молекулярної специфічності, дозволяючи білкам розпізнавати свої ліганди з високою точністю [120].

Термодинамічний внесок водневого зв’язку у вільну енергію зв’язування у водному середовищі є досить складним. Утворення сприятливого водневого зв’язку між білком та лігандом неминуче вимагає розриву наявних водневих зв’язків, які кожна з цих груп утворювала з молекулами води до зв’язування. Таким чином, чистий

ентальпійний виграш є різницею між енергією нового зв’язку та енергією розірваних зв’язків з водою. Цей процес десольватації часто є ентальпійно невигідним ($\Delta H > 0$). Проте, загальний процес зв’язування залишається сприятливим завдяки великому ентропійному виграшу від витіснення двох структурованих молекул води з сайту зв’язування. Це яскраво ілюструє явище ентальпійно-ентропійної компенсації та показує, що рушійною силою утворення водневого зв’язку в біологічних системах часто є не стільки його власна енергія, скільки сприятлива реорганізація системи білок-ліганд-розчинник в цілому. Для максимізації внеску водневого зв’язку в афінність, необхідно мінімізувати “штраф” за конкуренцію з водою. Це досягається, коли донор та акцептор є порівнянними за силою: або обидва значно сильніші, або обидва значно слабші у формуванні водневих зв’язків, ніж вода [121].

3.1.4. Галогенний зв’язок

Галогенний зв’язок - це тип нековалентної взаємодії, що відіграє важливу роль у молекулярному розпізнаванні та дизайні ліків. Він являє собою спрямовану, притягуючу взаємодію між електрофільною ділянкою на атомі галогену (Cl, Br, I), ковалентно зв’язаним з іншим атомом, та нуклеофільною ділянкою (атомом з вільною електронною парою, наприклад, киснем карбонільної групи). Фізична природа цієї взаємодії пояснюється анізотропним розподілом електронної густини навколо атома галогену. Уздовж осі ковалентного зв’язку D-X (D - атом, ковалентно зв’язаний з галогеном, X - галоген) утворюється ділянка збідненої електронної густини, так званий σ-отвір, який має позитивний електростатичний потенціал і може діяти як кислота Льюїса, притягуючи основу Льюїса [122].

Термодинамічно, галогенний зв’язок є переважно ентальпійно-керованим. Ізотермічна титраційна калориметрія (ITC) для інгібіторів фосфодіестерази 5 (PDE5) продемонструвала чітку залежність сили зв’язування від природи галогену: йод (-5.59 кДж/моль) утворює значно міцніший зв’язок, ніж бром (-3.09 кДж/моль) та хлор (-1.57 кДж/моль) [123]. Ця тенденція (I > Br > Cl) прямо корелює зі збільшенням розміру та поляризованості атома галогену, що призводить до

утворення більшого та більш позитивно зарядженого σ-отвору. Фтор, через свою високу електронегативність та низьку поляризованість, зазвичай не утворює значущих галогенних зв'язків.

Подібно до водневого зв'язку, галогенний зв'язок є високоспрямованим. Найбільш сприятлива геометрія досягається, коли кут D-X···A наближається до 180°, що дозволяє максимальне перекривання σ-отвору з вільною парою акцептора. Ця властивість часто використовується при дизайні ліків для точного позиціювання ліганду в активному центрі білка, підвищуючи як афінність, так і селективність [123].

3.1.5. Катіон-аніонні взаємодії

Катіон-аніонні взаємодії, широко відомі як сольові містки, є потужними нековалентними контактами, що утворюються між формально протилежно зарядженими групами. У біологічних системах вони найчастіше виникають між протонованими бічними ланцюгами лізину (R-NH_3^+) або аргініну (R-NH-$C(NH_2)_2^+$) та депротонованими карбоксильними групами аспартату або глутамату (R-COO^-), а також за участю заряджених функціональних груп у лігандах. Фізична природа сольового містка є комбінацією далекодіючої електростатичної (іонної) взаємодії, що описується законом Кулона:

$$E = k\frac{q_1 q_2}{\epsilon r}, \tag{3.2}$$

де E - енергія кулонівської взаємодії, q_1, q_2 - величини зарядів, r - відстань між ними, ϵ - діелектрична проникність середовища, а k - коефіцієнт пропорційності;

та короткодіючих, спрямованих водневих зв'язків, які часто утворюються між донорними N-H групами катіона та акцепторними атомами кисню аніона.

Енергія іонної взаємодії у вакуумі є дуже великою, однак у водному середовищі ситуація кардинально змінюється. Вода, як розчинник з високою діелектричною

проникністю ($\epsilon \approx 80$), ефективно екранує електростатичні взаємодії, значно їх послаблюючи. Щобільше, іони у водному розчині сильно гідратовані, оточені впорядкованими оболонками з молекул води. Процес утворення сольового містка вимагає руйнування цих гідратних оболонок (десольватації), що є ентальпійно дуже невигідним процесом ($\Delta H >> 0$). Цей ентальпійний "штраф" за десольватацію часто майже повністю або навіть повністю компенсує ентальпійний виграш від утворення іонної пари. В результаті, всупереч інтуїтивним очікуванням, утворення сольового містка у воді є процесом, який переважно керований ентропією. Рушійною силою є велике позитивне значення зміни ентропії ($\Delta S > 0$), що виникає внаслідок вивільнення великої кількості впорядкованих молекул води з гідратних оболонок іонів у менш впорядкований стан [124].

Сила сольового містка залежить від його оточення. Сольові містки, заховані всередині білка, де діелектрична проникність значно нижча ($\epsilon \approx 4$-10), є набагато міцнішими, ніж ті, що знаходяться на поверхні, експонованій до розчинника. Типовий внесок поверхневого сольового містка у стабільність білка або афінність зв'язування оцінюється в діапазоні 1-5 ккал/моль.

3.1.6. Взаємодії з ароматичними системами

Ароматичні π-системи, завдяки своєму унікальному електронному розподілу - електронно-збагаченій поверхні над та під площиною кільця та електронно-збідненому периметру - є надзвичайно універсальними партнерами у нековалентних взаємодіях. Вони здатні утворювати стабілізуючі контакти з катіонами, іншими ароматичними системами та навіть з полярними амідними групами поліпептидного остову, відіграючи важливу роль у молекулярному розпізнаванні.

π-π стекінг - це взаємодія між двома ароматичними кільцями, що є фундаментальною для стабілізації ДНК та зв'язування багатьох лігандів. Її природа є складною комбінацією електростатики (квадруполь-квадруполь) та дисперсійних

сил. Основними геометріями є паралельна зі зміщенням та Т-подібна (край до грані, перпендикулярна) конфігурації, які дозволяють сприятливу взаємодію між негативною поверхнею одного кільця та позитивним периметром іншого [125].

Катіон-π взаємодія виникає між катіоном білка (наприклад, бічними ланцюгами лізину або аргініну) чи ліганду та електроно-збагаченою поверхнею ароматичного кільця білка (фенілаланін, тирозин, триптофан) чи ліганду. Ця переважно електростатична взаємодія (монополь-квадруполь) є найсильнішою, коли катіон розташований над центром кільця. У біологічному контексті існує критична різниця між лізином та аргініном. Хоча у газовій фазі взаємодія з локалізованим зарядом лізину є сильнішою, у полярному середовищі білка вона значно послаблюється. Натомість взаємодія з делокалізованим зарядом аргініну має збалансований характер з рівними внесками електростатики та дисперсійних сил, що робить її набагато стійкішою до ефектів розчинника. Як наслідок, катіон-π контакти за участю аргініну зустрічаються частіше і роблять значніший внесок в афінність [126].

Амід-π взаємодія виникає між π-системою амідного зв’язку (з поліпептидного остову або аспарагіну чи глутаміну) та ароматичним кільцем. Цей контакт має електростатичну природу (диполь-квадруполь) з переважно Т-подібними геометріями [127].

3.2. Генерація штучних систем протеїн-мала молекула для розширення тренувальних даних моделей машинного навчання

3.2.1. Попередня обробка протеїнів та малих молекул в експериментальних комплексах

Python API хемоінформатичного пакета RDKit v. 2021.03 було використано для завантаження, обробки та генерації характеристик малих молекул. Було розроблено спеціальний модуль обробки білків для вилучення даних про білки з файлів PDB та генерації необхідних характеристик для тренування моделі. Цей модуль використовує назви атомів PDB для отримання ознак графа на рівні атомів, замість

того, щоб покладатися на стороннє програмне забезпечення для їх отримання. Такий підхід знижує частку виключення оброблених білків через часті артефакти у файлах PDB, що призводять до неможливості їх обробки за допомогою RDKit.

Для оцінки взаємодій білок-ліганд було використано пошук за підструктурами SMARTS (SMILES arbitrary target specification) для класифікації атомів ліганду або хімічних груп за такими категоріями: гідрофобні, ароматичні, амідні, донор, акцептор, катіон, аніон або галоген. Деталі класифікації та безпосередньо SMARTS репрезентації доступні у вихідному коді модуля попередньої обробки: https://github.com/vtarasv/rai-chem.git. Характеристики атомів білка отримувалися з використанням стандартних назв атомів PDB.

Перед оцінкою нековалентних взаємодій білки та ліганди протонували (додавали водні, відсутні у структурі). Протонування білків виконувалося аналогічно до EquiBind [60] та DiffDock [66] за допомогою програмного забезпечення reduce (https://github.com/rlabduke/reduce.git). Розглядались лише амінокислоти як об’єкти, що взаємодіють з лігандами, тоді як молекули води, іони та атоми металів, присутні в експериментальних структурах, не враховувалися.

3.2.2. Вибір нековалентних взаємодій

Між білком та лігандом оцінювалися, перелічені в таблиці 3.1. Також, було враховано енергетично невигідні (несприятливі) контакти, такі як пари атомів донор-донор на близькій відстані, щоб покращити загальну якість штучних комплексів шляхом виключення таких взаємодій. Вибір взаємодій та їх параметрів базується на компромісі між численними підходами в літературі [107], [128], [129], [130], загальновживаним хемоінформатичним програмним забезпеченням [94], [131], [132] та популярним інструментом молекулярного моделювання BIOVIA Discovery Studio Visualizer.

Таблиця 3.1. Нековалентні взаємодії та енергетично несприятливі контакти між білком і малою молекулою та їх параметри. Відстані для ароматичних кілець і амідних груп вимірюються відносно їх геометричних центроїдів. θ означає кут між нормалями ароматичної/амідної площини. Для водневих і галогенових зв'язків D є донором, ковалентно зв'язаним з воднем/галогеном; H/X - атом водню/галогену; A - акцептор; Y є ковалентно зв'язаним сусідом акцептора.

Тип	**Взаємодія**	**Орієнтація**	**Мін. відстань,** Å	**Макс. відстань,** Å	**Вимоги до кутів, °**
Нековалентні взаємодії					
Гідрофобна	Гідрофобні атоми (C, S) - Гідрофобні атоми (C, S, Cl, Br, I)			4.5	
Гідрофобна	Гідрофобні атоми (C, S) - Гідрофобні атоми (F)			3.9	
π-π стекінг	Ароматичне кільце - Ароматичне кільце	Паралельна		4.5	$\theta \in \{-30,30\}$
π-π стекінг	Ароматичне кільце - Ароматичне кільце	Перпендикулярна		5.5	$\theta \in \{60,120\}$
Амід-π	Ароматичне кільце - Амідна група	Паралельна		4.5	$\theta \in \{-30,30\}$
Амід-π	Ароматичне кільце - Амідна група	Перпендикулярна		5.0	$\theta \in \{60,120\}$
Водневий зв'язок	Донор - Акцептор		2.5	3.9	∠D-H-A ≥ 90 ∠H-A-Y ≥ 90
Галогенний зв'язок	Cl, Br, I - Акцептор			4.0	∠D-X-A ≥ 120 ∠X-A-Y ≥ 75
Електростатична	Катіон - Аніон		3.4	4.5	
Катіон-π	Катіон - Ароматичне кільце	Паралельна (над площиною кільця)	3.4	4.5	$\theta \in \{-40,40\}$

Таблиця 3.1. (Продовження)

Несприятливі контакти					
	Донор - Донор			3.4	
	Акцептор - Акцептор			3.2	
	Катіон - Катіон			5.2	
	Аніон - Аніон			5.2	

3.2.3. Отримання статистичних закономірностей нековалентних взаємодій з експериментально визначених структур систем протеїн-мала молекула

Комплекси протеїн-ліганд з відомими 3D-структурами були взяті з набору даних PDBbind [114]. Використовувалися лише ті комплекси, що задовольняли наступні критерії: наявність одного ліганду, роздільна здатність нижче 2.5 Å, та афінність менше ніж 10 μM. Всього було відібрано 10270 білок-лігандних комплексів. Серед них, 1805 файлів лігандів не обробилися RDKit. В результаті, були використані 8465 комплексів.

Для комплексів було отримано таку статистичну інформацію:

- Імовірність участі певної підструктури ліганду (конкретного типу атома, ароматичного кільця або амідної групи) у певній взаємодії білок-ліганд з таблиці 3.1.
- Розподіл кількості підструктур сайту зв’язування, які пов’язані з підструктурами ліганду через певний тип взаємодії.
- Розподіл амінокислот, залучених до певних типів взаємодій.
- Розподіли відстаней та кутів для певних типів взаємодій.

3.2.4. Алгоритм генерації штучних систем протеїн-мала молекула на основі статистичних закономірностей

Було використано програмне забезпечення PeptideBuilder [133] для створення колекції з 20 амінокислот у форматі PDB, а також 400 дипептидів, що представляють кожну можливу перестановку двох амінокислот. Амінокислоти та дипептиди були фланковані залишками амінокислоти гліцину (GLY) з обох боків і слугували основними будівельними блоками для штучних сайтів зв'язування. Включення фланкуючих залишків GLY допомагає генерувати правильні конформери пептидів, які обмежені пептидними зв'язками з кожного боку будівельних блоків.

Генерація штучного комплексу - це ітеративний процес розміщення будівельних блоків навколо довільного конформера малої молекули з метою формування реалістичної мережі нековалентних взаємодій та уникнення несприятливих контактів. Під час конструювання сайту враховувалося загалом 13 потенційних нековалентних взаємодій, як спрямованих, так і неспрямованих (Таблиця 3.2).

Таблиця 3.2. Ознаки (атоми/групи) молекул та відповідні нековалентні взаємодії, що враховувалися під час конструювання штучного комплексу білок-ліганд.

#	**Ознака білка**	**Ознака ліганд**	**Тип**
1	Ароматичне кільце	Ароматичне кільце	π-π стекінг
2	Амідна група	Ароматичне кільце	Амід-π
3	Ароматичне кільце	Амідна група	Амід-π
4	Ароматичне кільце	Атом катіона	Катіон-π
5	Донор	Акцептор	Водневий зв’язок
6	Акцептор	Донор	Водневий зв’язок
7	Акцептор	Атом галогену (Cl, Br, I)	Галогенний зв’язок
8	Атом катіона	Атом аніона	Електростатична
9	Атом аніона	Атом катіона	Електростатична
10	Атом катіона	Ароматичне кільце	Катіон-π
11	Атом C або S	Атом F	Гідрофобна
12	Атом C або S	Атом Cl, Br або I	Гідрофобна
13	Атом C або S	Атом C або S	Гідрофобна

Конструювання сайту починається з конкретного конформера малої органічної молекули. Для кожної з 13 взаємодій, перелічених в таблиці 3.2, виконуються наступні кроки:

1. Знайти всі ознаки ліганду L, які сумісні з поточним типом взаємодії i. Кожна така ознака позначається як F_{Li}.

2. Маючи експериментальну ймовірність P знаходження F_{Li} серед усіх взаємодій типу i та випадкове значення p ($0 \leq p \leq 1$), визначити, чи слід додавати нову взаємодію ($p \leq P$), або перейти до наступної взаємодії ($p > P$).

3. Вибрати пептидний будівельний блок B для розміщення в штучному сайті, взявши всі будівельні блоки з ознаками, сумісними з i, та випадковим чином обравши один з них відповідно до експериментально визначеної ймовірності участі відповідного залишку у взаємодії i.

4. Для обраного будівельного блоку B згенерувати випадковий конформер, враховуючи пептидні зв'язки з фланкуючими залишками GLY. Видалити фланкуючі залишки GLY.

5. Випадковим чином вибрати відстань d між F_{Li} та відповідною ознакою в B з експериментально визначених розподілів і розмістити будівельний блок B у визначеному місці.

6. Повторювати наступні кроки, доки не буде виявлено стеричних зіткнень або несприятливих контактів між будівельним блоком та лігандом і між самими будівельними блоками:

   6.1. Випадковим чином обертати та переміщувати будівельний блок B, зберігаючи відстань d.

   6.2. Визначити, чи виконуються вимоги до кутів при взаємодії i (якщо взаємодія передбачає такі вимоги). Якщо ні - повторити спробу з кроку 6.1.

6.3. Якщо досягнуто максимальної кількості спроб (2,000 за замовчуванням), будівельний блок пропускається.

За допомогою цього алгоритму було згенеровано 8465 штучних сайтів (по одній для кожного ліганду з комплексу з бази даних PDBbind). Були обчислені статистичні закономірності взаємодій у згенерованих комплексах та порівняні з експериментальними розподілами. Через достатній збіг розподілів, подальше налаштування алгоритму не було потрібне. Приклади випадково обраних штучних сайтів зображено на Рисунку Д1.

3.2.5. Порівняння штучних та експериментально визначених систем протеїн-мала молекула

В результаті аналіз набору експериментальних протеїн-лігандних комплексів PDBbind було виявлено 343784 міжмолекулярні нековалентні взаємодії. Як і очікувалося, гідрофобні контакти були найпоширенішим типом взаємодії, із загальною кількістю 267774 контактів. Переважна більшість цих взаємодій відбувалася між гідрофобними атомами вуглецю або сірки. Водневі зв’язки були другою за поширеністю формою взаємодії, трапляючись приблизно один раз на кожні п’ять гідрофобних контактів. Решта контактів становила менше 3% від загальної кількості (Рисунок 3.2 та Таблиця Е1).

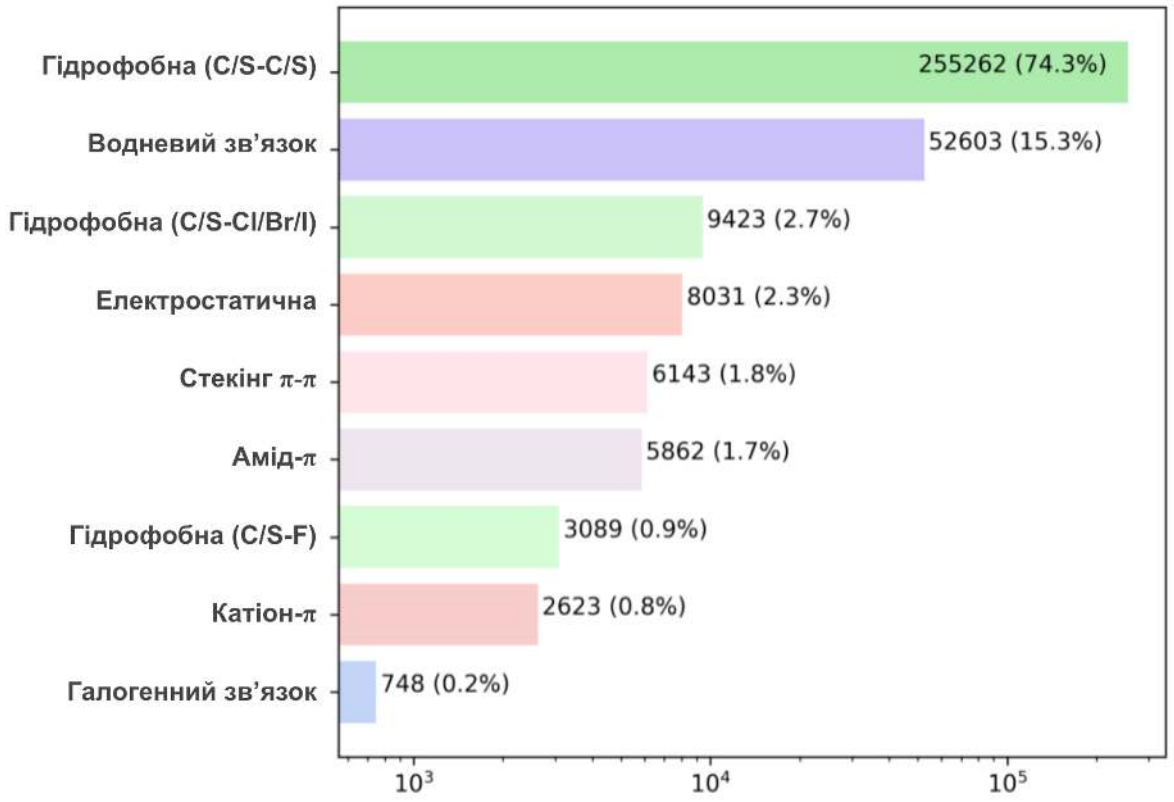

Рисунок 3.2. Кількість нековалентних білок-лігандних взаємодій у експериментальних комплексах PDBbind. Кожен стовпець позначений загальною кількістю знайдених взаємодій та їх часткою.

Загальна кількість взаємодій у штучних сайтах зв'язування, що включають ті самі ліганди, склала 453593. Ця значно більша кількість, порівняно з експериментальними комплексами, спричинена утворенням непередбачених взаємодій (переважно гідрофобних) під час розміщення будівельних блоків сайту у безпосередній близькості до ліганду. Алгоритм генерації комплексів не враховує внесок розчинника (води) у статистику експериментальних взаємодій. Це також, імовірно, призводить до утворення додаткових взаємодій з білком при генерації, які частково заміщують взаємодії ліганд-розчинник.

Далі порівно розподіли параметрів для кожного типу взаємодії між реальними та штучними комплексами. Для кожного типу взаємодії також обчислено ймовірність залучення конкретних амінокислот сайту зв'язування білка у взаємодію.

На Рисунку 3.3 показані статистичні закономірності гідрофобних взаємодій у реальних та штучних комплексах. Відповідність між розподілами задовільна, але розподіли відстаней є більш монотонними для штучних сайтів порівняно з експериментальними. Це артефакт, спричинений частим утворенням непередбачених гідрофобних контактів під час генерації інших типів взаємодій через велику кількість гідрофобних атомів як в амінокислотах, так і в малих молекулах. Розподіли взаємодій між амінокислотами дуже схожі, за винятком помітно збільшеного залучення PHE та TRP у штучних сайтах. Це пояснюється перекриттям між гідрофобними та π-π стекінг взаємодіями, яке не компенсується в алгоритмі задля обчислювальної ефективності.

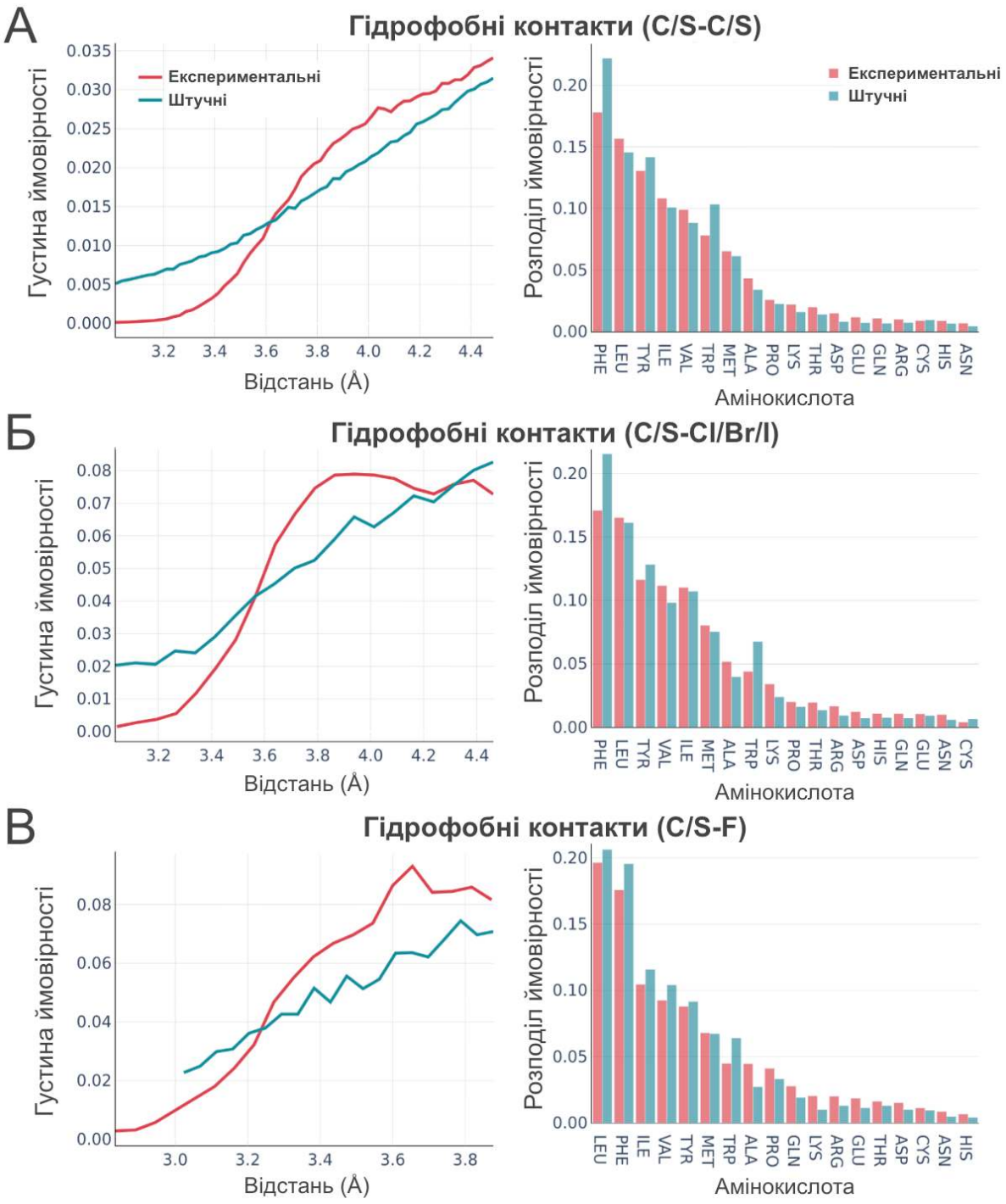

Рисунок 3.3. Розподіл відстаней гідрофобних взаємодій (ліворуч) та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах. Розділено взаємодії, що включають атоми C або S білка та ліганду (А); атоми Br, Cl або I ліганду (Б); атоми F ліганду (В).

На Рисунку 3.4 показані статистичні закономірності π-π стекінг взаємодій. Існує два типи цих взаємодій: паралельні та перпендикулярні (Т-подібні). Розподіли відстаней обох типів взаємодій близькі в реальних та штучних сайтах зв'язування, так само як і залучення різних ароматичних амінокислот. Однак, розподіли θ (кут між нормалями ароматичних площин) для паралельних взаємодій у штучних сайтах зміщені в бік 90° на 10-15° порівняно з експериментальними, оскільки алгоритм генерації перевіряє лише, чи знаходиться ароматичне кільце в межах заданого діапазону кутів.

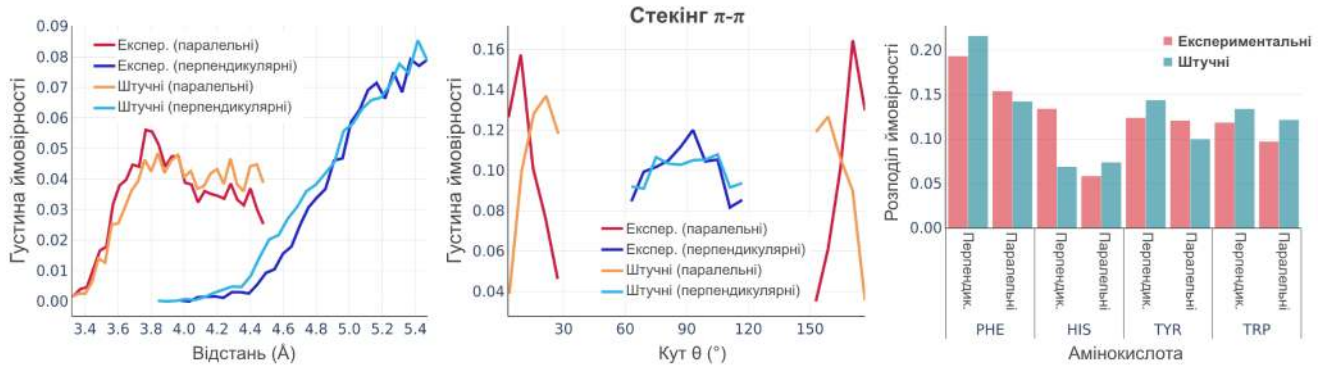

Рисунок 3.4. Розподіл відстаней π-π стекінг взаємодій між центроїдами (ліворуч), кутів θ між нормалями ароматичних кілець (посередині), та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах.

На Рисунку 3.5 показані статистичні закономірності амід-π взаємодій, які також можна класифікувати як паралельні та перпендикулярні. Подібно до π-π стекінгу, розподіли відстаней дуже схожі між реальними та штучними сайтами, тоді як θ у штучних сайтах дещо зміщені в бік 90°. Встановлено, що основними амінокислотними залишками, які слугують донорами амідної групи, є гліцин, який експонує пептидний зв'язок назовні (до ліганду) через відсутність бічного ланцюга, а також аспарагін та глутамін, які мають амідну групу на кінці своїх бічних ланцюгів. Найпоширенішим залишком, що несе ароматичне кільце в амід-π зв'язках, є HIS, на який припадає більшість перпендикулярних взаємодій в реальних комплексах. У штучних сайтах значно менше контактів амід-HIS, оскільки

несприятливі контакти донор-донор між нітрогенами кільця HIS та амідною групою були виключені під час генерації сайту.

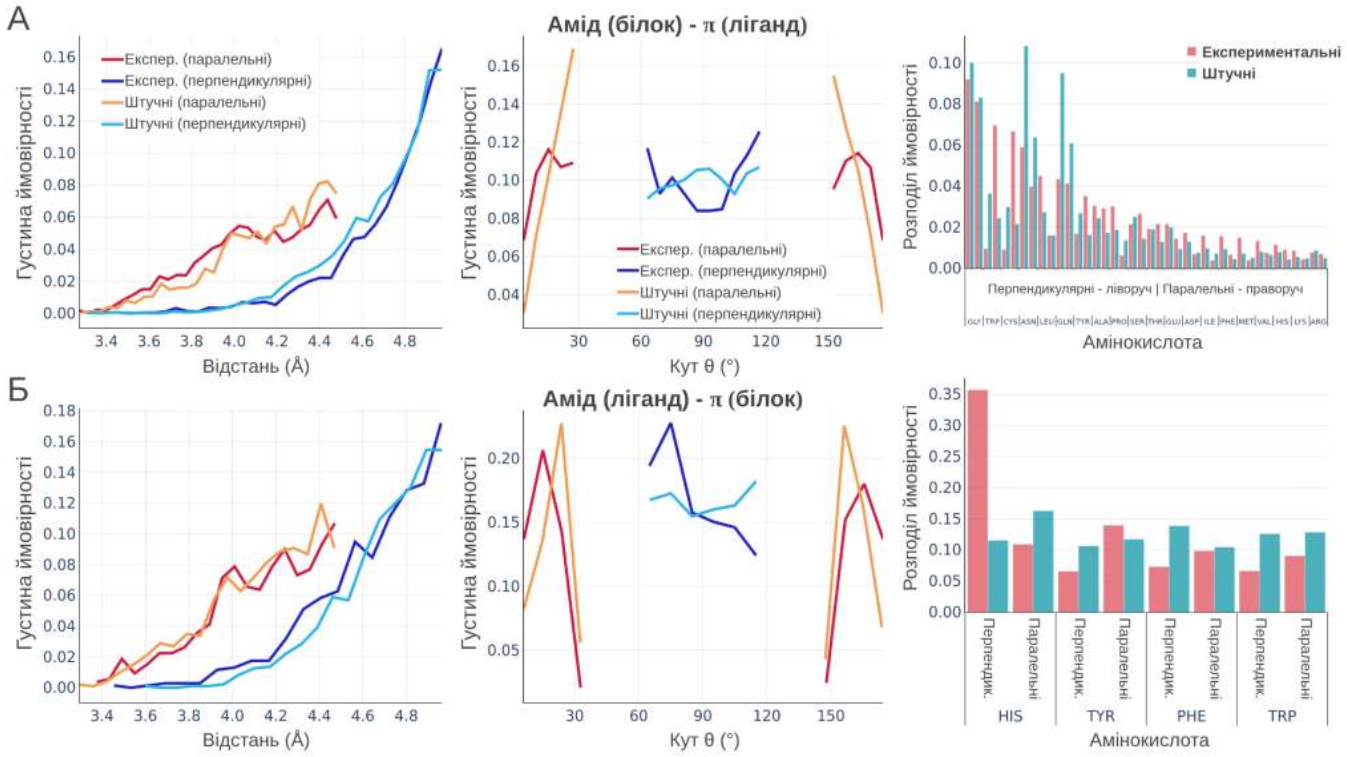

Рисунок 3.5. Розподіл відстаней амід-π взаємодій між центроїдами (ліворуч), кутів θ між нормалями ароматичного кільця та амідної площини (посередині), та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах. Розділено взаємодії, що відбуваються між амідом білка та ароматичним кільцем ліганду (А); амідом ліганду та ароматичним кільцем білка (Б).

Розподіли водневих зв’язків показані на Рисунку 3.6. Водневі зв’язки становлять найбільшу проблему з погляду генерації штучних комплексів, оскільки спричиняють найбільші розбіжності між реальними та штучними сайтами зв’язування, що особливо помітно для розподілів кутів. Загалом, алгоритм генерації має тенденцію недооцінювати кути D-H-A (важливість розподілу кутів була свідомо зменшена для прискорення алгоритму) і не вловлює пік відстані на 2.8-2.9 Å, який спостерігається в реальних комплексах. У штучних системах частіше спостерігаються віддалені контакти (короткі контакти, як правило, відкидаються під час генерації, оскільки

вони часто призводять до стеричних зіткнень між атомами, що не беруть участі у взаємодії). Ці компроміси дозволяють підтримувати достатню швидкість алгоритму для практичного використання (загалом, для подальшої оцінки алгоритму було згенеровано 640000 унікальних штучних комплексів). Водночас залучення різних амінокислот відтворюється добре у штучних сайтах.

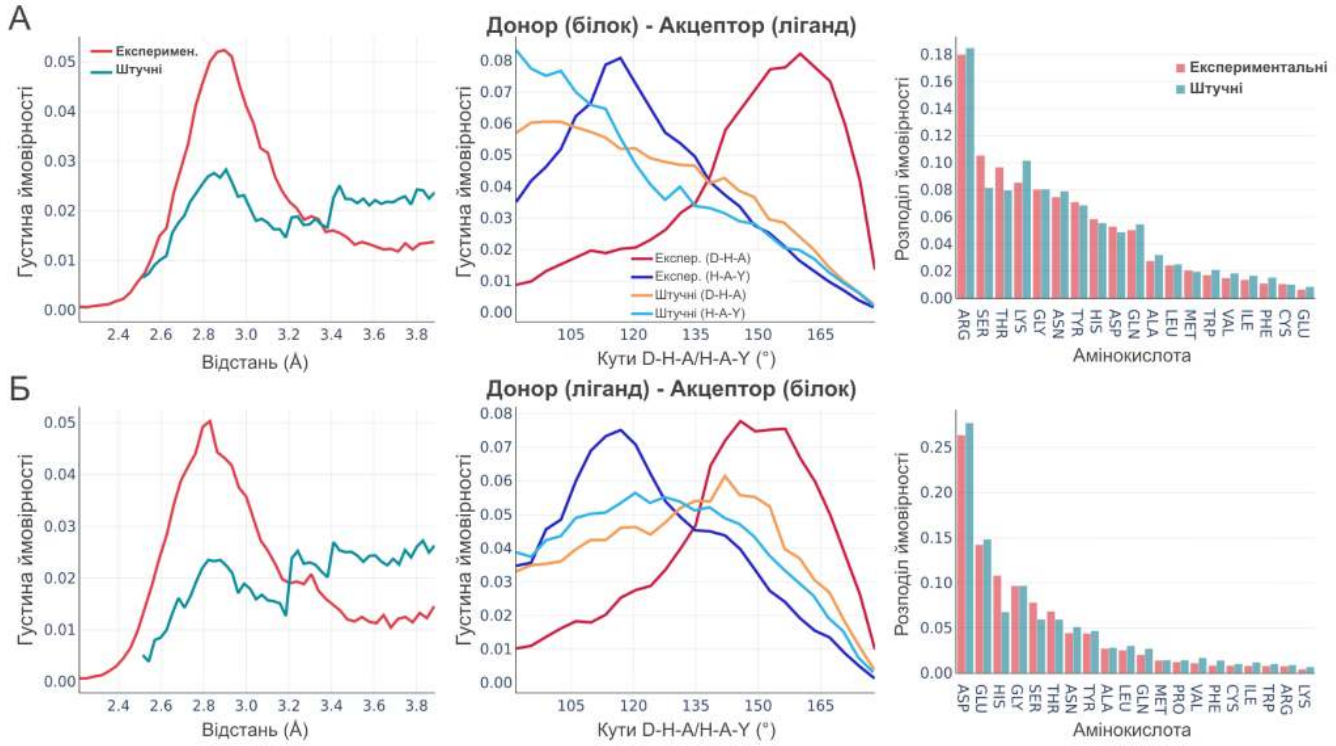

Рисунок 3.6. Розподіл відстаней водневих зв'язків (ліворуч), кутів D-H-A/H-A-Y (посередині), та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах. Розділено взаємодії, що відбуваються між донором білка та акцептором ліганду (А); донором ліганду та акцептором білка (Б). У D-H-A/H-A-Y кутах, D є донором, ковалентно зв'язаним з воднем; H - атом водню; A - акцептор; Y є ковалентно зв'язаним сусідом акцептора.

Галогенні зв'язки є найменш поширеним типом взаємодії в реальних системах з PDBbind. Беручи до уваги невелику кількість спостережуваних взаємодій цього типу, експериментальний та штучний розподіли є достатньо схожими (Рисунок 3.7).

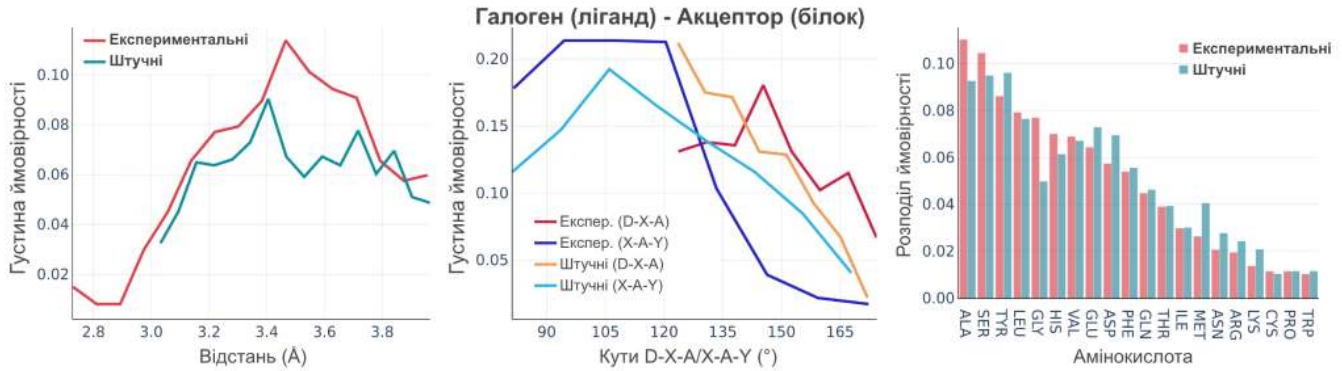

Рисунок 3.7. Розподіл відстаней галогенних зв’язків (ліворуч), кутів D-X-A/X-A-Y (посередині), та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах. У D-X-A/X-A-Y кутах, D є донором (аналогічно до водневого зв’язку), ковалентно зв’язаним з галогеном; X - атом галогену; A - акцептор галогенного зв’язку; Y є ковалентно зв’язаним сусідом акцептора.

Статистичні закономірності для електростатичних взаємодій між катіоном та аніоном зображена на Рисунку 3.8. Розподіл заряджених амінокислот відтворюється у штучних сайтах майже ідеально, тоді як розподіли відстаней дещо відрізняються в реальних та штучних комплексах. Штучні системи мають більше взаємодіючих пар на більших відстанях, ніж експериментальні, що можна пояснити підвищеною складністю розміщення амінокислотних залишків сайту в безпосередній близькості до малої молекули. Таке розміщення часто призводить до стеричних зіткнень і тому часто відкидається алгоритмом.

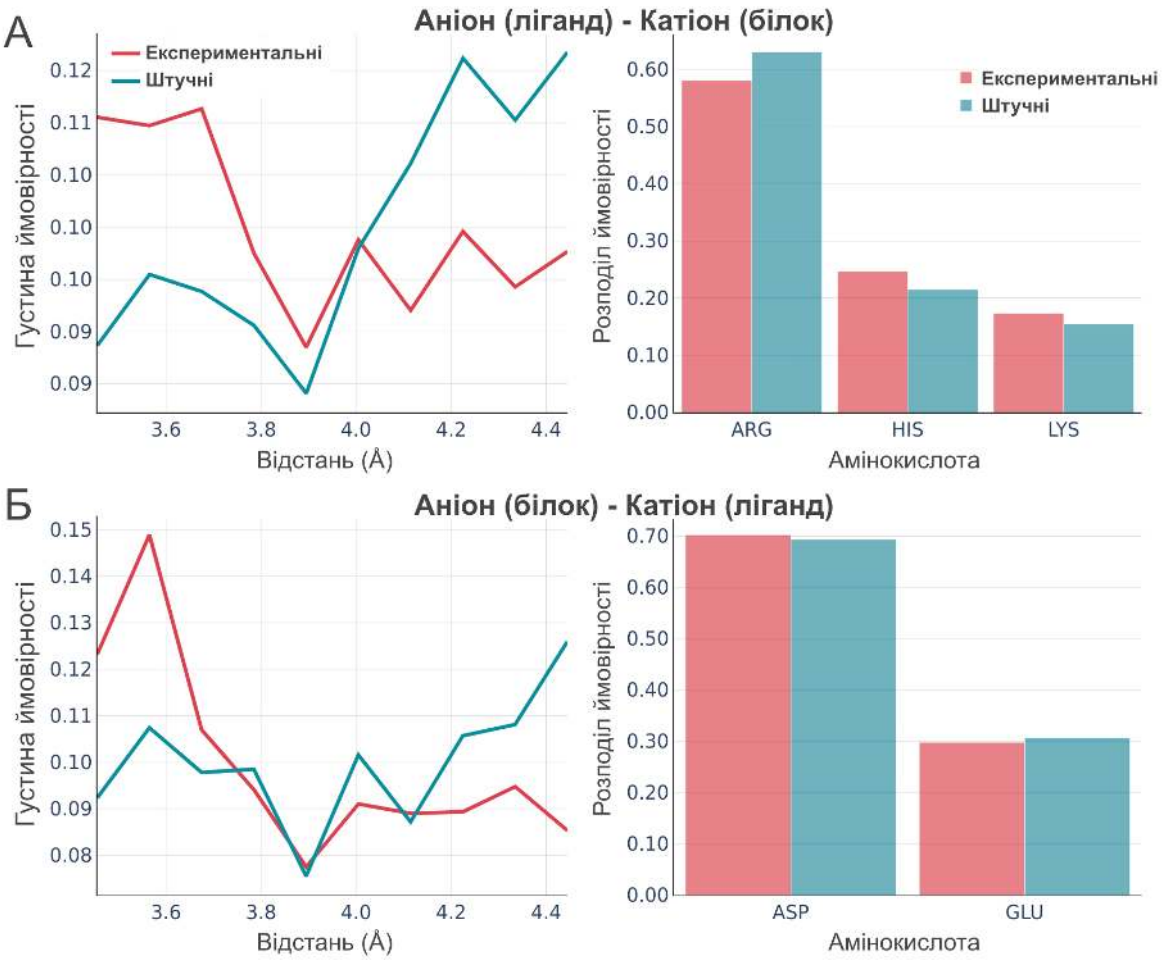

Рисунок 3.8. Розподіл відстаней катіон-аніонних взаємодій (ліворуч), та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах. Розділено взаємодії, що відбуваються між катіоном білка та аніоном ліганду (А); катіоном ліганду та аніоном білка (Б).

Останній тип взаємодії - це пари катіон-π (Рисунок 3.9). Вони є відносно рідкісними, і розподіли їхніх параметрів дуже схожі в реальних та штучних сайтах, без будь-яких суттєвих відмінностей.

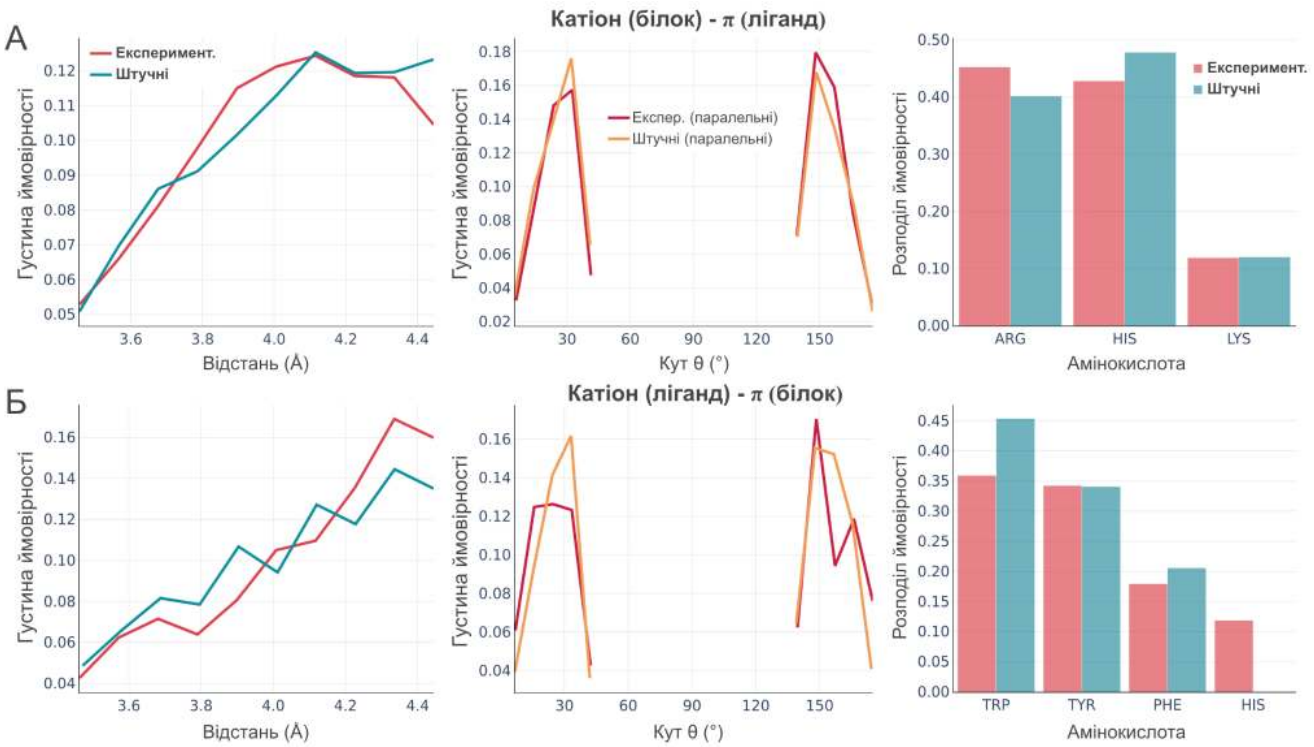

Рисунок 3.9. Розподіл відстаней катіон-π взаємодій між центроїдом ароматичного кільця та атомом катіона (ліворуч), кутів θ між нормаллю ароматичного кільця та катіоном (посередині), та частота амінокислот білка, що беруть участь у взаємодії (праворуч), між білком та малою молекулою в експериментальних та штучно згенерованих комплексах. Розділено взаємодії, що відбуваються між катіоном білка та ароматичним кільцем ліганду (А); катіоном ліганду та ароматичним кільцем білка (Б).

3.3. Використання штучних систем протеїн-мала молекула для оптимізації передбачення положення малих молекул

3.3.1. Набори даних для моделі машинного навчання

Щоб охопити максимальну різноманітність лігандів, було використано базу даних ZINC20 комерційно доступних органічних молекул, що широко використовується для віртуального скринінгу [134]. Було обрано категорію хімічних речовин In-Stock (в наявності), що включала 13 мільйонів молекул, представлених у вигляді SMILES. Сполуки були стандартизовані за допомогою ChEMBL Structure Pipeline [135].

Зокрема, були виключені дублікати, сполуки з менш ніж сімома атомами (не включаючи гідрогени) та великі молекули з молекулярною масою понад 750 г*моль$^{-1}$, що скоротило список до 9041707 сполук. Підмножина сполук була випадковим чином відібрана та використана для генерації штучних комплексів ліганд-білок, на яких надалі тренувалась дифузійна модель машинного навчання для молекулярного докінгу (деталізовано далі).

На додаток до штучних комплексів, до навчальної множини були додані експериментальні комплекси ліганд-сайт зв'язування білка з бази даних PDBbind [114]. Сайти зв'язування визначалися як амінокислотні залишки, що мають принаймні один атом у межах 5 Å від будь-якого атома ліганду.

Для тестування моделі була використана стандартна підмножина комплексів з PDBbind, яка раніше використовувалася в аналізі підходів молекулярного докінгу на основі машинного навчання, зокрема DiffDock [66]. Додатково, було видалено всі комплекси, що містять ліганди з понад 100 атомами, щоб зосередитися на оптимальних для розробки ліко-подібних (drug-like) молекулах. Остаточний тестовий набір даних включав 259 комплексів.

3.3.2. Критерії оцінювання

Аналогічно до EquiBind[60], TANKBind [113], та DiffDock [66], було використано 25-й, 50-й та 75-й квантилі кореня середньоквадратичного відхилення координат у тривимірному просторі (RMSD), скоригованого на внутрішньомолекулярну симетрію [136], між очікуваним та передбаченим положенням ліганду відносно білка. Також, було оцінено відсоток передбачень з RMSD нижче 5 Å або 2 Å. Відстані між центроїдами очікуваних та передбачених положень лігандів оцінювались за тими самими критеріями.

Використовувалися такі додаткові критерії якості передбачень на основі нековалентних взаємодій:

- Частка нековалентних взаємодій (F_{fav}) - загальна кількість взаємодій, поділена на кількість атомів ліганду. Внесок гідрофобних взаємодій враховувався з вагою 1/6 через їх високу відносну поширеність.

- Частка атомів, залучених до нековалентних взаємодій ($F_{fav\text{-}atoms}$) - частка атомів ліганду, що беруть участь принаймні в одній взаємодії.

- Частка несприятливих контактів (F_{unfav}) - загальна кількість несприятливих контактів, поділена на кількість атомів ліганду.

Також, для оцінки якості передбачених положень лігандів, визначалась частота стеричних зіткнень між атомами ліганду та білка. Зіткнення визначалося як відстань між атомами, менша за 70% від суми їхніх відповідних ван-дер-ваальсових радіусів.

3.3.3. Архітектура та тренування моделі машинного навчання

Основною архітектурою моделі, що використовувалася в цьому експерименті, була дифузійна генеративна нейронна мережа, адаптована з DiffDock [66]. Ця модель базується на SE(3)-еквіваріантних згорткових мережах, які працюють з хмарами точок [137], [138]. Модель використовує графові представлення сайту зв’язування білка та ліганду і враховує просторове розташування атомів. Вона генерує SE(3)-еквіваріантні вектори, що описують трансляції та обертання ліганду, а також SE(3)-інваріантні скаляри для кожного обертального зв’язку (торсійного кута) малої молекули. Положення випадкового конформера ліганду, що використовувався на вході до моделі, згодом змінюється на основі передбачень моделі (трансляції, обертання, торсійні кути), що призводить до остаточної конформації молекули в межах сайту зв’язування. Під час навчання позиція кожного вхідного ліганду в комплексі зазнає змін, спричинених трансляційним, обертальним та торсійним шумом, що можна назвати прямою дифузією. Потім модель навчається обертати процес дифузії. Цей метод дозволяє генерувати численні альтернативні положення лігандів під час передбачення.

Вхідні дані та архітектуру моделі DiffDock було адаптовано під експеримент наступним чином:

- Було використано граф сайту зв'язування на рівні атомів замість графа всього білка на рівні амінокислотних залишків. Сайти (замість всього білка) були використані для оптимізації швидкості навчання моделі та передбачення.
- Набір характеристик вершин ліганду був значно зменшений шляхом звуження кількості типів атомів, кількості сусідніх ковалентно зв'язаних атомів та зарядів атомів до тих, що очікуються в типових скринінгових базах даних малих молекул, зокрема ZINC20.

Моделі розроблялись на базі бібліотек для машинного навчання PyTorch [103] та PyTorch Geometric [104]. Темп навчання (learning rate) був зафіксований на рівні 0.0125. На кожній епосі моделі навчались протягом 10,000 ітерацій з розміром пакета (batch size) 4. Навчальний пакет (batch) завжди складався з ідентичних компонентів сайту та ліганду, тоді як рівень дифузійного шуму варіювався між зразками.

Три моделі навчалися паралельно протягом 25 епох з використанням лише штучних комплексів, лише експериментальних комплексів та їх комбінації. Під час навчання на штучних даних, 10000 лігандів випадковим чином вибиралися з попередньо обробленого набору даних ZINC20 на кожній епосі. Після цього для кожного ліганду створювалися штучні сайти. Таким чином, для навчання моделі було згенеровано 250000 унікальних штучних комплексів. Кожна епоха навчання на експериментальних білок-лігандних комплексах використовувала 10000 випадково відібраних систем з PDBbind (з 16379 комплексів у навчальній вибірці). Навчання на комбінованому наборі даних проводилося зі співвідношенням 4:1 штучних до експериментальних даних з 200000 унікальних штучних комплексів. Фінальна модель, доступна у вихідному коді (https://github.com/vtarasv/pocket-cfdm.git), навчалася протягом 80 епох з використанням комбінованого набору даних з 640000 унікальних штучних комплексів.

Враховуючи генеративну природу моделі, можливо створити нескінченну кількість альтернативних передбачених положень лігандів. Таким чином, реальна якість передбачень моделі є чутливою до функції оцінки, яка використовується для ранжування та вибору найкращих положень. Для швидкої розробки такої функції, було використано раніше описані критерії якості передбачень на основі нековалентних взаємодій. Функція оцінки *S* обчислюється наступним чином:

$$S = F_{fav} + F_{fav-atoms} - 2F_{unfav} - 2F_{unfav-atoms} - 10D_{clash}, \quad (3.3)$$

де $F_{unfav\text{-}atoms}$ - це частка атомів ліганду, що беруть участь принаймні в одному несприятливому контакті, D_{clash} - це сума всіх відстаней (Å), які є нижчими за поріг стеричних зіткнень. Коефіцієнти функції оцінки були підібрані емпірично шляхом візуального огляду передбачених білок-лігандних комплексів.

3.3.4. Вплив штучних систем на якість передбачень

Щоб оцінити потенційний вплив доповнення даних за допомогою штучних білок-лігандних комплексів на ефективність моделі, було порівняно три моделі. Перша модель навчалася виключно на експериментальних білок-лігандних комплексах, друга модель навчалася на штучно згенерованих сайтах з відповідними малими молекулами, а третя модель навчалася на комбінації обох. За аналогією до оригінального підходу оцінки DiffDock, метрики наводяться для найкращого передбаченого комплексу (топ 1) на основі функції оцінки *S* (рівняння 3.3) та передбачення з найнижчим RMSD серед п'яти найкращих передбачень (топ 5) на основі *S*.

Таблиця 3.3 ілюструє, що включення штучних або комбінованих даних у процес навчання призвело до покращень як метрик RMSD (до 10% підвищення частки передбачень з RMSD < 2 Å), так і відстані між центроїдами, порівняно з моделлю, навченою виключно на експериментальних комплексах. Цікаво, що навчання на штучному наборі даних виявилося кращим за комбінований набір даних за певними метриками, особливо коли генерувалось 40 положень для кожного сайту. Це може

бути спричинено ігноруванням внеску експозиції розчинника в алгоритмі генерації сайтів, що призводить до перебільшеної кількості білок-лігандних взаємодій та жорсткіших просторових обмежень у кінцевій позі ліганду. Крім того, експериментальні дані можуть сприяти дезорієнтації моделі щодо фізичних обмежень через численні зразки зі стеричними зіткненнями (наприклад, 1gj4, 5yr6) або навіть ковалентними зв’язками (наприклад, 3s3q, 6eyz) між білком та лігандом. Такі артефакти становлять майже 5% навчальних даних PDBBind.

Комбінований набір даних білок-лігандних комплексів продемонстрував вищу ефективність порівняно з двома іншими наборами даних з погляду кількості нековалентних взаємодій та несприятливих контактів, як показано в Таблиці 3.4.

Додатково спостерігалося, що використання комбінованих даних для навчання моделі призвело до зменшення частки передбачених зразків, що містять стеричні зіткнення, як показано в Таблиці 3.5.

Таблиця 3.3. RMSD та відстань між центроїдами між реальним та передбаченим положенням малої молекули для тестових комплексів з PDBbind (N=259). Моделі генерували 10 або 40 положень ліганду. Топ 1 та топ 5 положень обиралося на основі функції оцінки *S*. Найкраще з топ 5 положень обиралося на основі найнижчого RMSD. Найкращі значення виділено жирним шрифтом.

	RMSD					**Відстань між центроїдами**				
	Квантиль (Å)			**% нижче межі**		**Квантиль (Å)**			**% нижче межі**	
Тип даних	**25й**	**50й**	**75й**	**5Å**	**2Å**	**25й**	**50й**	**75й**	**5Å**	**2Å**
10 згенерованих положень, топ 1										
Експериментальні	4.08	5.51	7.45	40.93	5.02	1.02	1.75	2.7	94.59	57.92
Штучні	**3.69**	5.75	7.39	40.15	5.02	**1**	**1.46**	**2.19**	97.68	**69.88**
Комбіновані	3.76	**5.13**	**7.12**	**45.95**	**7.34**	1.05	1.69	2.33	**98.07**	63.32
10 згенерованих положень, найкраще з топ 5										
Експериментальні	2.68	3.72	4.64	82.24	11.2	0.95	1.37	1.97	**99.61**	75.29
Штучні	2.49	3.47	4.6	81.08	16.99	**0.78**	**1.24**	**1.86**	98.84	**80.31**
Комбіновані	**2.4**	**3.31**	**4.55**	**85.33**	**17.37**	0.81	1.29	1.88	99.23	77.99
40 згенерованих положень, топ 1										
Експериментальні	3.89	5.54	7.36	40.93	4.63	1.01	1.73	2.48	96.14	57.92
Штучні	**3.47**	5.46	7.25	**47.1**	**7.72**	**1**	**1.39**	**2.09**	**97.68**	**72.59**
Комбіновані	3.86	**5.2**	**7.16**	45.95	6.18	1.05	1.57	2.42	**97.68**	64.09
40 згенерованих положень, найкраще з топ 5										
Експериментальні	2.37	3.35	4.6	82.24	15.06	0.85	1.19	**1.69**	**99.61**	81.08
Штучні	**1.99**	**3.08**	**4.3**	**85.33**	**25.1**	**0.73**	**1.11**	**1.69**	**99.61**	**82.63**
Комбіновані	2.09	3.3	4.6	82.24	23.94	0.8	1.26	1.99	98.84	75.29

Таблиця 3.4. Частка нековалентних взаємодій (F_{fav}), частка атомів, залучених до нековалентних взаємодій ($F_{fav\text{-}atoms}$) та частка несприятливих контактів (F_{unfav}) у передбачених положеннях малої молекули для тестових комплексів з PDBbind (N=259). Моделі генерували 10 або 40 положень ліганду. Топ 1 та топ 5 положень обиралося на основі функції оцінки *S*. Найкраще з топ 5 положень обиралося на основі найнижчого RMSD. Найкращі значення виділено жирним шрифтом.

	F_{fav}			$F_{fav\text{-}atoms}$			F_{unfav}		
	Квантиль			**Квантиль**			**Квантиль**		
Тип даних	**25й**	**50й**	**75й**	**25й**	**50й**	**75й**	**25й**	**50й**	**75й**
10 згенерованих положень, топ 1									
Експериментальні	**0.27**	0.37	0.51	0.38	0.53	0.67	**0**	0.08	0.17
Штучні	**0.27**	**0.39**	0.53	0.38	0.53	0.67	**0**	**0.07**	**0.15**
Комбіновані	**0.27**	0.38	**0.54**	**0.41**	**0.54**	**0.68**	**0**	**0.07**	**0.15**
10 згенерованих положень, найкраще з топ 5									
Експериментальні	0.28	0.38	0.52	0.35	0.53	0.66	0.04	0.1	0.23
Штучні	**0.3**	0.39	0.52	0.38	0.52	**0.68**	0.04	0.1	0.21
Комбіновані	**0.3**	**0.4**	**0.56**	**0.39**	**0.55**	0.67	**0.03**	**0.09**	**0.18**
40 згенерованих положень, топ 1									
Експериментальні	0.29	0.37	0.54	0.4	0.54	**0.71**	**0**	0.06	0.14
Штучні	0.29	0.41	0.58	**0.42**	**0.58**	**0.71**	**0**	**0.04**	**0.12**
Комбіновані	**0.31**	**0.43**	**0.59**	**0.42**	**0.58**	**0.71**	**0**	0.05	**0.12**
40 згенерованих положень, найкраще з топ 5									
Експериментальні	0.26	0.4	0.54	0.39	0.53	0.69	0.03	0.09	0.18
Штучні	0.29	0.41	**0.55**	0.39	0.54	0.7	**0**	0.07	**0.14**
Комбіновані	**0.32**	**0.43**	**0.55**	**0.41**	**0.56**	**0.71**	**0**	**0.06**	**0.14**

Таблиця 3.5. Частка передбачених положень малої молекули, що утворюють стеричні зіткнення з білком у тестових комплексах PDBbind (N=259). Стеричне зіткнення було визначено як відстань між атомами, меншу за 70% суми їхніх відповідних ван-дер-ваальсових радіусів. Результати вказані для найкращої з топ 5-ти згенерованих положень малої молекули. Моделі генерували 40 положень ліганду. Топ 1 та топ 5 положень обиралося на основі функції оцінки *S*. Найкраще з топ 5 положень обиралося на основі найнижчого RMSD. Найкращі значення виділено жирним шрифтом.

	% зі стеричними зіткненнями	
Тип даних	**Топ 1**	**Топ 5**
Експериментальні	33.59	30.12
Штучні	30.50	26.64
Комбіновані	**28.19**	**23.94**

Фінальна модель PocketCFDM навчалася протягом 80 епох, доки втрати на валідаційній вибірці не досягли плато, з використанням високоякісних експериментальних даних (використаних для отримання статистики нековалентних взаємодій) та штучних зразків, подібно до вищезгаданих налаштувань комбінованого набору даних. Було досягнуто значного зниження частоти стеричних зіткнень білок-ліганд, з 19.31% та 13.90% для топ 1 та найкращого з топ 5 зразків, відповідно. Це майже на 10% менше порівняно з найоптимальнішою моделлю, навченою протягом 25 епох (Таблиця 3.5). Суттєвого покращення за іншими метриками не спостерігалося.

Була виявлена позитивна кореляція між кількістю атомів малої молекули та такими критеріями оцінювання, як RMSD та кількість стеричних зіткнень. Крім того, була виявлена негативна кореляція між кількістю атомів та часткою нековалентних

взаємодій, як зображено на Рисунку Д2. Це вказує на те, що якість передбачень моделі знижується зі збільшенням розміру ліганду.

3.3.5. Оцінка якості передбачень моделі машинного навчання

Було проведене детальне порівняння між фінальною версією PocketCFDM та DiffDock (Таблиця 3.6).

Таблиця 3.6. Порівняння між PocketCFDM та DiffDock на тестових комплексах PDBbind (N=259). Моделі генерували 40 положень ліганду. Кращі значення серед двох моделей виділено жирним шрифтом.

Критерій	**PocketCFDM**	**DiffDock**
Середній час передбачення (с)	**49**	90
RMSD (Å); медіана; топ 1	5.14	**2.8**
RMSD (Å); медіана; топ 5	3.02	**2.17**
Відстань між центроїдами (Å); медіана; топ 1	1.4	**1.01**
Відстань між центроїдами (Å); медіана; топ 5	1.08	**0.81**
F_{fav}; медіана; топ 1	**0.43**	0.37
F_{fav}; медіана; топ 5	**0.42**	0.39
$F_{fav\text{-}atoms}$; медіана; топ 1	**0.58**	0.47
$F_{fav\text{-}atoms}$; медіана; топ 5	**0.56**	0.5
F_{unfav}; медіана; топ 1	**0.02**	0.07
F_{unfav}; медіана; топ 5	**0.07**	0.08
% зі стеричними зіткненнями; топ 1	**19.31**	46.51
% зі стеричними зіткненнями; топ 5	**13.90**	25.97

DiffDock продемонстрував вищу точність з погляду RMSD та відстані між центроїдами. На противагу, PocketCFDM показав кращі результати з погляду нековалентних взаємодій та несприятливих контактів, а також значно нижчу частоту стеричних зіткнень. Це було очікувано через розбіжність в функціях оцінки, що використовуються для ранжування згенерованих зразків. Підхід DiffDock використовував додаткову модель оцінки, яка була навчена надавати пріоритет зразкам з нижчим RMSD до фактичної пози ліганду. Однак, у контексті PocketCFDM, підхід до оцінювання був більше зосереджений на кількості нековалентних контактів та відсутності стеричних зіткнень, не враховуючи специфічну конформацію пози ліганду. Ще однією помітною відмінністю, що стосується розбіжності в функціях оцінки, є різноманітність, що спостерігається серед передбачених положень ліганду. Медіанні значення RMSD між альтернативними топ 5 зразками в методах PocketCFDM та DiffDock становили 2.59 та 1.29, відповідно. Крім того, середній час передбачення для PocketCFDM був приблизно в 1.8 рази швидшим. Це пояснюється переважно зменшеним розміром графа білка та зменшенням кількості характеристик вершин графів.

PocketCFDM відрізняється від DiffDock як функцією оцінки, так і навчальним набором даних. DiffDock навчається на графах усього білка на рівні залишків, тоді як PocketCFDM використовує графи на атомному рівні, обмежені сайтами зв’язування (реальними або штучними). Навчальний набір даних для фінальної моделі PocketCFDM доповнюється штучно згенерованими комплексами, тоді як DiffDock навчається лише на експериментально визначених білок-лігандних комплексах. Таким чином, краща якість передбачень PocketCFDM за рядом критеріїв з Таблиці 3.6 походить від обох цих факторів.

3.4. Обмеження та перспективи розробленого методу

Основною сферою для вдосконалення моделі PocketCFDM є час передбачення, який наразі все ще завеликий для ефективного практичного застосування. Середній час, необхідний для прогнозування 40 положень одного ліганду, становить близько 50 с.

на графічному процесорі NVIDIA L4. Час передбачення можна було б значно покращити, замінивши обчислювально затратні блоки еквіваріантних графових нейронних мереж [112] більш ефективними альтернативами, такими як GVP [111]. Включення штучних комплексів у навчання навіть простіших технік на основі регресії, таких як EquiBind[60] та TANKBind [113], також може бути корисним, оскільки ці архітектури, як правило, більш чутливі до кількості окремих тренувальних зразків.

Ще одним помітним недоліком поточної реалізації, є можливість внутрішньомолекулярних стеричних зіткнень у передбачених положеннях ліганду. Включення внутрішньомолекулярних та міжмолекулярних зіткнень у функцію втрати під час процесу навчання могло б частково розв’язати цю проблему.

Також, включення більших лігандів у процес навчання потенційно дозволить розв’язати проблеми, що виникають з відносно великими сполуками. Набір даних ZINC20, який використовувався в цій роботі, містить лише 2.5% молекул, більших за 40 атомів (не включаючи гідрогени), що робить їх недостатньо представленими під час навчання моделі. Хоча медіанний розмір лігандів у цьому наборі даних становить 25 атомів, що є досить поширеним для наборів даних ліко-подібних молекул, вищий відсоток великих лігандів може покращити якість передбачень для більших сполук, зберігаючи ту саму якість для менших молекул.

Ще одним недоліком поточного алгоритму генерації сайтів є часте утворення непередбачених контактів (переважно гідрофобних) під час генерації інших типів взаємодій, що видно з дисбалансу між кількістю нековалентних взаємодій, знайдених в експериментальних та штучних комплексах. Коли додається будівельний блок сайту, окрім цільового контакту, випадково можуть утворюватися додаткові взаємодії.

Хоча цей розділ зосереджений на оптимізації прогнозування положення лігандів, необхідно зазначити, що прогнозування положення є лише частиною точного прогнозування білок-лігандного зв’язування. Двома іншими важливими

компонентами є точна оцінка положення з погляду афінності/енергії зв’язування та врахування гнучкості та динаміки білка. Перша проблема наразі вирішується не лише за допомогою традиційних функцій оцінки докінгу, але й за допомогою технік на основі ML [139]. Другу проблему можна вирішити за допомогою багатьох підходів, включаючи гнучкий (flexible) та ансамблевий (ensemble) докінг. Врахування гнучкості білка та ансамблів білкових конформацій є одними з майбутніх напрямків удосконалення техніки генерації штучних даних.

3.5. Висновки розділу

У цьому розділі розроблено алгоритм генерації штучних систем протеїн-мала молекула для розширення тренувальних даних моделей прогнозування положень лігандів у сайтах зв'язування протеїнів. Алгоритм базується на тих самих статистичних патернах нековалентних взаємодій, що було знайдено в експериментальних даних. Для побудови таких штучних комплексів, було оцінено статистичні характеристики нековалентних взаємодій у реальних протеїн-лігандних комплексах та розроблено алгоритмічний підхід, який відтворює їх у штучних сайтах, що будуються навколо довільних конформерів малих органічних молекул.

Ефективність розробленого підходу була оцінена шляхом порівняння моделей, що базуються на підході DiffDock, навчених лише на експериментальних даних, лише на штучних даних та на їх комбінації.

Тренування моделей на штучних комплексах дозволяє підвищити частку правильних передбачень до 10% за критерієм RMSD між передбаченими та реальними положеннями ліганду. Також показано, що включення штучних комплексів в навчання моделі призвело до підвищення якості та ефективності її передбачень. Зокрема, PocketCFDM перевершує DiffDock за кількістю нековалентних взаємодій, кількістю стеричних зіткнень та швидкістю передбачення.

Код для передбачення, фінальні ваги моделі та приклади передбачень моделі доступні у репозиторії GitHub (https://github.com/vtarasv/pocket-cfdm.git).

4. ПЕРЕДБАЧЕННЯ ПОЛОЖЕННЯ МАЛИХ МОЛЕКУЛ У МІСЦІ ЗВ'ЯЗУВАННЯ З ПРОТЕЇНОМ

Як було згадано в попередньому розділі, класичний, або фізичний, докінг значною мірою досяг плато своєї практичної точності. Це стимулювало розробку альтернативних підходів на основі машинного навчання. На відміну від класичного докінгу, який базується на мінімізації певної фізично обґрунтованої функції оцінки, ці методи використовують підхід, повністю орієнтований на дані, навчаючись на експериментально визначених комплексах білок-ліганд.

Сучасні генеративні архітектури моделей для докінгу на основі ML, такі як раніше розглянута дифузійна модель DiffDock [66], або методи спільного згортання (ко-фолдингу) RoseTTAFold All-Atom [140], NeuralPLexer3 [141], AlphaFold3 [65], Chai-1 [142], Boltz-1 [70] та Protenix [143], часто дозволяють отримати вищу точність порівняно з класичним докінгом, яка, однак, досягається ціною складних архітектур, великих розмірів моделей, а також повільного навчання та передбачення.

Крім того, точна та неупереджена оцінка методів ML для докінгу залишається складною через відсутність загальноприйнятих, надійних наборів даних. Більшість поточних оцінок не враховують витік даних (data leakage) між тренувальними та тестовими наборами даних (наявність схожих даних в обох наборах), що може призвести до надмірно оптимістичних показників точності [144], [145]. Зокрема, ефективність інструментів ко-фолдингу, які часто демонструють найвищу точність, виявляється сильно корельованою зі схожістю між їхніми тестовими та тренувальними молекулярними структурами [145]. У результаті, наявні ML інструменти не в змозі замінити класичні у практичних завданнях віртуального скринінгу з високою пропускною здатністю з двох основних причин: несприятливе співвідношення точності та швидкості та відсутність належної оцінки для ML підходів.

Нещодавно опублікований набір даних PLINDER [146] розв'язує описані проблеми оцінки підходів. Дані містять великий та різноманітний попередньо визначений

тренувальний піднабір з PDB [36]. Процес розділення на тренувальний та тестовий набори базується на кластеризації, що мінімізує як витік даних між наборами, так і наявність схожих поширених структур у тестовому наборі. Розділення враховує як схожість послідовностей білків, так і схожість білок-лігандних взаємодій, водночас надаючи пріоритет включенню біологічно релевантних, високоякісних структур у тестовий набір.

У цьому розділі описано розробку та оцінку нового інструменту для білок-лігандного докінгу на основі DL, ArtiDock [3] (Рисунок 4.1), що передбачає міжмолекулярну матрицю відстаней білок-ліганд та конвертує її у хімічно валідне положення ліганду. Розроблений інструмент порівняно з провідними відкритими та комерційними програмами класичного докінгу, такими як AutoDock4 (надалі AutoDock-CPU) [147], [148], AutoDock-GPU [149], Vina [150], [151] та Glide [152], [153], [154], [155]. Основною метою була оцінка цих методів у контексті високопродуктивного жорсткого докінгу (rigid docking, з зафіксованою структурою білка) у відомий, попередньо визначений сайт зв’язування білка, що відображає найпоширеніший випадок використання докінгу у реальних проєктах розробки ліків на ранніх стадіях.

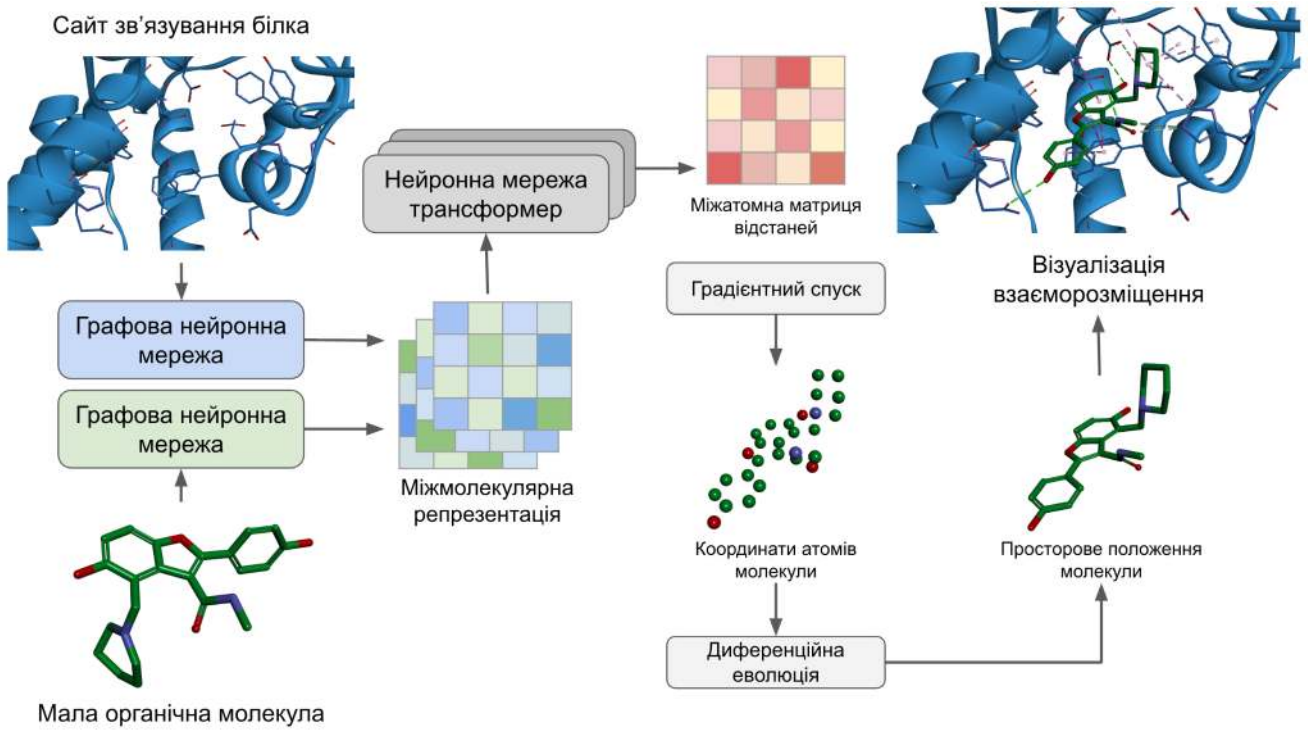

Рисунок 4.1. Загальна схема архітектури моделі ArtiDock та післяобробка передбачень моделі.

Щоб забезпечити неупереджену оцінку ArtiDock, модель навчалась на тренувальному наборі PLINDER, а оцінка її фінальної ефективності проводилася на тестовому наборі PLINDER, разом із класичними підходами. Оцінка охоплює точність докінгу, хімічну та геометричну коректність передбачених положень, а також обчислювальну ефективність. Тестові випадки також представляли сценарії, коли вхідна структура білка є передбаченою ML методами або отримана з незв'язаного з лігандом стану (апо), а також випадки з органічними кофакторами, іонами або молекулами води в сайтах зв'язування.

У більшості порівнянь між традиційними методами докінгу та ML підходами, для класичних інструментів використовуються стандартні субоптимальні параметри, або ж вибір параметрів не враховує обчислювальні витрати [156], [157]. Два методи зі схожою точністю можуть мати дуже різну швидкість і, таким чином, різну корисність у реальних сценаріях застосування. Тому, було проведено серію експериментів з метою оптимізації згаданих традиційних методів, що порівнювались

з ArtiDock, для досягнення компромісу між точністю та часом обчислень, забезпечуючи справедливі та практичні налаштування для подальшого порівняння.

Додатково, було використано набір даних PoseX [61], де розділ на тестовий та тренувальний піднабори визначається за часом публікації даних, щоб оцінити ефективність ArtiDock у порівнянні з популярними ML рішеннями без необхідності перенавчати складні архітектури на наборі PLINDER. Цей набір даних був створений для оцінки як селф-докінгу (self-docking, голо, структура білка отримується безпосередньо з відомої структури комплексу), так і крос-докінгу (cross-docking, апо, структура білка отримується з незв'язаного з лігандом стану), охоплюючи широкий спектр класичних методів докінгу [150], [152], [158], [159], [160], підходів ко-фолдингу [65], [70], [140], [142], [143], [161], та інших підходів машинного навчання [60], [64], [66], [68], [69], [113], [162], [163], [164], [165], [166] (загалом 23 методи).

Отримані результати демонструють, що ArtiDock значно перевершує всі протестовані традиційні методи та конкурує з найновітнішими на момент написання ML моделями з погляду точності, водночас пропонуючи пропускну здатність на рівні найшвидших інструментів. Таким чином, розроблений підхід для молекулярного докінгу може застосовуватися в реальних сценаріях віртуального скринінгу. У розділі описано архітектуру моделі ArtiDock, післяобробку її передбачень, підготовку тренувальних даних та процедуру навчання; окреслено вибір оптимальних параметрів класичного докінгу; надано всебічне порівняння точності та якості передбачень; обговорено обмеження й напрямки вдосконалення ArtiDock.

4.1. Інструменти для порівняння методів передбачення положення малих молекул

4.1.1. Набори даних для оцінки методів

Для експериментів з параметризації класичного докінгу було використано набір даних PoseBusters [156], що складається з 308 експериментально визначених комплексів з PDB, кожен з яких представляє унікальну пару білок-ліганд (PLP, Protein-Ligand Pair). Було видалено 2 PLP, які не вдалося обробити програмою для оцінювання (PLINDER Python API: https://github.com/plinder-org/plinder.git).

Тестовий набір даних PLINDER (реліз 2024-06, ітерація v2) [146] було обрано для фінального порівняльного аналізу підходів класичного докінгу та ArtiDock. Після попередньої обробки та фільтрації за якістю (Додаток А), тестові дані містили 1035 PLP з 676 унікальними кодами словника хімічних компонентів (CCD, Chemical Component Dictionary) лігандів. Було залишено один ліганд на CCD, який утворює максимальну кількість нековалентних взаємодій з білком. У результаті фінальна підмножина PLINDER складалася з 676 PLP, отриманих з 656 записів PDB. Було видалено одну PLP, яку не вдалося обробити програмою для оцінювання. Усі небілкові молекули та атоми (гетероатоми) були збережені в тестових системах для оцінки ефективності методів на сайтах, що містять кофактори, іони та молекули води.

Також було використано підмножину тестових систем PLINDER, запропоновану в Machine Learning in Structural Biology (MLSB) 2024 (https://www.mlsb.io/). Тестовий набір PLINDER MLSB містить 346 систем ліганд-білок. Головними особливостями цього набору даних є вхідні структури білків з іншої системи (апо конформації) та структури білків, передбачені методами машинного навчання (https://www.plinder.sh/blog/updates). Цей набір даних представляє реалістичний сценарій докінгу, в якому оптимальна (голо) для зв'язування конформація білка невідома. Щоб забезпечити узгоджені визначення сайтів зв'язування як в апо, так і у голо структурах, апо/передбачені білки було накладено на їхні голо аналоги за

допомогою бібліотеки молекулярного моделювання MolAR [167], вирівнюючи сайти зв’язування лігандів, визначені як усі амінокислотні залишки в межах 8 Å від ліганду. Після попередньої обробки фінальний тестовий набір PLINDER MLSB містив 344 PLP.

Оскільки ArtiDock навчався на тренувальному наборі PLINDER (ArtiDock-PL), витік даних між наборами зведений до мінімуму, що дозволяє уникнути потенційного перенавчання на будь-яких тестових зразках і забезпечує об’єктивну оцінку можливостей моделі. Однак це не дозволяє включити інші методи докінгу на основі ML у порівняння, оскільки їх також потрібно було б перенавчати на тому самому розбитті, що технічно неможливо через значні обчислювальні затрати.

З усім тим, порівняння ArtiDock з рішеннями на основі ML було проведено, використовуючи набір даних PoseX [61]. Набір містить 718 пар білок-ліганд для селф-докінгу (голо конформації) та 1312 пар для крос-докінгу (апо конформації), що охоплюють 109 білкових мішеней з 371 запису PDB та 362 унікальні малі молекули. Важливо, що PoseX побудований за часовим поділом даних, у якому зберігаються лише комплекси PDB, випущені між 01.01.2022 та 01.01.2025. Це часове розділення є ефективним для порівняння готових інструментів без перенавчання. Зауважимо, що таке розділення не є оптимальним для порівняння ML методів, де фіксовані розділення на тренувальну/валідаційну/тестову вибірки на основі структури були б кращими за наявності достатніх обчислювальних ресурсів для перенавчання. Для оцінювання на PoseX модель ArtiDock було перенавчено на PLP з PLINDER, випущених до 01.01.2022 (ArtiDock-PX).

4.1.2. Критерії оцінювання

Для оцінки передбачених положень лігандів було використано три типи оцінок:

- Корінь середньоквадратичного відхилення координат у тривимірному просторі (RMSD), скоригований на внутрішньомолекулярну симетрію. Цей критерій вже використовувався у попередньому розділі та представляє абсолютне

відхилення між експериментально визначеним та передбаченим положеннями ліганду [168].

- Тест різниці локальних відстаней для взаємодій білок-ліганд (lDDT-PLI). Частка відтворених взаємодій між білком та малою органічною молекулою, присутніх в експериментальних комплексах (більше - краще) [168]. Для кожної пари атомів білок-ліганд у радіусі 15 Å перевіряється, чи різниця між передбаченою та референсною міжатомною відстанню вкладається у чотири порогові значення (0.5, 1, 2 та 4 Å), а фінальне значення є середнім за всіма такими парами та порогами.
- Перевірки якості за PoseBusters, розроблені для вимірювання фізичної та хімічної коректності передбачених положень лігандів [156]. Набір тестів організований у три групи: перша перевіряє хімічну валідність молекули (валентність, ароматичність, стереохімія, сумісність із вхідною структурою), друга - внутрішньомолекулярні властивості (довжини та кути зв'язків, планарність ароматичних кілець, енергія конформації), третя - міжмолекулярні взаємодії, зокрема стеричні зіткнення ліганду з білком.

Оцінки розраховувалися за допомогою Python API PLINDER, що, своєю чергою, використовував пакети OpenStructure v.2.8.0 [169], [170] та PoseBusters v.0.3.1 7 [156].

Для оцінки методів на тестових даних PLINDER обиралося найкраще передбачення на основі внутрішньої функції оцінки кожного методу, і порівнювались медіана RMSD, частка зразків з RMSD < 2Å, медіана lDDT-PLI, частка зразків з lDDT-PLI > 0.5 та частка зразків, що пройшли всі перевірки якості PoseBusters (PB-Валідні).

Порівняння на даних PoseX включало частку зразків з RMSD < 2Å та частку зразків, що одночасно мають RMSD < 2Å та PB-Валідні для найкращих передбачень на основі внутрішньої функції оцінки. Для сценарію крос-докінгу враховувалась

усереднена частка на рівні мішені через нерівномірний розподіл кількості комплексів на мішень [61].

4.2. Оптимізація традиційних методів для передбачення положення малих молекул

4.2.1. Програмне забезпечення

Програми AutoDock-CPU, AutoDock-GPU та Vina потребують вхідних структур протеїну та ліганду у форматі PDBQT. Цей самий формат використовується для представлення вихідних положень лігандів. Тому було застосовано уніфікований алгоритм попередньої обробки та післяобробки даних для цих інструментів докінгу. Підготовка лігандів та обробка результатів після докінгу виконувалися за допомогою Python API бібліотеки Meeko v.0.6.1 (https://github.com/forlilab/Meeko). Для AutoDock-CPU та AutoDock-GPU макроцикли лігандів були зафіксовані, щоб уникнути хімічно невалідних структур у вихідних даних.

Файли PDBQT білків були отримані за допомогою MGLTools v.1.5.7 (https://ccsb.scripps.edu/mgltools/) з використанням опцій для протонування молекул та об'єднання зарядів неполярних воднів і неподілених пар. Було помічено, що цей інструмент не може призначити заряди молекулам води та деяким металам. Щоб виправити це, до фінальних файлів було додано поширені ступені окиснення [171] для цих атомів/молекул.

Слід зазначити, що існували певні відмінності у підготовці білків та лігандів для експериментів з параметризації докінгу та для фінального бенчмаркінгу PLINDER. У початкових експериментах на наборі даних PoseBusters [156] було використано MGLTools для підготовки як білків, так і лігандів для AutoDock-CPU/GPU та конвертовано вихідні дані докінгу у формат SDF за допомогою Open Babel v.3.1.1 [172]. Meeko використовувався для обробки структур білка та ліганду для Vina, як рекомендовано в офіційній документації.

Однак було помічено, що значна частина положень лігандів з тестового набору PLINDER, згенерованих за допомогою MGLTools та Open Babel, не зчитувалася RDKit, який є необхідним для подальшої оцінки. Цю проблему було вирішено шляхом переходу на Meeko для обробки лігандів. Що стосується білка, то дизайн модуля Meeko передбачає високоякісні дані, де різні структурні проблеми виправляються вручну, що є неможливим при роботі з масштабними наборами даних, які включають сотні білків. Тому для обробки білків пріоритет було надано MGLTools замість Meeko. Загалом, комбіноване використання Meeko та MGLTools дозволило зберегти до 20% передбачень для подальшої оцінки.

Для AutoDock-CPU та AutoDock-GPU файли параметрів сітки (GPF) створювалися за допомогою MGLTools, а для генерації сітки використовувався AutoGrid v.4.2.6. Докінг за допомогою AutoDock-CPU запускався з використанням AutoDock v.4.2.6 (http://autodock.scripps.edu/) та згенерованого в MGLTools файлу параметрів докінгу (DPF). AutoDock-GPU v.1.6 був скомпільований з вихідного коду (https://github.com/ccsb-scripps/AutoDock-GPU) з параметрами DEVICE=GPU та NUMWI=256.

Python API Vina v.1.2.6 (https://github.com/ccsb-scripps/AutoDock-Vina) використовувався для обчислення карт афінності та виконання докінгу. Через варіативність вихідних даних AutoDock-CPU, AutoDock-GPU та Vina між запусками, усі результати наводяться як середнє значення трьох незалежних запусків докінгу.

Було використано інтерфейс командного рядка Schrödinger Suite v.2024-3 для підготовки вхідних даних та запуску Glide докінгу. Молекули лігандів оброблялися за допомогою вбудованого у програму інструменту LigPrep зі стандартними параметрами. Білки готувалися за допомогою вбудованого PrepWizard, при цьому в оброблених структурах білків зберігалися всі молекули води.

4.2.2. Підбір параметрів

Щоб отримати оптимальне налаштування докінгу з погляду компромісу часу та якості, було запущено ряд експериментів зі змінними параметрами докінгу для кожного з методів. Крім того, було проведено експерименти з представленням сайту зв'язування білка, щоб зрозуміти, як розмір і форма сайту впливають на фінальну ефективність. Множини перевірених значень параметрів кожного з підходів докінгу доступні в Таблицях E2-E5.

Протестований простір параметрів генетичного алгоритму (GA) AutoDock-CPU включав:

- ga_num_evals: максимальна кількість оцінок енергії, дозволених під час запуску GA. Це визначає тривалість пошуку. Як тільки цей ліміт досягнуто, запуск зупиняється.
- ga_run: кількість незалежних запусків GA для виконання. Кожен запуск створює одне передбачене положення зв'язування. Більша кількість запусків зазвичай збільшує шанси знайти глобальний мінімум.
- ga_pop_size: кількість рішень у популяції протягом кожного покоління.
- ga_num_generations: максимальна кількість поколінь, що запускаються на одну симуляцію докінгу.
- ga_elitism: кількість найкращих індивідів (найкращих рішень), що виживають до наступного покоління.

Параметри GA включалися у DPF перед докінгом AutoDock-CPU.

В експериментах з AutoDock-GPU було протестовано параметри nev, nrun, psize та ngen, які є еквівалентними параметрам AutoDock-CPU ga_num_evals, ga_run, ga_pop_size та ga_num_generations відповідно.

В експериментах з Vina було протестовано два основні параметри, що відповідають за компроміс часу та якості: exhaustiveness (кількість спроб незалежного пошуку) та max_evals (кількість оцінок функції оцінки).

На додаток до параметрів докінгу, було досліджено, чи впливає роздільна здатність сітки, визначена параметром spacing, на ефективність AutoDock-CPU, AutoDock-GPU та Vina.

Для докінгу Glide було досліджено такі параметри:

- PRECISION (режим точності): HTVS (High Throughput Virtual Screening - висока швидкість ціною точності), SP (Standard Precision - баланс між швидкістю та точністю) або XP (Extra Precision - висока точність ціною швидкості).
- MAXKEEP: кількість положень зв’язування для збереження на початковій фазі докінгу.
- MAXREF: кількість положень зв’язування для збереження для мінімізації енергії.
- MAX_ITERATIONS: максимальна кількість кроків методу спряжених градієнтів під час пост-докінгової мінімізації.
- POSTDOCK_NPOSE: кількість положень для використання у пост-докінговій мінімізації.

Протестовані представлення розміру та форми сайту зв’язування білка включали:

- Куб з довжиною сторін, що дорівнює 15, 20 або 25 Å.
- Обмежувальну рамку, вирівняну за лігандом (Ligand-aligned bounding box, LABB), з довжинами сторін, що дорівнюють максимальній протяжності відомого положення ліганду вздовж кожної осі. Для експериментів було додано 10, 15 або 20 Å до кожного виміру.

Центроїди сайті були розташовані в центрі експериментально визначеного положення ліганду. Сайт зв’язування у формі куба представляє сценарій, коли точна форма сайту невідома або неважлива. LABB передбачає наявність певних попередніх знань про форму зв’язаного ліганду. У налаштуваннях докінгу параметри сайту конвертувалися у кількість точок сітки під час генерації сітки (параметр GPF npts) для AutoDock-CPU та AutoDock-GPU, передавалися безпосередньо як параметр box_size для генерації карт афінності Vina, або передавалися безпосередньо як параметр OUTERBOX для генерації сітки Glide.

4.2.3. Результати підбору параметрів

Експерименти з AutoDock-CPU (Таблиця E2) виявили два параметри зі значним впливом на співвідношення часу та якості докінгу: максимальна кількість оцінок енергії та кількість незалежних запусків генетичного алгоритму. Обидва параметри прямо пропорційні обчислювальним витратам. Зменшення максимальної кількості оцінок енергії зі стандартних 2.5 мільйонів до 350000 дозволило досягти ефективності, близької до максимальної, при приблизно 7-кратному прискоренні докінгу. Подвоєння стандартної кількості запусків (з 10 до 20) значно збільшило частку передбачень з RMSD < 2Å з 0.42 до 0.46 (медіана RMSD зменшилася на 0.2Å). Як результат, для фінальної оцінки AutoDock-CPU на тестах PLINDER було використано до 350000 оцінок енергії та 20 запусків генетичного алгоритму.

Експерименти з AutoDock-GPU (Таблиця E3) продемонстрували незначні зміни як в обчислювальних витратах, так і в точності при варіюванні параметрів докінгу. Це пов’язується з добре налаштованими евристиками на основі ліганду для визначення максимальної кількості оцінок енергії та алгоритмом ранньої зупинки, реалізованими в інструменті докінгу. Отже, AutoDock-GPU запускався зі стандартними параметрами для подальших тестів.

Подібно до AutoDock-GPU, Vina використовує евристики для визначення максимальної кількості оцінок. Заміна цього параметра на заздалегідь визначене значення не призвела до покращення ефективності з погляду співвідношення

швидкості та якості (Таблиця E4). Параметр кількість спроб незалежного пошуку, який лінійно масштабує обчислювальні витрати (без нативної паралелізації), досяг точності, близької до пікової, при значенні 16 (стандартне значення - 8). Внаслідок цього для бенчмаркінгу PLINDER було застосовано 16 спроб пошуку незалежного пошуку.

Серед трьох вище описаних режимів точності Glide, SP продемонстрував найкращу ефективність (Таблиця E5). Він забезпечив 49% передбачень з RMSD < 2Å, суттєво випередивши швидший режим HTVS (35%) і впритул наблизившись до показника 51%, досягнутого майже в 5 разів повільнішим режимом XP. Однак важливо зазначити, що за метрикою lDDT-PLI, яка оцінює частку відтворення взаємодій білок-ліганд, режим XP значно перевершив SP, демонструючи медіану 0.65 проти 0.55. Тому режим XP може бути оптимальним варіантом, коли обчислювальні витрати не є лімітуючим фактором.

При використанні режиму SP збільшення кількості положень ліганду для використання у пост-докінговій мінімізації помітно покращує якість передбачення, додаючи лише незначні часові витрати. Збільшення кількості положень зі стандартного значення 5 до 80 зменшило медіану RMSD з 2.1Å до 1.8Å та покращило медіану lDDT-PLI з 0.55 до 0.66, що є точнішим за режим XP. Подальше збільшення значення цього параметра не вплинуло на результати. З огляду на ці спостереження, для оцінки на PLINDER програма Glide запускалася в режимі SP з кількістю положень у пост-докінговій мінімізації, що дорівнює 80.

Щодо представлення сайту зв’язування білка, обмежувальна рамка, вирівняна за лігандом (LABB), що включає попередні знання про форму зв’язаного ліганду, не перевершила просту кубічну репрезентацію в жодному з протестованих методів класичного докінгу. Всупереч очікуванням, кубічна репрезентація продемонструвала навіть кращу ефективність в експериментах з AutoDock-CPU та Glide (Рисунок 4.2, Таблиці E2-E5). Загалом, зменшення розміру сайту білка для докінгу або покращувало, або не впливало на точність усіх інструментів, за винятком Glide, який

досяг пікової точності за lDDT-PLI при використанні кубічної репрезентації розміром 20Å. Такий результат є очікуваним, оскільки менші сайти, центровані на експериментально визначеному ліганді, зазвичай дозволяють проводити більш вичерпний пошук у правильній області та зменшують ймовірність відбору високооцінених положень зв'язування, розташованих далеко від справжнього сайту зв'язування.

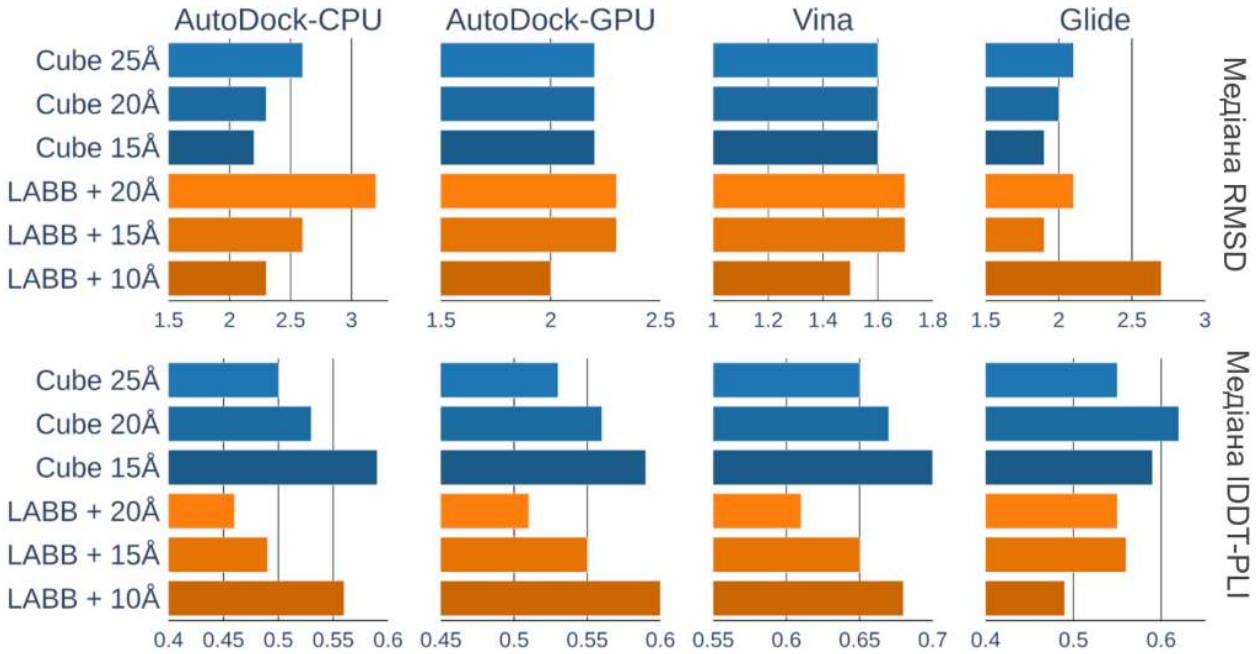

Рисунок 4.2. Медіана RMSD (Å, менше - краще) та медіана lDDT-PLI (більше - краще) на наборі даних PoseBusters (N=306) при використанні вхідного сайту зв'язування білка як кубічної (Cube) репрезентації різного розміру або обмежувальної рамки, вирівняної за лігандом (LABB), з різним приростом розмірів до кожного виміру.

Однак це підкреслює важливий компроміс між точністю докінгу та необхідністю точного визначення сайту. У реальних застосуваннях точне положення ліганду невідоме, що зумовлює необхідність використання більших сайтів для охоплення всіх ймовірних положень зв'язування. Щоб відобразити це в тестуванні підходів, варто використовувати незалежне від форми, достатньо велике представлення сайту, яке при цьому зберігає високу точність передбачення. Як результат, було обрано

кубічну репрезентацію розміром 20 Å як стандартизовані вхідні дані для всіх методів, включаючи ArtiDock, в оцінюванні на наборі даних PLINDER.

4.3. Розробка методу машинного навчання для передбачення положення малих молекул

4.3.1. Представлення малих органічних молекул та протеїнів

Ліганди були представлені як молекулярні графи на атомному рівні. Ознаки вершин включали унітарний код (one-hot) групи та періоду атома з періодичної таблиці. Ребра графа представляли ковалентні зв'язки між атомами та унітарний код за типом ковалентного зв'язку.

Ознаки вершин сайту зв'язування білка визначалися для кожного атома (окрім гідрогенів) та включали скалярні (унітарний код за групою та періодом атома, назвою амінокислоти та атома за стандартами PDB) та векторні (вектор між центроїдом сайту та атомом) компоненти. Ребра графа формувалися між вершиною та її 30 найближчими сусідами. Скалярні характеристики ребер представляли позиційну репрезентацію, розраховану за допомогою радіальної базисної функції відстані [173]. Вектори між вершинами, з'єднаними ребрами, розглядалися як векторні ознаки ребер.

4.3.2. Архітектура моделі машинного навчання

ArtiDock базується на легкій обчислювально ефективній DL архітектурі Trigonometry-Aware Neural Networks [113]. Модель було побудовано з використанням бібліотеки для машинного навчання PyTorch [103] та бібліотеки графових нейронних мереж PyTorch Geometric [104].

Графи ліганду та сайту зв'язування на атомному рівні кодувалися за допомогою GNN. Граф ліганду оброблявся за допомогою Graph Isomorphism Network з урахуванням характеристик ребер (GINE) [101]. Граф сайту пропускався через графову нейронну мережу на основі геометричного векторного перцептрона

(GVP-GNN) [111], яка продемонструвала високу ефективність у низці завдань, що передбачають навчання на структурі білка. GVP-GNN використовує як вхідні дані як скалярні, так і векторні характеристики графа. Отримані скалярні та векторні вихідні дані є інваріантними та еквіваріантними, відповідно, для будь-якого обертання або відображення сайту білка у тривимірному Евклідовому просторі. Це забезпечує фізичну коректність моделі: скалярні властивості не залежать від довільної орієнтації структури у просторі, тоді як векторні величини закономірно обертаються разом зі структурою, що дозволяє моделі навчатися на білках із різною орієнтацією без необхідності їх попереднього вирівнювання.

Після кодування графами, репрезентації вершин сайту та ліганду об'єднувалися у міжмолекулярну репрезентацію сайт зв'язування-ліганд $z \in \mathbb{R}^{p \times c \times d}$, де p - кількість вершин сайту; c - кількість вершин ліганду; d - розмір репрезентації на пару вершин.

Міжмолекулярна репрезентація оновлювалася послідовністю блоків, раніше описаних в [113], кожен з яких складався з:

- оновлення інформацією про внутрішньомолекулярної відстані сайту зв'язування $z^p \in \mathbb{R}^{p \times p \times d}$ та ліганду $z^c \in \mathbb{R}^{c \times c \times d}$;
- багатоголового механізму уваги (multi-head self-attention, подібно до архітектур на основі трансформерів);
- нелінійного переходу за допомогою щільної нейронної мережі.

Наприкінці міжмолекулярна репрезентація перетворювалася на матрицю відстаней сайт-ліганд $DM^{pc} \in \mathbb{R}^{p \times c \times 1}$ шляхом лінійного перетворення з подальшою нормалізацією (зведення в межі 0-10 Å).

4.3.3. Післяобробка передбачень

Модель передбачає матрицю міжмолекулярних відстаней сайт зв'язування-ліганд DM^{pc} для заданого сайту протеїну та малої молекули. Таким чином, потрібен

додатковий алгоритм перетворення матриці відстаней у положення для конвертації матриці в координати атомів ліганду (фактичне положення зв'язування). ArtiDock використовує алгоритм, який спочатку генерує 3D хмару точок з матриці відстаней, а потім вирівнює випадковий конформер ліганду по згенерованій хмарі.

Для генерації хмари точок було об'єднано та вдосконалено підходи до генерації положень лігандів з TankBind [113], [174], [175] та Uni-Mol [163]. Після отримання передбаченої матриці DM^{pc}, 3D хмара точок $C' \in \mathbb{R}^{c\times 3}$ ініціалізувалась нулями та оптимізувалась через метод зворотного поширення помилки (backpropagation) за допомогою оптимізатора Adam. Вибір Adam обумовлений стійкістю до вибору гіперпараметрів та адаптивним масштабуванням кроку для кожного параметра окремо; у режимі детермінованого градієнта він функціонує як адаптивний метод першого порядку з моментом. Оптимізація зупинялася, як тільки функція втрат $\mathcal{L}$ досягала плато.

Функція втрат $\mathcal{L}$ враховувала внутрішньомолекулярні та міжмолекулярні похибки відстаней і внески стеричних зіткнень, включаючи 4 компоненти:

- Міжмолекулярні відстані:

$$\mathcal{L}_{inter} = \sum_{i}^{p}\sum_{j}^{c}\left(\left|DM_{ij}^{pc'} - DM_{ij}^{pc}\right|\right), \tag{4.1}$$

де $DM_{ij}^{pc'}$ - матриця відстаней між атомами сайту зв'язування та C'. Це найбільш фундаментальний компонент, що відповідає за міжмолекулярне розташування атомів.

- Внутрішньомолекулярної відстані:

$$\mathcal{L}_{intra} = \sum_{i}^{c}\sum_{j}^{c}\left(\left|DM_{ij}^{c'c'} - DM_{ij}^{cc}\right|\right), \tag{4.2}$$

для будь-якої пари атомів ліганду $ij \in LAS$, де *LAS* (локальні атомні структури) включає пари атомів, з'єднані ковалентними зв'язками, або через 2 зв'язки, або в одній кільцевій структурі [60], [150]; $DM_{ij}^{c'c'}$ - матриця внутрішньомолекулярних відстаней, розрахована з C'; DM_{ij}^{cc} - матриця внутрішньомолекулярних відстаней випадкового конформера ліганду. Врахування лише *LAS* пар уможливлює гнучкість ліганду під час оптимізації C', зберігаючи при цьому обмеження локальних внутрішньомолекулярних відстаней.

- Міжмолекулярні стеричні зіткнення:

$$\mathcal{L}_{inter-clash} = \sum_{i}^{p}\sum_{j}^{c}\left(ReLU\left(VdW_{ij}^{pc} - DM_{ij}^{pc'}\right) + 1\right)^2, \tag{4.3}$$

для будь-якого атома сайту зв'язування i та атома ліганду j, якщо $VdW_{ij}^{pc} > DM_{ij}^{pc'}$, де $ReLU$ - функція ReLU (rectified linear unit, зрізаний лінійний вузол); $VdW^{pc} \in \mathbb{R}^{p\times c\times 1}$ - міжмолекулярна сума радіусів ван-дер-ваальса атомів, помножена на $T_{VdW} \in (0, 1)$; T_{VdW} визначає частку суми радіусів ван-дер-ваальса, нижче якої утворюється стеричне зіткнення між атомами. Цей компонент використовується для запобігання стеричних зіткнень між білком та лігандом.

- Внутрішньомолекулярні стеричні зіткнення:

$$\mathcal{L}_{intra-clash} = \sum_{i}^{c}\sum_{j}^{c}\left(ReLU\left(VdW_{ij}^{cc} - DM_{ij}^{c'c'}\right) + 1\right)^2, \tag{4.4}$$

для будь-якої пари атомів ліганду ij, якщо $VdW_{ij}^{cc} > DM_{ij}^{c'c'}$ та $ij \notin LAS$, та $i \neq j$, де $VdW^{cc} \in \mathbb{R}^{c\times c\times 1}$ - внутрішньомолекулярна сума радіусів ван-дер-ваальса

атомів, помножена на T_{VdW}. Цей компонент використовується для запобігання стеричних зіткнень всередині ліганду.

Фінальна функція втрат розраховувалася як зважена сума вищезгаданих компонентів:

$$\mathcal{L} = w_{inter} * \mathcal{L}_{inter} + w_{intra} * \mathcal{L}_{intra} + w_{inter-clash} * \mathcal{L}_{inter-clash} + w_{intra-clash} * \mathcal{L}_{intra-clash}, \quad (4.5)$$

де w_{inter}, w_{intra}, $w_{inter-clash}$, $w_{intra-clash}$ - відповідні ваги компонентів.

Попри врахування обмежень міжатомних відстаней під час трансформації матриці відстаней у хмару точок C', отримана C' все ще містить артефакти та порушенням геометричних критеріїв якості PoseBusters. Ці проблеми добре відомі та про них повідомлялося для більшості методів докінгу на основі ML [156].

Проблему було частково вирішено шляхом застосування техніки диференціальної еволюції як фінального кроку при перетворенні матриці відстаней у положення ліганду. Диференціальна еволюція - це стохастичний підхід, який не використовує градієнтні методи для пошуку мінімуму та може виконувати пошук у великих областях конформаційного простору [176]. Було використано модифіковану версію алгоритму для оптимізації положення ліганду [162]. Отримавши хмару точок C' та випадковий вхідний конформер ліганду, було встановлено цільову функцію диференціальної еволюції на мінімізацію RMSD між координатами C' та конформера шляхом модифікації обертальних зв'язків конформера та вирівнювання його абсолютних координат по C'. У результаті, отримується оптимізований конформер ліганду з атомними координатами, близькими до C', який має правильну тетраедричну хіральність, стереохімію подвійних зв'язків, довжини та кути ковалентних зв'язків, планарність ароматичних кілець та планарність подвійних зв'язків.

Щоб мінімізувати кількість стеричних зіткнень ліганду після диференціальної еволюції, до цільової функції було включено додаткові внески між- та

внутрішньомолекулярних зіткнень, описані вище, використовуючи координати оптимізованого конформера замість C' для визначення зіткнень. В результаті, цільова функція була визначена наступним чином:

$$\mathcal{L}^{DE} = RMSD + w^{DE}_{inter-clash} * \mathcal{L}^{DE}_{inter-clash} + w^{DE}_{intra-clash} * \mathcal{L}^{DE}_{intra-clash}, \quad (4.6)$$

де $w^{DE}_{inter-clash}$ та $w^{DE}_{intra-clash}$ - відповідні ваги компонентів.

Ваги компонентів при генерації C' та цільової функції диференціальної еволюції ($w_{inter} = 1.11$, $w_{intra} = 7.57$, $w_{inter-clash} = 1.31$, $w_{intra-clash} = 0$, $w^{DE}_{inter-clash} = 6.7$, $w^{DE}_{intra-clash} = 2.7$) разом з $T_{VdW} = 0.762$ підбиралися з використанням перевірок якості PoseBusters на 300 PLP з унікальними лігандами з валідаційної вибірки даних PLINDER. Підбір оптимальних параметрів виконувався протягом 200 ітерацій за допомогою програмного забезпечення для оптимізації параметрів Optuna v.3.5.0 [105] з використанням алгоритму деревоподібного оцінювача Парзена. Найкращі параметри обиралися на основі найвищої частки передбачених положень лігандів з RMSD < 2Å, що пройшли всі перевірки якості PoseBusters (PB-Валідні).

Додатково, з метою максимізації валідності результатів передбачень ArtiDock та післяобробки за критеріями PoseBusters, розглядалися дві стратегії швидкої фізично-обґрунтованої локальної мінімізації положень лігандів:

- Локальна оптимізація енергії за допомогою Vina v.1.2.6 до 500 кроків з використанням методу Бройдена-Флетчера-Голдфарба-Шанно (BFGS) [177].
- Мінімізація методом BFGS за допомогою універсального силового поля (UFF) [178] до 500 ітерацій з використанням RDKit v.2024.9.5. Усі атоми сайту були зафіксовані, і дозволявся рух лише молекули ліганду.

4.3.4. Підготовка даних для тренування моделей машинного навчання

Тренувальна вибірка даних PLINDER була використана як навчальний набір даних для моделі ArtiDock-PL. Сирий набір даних містить 344033 пари білок-ліганд (PLP) із 76901 експериментально визначених структур PDB, що охоплюють 34103 унікальних ліганди, визначених за їхніми CCD.

Після попередньої обробки та фільтрації за якістю (Додаток А), тренувальний набір даних складався з 227066 PLP із 67888 ідентифікаторів PDB. Комплекси містили 29440 унікальних кодів CCD лігандів. 34% PLP містили ліганди, анотовані як кофактори. 38% PLP включали ліганди, що відповідають правилу п’яти Ліпінського [179]. Варто зазначити, що, згідно з метаданими PLINDER, 9% відфільтрованих PLP включають ліганди, ковалентно зв’язані зі своїми мішенями. Однак, вони були збережені в наборі даних, оскільки відповідали критеріям фільтрації за якістю.

Оскільки в комплексах було збережено всі атоми та молекули, кожен PLP може містити інші ліганди, іони та воду як частину сайту зв’язування білка. У відфільтрованих даних 12% PLP містили інші ліганди (не кофактори) у сайтах; 8% містили кофактори; 16% включали принаймні один іон. 48% лігандів мали принаймні одну нековалентну взаємодію з молекулою води у комплексі (водний місток).

Для тестування на наборі даних PoseX, модель ArtiDock-PX тренували на 275117 відфільтрованих за якістю PLP з PLINDER (з усіх вибірок: тренувальна, валідаційна, тестова), випущених до 01.01.2022. Дані містили 248910 систем, 84061 ідентифікатор PDB та 34460 унікальних кодів CCD лігандів.

Багато молекулярних структур природним чином надмірно представлені в PDB як популярні мішені для ліків, модельні білки, поширені ліко-подібні ліганди, кофактори тощо. Наприклад, серед PLP PLINDER 3.6% сайтів зв’язування

представлені глікопротеїном оболонки ВІЛ GP120. Крім того, майже 18% PLP містять ліганди з нуклеозидним ядром (приклади CCD: ATP, UDP, NAD).

Щоб усунути надлишковість певних тренувальних даних, PLP було кластеризовано за структурою ліганду, схожістю послідовності білкового сайту та схожістю взаємодії білок-ліганд, як описано в Додатку Б. У результаті відфільтровані за якістю PLP було згруповано у 40148 (тренувальний набір PLINDER) або 53205 (часове розбиття для набору даних PoseX) кластерів.

4.3.5. Особливості тренування моделей машинного навчання

Навчання моделі базувалося на мінімізації різниці між передбаченою та реальною матрицями міжмолекулярних відстаней сайт зв'язування білка-ліганд з використанням функції втрат MSE (середньоквадратичної похибки). Реальна матриця відстаней була отримана з координат експериментально визначеного положення ліганду (координати частини атомів могли бути відсутні через експериментальні обмеження), тоді як випадково згенерований конформер повного ліганду використовувався як вхідні дані моделі для отримання передбаченої матриці відстаней. Відстані були обмежені 10 Å, щоб зосередити місткість моделі на передбаченні міжатомних контактів на близькій відстані.

Під час розрахунку втрат враховувалися відмінності в порядку атомів ліганду між експериментально визначеною та вхідною структурами (вхідні структури генерувались зі SMILES), потенційна відсутність атомів в експериментальній структурі та можливі внутрішньомолекулярні симетрії. Модель навчалася протягом 200 епох (10 мільйонів ітерацій) з оптимізатором Adam і темпом навчання (learning rate) 0.0001. Використовувався малий розмір пакета (batch size) 2 через обмеження пам'яті графічного процесора. Для навчання моделі ArtiDock було використано робочі станції з графічним процесором NVIDIA GeForce RTX 4090 та 16-ядерним процесором AMD Ryzen 9 7950X3D.

Для навчання моделі використовувалися лише PLP, що пройшли фільтри якості. Для кожної епохи тренування випадковим чином відбиралася лише одна PLP з кожного кластера. Проте було помічено, що певні коди CCD лігандів залишалися надмірно представленими, оскільки вони з'являлися у багатьох способах зв'язування з широким спектром білків (наприклад, HSR, ADP, HEM). Тому вони випадковим чином виключалися з вибірки, щоб брати участь не більше ніж у 0.1% відібраних PLP.

Додатково, під час навчання моделі було впроваджено розширення даних на основі сайту білка, як детально описано в Додатку В.

4.4. Результати розробки та порівняння методу машинного навчання для передбачення положення малих молекул

4.4.1. Порівняння з традиційними підходами

Модель ArtiDock-PL, навчена на тренувальному наборі PLINDER для запобігання витоку даних, перевершує класичні методи докінгу на тестовому наборі PLINDER (Таблиця 4.1). Підхід на основі ML передбачає положення лігандів із медіаною RMSD 2.0 Å та медіаною lDDT-PLI 0.71, у порівнянні з 2.8 Å та 0.64 відповідно для найефективнішого традиційного методу, Glide. Серед класичних інструментів докінгу Glide стабільно демонструє найкращу точність за більшістю критеріїв, а Vina показує близький результат. Натомість AutoDock-CPU демонструє найнижчу ефективність із медіаною RMSD 3.5 Å.

Таблиця 4.1. Точність та якість передбачень методів докінгу за критеріями оцінювання на тестовому наборі даних PLINDER (N=675). Зеленим кольором позначено відносну ефективність (темніше - краще). Найкращі значення виділено жирним шрифтом.

Метод	RMSD медіана (Å)	RMSD < 2Å	lDDT-PLI медіана	lDDT-PLI > 0.5	PB-Валідні	RMSD < 2Å і PB-Валідні	Зразків втрачено
AutoDock-CPU	3.5	0.32	0.56	0.56	0.90	0.30	18
AutoDock-GPU	3.2	0.35	0.60	0.60	0.91	0.33	17
Vina	2.9	0.38	0.62	0.59	**0.95**	0.36	22
Glide	2.8	0.39	0.64	0.62	0.91	0.36	66
ArtiDock	**2.0**	**0.48**	**0.71**	**0.79**	0.72	0.38	0
ArtiDock + Vina	2.2	0.45	0.70	0.74	0.84	**0.39**	19
ArtiDock + UFF	2.3	0.41	0.69	0.76	**0.95**	**0.39**	14

Як видно з Рисунка 4.3, традиційні методи докінгу мають значно ширший розподіл RMSD, ніж ArtiDock, з вищими піковими значеннями lDDT-PLI (приблизно 0.9 проти 0.75). Це свідчить про те, що вони іноді здатні передбачати дуже точні положення, але водночас створюють помітну кількість низькоякісних результатів. На відміну від них, ArtiDock генерує значно менше низькоякісних передбачень, що призводить до кращої загальної точності на цьому наборі даних.

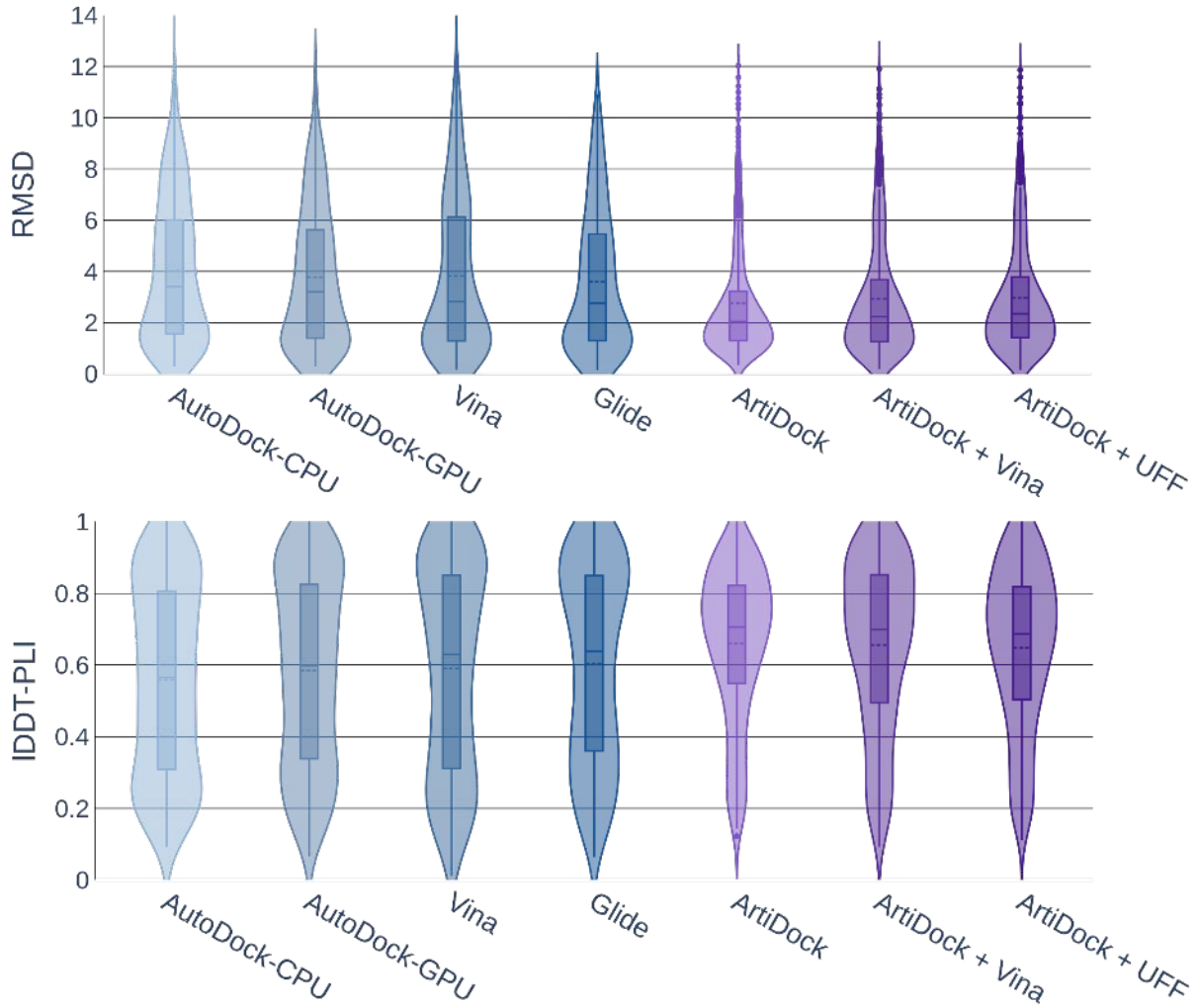

Рисунок 4.3. Розподіл RMSD (Å, менше - краще) та lDDT-PLI (більше - краще) на тестовому наборі даних PLINDER (N=675). Межі смуг представляють перший та третій квантилі, суцільні та пунктирні лінії всередині смуг вказують медіанні та середні значення відповідно. Форма представляє розподіл значень (графік густини розподілу).

Проблеми якості PoseBusters для всіх протестованих інструментів переважно виникають через занадто високу внутрішню енергію передбаченого положення ліганду або стеричні зіткнення з атомами сайту зв'язування (включаючи гетероатоми), як показано в Таблиці E6. Якщо розглядати класичні підходи, відсоток результатів із проблемами внутрішньої енергії коливається від 5% в AutoDock-CPU/GPU до 1% у Glide. Натомість Glide не вдається уникнути стеричних зіткнень у до 7% передбачень, тоді як Vina створює зіткнення лише в 1% випадків.

Частка передбачень ArtiDock, які проходять критерії PB-Валідності, є нижчою, ніж в інших інструментів. Однак додаткове застосування швидкої оптимізації положення ліганду після передбачення за допомогою Vina або RDKit UFF покращує цей показник ціною незначного зниження точності. Зокрема, мінімізація UFF збільшує частку хімічно та геометрично правильних положень на 22%, зрівнюючись із показниками PB-Валідності для Vina. Навіть після цієї корекції, ArtiDock продовжує перевершувати класичні методи докінгу за всіма іншими критеріями.

Щодо обчислювальної ефективності методів, то вартість обчислень на робочих станція з використанням ArtiDock приблизно у 2 рази нижча порівняно з найшвидшим класичним інструментом. Деталі наведено в Додатку Г.

Для деяких пар білок-ліганд не вдалося отримати передбачень, як показано в стовпці “Зразків втрачено” Таблиці 4.1. Більшість невдалих передбачень від AutoDock-CPU/GPU та Vina пов’язані з проблемами під час попередньої обробки білка або ліганду, наприклад, неможливістю обробити молекули, що містять селен або атоми металів. З 66 помилок Glide, 11 пов’язані з подібними помилками попередньої обробки лігандів, тоді як решта помилок переважно зумовлені відсутністю прийнятних способів зв’язування або відсіюванням усіх положень під час мінімізації енергії. Помилки під час мінімізації енергії за UFF зазвичай виникають, коли RDKit не може коректно обробити білковий сайт зв’язування. Важливо зазначити, що жоден метод не “штрафувався” за невдалі передбачення. Усі наведені результати базуються на підмножині успішно передбачених положень лігандів.

4.4.2. Вплив кофакторів, іонів та молекул води на ефективність підходів

Структурні дані PLINDER зберігають небілкові молекули (гетероатоми), присутні в оригінальних структурах PDB, такі як кофактори, іони та зв’язані з лігандом молекули води (Додаток А). Усі гетероатоми було включено до вхідного сайту зв’язування для докінгу, а ефективність докінгу відстежувалася на піднаборах даних,

що містять різні типи небілкових молекул. Загалом, найвища точність докінгу спостерігалася в сайтах, що містять органічні кофактори (Рисунок 4.4), хоча на це міг вплинути відносно невеликий розмір вибірки - всього 23 PLP.

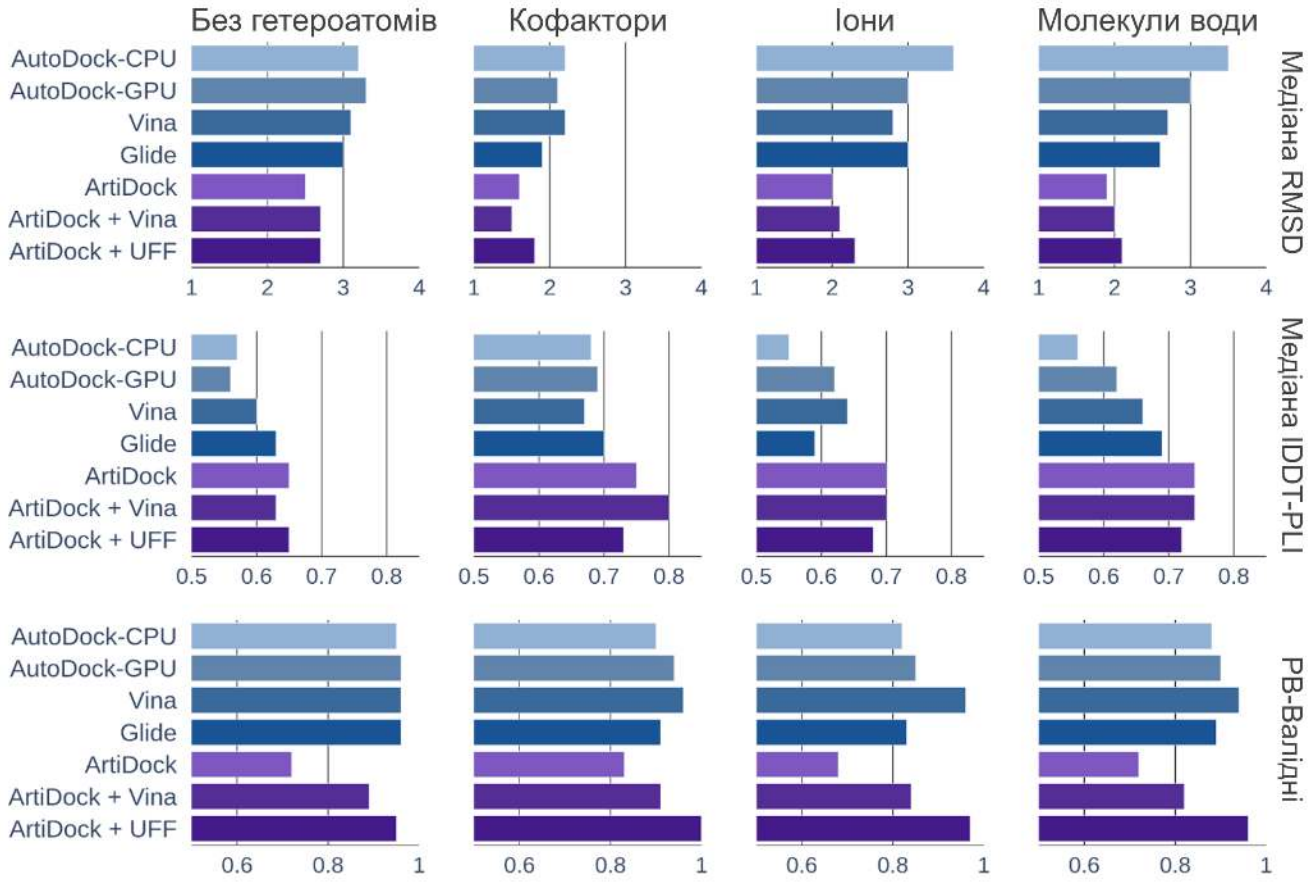

Рисунок 4.4. Медіана RMSD (Å, менше - краще), медіана lDDT-PLI (більше - краще) та частка PB-Валідних передбачень (більше - краще) на піднаборах тестових даних PLINDER без гетероатомів (N=189), з органічними кофакторами (N=23), з іонами (N=123) або зв’язаними з лігандом молекулами води (N=403) у сайті зв’язування.

ArtiDock-PL стабільно перевершує класичні методи докінгу за точністю (RMSD та lDDT-PLI) в усіх піднаборах, демонструючи найбільші покращення в сайтах, що містять іони та воду. Зокрема, ArtiDock досяг медіани RMSD 2.0 Å у сайтах з іонами (проти 2.8 Å для найкращого класичного методу) та 1.9 Å у сайтах з водою (проти 2.6 Å для найкращого класичного методу). Згідно з перевірками якості PoseBusters, оптимізація передбачень ArtiDock на основі UFF знаходиться на одному рівні або

перевершує найкращий класичний підхід, тоді як мінімізація на основі Vina дає гірші результати.

Серед традиційних інструментів, Glide демонструє найкращу точність у сайтах зв'язування, що містять кофактори, воду або лише білок. Vina є найточнішим методом у сайтах, що містять іони, і демонструє найкращу хімічну валідність у всіх піднаборах, з часткою PB-Валідних передбачень від 0.94 до 0.96 (Рисунок 4.4).

4.4.3. Вплив незв'язаних конформацій протеїнів на ефективність підходів

На відміну від тесту PLINDER, який використовує зв'язані (голо) білкові структури, піднабір PLINDER MLSB розроблено для оцінки інструментів докінгу в більш реалістичних умовах, використовуючи апо або передбачені білкові структури. Ця зміна призводить до суттєвого зниження точності, більшість методів демонструють двократне збільшення медіани RMSD (Таблиця 4.2).

Таблиця 4.2. Точність та якість передбачень методів докінгу за критеріями оцінювання на тестовому наборі даних PLINDER MLSB (N=344). Зеленим кольором позначено відносну ефективність (темніше - краще). Найкращі значення виділено жирним шрифтом.

Метод	RMSD медіана (Å)	RMSD < 2Å	lDDT-PLI медіана	lDDT-PLI > 0.5	PB-Валідні	RMSD < 2Å і PB-Валідні	Зразків втрачено
AutoDock-CPU	5.9	0.10	0.30	0.23	0.91	0.10	6
AutoDock-GPU	6.2	0.09	0.28	0.21	0.91	0.09	5
Vina	6.5	0.08	0.27	0.17	**0.94**	0.07	5
Glide	5.5	0.13	0.33	0.26	**0.94**	0.13	54
ArtiDock	**3.4**	**0.23**	**0.51**	**0.52**	0.65	**0.18**	0
ArtiDock + Vina	4.2	0.19	0.46	0.43	0.80	**0.18**	5
ArtiDock + UFF	3.9	0.16	0.47	0.45	0.89	0.15	6

Як і для голо конформацій білків, ArtiDock-PL стабільно перевершує всі класичні методи, досягаючи найнижчої медіани RMSD (3.4Å) та найвищої медіани lDDT-PLI (0.51). Розподіл критеріїв точності в тесті PLINDER MLSB показує, що ArtiDock передбачає більшу кількість високоякісних положень порівняно з іншими методами. Зокрема, розподіл lDDT-PLI для ArtiDock має пік близько 0.55, тоді як Glide має пік приблизно на 0.25, з ще нижчими значеннями для інших методів (Рисунок 4.5).

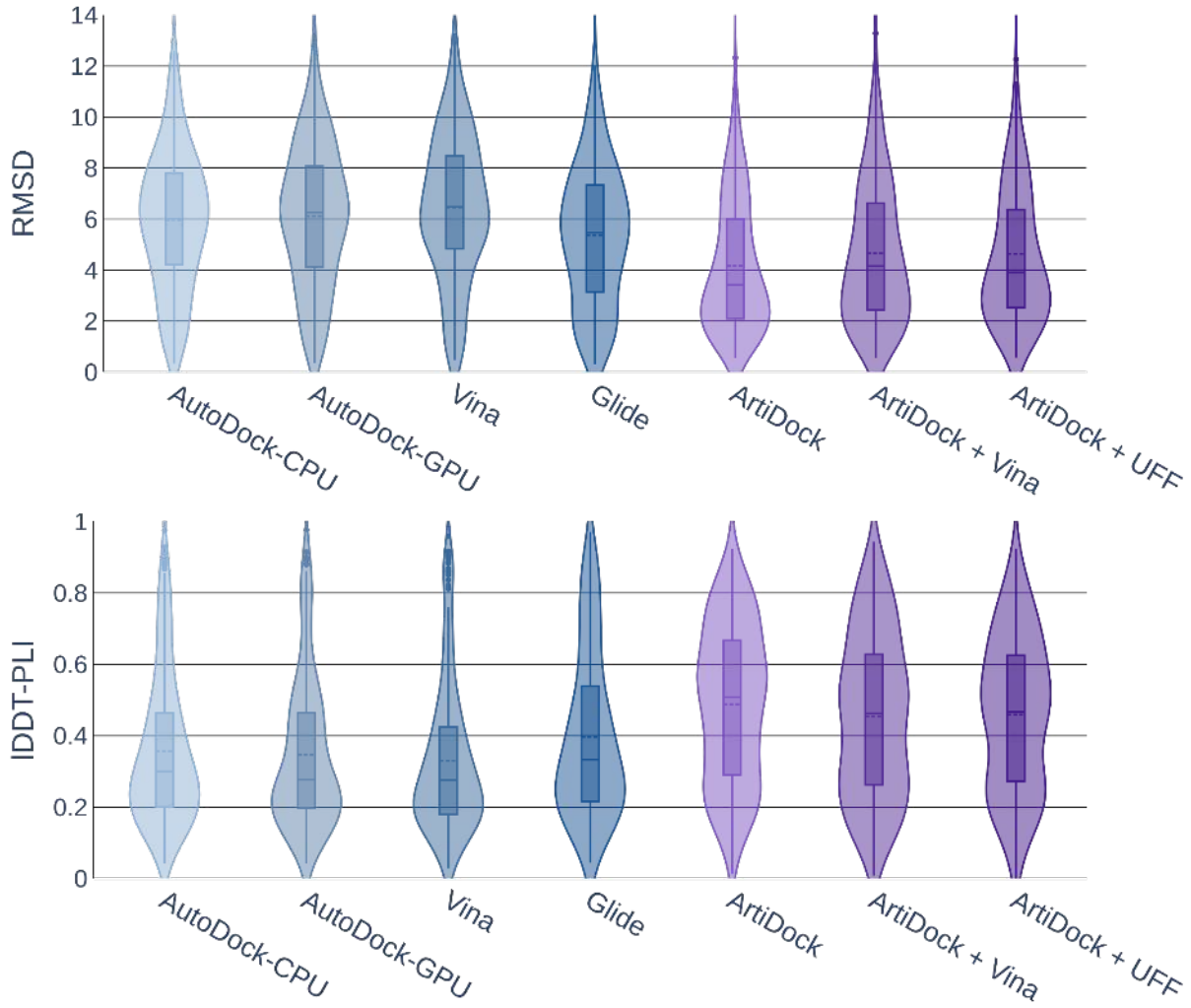

Рисунок 4.5. Розподіл RMSD (Å, менше - краще) та lDDT-PLI (більше - краще) на тестовому наборі даних PLINDER MLSB (N=344). Межі смуг представляють перший та третій квантилі, суцільні та пунктирні лінії всередині смуг вказують медіанні та середні значення відповідно. Форма представляє розподіл значень (графік густини розподілу).

Серед традиційних інструментів Glide зберігає найкращу загальну ефективність. Vina демонструє найбільш різке падіння точності, опустившись з другого найкращого традиційного методу в тесті PLINDER до найгіршого в PLINDER MLSB (Таблиця 4.2). Низька ефективність Vina на незв'язаних структурах частково пояснюється потенціалами на основі знань, що використовуються у функції оцінки [150], які можуть спричинити зсув точності методу в бік голо конформацій.

Показник PB-Валідності для ArtiDock у тесті MLSB до 7 відсоткових пунктів нижчий, ніж у тесті зі зв'язаними конформаціями, незалежно від мінімізації після передбачення. Це зниження пояснюється випадками, коли ліганд не може фізично вміститися в незв'язану структуру через значні зміни геометрії сайту зв'язування. У той час як класичні методи докінгу схильні зміщувати ліганд за межі сайту у відповідь на стеричні зіткнення, ArtiDock передбачає матрицю відстаней, яка може не повністю враховувати такі просторові несумісності, але точніше відображає істинний спосіб зв'язування.

Оскільки післяобробка передбачень ArtiDock надає пріоритет збереженню передбачених міжатомних відстаней білок-ліганд, отримані положення в апо або передбачених структурах з більшою ймовірністю демонструватимуть стеричні проблеми. Ця тенденція додатково підтверджується аналізом, що включає RMSD між накладеними атомами остова білка Cα у голо- та апо/передбачених сайтах ($RMSD_{C\alpha}^{holo-apo}$). Як показано на Рисунку Д3, частка результатів ArtiDock до фізичної мінімізації, що відповідає критеріям PB-Валідності, зменшується з 0.81 у підмножині з $RMSD_{C\alpha}^{holo-apo} < 0.1$ Å до 0.53 у підмножині з $RMSD_{C\alpha}^{holo-apo} > 0.5$ Å. Натомість класичні методи не демонструють значних змін у валідності в цих підмножинах.

При розгляді критеріїв точності, не було виявлено кореляції між $RMSD_{C\alpha}^{holo-apo}$ та RMSD між реальним та передбаченим положенням ліганду або lDDT-PLI (Таблиця Е7). Не спостерігалося помітного зниження точності до порогового значення

$RMSD_{C\alpha}^{holo-apo} = 0.5$ Å. Проте, більші відхилення за межами 0.5 Å призводять до помітного падіння точності (Рисунок Д3). Варто зауважити, що частка даних з $RMSD_{C\alpha}^{holo-apo} > 0.5$ Å становить лише 11.6%. Щоб повністю оцінити вплив відхилень незв'язаних структур на ефективність докінгу, необхідні додаткові тестування з більш структурно відмінними апо/передбаченими білками.

4.4.4. Порівняння з іншими методами машинного навчання

Тестовий набір даних PoseX для селф-докінгу (голо) продемонстрував, що ArtiDock-PX працює на рівні з найкращими методами за RMSD (Рисунок 4.6), демонструючи при цьому вищу обчислювальну ефективність (Додаток Г). ArtiDock успішно відтворив 73.5% комплексів з RMSD < 2 Å, майже зрівнявшись із найточнішим у порівнянні методом SurfDock [68] (77.6%). Коли також враховувалась якість передбачень (RMSD < 2 Å та PB-Валідні), ArtiDock досяг найкращих результатів серед всіх порівняних підходів з 63.9% успішних передбачень. Швидка оптимізація після передбачення за допомогою Vina або UFF додатково покращила цей показник до 66.2% або 66.4% відповідно.

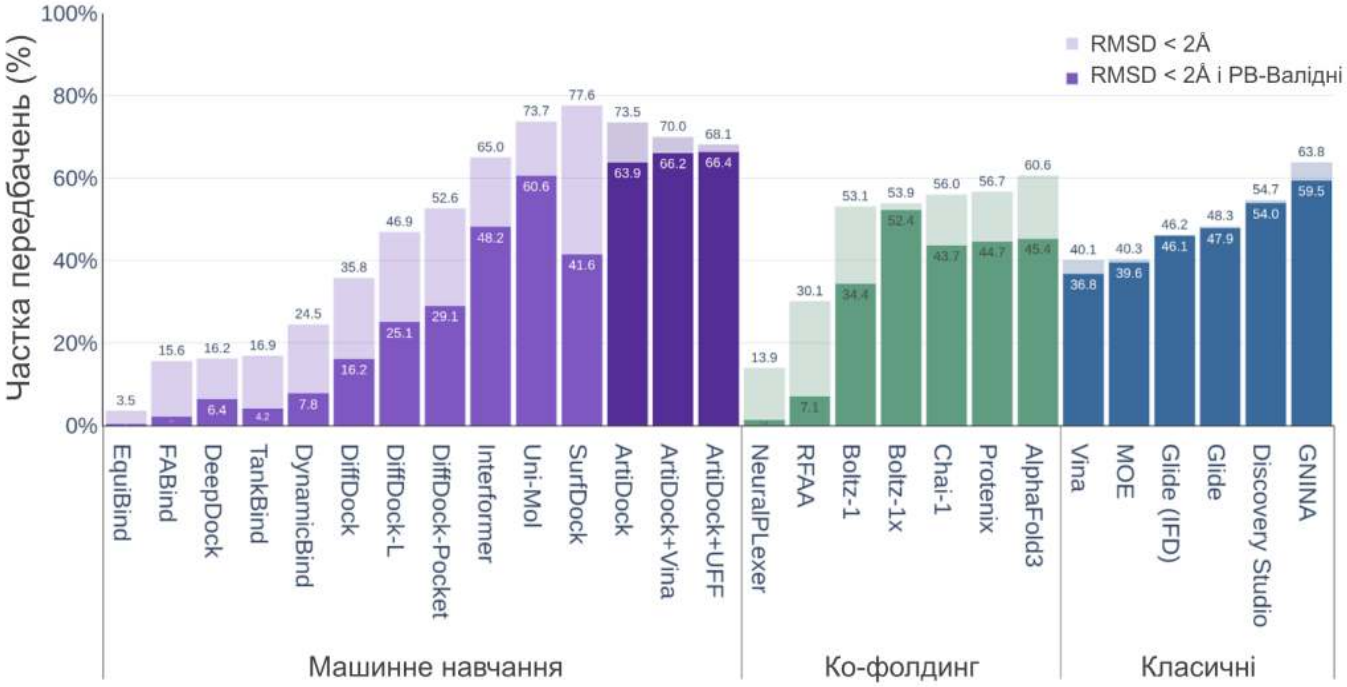

Рисунок 4.6. Частка передбачень з RMSD < 2Å та RMSD < 2Å і PB-Валідністю для методів машинного навчання (окрім ко-фолдингу), ко-фолдингу та класичного докінгу на наборі даних PoseX зі зв’язаними вхідними структурами білків (селф-докінг, N=718).

Методи ко-фолдингу продемонстрували відносний приріст ефективності в сценарії крос-докінгу, оскільки вони не залежать від жорсткого представлення сайту субоптимальних для зв’язування апо конформацій білків (Рисунок 4.7). З усім тим, жодна з моделей ко-фолдингу суттєво не перевершила ArtiDock за часткою передбачень з RMSD < 2 Å. ArtiDock зберіг найкращі результати, досягнувши 62.4% комплексів, які були одночасно точними та валідними за PoseBusters, демонструючи близьку ефективність до Boltz-1x [70] (60.2%). Цікаво, що оптимізація передбачень за допомогою Vina або UFF дещо погіршила ефективність ArtiDock на цьому наборі даних, ймовірно, через неможливість для фізично обґрунтованих алгоритмів визначити енергетично вигідні положення у деяких сайтах незв’язаних білків.

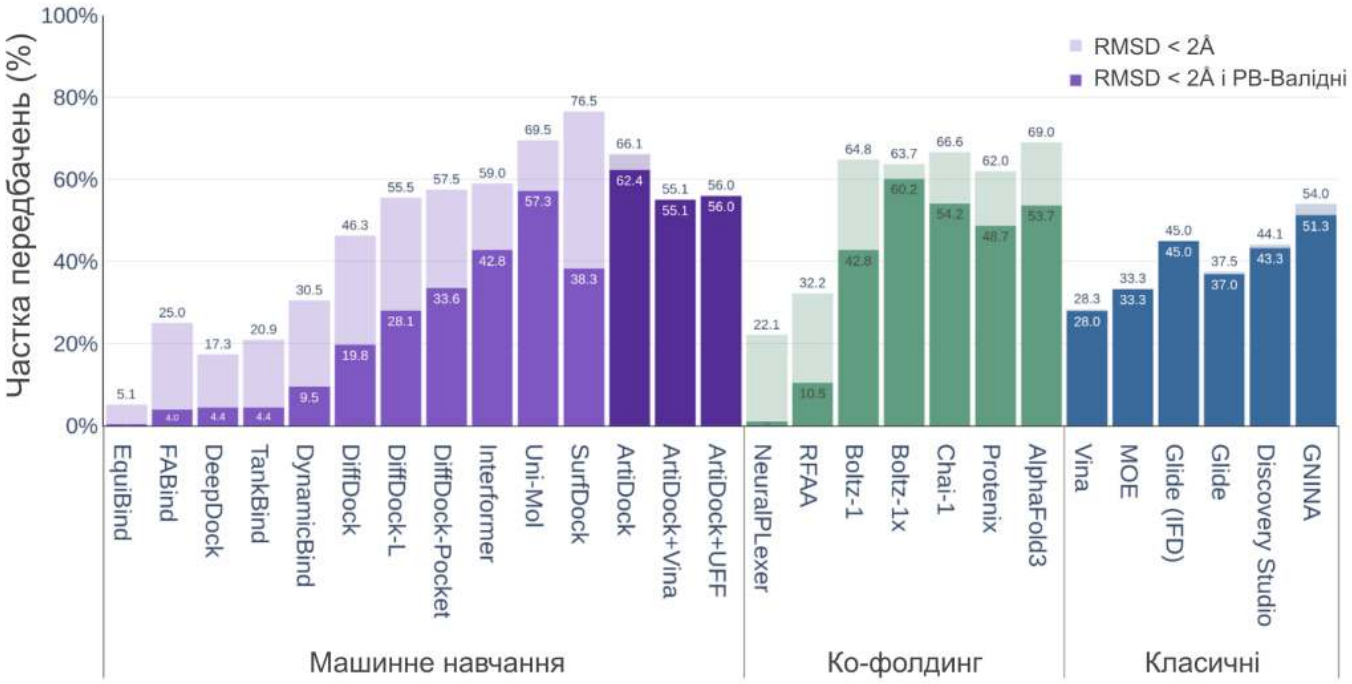

Рисунок 4.7. Частка передбачень з RMSD < 2Å та RMSD < 2Å і PB-Валідністю для методів машинного навчання (окрім ко-фолдингу), ко-фолдингу та класичного докінгу на наборі даних PoseX з незв’язаними вхідними структурами білків (крос-докінг, N=1312).

4.5. Обмеження та перспективи розробленого методу

Поточне оцінювання ArtiDock на наборі даних PLINDER було проведено у порівнянні з традиційними інструментами докінгу. Було забезпечено суворі стандарти оцінювання, які вимагають перенавчання ML моделей на тренувальному піднаборі PLINDER. Це обмежує можливість широкого тестування альтернативних методів докінгу на основі ML, особливо великих дифузійних моделей та архітектур ко-фолдингу.

Надалі може бути доцільною оцінка ширшого спектра класичних інструментів докінгу, таких як комерційно доступні CCDC GOLD [180] та DOCK [181], нові рішення з відкритим кодом, такі як PandaDock (https://github.com/pritampanda15/PandaDock.git), та різноманітні альтернативи на основі Vina [182], [183], [184], [185], щоб охопити більшість фізично обґрунтованих

методів докінгу та визначити сфери для покращення їхніх аналогів на основі машинного навчання.

Хімічна валідність залишається проблемою для більшості підходів докінгу на основі ML, включаючи ArtiDock. Попри високу точність передбачення, результати ArtiDock часто потребують додаткових етапів післяобробки для забезпечення прийнятної якості. Майбутні зусилля слід спрямувати на вдосконалення внутрішніх механізмів ArtiDock для прямого створення хімічно валідних положень лігандів, зменшуючи або усуваючи таким чином залежність від зовнішніх коригувань.

ArtiDock передбачає лише одне положення на пару білок-ліганд та не потребує використання функції оцінки. Тому, метод не підходить для пріоритезації альтернативних положень одного й того самого ліганду або порівняння різних лігандів за силою зв’язування, як це роблять традиційні методи докінгу. Положення, згенеровані ArtiDock, є оптимальними для даного ліганду та даної конформації сайту зв’язування, але їх оцінювання має виконуватися зовнішніми засобами за допомогою алгоритмічних функцій оцінки або відповідних ML моделей. Для оптимізації та підвищення точності передбачення в сценаріях віртуального скринінгу з високою пропускною здатністю необхідна розробка спеціальної функції оцінки на основі ML, тісно інтегрованої з ArtiDock. Такий підхід до оцінювання, ймовірно, забезпечить більш ефективне застосування методу в реальних завданнях з розробки ліків.

4.6. Висновки розділу

У цьому розділі представлено ArtiDock - підхід до протеїн-лігандного докінгу на основі DL, оптимізований для сценаріїв високопродуктивного віртуального скринінгу. Оцінювання, проведене на тестовому наборі даних PLINDER, демонструє, що ArtiDock передбачає положення лігандів значно точніше порівняно з традиційними методами докінгу, досягаючи медіани RMSD 2.0 Å та медіани lDDT-PLI 0.71 проти 2.8 Å та 0.64 відповідно для найефективнішого класичного методу, Glide.

Обчислювально ефективна архітектура моделі ArtiDock та оптимізовані алгоритми післяобробки дозволяють досягти пропускної здатності докінгу, яка знаходиться на рівні або є кращою за найшвидші традиційні інструменти.

ArtiDock є точнішим за традиційні методи в різноманітних реальних сценаріях докінгу, включаючи білкові сайти, що містять кофактори, іони та молекули води. Підхід зберігає перевагу на наборі даних PLINDER MLSB, який включає структури білків, не зв'язані з лігандом, що відхиляються від оптимальних для зв'язування конформацій.

Інтеграція швидкого етапу фізично обґрунтованої обробки передбачень суттєво підвищує частку хімічно валідних положень до рівня традиційних підходів докінгу, без суттєвого погіршення точності передбачення або значного ускладнення обчислень.

У розділеному за часом тестовому наборі даних PoseX ArtiDock демонструє конкурентоспроможну точність порівняно з загалом повільнішими найефективнішими ML інструментами та досягає найкращих результатів, коли враховується валідність за PoseBusters.

Таким чином, ArtiDock сприяє зміні парадигми докінгу в бік інструментів на основі ML, ефективно поєднуючи високу пропускну здатність, точність та адаптивність до реалістичних сценаріїв докінгу.

Майбутня робота має бути пов'язана з включенням інших класичних підходів докінгу та моделей машинного навчання до поточного порівняння на наборі даних PLINDER , покращенням хімічної та геометричної валідності результатів ArtiDock, а також реалізацією власної функції оцінки передбачень ArtiDock.

Код для налаштування класичного докінгу, обробки вхідних даних, генерації результатів та оцінювання доступний у репозиторії GitHub: https://github.com/receptor-ai/dock-eval.git. Попередньо оброблені підмножини тестових даних PLINDER, передбачені положення лігандів від усіх порівнюваних

інструментів та таблиці з оцінками для кожного положення зберігаються у репозиторії Zenodo: https://zenodo.org/records/17937766. Підхід ArtiDock є комерційно доступним продуктом компанії Receptor.AI: https://receptor.ai/.

ВИСНОВКИ

У дисертаційній роботі вирішено завдання підвищення точності та швидкості передбачення параметрів взаємодії протеїнів з малими органічними молекулами шляхом поєднання фізичних принципів молекулярного розпізнавання в біологічних системах з методами машинного навчання.

1. Доведено, що інтеграція просторової інформації про структуру протеїну (отриманої за допомогою AlphaFold) у вигляді графів разом із фізико-хімічними властивостями амінокислот та еволюційними характеристиками дозволяє підвищити точність передбачення афінності зв’язування в системі протеїн-ліганд. Це підтверджується створеною моделлю машинного навчання 3DProtDTA, яка перевершує наявні аналоги на 1-20% за загальноприйнятими критеріями, такими як середньоквадратичне відхилення, коефіцієнт конкордації, та коефіцієнт кореляції між реальними та передбаченими значеннями афінності.

2. Встановлено, що використання детального статистичного аналізу геометричних параметрів нековалентних взаємодій (гідрофобних контактів, π-стекінгу, водневих та галогенних зв’язків, а також електростатичних контактів) як фізичного підґрунтя для алгоритму генерації штучних систем протеїн-ліганд дозволяє покращити якість тренувальних даних. Застосування цього алгоритму для аугментації тренувальних даних вдосконаленої моделі PocketCFDM дозволило підвищити частку правильних передбачень до 10% за критерієм RMSD між передбаченими та реальними положеннями ліганду та знизити частку передбачень зі стеричними зіткненнями на 3-6%.

3. Доведено, що застосування обчислювально ефективної архітектури штучних нейронних мереж, в межах підходу ArtiDock, для високопродуктивного молекулярного докінгу з фізично обґрунтованою післяобробкою передбачень дозволяє перевершити класичні підходи та інші методи машинного навчання за точністю та швидкістю. Висока точність зберігається в різноманітних

реальних сценаріях докінгу, включаючи сайти зв’язування протеїнів, що містять кофактори, іони та молекули води, а також на даних, що включають структури протеїнів, не зв’язані з лігандом, що відхиляються від оптимальних для зв’язування конформацій. Створений метод ArtiDock досягає медіани RMSD між передбаченими та реальними положеннями ліганду 2.0 Å, що значно точніше за класичні методи докінгу (Glide - 2.8 Å, Vina - 2.9 Å). Порівнюючи з іншими підходами машинного навчання, ArtiDock демонструє конкурентоспроможну точність відносно загалом повільніших найефективніших підходів та досягає найкращих результатів, коли враховується фізична та хімічна коректність передбачень (66.4% коректних передбачень порівняно з 60.6% у найближчого конкурента).

Розроблені методи реалізують підхід до апроксимації складної багатовимірної залежності між координатами та типами атомів системи протеїн-ліганд, з одного боку, та її кількісними характеристиками зв'язування - з іншого. Цю апроксимацію досягнуто шляхом використання у кожній із трьох задач інформації зі статистичних розподілів відповідного порядку - від одновимірних статистик геометричних параметрів нековалентних взаємодій до багатовимірних розподілів конформацій протеїнів та матриць міжмолекулярних міжатомних відстаней.

Результати роботи, зокрема методи ArtiDock та 3DProtDTA, було успішно застосовано для розв’язання проблем у галузі розробки ліків, а саме для високопродуктивного віртуального скринінгу великих бібліотек хімічних сполук. Це дозволяє відсіювати неефективних кандидатів на ранніх етапах та заощаджувати ресурси на синтез та біологічні експерименти. Розроблені інструменти впроваджено у платформу компанії Receptor.AI, де вони використовуються у проєктах пошуку нових молекул із терапевтичними властивостями.

Основними напрямками розвитку подальших досліджень є інтеграція фізично обґрунтованих функцій оцінки афінності зв’язування безпосередньо в процес навчання нейронних мереж та врахування динаміки протеїну в моделях докінгу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] T. Voitsitskyi *et al.*, “3DProtDTA: a deep learning model for drug-target affinity prediction based on residue-level protein graphs,” *RSC Adv.*, vol. 13, no. 15, pp. 10261–10272, Mar. 2023, doi: 10.1039/D3RA00281K.

[2] T. Voitsitskyi *et al.*, “Augmenting a training dataset of the generative diffusion model for molecular docking with artificial binding pockets,” *RSC Adv.*, vol. 14, no. 2, pp. 1341–1353, Jan. 2024, doi: 10.1039/D3RA08147H.

[3] T. Voitsitskyi *et al.*, “ArtiDock: Accurate Machine Learning Approach to Protein–Ligand Docking Optimized for High-Throughput Virtual Screening,” *J. Chem. Inf. Model.*, vol. 66, no. 4, pp. 2117–2128, Feb. 2026, doi: 10.1021/acs.jcim.5c02777.

[4] X. Du *et al.*, “Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods,” *Int. J. Mol. Sci.*, vol. 17, no. 2, Art. no. 2, Feb. 2016, doi: 10.3390/ijms17020144.

[5] M. K. Gilson and H.-X. Zhou, “Calculation of Protein-Ligand Binding Affinities,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, no. 1, pp. 21–42, Jun. 2007, doi: 10.1146/annurev.biophys.36.040306.132550.

[6] H. Li, Y. Xie, C. Liu, and S. Liu, “Physicochemical bases for protein folding, dynamics, and protein-ligand binding,” *Sci. China Life Sci.*, vol. 57, no. 3, pp. 287–302, Mar. 2014, doi: 10.1007/s11427-014-4617-2.

[7] R. Perozzo, G. Folkers, and L. Scapozza, “Thermodynamics of Protein–Ligand Interactions: History, Presence, and Future Aspects,” *J. Recept. Signal Transduct.*, vol. 24, no. 1–2, pp. 1–52, Jan. 2004, doi: 10.1081/RRS-120037896.

[8] C. A. MacRaild, A. H. Daranas, A. Bronowska, and S. W. Homans, “Global Changes in Local Protein Dynamics Reduce the Entropic Cost of Carbohydrate Binding in the Arabinose-binding Protein,” *J. Mol. Biol.*, vol. 368, no. 3, pp. 822–832, May 2007, doi: 10.1016/j.jmb.2007.02.055.

[9] L. M. Amzel, “Calculation of entropy changes in biological processes: Folding, binding, and oligomerization,” in *Methods in Enzymology*, vol. 323, Elsevier, 2000, pp. 167–177. doi: 10.1016/S0076-6879(00)23366-1.

[10] B. Ma, S. Kumar, C.-J. Tsai, and R. Nussinov, “Folding funnels and binding mechanisms,” *Protein Eng. Des. Sel.*, vol. 12, no. 9, pp. 713–720, Sep. 1999, doi: 10.1093/protein/12.9.713.

[11] C. Tsai, S. Kumar, B. Ma, and R. Nussinov, “Folding funnels, binding funnels, and protein function,” *Protein Sci.*, vol. 8, no. 6, pp. 1181–1190, Jan. 1999, doi: 10.1110/ps.8.6.1181.

[12] Y.-H. Xie, P. Sang, Y. Tao, and S.-Q. Liu, “154 Protein folding and binding funnels: a common driving force and a common mechanism,” *J. Biomol. Struct. Dyn.*, vol. 31, no. sup1, pp. 100–101, Jan. 2013, doi: 10.1080/07391102.2013.786396.

[13] J. D. Dunitz, “Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions,” *Chem. Biol.*, vol. 2, no. 11, pp. 709–712, Nov. 1995, doi: 10.1016/1074-5521(95)90097-7.

[14] J. D. Chodera and D. L. Mobley, “Entropy-Enthalpy Compensation: Role and

Ramifications in Biomolecular Ligand Recognition and Design," *Annu. Rev. Biophys.*, vol. 42, no. 1, pp. 121–142, May 2013, doi: 10.1146/annurev-biophys-083012-130318.
[15] U. Ryde, "A fundamental view of enthalpy–entropy compensation," *Med Chem Commun*, vol. 5, no. 9, pp. 1324–1336, 2014, doi: 10.1039/C4MD00057A.
[16] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme," *Berichte Dtsch. Chem. Ges.*, vol. 27, no. 3, pp. 2985–2993, Oct. 1894, doi: 10.1002/cber.18940270364.
[17] D. E. Koshland, "Application of a Theory of Enzyme Specificity to Protein Synthesis," *Proc. Natl. Acad. Sci.*, vol. 44, no. 2, pp. 98–104, Feb. 1958, doi: 10.1073/pnas.44.2.98.
[18] D. Tobi and I. Bahar, "Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state," *Proc. Natl. Acad. Sci.*, vol. 102, no. 52, pp. 18908–18913, Dec. 2005, doi: 10.1073/pnas.0507603102.
[19] P. Bongrand, "Ligand-receptor interactions," *Rep. Prog. Phys.*, vol. 62, no. 6, pp. 921–968, Jun. 1999, doi: 10.1088/0034-4885/62/6/202.
[20] Y.-H. Xie, Y. Tao, and S.-Q. Liu, "153 Wonderful roles of the entropy in protein dynamics, binding and folding," *J. Biomol. Struct. Dyn.*, vol. 31, no. sup1, pp. 98–100, Jan. 2013, doi: 10.1080/07391102.2013.786395.
[21] M. Held, P. Metzner, J.-H. Prinz, and F. Noé, "Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations," *Biophys. J.*, vol. 100, no. 3, pp. 701–710, Feb. 2011, doi: 10.1016/j.bpj.2010.12.3699.
[22] J. Schluttig, D. Alamanova, V. Helms, and U. S. Schwarz, "Dynamics of protein-protein encounter: A Langevin equation approach with reaction patches," *J. Chem. Phys.*, vol. 129, no. 15, p. 155106, Oct. 2008, doi: 10.1063/1.2996082.
[23] G. Odriozola and M. Lozada-Cassou, "Statistical Mechanics Approach to Lock-Key Supramolecular Chemistry Interactions," *Phys. Rev. Lett.*, vol. 110, no. 10, p. 105701, Mar. 2013, doi: 10.1103/PhysRevLett.110.105701.
[24] A. Oleinikova, N. Smolin, I. Brovchenko, A. Geiger, and R. Winter, "Formation of Spanning Water Networks on Protein Surfaces via 2D Percolation Transition," *J. Phys. Chem. B*, vol. 109, no. 5, pp. 1988–1998, Feb. 2005, doi: 10.1021/jp045903j.
[25] P. T. Corbett, L. H. Tong, J. K. M. Sanders, and S. Otto, "Diastereoselective Amplification of an Induced-Fit Receptor from a Dynamic Combinatorial Library," *J. Am. Chem. Soc.*, vol. 127, no. 25, pp. 8902–8903, Jun. 2005, doi: 10.1021/ja050790i.
[26] H. Kumar, V. Kasho, I. Smirnova, J. S. Finer-Moore, H. R. Kaback, and R. M. Stroud, "Structure of sugar-bound LacY," *Proc. Natl. Acad. Sci.*, vol. 111, no. 5, pp. 1784–1788, Feb. 2014, doi: 10.1073/pnas.1324141111.
[27] P. Hariharan and L. Guan, "Insights into the Inhibitory Mechanisms of the Regulatory Protein IIAGlc on Melibiose Permease Activity," *J. Biol. Chem.*, vol. 289, no. 47, pp. 33012–33019, Nov. 2014, doi: 10.1074/jbc.M114.609255.
[28] G. G. Hammes, Y.-C. Chang, and T. G. Oas, "Conformational selection or induced fit: A flux description of reaction mechanism," *Proc. Natl. Acad. Sci.*, vol. 106, no. 33, pp. 13737–13741, Aug. 2009, doi: 10.1073/pnas.0907195106.

[29] S. Kumar, B. Ma, C. Tsai, N. Sinha, and R. Nussinov, “Folding and binding cascades: Dynamic landscapes and population shifts,” *Protein Sci.*, vol. 9, no. 1, pp. 10–19, Jan. 2000, doi: 10.1110/ps.9.1.10.
[30] P. Csermely, R. Palotai, and R. Nussinov, “Induced fit, conformational selection and independent dynamic segments: an extended view of binding events,” *Trends Biochem. Sci.*, vol. 35, no. 10, pp. 539–546, Oct. 2010, doi: 10.1016/j.tibs.2010.04.009.
[31] D. D. Boehr, R. Nussinov, and P. E. Wright, “The role of dynamic conformational ensembles in biomolecular recognition,” *Nat. Chem. Biol.*, vol. 5, no. 11, pp. 789–796, Nov. 2009, doi: 10.1038/nchembio.232.
[32] R. Nussinov, B. Ma, and C.-J. Tsai, “Multiple conformational selection and induced fit events take place in allosteric propagation,” *Biophys. Chem.*, vol. 186, pp. 22–30, Feb. 2014, doi: 10.1016/j.bpc.2013.10.002.
[33] Z. Zhao, L. Zhao, C. Kong, J. Zhou, and F. Zhou, “A review of biophysical strategies to investigate protein-ligand binding: What have we employed?,” *Int. J. Biol. Macromol.*, vol. 276, no. Pt 2, p. 133973, Sep. 2024, doi: 10.1016/j.ijbiomac.2024.133973.
[34] H.-W. Wang and J.-W. Wang, “How cryo-electron microscopy and X-ray crystallography complement each other,” *Protein Sci. Publ. Protein Soc.*, vol. 26, no. 1, pp. 32–39, Jan. 2017, doi: 10.1002/pro.3022.
[35] M. S. Smyth and J. H. J. Martin, “x Ray crystallography,” *Mol. Pathol.*, vol. 53, no. 1, pp. 8–14, Feb. 2000, doi: 10.1136/mp.53.1.8.
[36] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.
[37] W. S. Veeman, “Nuclear magnetic resonance, a simple introduction to the principles and applications,” *Geoderma*, vol. 80, no. 3, pp. 225–242, Nov. 1997, doi: 10.1016/S0016-7061(97)00054-2.
[38] D. M. Dias and A. Ciulli, “NMR approaches in structure-based lead discovery: Recent developments and new frontiers for targeting multi-protein complexes,” *Prog. Biophys. Mol. Biol.*, vol. 116, no. 2–3, pp. 101–112, Nov. 2014, doi: 10.1016/j.pbiomolbio.2014.08.012.
[39] M. Carroni and H. R. Saibil, “Cryo electron microscopy to determine the structure of macromolecular complexes,” *Methods San Diego Calif*, vol. 95, pp. 78–85, Feb. 2016, doi: 10.1016/j.ymeth.2015.11.023.
[40] K.-F. Zhu *et al.*, “Applications and prospects of cryo-EM in drug discovery,” *Mil. Med. Res.*, vol. 10, no. 1, p. 10, Mar. 2023, doi: 10.1186/s40779-023-00446-y.
[41] S. Natsuka, “Equilibrium Dialysis,” in *Experimental Glycoscience: Glycochemistry*, N. Taniguchi, A. Suzuki, Y. Ito, H. Narimatsu, T. Kawasaki, and S. Hase, Eds., Tokyo: Springer Japan, 2008, pp. 132–133. doi: 10.1007/978-4-431-77924-7_35.
[42] G. C.-H. Leung, S. S.-P. Fung, N. R. B. Dovey, E. L. Raven, and A. J. Hudson, “Precise determination of heme binding affinity in proteins,” *Anal. Biochem.*, vol. 572, pp. 45–51, May 2019, doi: 10.1016/j.ab.2019.02.021.
[43] J. J. Maguire, R. E. Kuc, and A. P. Davenport, “Radioligand binding assays and their analysis,” *Methods Mol. Biol. Clifton NJ*, vol. 897, pp. 31–77, 2012, doi:

10.1007/978-1-61779-909-9_3.
[44] C. S. Schneider *et al.*, “Surface plasmon resonance as a high throughput method to evaluate specific and non-specific binding of nanotherapeutics,” *J. Control. Release Off. J. Control. Release Soc.*, vol. 219, pp. 331–344, Dec. 2015, doi: 10.1016/j.jconrel.2015.09.048.
[45] B. Yan and J. Bunch, “A straightforward method for measuring binding affinities of ligands to proteins of unknown concentration in biological tissues,” *Chem. Sci.*, vol. 16, no. 20, pp. 8673–8681, May 2025, doi: 10.1039/D5SC02460A.
[46] E. Monteagudo-Cascales, M. Cano-Muñoz, R. Genova, J. J. Cabrera, M. A. Matilla, and T. Krell, “Thermal shift assay to identify ligands for bacterial sensor proteins,” *FEMS Microbiol. Rev.*, vol. 49, p. fuaf033, Jul. 2025, doi: 10.1093/femsre/fuaf033.
[47] V. Petrauskas, E. Kazlauskas, M. Gedgaudas, L. Baranauskienė, A. Zubrienė, and D. Matulis, “Thermal shift assay for protein–ligand dissociation constant determination,” *TrAC Trends Anal. Chem.*, vol. 170, p. 117417, Jan. 2024, doi: 10.1016/j.trac.2023.117417.
[48] H. K. Nivatya *et al.*, “Assessing molecular docking tools: understanding drug discovery and design,” *Future J. Pharm. Sci.*, vol. 11, no. 1, p. 111, Aug. 2025, doi: 10.1186/s43094-025-00862-y.
[49] S. Amari, R. Kataoka, T. Ikegami, and N. Hirayama, “HLA-Modeler: Automated Homology Modeling of Human Leukocyte Antigens,” *Int. J. Med. Chem.*, vol. 2013, p. 690513, 2013, doi: 10.1155/2013/690513.
[50] H. Wei and J. A. McCammon, “Structure and dynamics in drug discovery,” *Npj Drug Discov.*, vol. 1, no. 1, p. 1, Nov. 2024, doi: 10.1038/s44386-024-00001-2.
[51] S. Grewal, G. Deswal, A. S. Grewal, and K. Guarve, “Molecular dynamics simulations: Insights into protein and protein ligand interactions,” *Adv. Pharmacol. San Diego Calif*, vol. 103, pp. 139–162, 2025, doi: 10.1016/bs.apha.2025.01.007.
[52] D. Yan, Y. Ma, X. Chen, S. Deng, and Q. Wang, “Molecular dynamics-driven drug discovery,” *Phys. Chem. Chem. Phys.*, vol. 27, no. 24, pp. 12633–12651, Jun. 2025, doi: 10.1039/D5CP00380F.
[53] P. A. Kollman *et al.*, “Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models,” *Acc. Chem. Res.*, vol. 33, no. 12, pp. 889–897, Dec. 2000, doi: 10.1021/ar000033j.
[54] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo, “Recent Developments and Applications of the MMPBSA Method,” *Front. Mol. Biosci.*, vol. 4, p. 87, Jan. 2018, doi: 10.3389/fmolb.2017.00087.
[55] R. Qian, J. Xue, Y. Xu, and J. Huang, “Alchemical Transformations and Beyond: Recent Advances and Real-World Applications of Free Energy Calculations in Drug Discovery,” *J. Chem. Inf. Model.*, vol. 64, no. 19, pp. 7214–7237, Oct. 2024, doi: 10.1021/acs.jcim.4c01024.
[56] G. A. Ross *et al.*, “The maximal and current accuracy of rigorous protein-ligand binding free energy calculations,” *Commun. Chem.*, vol. 6, no. 1, p. 222, Oct. 2023, doi: 10.1038/s42004-023-01019-9.
[57] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, “GraphDTA: predicting drug–target binding affinity with graph neural networks,” *Bioinformatics*,

vol. 37, no. 8, pp. 1140–1147, May 2021, doi: 10.1093/bioinformatics/btaa921.
[58] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
[59] O. Zhang *et al.*, "Graph Neural Networks in Modern AI-Aided Drug Discovery," *Chem. Rev.*, vol. 125, no. 20, pp. 10001–10103, Oct. 2025, doi: 10.1021/acs.chemrev.5c00461.
[60] H. Stärk, O.-E. Ganea, L. Pattanaik, R. Barzilay, and T. Jaakkola, "EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction," Jun. 04, 2022, *arXiv*: arXiv:2202.05146. doi: 10.48550/arXiv.2202.05146.
[61] Y. Jiang *et al.*, "PoseX: AI Defeats Physics Approaches on Protein-Ligand Cross Docking," May 21, 2025, *arXiv*: arXiv:2505.01700. doi: 10.48550/arXiv.2505.01700.
[62] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
[63] A. Rives *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 15, p. e2016239118, Apr. 2021, doi: 10.1073/pnas.2016239118.
[64] H. Lai *et al.*, "Interformer: an interaction-aware model for protein-ligand docking and affinity prediction," *Nat. Commun.*, vol. 15, no. 1, p. 10223, Nov. 2024, doi: 10.1038/s41467-024-54440-6.
[65] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024, doi: 10.1038/s41586-024-07487-w.
[66] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, "DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking," Feb. 11, 2023, *arXiv*: arXiv:2210.01776. doi: 10.48550/arXiv.2210.01776.
[67] M. A. Ketata *et al.*, "DiffDock-PP: Rigid Protein-Protein Docking with Diffusion Models," Apr. 08, 2023, *arXiv*: arXiv:2304.03889. doi: 10.48550/arXiv.2304.03889.
[68] D. Cao *et al.*, "SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction," *Nat. Methods*, vol. 22, no. 2, pp. 310–322, Feb. 2025, doi: 10.1038/s41592-024-02516-y.
[69] W. Lu *et al.*, "DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model," *Nat. Commun.*, vol. 15, no. 1, p. 1071, Feb. 2024, doi: 10.1038/s41467-024-45461-2.
[70] J. Wohlwend *et al.*, "Boltz-1 Democratizing Biomolecular Interaction Modeling," May 06, 2025, *bioRxiv*. doi: 10.1101/2024.11.19.624167.
[71] A. D. Roses, "Pharmacogenetics in drug discovery and development: a translational perspective," *Nat. Rev. Drug Discov.*, vol. 7, no. 10, pp. 807–817, Oct. 2008, doi: 10.1038/nrd2593.
[72] G. A. Van Norman, "Phase II Trials in Drug Development and Adaptive Trial Design," *JACC Basic Transl. Sci.*, vol. 4, no. 3, pp. 428–437, Jun. 2019, doi: 10.1016/j.jacbts.2019.02.005.
[73] J. Arrowsmith, "Trial watch: Phase II failures: 2008-2010," *Nat. Rev. Drug Discov.*,

vol. 10, no. 5, pp. 328–329, May 2011, doi: 10.1038/nrd3439.
[74] M. Thafar, A. B. Raies, S. Albaradei, M. Essack, and V. B. Bajic, “Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities,” *Front. Chem.*, vol. 7, p. 782, 2019, doi: 10.3389/fchem.2019.00782.
[75] J. Li, A. Fu, and L. Zhang, “An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking,” *Interdiscip. Sci. Comput. Life Sci.*, vol. 11, no. 2, pp. 320–328, Jun. 2019, doi: 10.1007/s12539-019-00327-w.
[76] T. Pahikkala *et al.*, “Toward more realistic drug-target interaction predictions,” *Brief. Bioinform.*, vol. 16, no. 2, pp. 325–337, Mar. 2015, doi: 10.1093/bib/bbu010.
[77] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester, “SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines,” *J. Cheminformatics*, vol. 9, no. 1, p. 24, Apr. 2017, doi: 10.1186/s13321-017-0209-z.
[78] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, and A. Masoudi-Nejad, “DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks,” *Bioinforma. Oxf. Engl.*, vol. 36, no. 17, pp. 4633–4642, Nov. 2020, doi: 10.1093/bioinformatics/btaa544.
[79] H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: deep drug-target binding affinity prediction,” *Bioinforma. Oxf. Engl.*, vol. 34, no. 17, pp. i821–i829, Sep. 2018, doi: 10.1093/bioinformatics/bty593.
[80] L. Zhao, J. Wang, L. Pang, Y. Liu, and J. Zhang, “GANsDTA: Predicting Drug-Target Binding Affinity Using GANs,” *Front. Genet.*, vol. 10, p. 1243, 2019, doi: 10.3389/fgene.2019.01243.
[81] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, Mar. 1981, doi: 10.1016/0022-2836(81)90087-5.
[82] M. I. Davis *et al.*, “Comprehensive analysis of kinase inhibitor selectivity,” *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, Oct. 2011, doi: 10.1038/nbt.1990.
[83] J. Tang *et al.*, “Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis,” *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 735–743, Mar. 2014, doi: 10.1021/ci400709d.
[84] UniProt Consortium, “UniProt: a worldwide hub of protein knowledge,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.
[85] M. Gönen and G. Heller, “Concordance probability and discriminatory power in proportional hazards regression,” *Biometrika*, vol. 92, no. 4, pp. 965–970, Dec. 2005, doi: 10.1093/biomet/92.4.965.
[86] K. Roy, P. Chakraborty, I. Mitra, P. K. Ojha, S. Kar, and R. N. Das, “Some case studies on application of ‘r(m)2’ metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data,” *J. Comput. Chem.*, vol. 34, no. 12, pp. 1071–1082, May 2013, doi: 10.1002/jcc.23231.
[87] K. Liu *et al.*, “Chemi-Net: A Molecular Graph Convolutional Network for Accurate Drug Property Prediction,” *Int. J. Mol. Sci.*, vol. 20, no. 14, p. 3389, Jul. 2019, doi: 10.3390/ijms20143389.

[88] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/ci100050t.
[89] A. Gobbi and D. Poppinger, “Genetic optimization of combinatorial libraries,” *Biotechnol. Bioeng.*, vol. 61, no. 1, pp. 47–54, 1998, doi: 10.1002/(sici)1097-0290(199824)61:1<47::aid-bit9>3.0.co;2-z.
[90] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, “ColabFold: making protein folding accessible to all,” *Nat. Methods*, vol. 19, no. 6, pp. 679–682, Jun. 2022, doi: 10.1038/s41592-022-01488-1.
[91] P. J. A. Cock *et al.*, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.
[92] S. O. Yesylevskyy, “Pteros: fast and easy to use open-source C++ library for molecular analysis,” *J. Comput. Chem.*, vol. 33, no. 19, pp. 1632–1636, Jul. 2012, doi: 10.1002/jcc.22989.
[93] S. O. Yesylevskyy, “Pteros 2.0: Evolution of the fast parallel molecular analysis library for C++ and python,” *J. Comput. Chem.*, vol. 36, no. 19, pp. 1480–1488, Jul. 2015, doi: 10.1002/jcc.23943.
[94] M. Wójcikowski, P. Zielenkiewicz, and P. Siedlecki, “Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field,” *J. Cheminformatics*, vol. 7, no. 1, p. 26, Jun. 2015, doi: 10.1186/s13321-015-0078-2.
[95] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, “Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks,” *Mol. Model. Annu.*, vol. 7, no. 9, pp. 360–369, Sep. 2001, doi: 10.1007/s008940100038.
[96] D. W. Mount, “Using BLOSUM in Sequence Alignments,” *Cold Spring Harb. Protoc.*, vol. 2008, no. 6, p. pdb.top39, Jun. 2008, doi: 10.1101/pdb.top39.
[97] J. Chen, S. Zheng, H. Zhao, and Y. Yang, “Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map,” *J. Cheminformatics*, vol. 13, no. 1, p. 7, Feb. 2021, doi: 10.1186/s13321-021-00488-1.
[98] S. Brody, U. Alon, and E. Yahav, “How Attentive are Graph Attention Networks?,” Jan. 31, 2022, *arXiv*: arXiv:2105.14491. doi: 10.48550/arXiv.2105.14491.
[99] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” Feb. 22, 2017, *arXiv*: arXiv:1609.02907. doi: 10.48550/arXiv.1609.02907.
[100] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How Powerful are Graph Neural Networks?,” Feb. 22, 2019, *arXiv*: arXiv:1810.00826. doi: 10.48550/arXiv.1810.00826.
[101] W. Hu *et al.*, “Strategies for Pre-training Graph Neural Networks,” Feb. 18, 2020, *arXiv*: arXiv:1905.12265. doi: 10.48550/arXiv.1905.12265.
[102] D. Duvenaud *et al.*, “Convolutional Networks on Graphs for Learning Molecular Fingerprints,” Nov. 03, 2015, *arXiv*: arXiv:1509.09292. doi: 10.48550/arXiv.1509.09292.
[103] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning

Library," Dec. 03, 2019, *arXiv*: arXiv:1912.01703. doi: 10.48550/arXiv.1912.01703.
[104] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," Apr. 25, 2019, *arXiv*: arXiv:1903.02428. doi: 10.48550/arXiv.1903.02428.
[105] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 25, 2019, *arXiv*: arXiv:1907.10902. doi: 10.48550/arXiv.1907.10902.
[106] M. Blum *et al.*, "The InterPro protein families and domains database: 20 years on," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D344–D354, Jan. 2021, doi: 10.1093/nar/gkaa977.
[107] C. Bissantz, B. Kuhn, and M. Stahl, "A Medicinal Chemist's Guide to Molecular Interactions," *J. Med. Chem.*, vol. 53, no. 14, pp. 5061–5084, Jul. 2010, doi: 10.1021/jm100112j.
[108] J. Schiebel *et al.*, "Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes," *Nat. Commun.*, vol. 9, no. 1, p. 3559, Sep. 2018, doi: 10.1038/s41467-018-05769-2.
[109] D. Chen, N. Oezguen, P. Urvil, C. Ferguson, S. M. Dann, and T. C. Savidge, "Regulation of protein-ligand binding affinity by hydrogen bond pairing," *Sci. Adv.*, vol. 2, no. 3, p. e1501240, Mar. 2016, doi: 10.1126/sciadv.1501240.
[110] X. Li, Y. Li, T. Cheng, Z. Liu, and R. Wang, "Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes," *J. Comput. Chem.*, vol. 31, no. 11, pp. 2109–2125, 2010, doi: 10.1002/jcc.21498.
[111] B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, and R. Dror, "Learning from Protein Structure with Geometric Vector Perceptrons," May 16, 2021, *arXiv*: arXiv:2009.01411. doi: 10.48550/arXiv.2009.01411.
[112] B. Jing, S. Eismann, P. N. Soni, and R. O. Dror, "Equivariant Graph Neural Networks for 3D Macromolecular Structure," Jul. 13, 2021, *arXiv*: arXiv:2106.03843. doi: 10.48550/arXiv.2106.03843.
[113] W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li, and S. Zheng, "TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction," Jun. 06, 2022, *bioRxiv*. doi: 10.1101/2022.06.06.495043.
[114] Z. Liu *et al.*, "PDB-wide collection of binding data: current status of the PDBbind database," *Bioinformatics*, vol. 31, no. 3, pp. 405–412, Feb. 2015, doi: 10.1093/bioinformatics/btu626.
[115] Ž. H. Petrovski, B. Hribar-Lee, and Z. Bosnić, "CAT-Site: Predicting Protein Binding Sites Using a Convolutional Neural Network," *Pharmaceutics*, vol. 15, no. 1, p. 119, Jan. 2023, doi: 10.3390/pharmaceutics15010119.
[116] J. Kandel, H. Tayara, and K. T. Chong, "PUResNet: prediction of protein-ligand binding sites using deep residual neural network," *J. Cheminformatics*, vol. 13, no. 1, p. 65, Sep. 2021, doi: 10.1186/s13321-021-00547-7.
[117] J.-L. Reymond, R. van Deursen, L. C. Blum, and L. Ruddigkeit, "Chemical space as a source for new drugs," *MedChemComm*, vol. 1, no. 1, pp. 30–38, Jul. 2010, doi: 10.1039/C0MD00020E.
[118] P. W. Snyder *et al.*, "Mechanism of the hydrophobic effect in the biomolecular

recognition of arylsulfonamides by carbonic anhydrase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 44, pp. 17889–17894, Nov. 2011, doi: 10.1073/pnas.1114107108.
[119] G. Bitencourt-Ferreira, M. Veit-Acosta, and W. F. de Azevedo, "Van der Waals Potential in Protein Complexes," *Methods Mol. Biol. Clifton NJ*, vol. 2053, pp. 79–91, 2019, doi: 10.1007/978-1-4939-9752-7_6.
[120] A. Madushanka, R. T. Moura, N. Verma, and E. Kraka, "Quantum Mechanical Assessment of Protein–Ligand Hydrogen Bond Strength Patterns: Insights from Semiempirical Tight-Binding and Local Vibrational Mode Theory," *Int. J. Mol. Sci.*, vol. 24, no. 7, p. 6311, Jan. 2023, doi: 10.3390/ijms24076311.
[121] P. R. Connelly *et al.*, "Enthalpy of hydrogen bond formation in a protein-ligand binding reaction.," *Proc. Natl. Acad. Sci.*, vol. 91, no. 5, pp. 1964–1968, Mar. 1994, doi: 10.1073/pnas.91.5.1964.
[122] M. L. Verteramo *et al.*, "Interplay of halogen bonding and solvation in protein–ligand binding," *iScience*, vol. 27, no. 4, p. 109636, Mar. 2024, doi: 10.1016/j.isci.2024.109636.
[123] J. Ren *et al.*, "Thermodynamic and Structural Characterization of Halogen Bonding in Protein–Ligand Interactions: A Case Study of PDE5 and Its Inhibitors," *J. Med. Chem.*, vol. 57, no. 8, pp. 3588–3593, Apr. 2014, doi: 10.1021/jm5002315.
[124] N. Zhang *et al.*, "Entropy Drives the Formation of Salt Bridges in the Protein GB3," *Angew. Chem.*, vol. 129, no. 26, pp. 7709–7712, 2017, doi: 10.1002/ange.201702968.
[125] D. Paik, H. Lee, H. Kim, and J.-M. Choi, "Thermodynamics of π–π Interactions of Benzene and Phenol in Water," *Int. J. Mol. Sci.*, vol. 23, no. 17, p. 9811, Jan. 2022, doi: 10.3390/ijms23179811.
[126] J. P. Gallivan and D. A. Dougherty, "Cation-π interactions in structural biology," *Proc. Natl. Acad. Sci.*, vol. 96, no. 17, pp. 9459–9464, Aug. 1999, doi: 10.1073/pnas.96.17.9459.
[127] L. M. Breberina, M. V. Zlatović, S. Đ. Stojanović, and M. R. Nikolić, "Amide–π Interactions in the Structural Stability of Proteins: Role in the Oligomeric Phycocyanins," *Computation*, vol. 12, no. 9, p. 172, Sep. 2024, doi: 10.3390/computation12090172.
[128] B. Kuhn, E. Gilberg, R. Taylor, J. Cole, and O. Korb, "How Significant Are Unusual Protein–Ligand Interactions? Insights from Database Mining," *J. Med. Chem.*, vol. 62, no. 22, pp. 10441–10455, Nov. 2019, doi: 10.1021/acs.jmedchem.9b01545.
[129] R. F. de Freitas and M. Schapira, "A systematic analysis of atomic protein–ligand interactions in the PDB," *MedChemComm*, vol. 8, no. 10, pp. 1970–1981, Oct. 2017, doi: 10.1039/C7MD00381A.
[130] R. Wilcken, M. O. Zimmermann, A. Lange, A. C. Joerger, and F. M. Boeckler, "Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology," *J. Med. Chem.*, vol. 56, no. 4, pp. 1363–1388, Feb. 2013, doi: 10.1021/jm3012068.
[131] M. F. Adasme *et al.*, "PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W530–W534, Jul. 2021, doi: 10.1093/nar/gkab294.

[132] H. C. Jubb, A. P. Higueruelo, B. Ochoa-Montaño, W. R. Pitt, D. B. Ascher, and T. L. Blundell, “Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures,” *J. Mol. Biol.*, vol. 429, no. 3, pp. 365–371, Feb. 2017, doi: 10.1016/j.jmb.2016.12.004.
[133] M. Z. Tien, D. K. Sydykova, A. G. Meyer, and C. O. Wilke, “PeptideBuilder: A simple Python library to generate model peptides,” *PeerJ*, vol. 1, p. e80, May 2013, doi: 10.7717/peerj.80.
[134] J. J. Irwin *et al.*, “ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery,” *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 6065–6073, Dec. 2020, doi: 10.1021/acs.jcim.0c00675.
[135] A. P. Bento *et al.*, “An open source chemical structure curation pipeline using RDKit,” *J. Cheminformatics*, vol. 12, no. 1, p. 51, Sep. 2020, doi: 10.1186/s13321-020-00456-1.
[136] R. Meli and P. C. Biggin, “spyrmsd: symmetry-corrected RMSD calculations in Python,” *J. Cheminformatics*, vol. 12, no. 1, p. 49, Aug. 2020, doi: 10.1186/s13321-020-00455-2.
[137] M. Geiger and T. Smidt, “e3nn: Euclidean Neural Networks,” Jul. 18, 2022, *arXiv*: arXiv:2207.09453. doi: 10.48550/arXiv.2207.09453.
[138] N. Thomas *et al.*, “Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds,” 2018, doi: 10.48550/ARXIV.1802.08219.
[139] C. Yang, E. A. Chen, and Y. Zhang, “Protein–Ligand Docking in the Machine-Learning Era,” *Molecules*, vol. 27, no. 14, p. 4568, Jan. 2022, doi: 10.3390/molecules27144568.
[140] R. Krishna *et al.*, “Generalized biomolecular modeling and design with RoseTTAFold All-Atom,” *Science*, vol. 384, no. 6693, p. eadl2528, Mar. 2024, doi: 10.1126/science.adl2528.
[141] Z. Qiao *et al.*, “NeuralPLexer3: Accurate Biomolecular Complex Structure Prediction with Flow Models,” Dec. 18, 2024, *arXiv*: arXiv:2412.10743. doi: 10.48550/arXiv.2412.10743.
[142] C. Discovery *et al.*, “Chai-1: Decoding the molecular interactions of life,” Oct. 15, 2024, *bioRxiv*. doi: 10.1101/2024.10.10.615955.
[143] B. A. A. Team *et al.*, “Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction,” Jan. 11, 2025, *bioRxiv*. doi: 10.1101/2025.01.08.631967.
[144] M. Leemann, A. Sagasta, J. Eberhardt, T. Schwede, X. Robin, and J. Durairaj, “Automated benchmarking of combined protein structure and ligand conformation prediction,” *Proteins Struct. Funct. Bioinforma.*, vol. 91, no. 12, pp. 1912–1924, 2023, doi: 10.1002/prot.26605.
[145] P. Škrinjar, J. Eberhardt, J. Durairaj, and T. Schwede, “Have protein-ligand co-folding methods moved beyond memorisation?,” Feb. 10, 2025, *bioRxiv*. doi: 10.1101/2025.02.03.636309.
[146] J. Durairaj *et al.*, “PLINDER: The protein-ligand interactions dataset and evaluation resource,” Jul. 19, 2024, *bioRxiv*. doi: 10.1101/2024.07.17.603955.
[147] G. M. Morris *et al.*, “AutoDock4 and AutoDockTools4: Automated docking with

selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, 2009, doi: 10.1002/jcc.21256.
[148] D. S. Goodsell, M. F. Sanner, A. J. Olson, and S. Forli, "The AutoDock suite at 30," *Protein Sci.*, vol. 30, no. 1, pp. 31–43, 2021, doi: 10.1002/pro.3934.
[149] D. Santos-Martins, L. Solis-Vasquez, A. F. Tillack, M. F. Sanner, A. Koch, and S. Forli, "Accelerating AutoDock4 with GPUs and Gradient-Based Local Search," *J. Chem. Theory Comput.*, vol. 17, no. 2, pp. 1060–1073, Feb. 2021, doi: 10.1021/acs.jctc.0c01006.
[150] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, Jan. 2010, doi: 10.1002/jcc.21334.
[151] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, "AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings," *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 3891–3898, Aug. 2021, doi: 10.1021/acs.jcim.1c00203.
[152] R. A. Friesner *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004, doi: 10.1021/jm0306430.
[153] T. A. Halgren *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening," *J. Med. Chem.*, vol. 47, no. 7, pp. 1750–1759, Mar. 2004, doi: 10.1021/jm030644s.
[154] R. A. Friesner *et al.*, "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes," *J. Med. Chem.*, vol. 49, no. 21, pp. 6177–6196, Oct. 2006, doi: 10.1021/jm051256o.
[155] Y. Yang *et al.*, "Efficient Exploration of Chemical Space with Docking and Deep Learning," *J. Chem. Theory Comput.*, vol. 17, no. 11, pp. 7106–7119, Nov. 2021, doi: 10.1021/acs.jctc.1c00810.
[156] M. Buttenschoen, G. M. Morris, and C. M. Deane, "PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences," *Chem. Sci.*, vol. 15, no. 9, pp. 3130–3139, Feb. 2024, doi: 10.1039/D3SC04185A.
[157] A. N. Jain, A. E. Cleves, and W. P. Walters, "Deep-Learning Based Docking Methods: Fair Comparisons to Conventional Docking Workflows," Dec. 09, 2024, *arXiv*: arXiv:2412.02889. doi: 10.48550/arXiv.2412.02889.
[158] S. S. Pawar and S. H. Rohane, "Review on Discovery Studio: An important Tool for Molecular Docking," *Asian J. Res. Chem.*, vol. 14, no. 1, pp. 1–3, 2021, doi: 10.5958/0974-4150.2021.00014.6.
[159] S. Vilar, G. Cozza, and S. Moro, "Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery," *Curr. Top. Med. Chem.*, vol. 8, no. 18, pp. 1555–1572, 2008, doi: 10.2174/156802608786786624.
[160] A. T. McNutt *et al.*, "GNINA 1.0: molecular docking with deep learning," *J. Cheminformatics*, vol. 13, no. 1, p. 43, Jun. 2021, doi: 10.1186/s13321-021-00522-2.
[161] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller, and A. Anandkumar, "State-specific protein–ligand complex structure prediction with a multiscale deep generative model," *Nat. Mach. Intell.*, vol. 6, no. 2, pp. 195–208, Feb. 2024, doi:

10.1038/s42256-024-00792-z.
[162] O. Méndez-Lucio, M. Ahmad, E. A. del Rio-Chanona, and J. K. Wegner, “A geometric deep learning approach to predict binding conformations of bioactive molecules,” *Nat. Mach. Intell.*, vol. 3, no. 12, pp. 1033–1039, Dec. 2021, doi: 10.1038/s42256-021-00409-9.
[163] G. Zhou *et al.*, “Uni-Mol: A Universal 3D Molecular Representation Learning Framework,” Mar. 07, 2023, *ChemRxiv*. doi: 10.26434/chemrxiv-2022-jjm0j-v4.
[164] G. Corso, A. Deng, B. Fry, N. Polizzi, R. Barzilay, and T. Jaakkola, “Deep Confident Steps to New Pockets: Strategies for Docking Generalization,” Feb. 28, 2024, *arXiv*: arXiv:2402.18396. doi: 10.48550/arXiv.2402.18396.
[165] Q. Pei *et al.*, “FABind: Fast and Accurate Protein-Ligand Binding,” Jan. 09, 2024, *arXiv*: arXiv:2310.06763. doi: 10.48550/arXiv.2310.06763.
[166] M. Plainer *et al.*, “DiffDock-Pocket: Diffusion for Pocket-Level Docking with Sidechain Flexibility,” Oct. 2023, Accessed: Oct. 14, 2025. [Online]. Available: https://openreview.net/forum?id=1IaoWBqB6K
[167] S. Yesylevskyy, “MolAR: Memory-Safe Library for Analysis of MD Simulations Written in Rust,” *J. Comput. Chem.*, vol. 46, no. 1, p. e27536, 2025, doi: 10.1002/jcc.27536.
[168] X. Robin, G. Studer, J. Durairaj, J. Eberhardt, T. Schwede, and W. P. Walters, “Assessment of protein–ligand complexes in CASP15,” *Proteins Struct. Funct. Bioinforma.*, vol. 91, no. 12, pp. 1811–1821, Dec. 2023, doi: 10.1002/prot.26601.
[169] M. Biasini *et al.*, “OpenStructure: an integrated software framework for computational structural biology,” *Acta Crystallogr. D Biol. Crystallogr.*, vol. 69, no. 5, pp. 701–709, May 2013, doi: 10.1107/s0907444913007051.
[170] M. Bertoni, F. Kiefer, M. Biasini, L. Bordoli, and T. Schwede, “Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology,” *Sci. Rep.*, vol. 7, no. 1, Sep. 2017, doi: 10.1038/s41598-017-09654-8.
[171] P. Karen, P. McArdle, and J. Takats, “Comprehensive definition of oxidation state (IUPAC Recommendations 2016),” *Pure Appl. Chem.*, vol. 88, no. 8, pp. 831–839, Aug. 2016, doi: 10.1515/pac-2015-1204.
[172] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, “Open Babel: An open chemical toolbox,” *J. Cheminformatics*, vol. 3, no. 1, p. 33, Oct. 2011, doi: 10.1186/1758-2946-3-33.
[173] J. Ingraham, V. Garg, R. Barzilay, and T. Jaakkola, “Generative Models for Graph-Based Protein Design,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: May 15, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html
[174] M. R. Masters, A. H. Mahmoud, Y. Wei, and M. A. Lill, “Deep Learning Model for Efficient Protein–Ligand Docking with Implicit Side-Chain Flexibility,” *J. Chem. Inf. Model.*, vol. 63, no. 6, pp. 1695–1707, Mar. 2023, doi: 10.1021/acs.jcim.2c01436.
[175] Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad, and A. P. Johnson, “eHiTS: A new fast,

exhaustive flexible ligand docking system," *J. Mol. Graph. Model.*, vol. 26, no. 1, pp. 198–212, Jul. 2007, doi: 10.1016/j.jmgm.2006.06.002.
[176] R. Storn and K. Price, "Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces," *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, Dec. 1997, doi: 10.1023/A:1008202821328.
[177] J. Nocedal and S. J. Wright, Eds., "Sequential Quadratic Programming," in *Numerical Optimization*, New York, NY: Springer, 1999, pp. 526–573. doi: 10.1007/0-387-22742-3_18.
[178] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *J. Am. Chem. Soc.*, vol. 114, no. 25, pp. 10024–10035, Dec. 1992, doi: 10.1021/ja00051a040.
[179] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Deliv. Rev.*, vol. 46, no. 1–3, pp. 3–26, Mar. 2001, doi: 10.1016/S0169-409X(00)00129-0.
[180] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor, "Improved protein-ligand docking using GOLD," *Proteins*, vol. 52, no. 4, pp. 609–623, Sep. 2003, doi: 10.1002/prot.10465.
[181] W. J. Allen *et al.*, "DOCK 6: Impact of new features and current docking performance," *J. Comput. Chem.*, vol. 36, no. 15, pp. 1132–1156, 2015, doi: 10.1002/jcc.23905.
[182] A. T. McNutt *et al.*, "GNINA 1.0: molecular docking with deep learning," *J. Cheminformatics*, vol. 13, no. 1, p. 43, Jun. 2021, doi: 10.1186/s13321-021-00522-2.
[183] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, "Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1893–1904, Aug. 2013, doi: 10.1021/ci300604z.
[184] A. Alhossary, S. D. Handoko, Y. Mu, and C.-K. Kwoh, "Fast, accurate, and reliable molecular docking with QuickVina 2," *Bioinformatics*, vol. 31, no. 13, pp. 2214–2216, Jul. 2015, doi: 10.1093/bioinformatics/btv082.
[185] N. M. Hassan, A. A. Alhossary, Y. Mu, and C.-K. Kwoh, "Protein-Ligand Blind Docking Using QuickVina-W With Inter-Process Spatio-Temporal Integration," *Sci. Rep.*, vol. 7, no. 1, p. 15451, Nov. 2017, doi: 10.1038/s41598-017-15571-7.

ДОДАТКИ

Додаток А. Підготовка даних PLINDER

Кожна система (комплекс) PLINDER [146] може містити до 5 власних молекул лігандів, включаючи кофактори. Усі іони та експериментальні артефакти, а також молекули води, що утворюють водні містки з будь-якою небілковою молекулою, зберігаються в комплексі.

На початку попередньої обробки набору даних було видалено всі атоми водню з усіх молекул системи PLINDER. Тому всі значення, пов’язані з кількістю атомів або міжатомними відстанями, не враховують атомів водню. Пари білок-ліганд (PLP) низької якості відфільтровувалися за такими критеріями:

- Базові фільтри:
    - Кількість амінокислотних залишків білка менша за 5.
    - Кількість атомів ліганду менша за 5 або більша за 100.
    - Кількість обертальних ковалентних зв’язків у ліганді перевищує 50.
    - Положення половини чи більше атомів ліганду відсутні у структурі.
- Фільтри відстані:
    - Мінімальна відстань між атомами ліганду та іншими молекулами/атомами в комплексі менша за 1 Å. Ця відстань є занадто малою навіть для ковалентних зв’язків. Потенційні артефакти структур з низькою роздільною здатністю або помилки представлення даних.
    - Відношення кількості стеричних зіткнень ліганду з іншими молекулами/атомами в комплексі до кількості атомів ліганду перевищує 0.1. Стеричне зіткнення визначалося як відстань між атомами, менша за 70% від суми їхніх відповідних ван-дер-ваальсових радіусів.
    - Середнє значення або стандартне відхилення мінімальних відстаней (Å) від кожного атома ліганду до інших молекул/атомів у комплексі становить більше 4.65 або 1.7 відповідно. Цей фільтр допомагає уникнути комплексів, у яких значна частина ліганду знаходиться занадто

далеко від сайту зв’язування. Порогові значення коригувалися емпірично шляхом візуального огляду комплексів.

- Фільтри взаємодії:
    - Кількість нековалентних взаємодій білок-ліганд менша за 3. Взаємодії попередньо обчислені в метаданих PLINDER за допомогою методу PLIP [131].
    - Відношення кількості нековалентних взаємодій білок-ліганд до кількості атомів ліганду менша за 0.1.

Частка PLP з PLINDER, що пройшли фільтри якості, підсумована в таблиці:

Набір даних	**Базові фільтри**	**Фільтри відстані**	**Фільтри взаємодії**
Тренувальний	0.99	0.82	0.75
Валідаційний	0.99	0.89	0.95
Тестовий	1.00	0.94	0.98

Додаток Б. Кластеризація даних

Для розв’язання проблеми зміщення тренувальних даних у бік надмірно представлених молекулярних структур в PDB, PLP кластеризували відповідно до:

- Структури ліганду. Було використано фінгерпринти Моргана з радіусом 2 для призначення кожного унікального ліганду до кластера з порогом відстані 0.5 за коефіцієнтом Танімото.
- Схожість послідовності білкового сайту зв’язування між системами. Кластери отримані за допомогою алгоритму виявлення спільнот (community detection) з порогом подібності 0.7. Кластери були попередньо обчислені в метаданих PLINDER та доступні під назвою pocket_fident_qcov__70__community.
- Схожість взаємодії білок-ліганд між системами. Кластери отримані за допомогою алгоритму виявлення спільнот з порогом подібності 0.7. Кластери були попередньо обчислені в метаданих PLINDER та доступні під назвою pli_qcov__70__community.

Підходи до кластеризації та порогові значення обиралися емпірично на основі візуального огляду комплексів, згрупованих у кожному кластері. Перевага надавалася методам, які створювали велику кількість репрезентативних кластерів, мінімізуючи при цьому кількість одноелементних кластерів (кластерів, що містять лише один комплекс). Фінальні кластери PLP визначалися як об’єднання трьох раніше описаних типів кластерів.

Додаток В. Розширення даних на основі сайту зв’язування білка

У кожній ітерації навчання сайт білка визначався випадковим чином для PLP як усі амінокислотні залишки/атоми білка:

- у межах порогової відстані від атомів ліганду;
- всередині сфери/куба з центром у центроїді ліганду;
- всередині сфероїда/кубоїда, що відтворює форму ліганду.

Порогові відстані та розміри вибиралися випадковим чином, причому мінімальні та максимальні значення обиралися так, щоб відображати розподіл кількості атомів сайту, отриманий шляхом включення амінокислотних залишків у межах від 5 Å (мінімум) до 8 Å (максимум) від експериментально визначеного положення ліганду. Усі небілкові молекули, що потрапляли в сайт, зберігалися.

Основна ідея навчання на різних представленнях сайтів зв’язування полягає в тому, щоб представити моделі можливі варіації сайтів зв’язування, які можуть використовуватися як вхідні дані кінцевим користувачем. На практиці, сайт зазвичай визначається як набір залишків амінокислот, про які відомо/передбачено, що вони знаходяться в сайті зв’язування, та/або залишків/атомів навколо іншого ліганду, положення зв'язування якого відоме. Відповідно, існує значна варіативність у потенційних представленнях сайтів зв’язування, що враховується під час навчання шляхом розширення представлень цих сайтів.

Додаток Г. Оцінка обчислювальної ефективності

Для оцінювання методів докінгу на основі CPU (AutoDock-CPU, Vina та Glide) була використана робоча станція з 48-ядерним процесором AMD EPYC 9454P. Обчислення були розпаралелені з використанням модуля Python multiprocessing. Обчислювальні витрати методів, залежних від графічного процесора (GPU), оцінювалися на робочій станції з 16-ядерним процесором AMD Ryzen 9 7950X3D та відеокартою NVIDIA GeForce RTX 3060 на платформі CUDA 12.4.

Glide є найшвидшим методом на основі CPU, працюючи приблизно в 3 рази швидше, ніж AutoDock-CPU, і в 15 разів швидше, ніж Vina. Хоча Vina за своєю природою є багатопотоковою, що дозволяє скоротити час виконання для налаштувань високої вичерпності (exhaustiveness) за наявності достатньої кількості ядер CPU, її було запущено в однопроцесорному режимі, а розпаралелювання оброблялося зовнішніми засобами. Такий підхід забезпечив узгоджений розподіл ресурсів для всіх методів на основі CPU.

У режимі віртуального скринінгу ArtiDock працює більш ніж у 2 рази швидше, ніж інший метод, залежний від GPU, - AutoDock-GPU. ArtiDock може зменшити загальний час виконання шляхом одночасного передбачення матриці відстаней на основі GPU та післяобробки на основі CPU для вже передбачених матриць. Післяобробка ArtiDock, залежна від CPU, також може бути ефективно розпаралелена і не додавати значного часу до загального часу виконання на сучасних багатоядерних робочих станціях.

Для інструментів класичного докінгу обчислювальні витрати відображають оптимізовані налаштування параметрів, спрямовані на досягнення точності, близької до пікової, з мінімальним часом виконання. На основі середнього часу виконання на один тестовий ліганд PLINDER та погодинної вартості використаних робочих станцій, за приблизними оцінками, докінг одного мільйона малих молекул коштуватиме приблизно $5 при використанні ArtiDock та $12 при використанні AutoDock-GPU. Далі йдуть Glide ($15), AutoDock-CPU ($47) та Vina ($248).

Необхідно зазначити, що фактичні витрати можуть змінюватися залежно від багатьох факторів, таких як конфігурація обладнання, тарифний план провайдера, розмір сайту зв’язування білка, а також розмір і складність лігандів.

У наборі даних PoseX [61] наведений середній час виконання на зразок перевищує 10 с. для конкурентних інструментів докінгу на основі машинного навчання і досягає хвилин для підходів ко-фолдингу. Протоколи запуску підходів в PoseX вказують на те, що цей час може включати підготовку даних та ініціалізацію моделі, які слід виключати при оцінюванні пропускної здатності віртуального скринінгу. Порівняння додатково ускладнюються оцінюваннями, проведеними на різнорідному обладнанні. За порівнянних умов, ArtiDock передбачатиме одне положення ліганду за 0.2-2 с, залежно від робочої станції з GPU, що використовується в PoseX. Навіть якщо включити ініціалізацію моделі та підготовку даних, середній час виконання ArtiDock повинен залишатися нижче 5 с.

Додаток Д. Додаткові рисунки

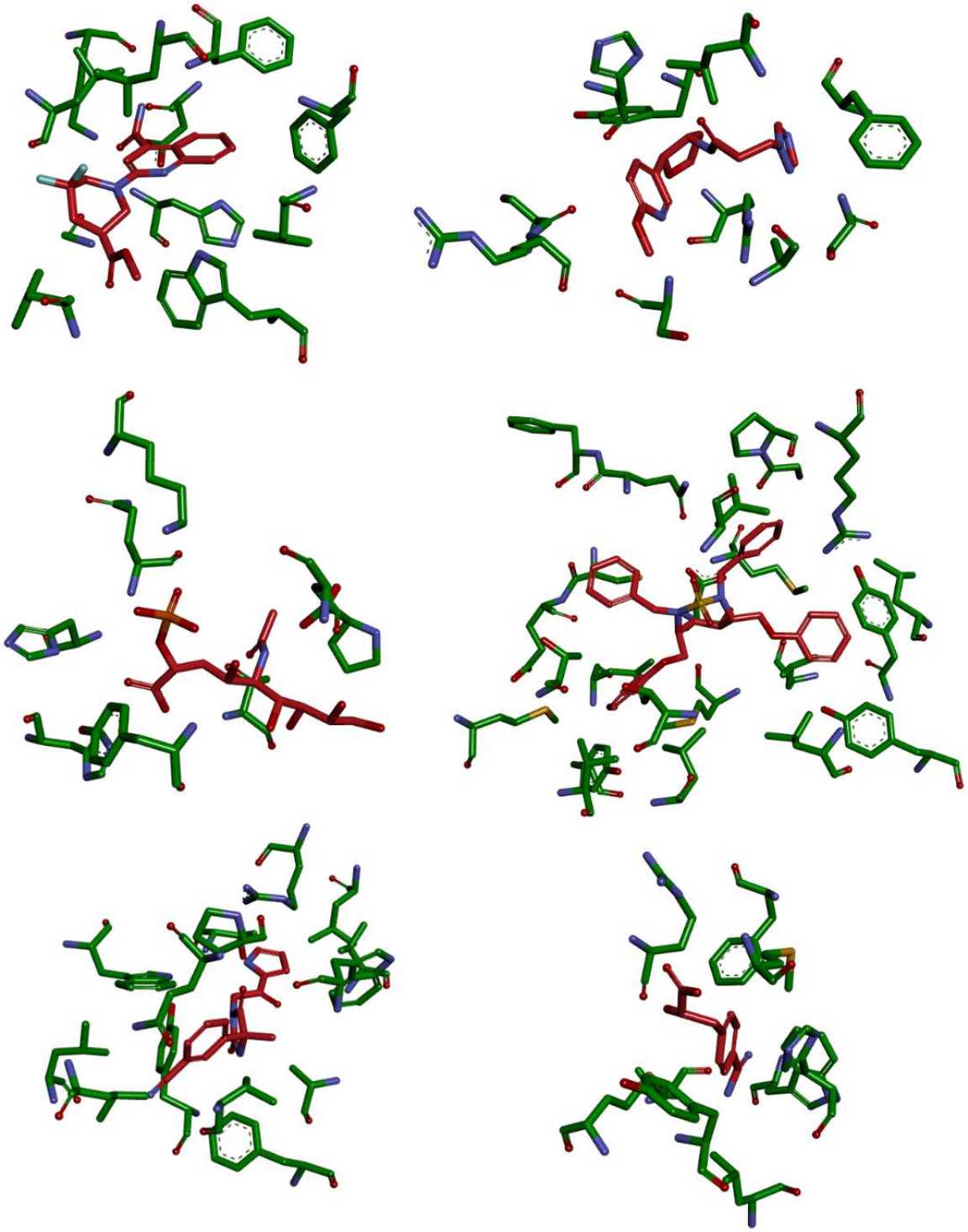

Рисунок Д1. Приклади випадково обраних штучних комплексів білок-ліганд, що складаються з сайту зв’язування білка (зелена) та довільного конформера малої органічної молекули (червоний).

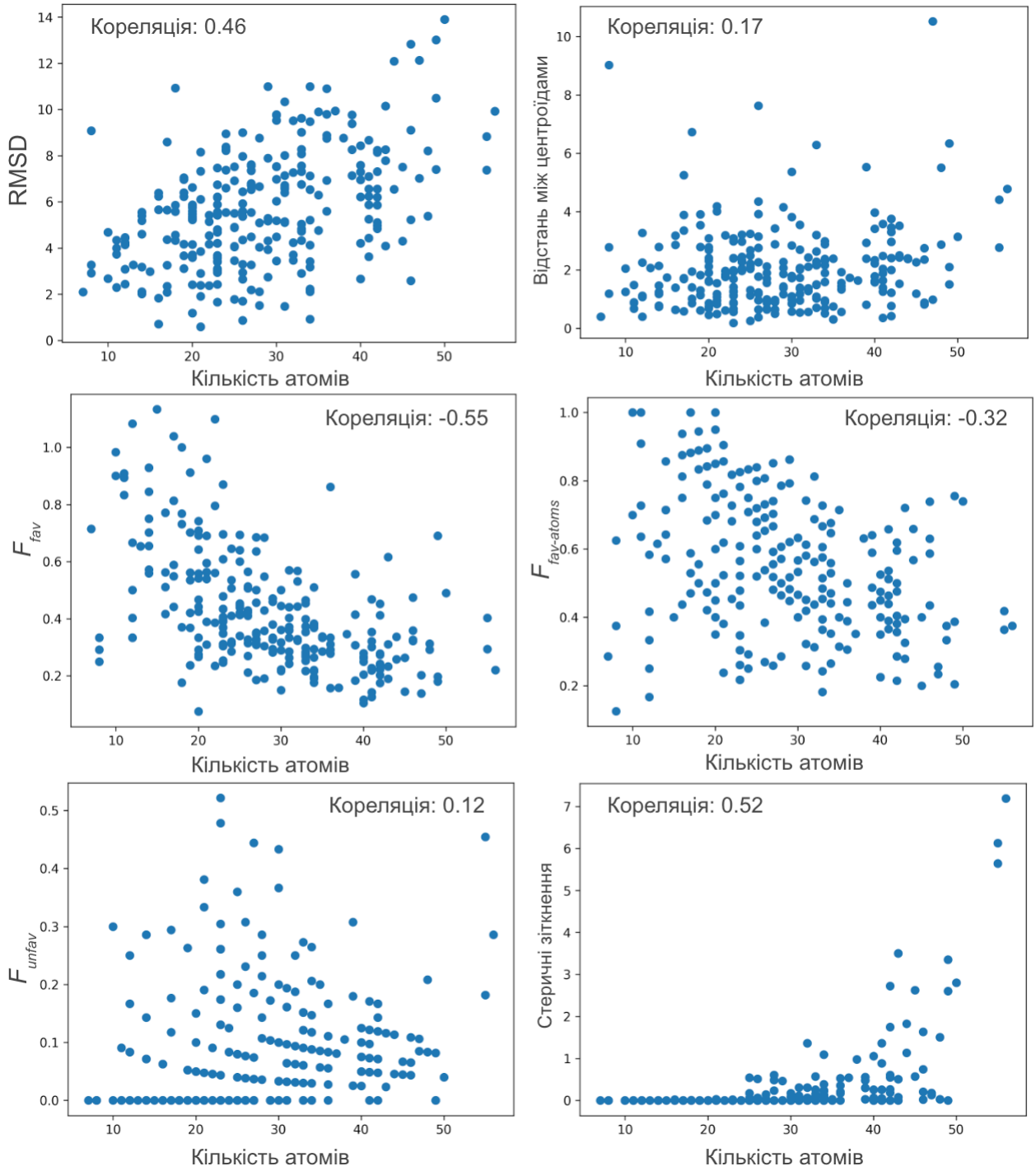

Рисунок Д2. Коефіцієнт кореляції Пірсона між кількістю атомів ліганду (не включаючи гідрогени) та критеріями оцінювання якості передбачень для топ 1 згенерованого положення ліганду у тестових комплексах PDBbind (N=259). Модель генерувала 40 положень ліганду. Топ 1 положення обиралося на основі функції оцінки S.

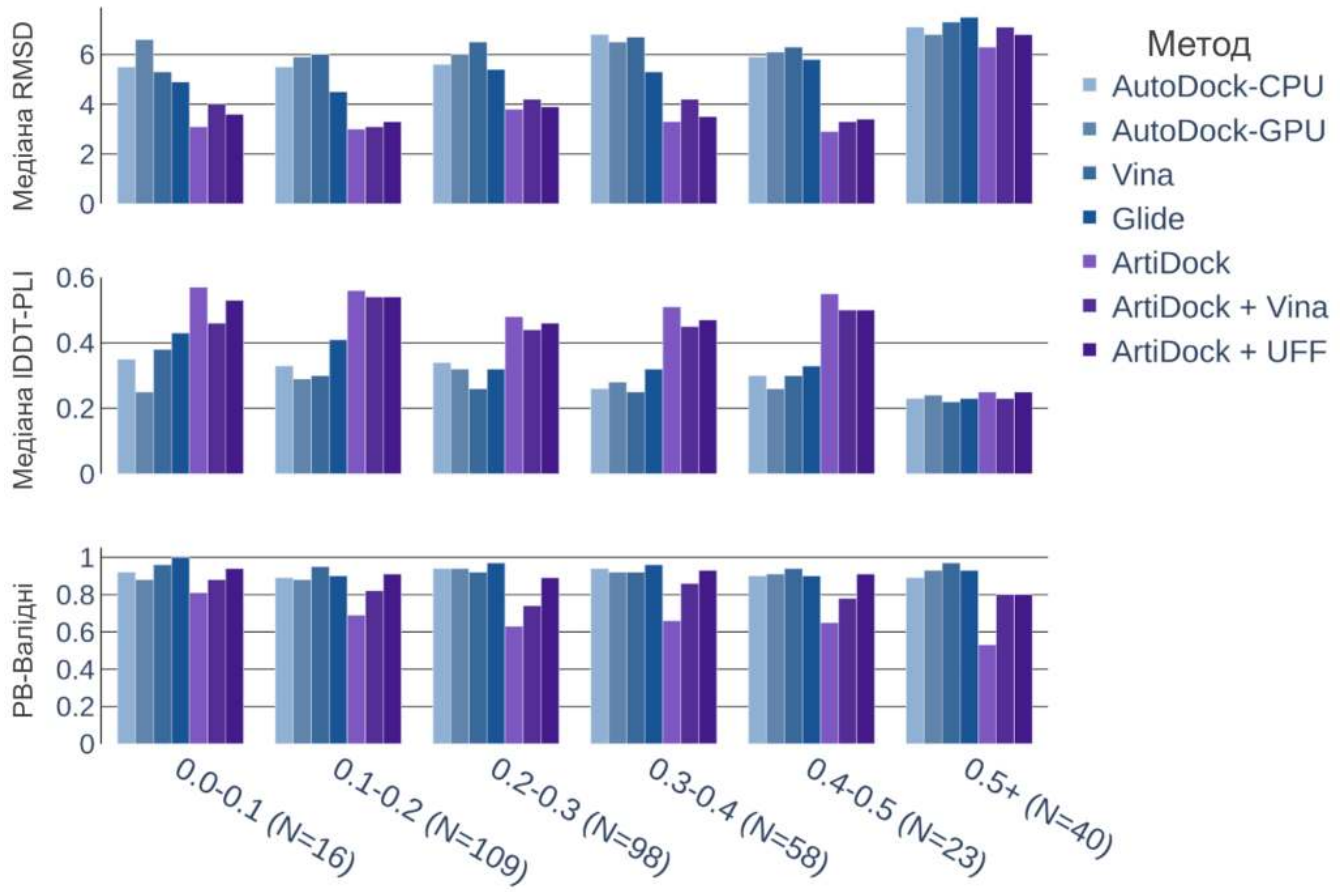

Рисунок Д3. Медіана RMSD (Å, менше - краще), медіана lDDT-PLI (більше - краще) та частка PB-Валідних передбачень (більше - краще) на піднаборах тестових даних PLINDER MLSB при різних порогових значеннях Cα RMSD (Å) між накладеними голо- та апо/передбаченими сайтами зв’язування.

Додаток Е. Додаткові таблиці

Таблиця Е1. Участь груп ліганду (атомів, ароматичних кілець або амідних груп) в нековалентних взаємодіях у комплексах PDBbind.

Білок	**Ліганд**	**Частка**	**N взаємодій**	**N унікальних груп ліганду, що взаємодіють**	**N унікальних груп ліганду, що можуть взаємодіяти**	**Ймовірність взаємодії групи ліганду**	**Середня N взаємодій ліганду на групу**
Атом C або S	Атом C або S	74.25%	255262	81217	98390	0.83	3.14
Атом C або S	Атом Cl, Br або I	2.74%	9423	1932	2043	0.95	4.88
Атом C або S	Атом F	0.90%	3089	1339	1737	0.77	2.31
Ароматичне кільце	Ароматичне кільце	1.79%	6143	4999	18463	0.27	1.23
Амідна група	Ароматичне кільце	1.39%	4775	3847	18463	0.21	1.24
Ароматичне кільце	Амідна група	0.32%	1087	930	6982	0.13	1.17
Акцептор	Донор	6.65%	22861	14692	23068	0.64	1.56
Донор	Акцептор	8.65%	29742	20889	45843	0.46	1.42
Акцептор	Атом галогену (Cl, Br, I)	0.22%	748	670	2043	0.33	1.12
Атом аніона	Атом катіона	0.66%	2274	1802	4206	0.43	1.26
Атом катіона	Атом аніона	1.67%	5757	4024	8427	0.48	1.43
Ароматичне кільце	Атом катіона	0.16%	546	429	4206	0.10	1.27
Атом катіона	Ароматичне кільце	0.60%	2077	1507	18463	0.08	1.38

Таблиця Е2. Результати підбору параметрів AutoDock-CPU.

	Параметри						
№ рядка	**ga_num_evals**	**ga_run**	**ga_pop_size**	**ga_num_generations**	**ga_elitism**	**spacing**	**grid_size**
1	2,500,000	10	150	27,000	1	0.375	cube 25
2	700,000	10	150	27,000	1	0.375	cube 25
3	350,000	10	150	27,000	1	0.375	cube 25
4	100,000	10	150	27,000	1	0.375	cube 25
5	700,000	20	150	27,000	1	0.375	cube 25
6	700,000	40	150	27,000	1	0.375	cube 25
7	700,000	10	300	27,000	1	0.375	cube 25
8	700,000	10	600	27,000	1	0.375	cube 25
9	700,000	10	150	50,000	1	0.375	cube 25
10	700,000	10	150	100,000	1	0.375	cube 25
11	700,000	10	150	27,000	2	0.375	cube 25
12	700,000	10	150	27,000	4	0.375	cube 25
13	700,000	10	150	27,000	1	0.2	cube 25
14	700,000	10	150	27,000	1	0.1	cube 25
15	700,000	10	150	27,000	1	0.375	cube 25
16	700,000	10	150	27,000	1	0.375	cube 20
17	700,000	10	150	27,000	1	0.375	cube 15
18	700,000	10	150	27,000	1	0.375	LABB + 20
19	700,000	10	150	27,000	1	0.375	LABB + 15
20	700,000	10	150	27,000	1	0.375	LABB + 10

Таблиця Е2. (Продовження)

	Час докінгу (с)						
№ рядка	**Середнє**	**Медіана**	**RMSD<2A**	**PB-Valid**	**lDDT-PLI>0.5**	**RMSD медіана**	**lDDT-PLI медіана**
1	172	148	0.43	0.96	0.50	2.5	0.50
2	49	40	0.42	0.95	0.50	2.6	0.50
3	24	20	0.41	0.94	0.49	2.6	0.49
4	7	6	0.34	0.90	0.44	3.3	0.44
5	98	81	0.46	0.96	0.51	2.4	0.51
6	196	161	0.46	0.96	0.52	2.3	0.52
7	49	40	0.42	0.94	0.50	2.5	0.51
8	49	40	0.42	0.92	0.48	2.7	0.48
9	49	40	0.43	0.94	0.51	2.5	0.50
10	49	40	0.42	0.95	0.49	2.6	0.50
11	49	40	0.42	0.95	0.48	2.6	0.49
12	49	40	0.43	0.96	0.49	2.6	0.49
13	55	46	0.44	0.94	0.50	2.5	0.51
14	80	74	0.42	0.96	0.49	2.6	0.50
15	49	40	0.42	0.95	0.50	2.6	0.50
16	49	40	0.47	0.94	0.53	2.3	0.53
17	49	40	0.48	0.93	0.57	2.2	0.59
18	49	40	0.38	0.94	0.46	3.2	0.46
19	49	40	0.42	0.96	0.48	2.6	0.49
20	49	40	0.46	0.94	0.55	2.3	0.56

Таблиця Е3. Результати підбору параметрів AutoDock-GPU.

	Параметри						
№ рядка	**nev**	**nrun**	**psize**	**ngen**	**autostop**	**spacing**	**grid_size**
1	-	20	150	42,000	1	0.375	cube 25
2	2,500,000	20	150	42,000	1	0.375	cube 25
3	700,000	20	150	42,000	1	0.375	cube 25
4	350,000	20	150	42,000	1	0.375	cube 25
5	100,000	20	150	42,000	1	0.375	cube 25
6	-	40	150	42,000	1	0.375	cube 25
7	-	80	150	42,000	1	0.375	cube 25
8	-	20	300	42,000	1	0.375	cube 25
9	-	20	600	42,000	1	0.375	cube 25
10	-	20	150	100,000	1	0.375	cube 25
11	-	20	150	200,000	1	0.375	cube 25
12	-	20	150	42,000	0	0.375	cube 25
13	-	20	150	42,000	1	0.2	cube 25
14	-	20	150	42,000	1	0.375	cube 25
15	-	20	150	42,000	1	0.375	cube 20
16	-	20	150	42,000	1	0.375	cube 15
17	-	20	150	42,000	1	0.375	LABB + 20
18	-	20	150	42,000	1	0.375	LABB + 15
19	-	20	150	42,000	1	0.375	LABB + 10

Таблиця Е3. (Продовження)

	Час докінгу (с)						
№ рядка	**Середнє**	**Медіана**	**RMSD<2A**	**PB-Valid**	**lDDT-PLI>0.5**	**RMSD медіана**	**lDDT-PLI медіана**
1	1.1	0.6	0.46	0.96	0.52	2.2	0.53
2	0.8	0.6	0.46	0.95	0.52	2.3	0.53
3	0.6	0.5	0.45	0.96	0.52	2.3	0.53
4	0.5	0.4	0.44	0.96	0.53	2.4	0.54
5	0.4	0.4	0.45	0.95	0.53	2.3	0.54
6	1.4	0.7	0.46	0.95	0.52	2.2	0.53
7	1.7	0.9	0.47	0.95	0.52	2.3	0.54
8	1.4	0.7	0.45	0.96	0.52	2.3	0.54
9	1.6	0.9	0.45	0.96	0.53	2.3	0.54
10	1.1	0.6	0.46	0.95	0.52	2.3	0.53
11	1.0	0.6	0.45	0.95	0.51	2.3	0.53
12	1.7	0.9	0.44	0.95	0.51	2.3	0.53
13	1.5	1.1	0.43	0.94	0.50	2.4	0.51
14	1.1	0.6	0.46	0.96	0.52	2.2	0.53
15	1.0	0.5	0.46	0.96	0.54	2.2	0.56
16	1.0	0.5	0.47	0.95	0.56	2.2	0.59
17	1.1	0.7	0.44	0.96	0.51	2.3	0.51
18	1.0	0.6	0.45	0.96	0.53	2.3	0.55
19	1.0	0.5	0.49	0.96	0.56	2.0	0.60

Таблиця Е4. Результати підбору параметрів Vina.

	Параметри			
№ рядка	**exhaustiveness**	**max_evals**	**spacing**	**grid_size**
1	4	-	0.375	cube 25
2	8	-	0.375	cube 25
3	16	-	0.375	cube 25
4	32	-	0.375	cube 25
5	64	-	0.375	cube 25
6	16	10,000	0.375	cube 25
7	16	100,000	0.375	cube 25
8	16	500,000	0.375	cube 25
9	16	1,000,000	0.375	cube 25
10	16	10,000,000	0.375	cube 25
11	16	-	0.2	cube 25
12	16	-	0.375	cube 25
13	16	-	0.375	cube 20
14	16	-	0.375	cube 15
15	16	-	0.375	LABB + 20
16	16	-	0.375	LABB + 15
17	16	-	0.375	LABB + 10

Таблиця E4. (Продовження)

	Час докінгу (с)						
№ рядка	**Середнє**	**Медіана**	**RMSD<2A**	**PB-Valid**	**lDDT-PLI>0.5**	**RMSD медіана**	**lDDT-PLI медіана**
1	48	25	0.49	0.93	0.53	2.1	0.55
2	96	50	0.53	0.94	0.56	1.8	0.60
3	190	100	0.57	0.94	0.59	1.6	0.65
4	369	188	0.58	0.94	0.61	1.6	0.66
5	603	352	0.59	0.95	0.61	1.5	0.69
6	1	1	0.38	0.92	0.43	3.2	0.44
7	13	10	0.51	0.92	0.54	1.9	0.56
8	63	49	0.53	0.93	0.57	1.8	0.59
9	117	89	0.56	0.94	0.58	1.6	0.64
10	193	100	0.56	0.94	0.60	1.7	0.65
11	198	108	0.56	0.94	0.62	1.6	0.66
12	190	100	0.57	0.94	0.59	1.6	0.65
13	208	97	0.57	0.95	0.61	1.6	0.67
14	280	113	0.60	0.95	0.63	1.6	0.70
15	186	98	0.55	0.94	0.58	1.7	0.61
16	186	99	0.56	0.94	0.60	1.7	0.65
17	198	94	0.59	0.95	0.63	1.5	0.68

Таблиця Е5. Результати підбору параметрів Glide.

	Параметри					
№ рядка	**PRECISION**	**MAXKEEP**	**MAXREF**	**MAX_ ITERATIONS**	**POSTDOCK_ NPOSE**	**OUTERBOX**
1	HTVS					cube 25
2	XP					cube 25
3	SP	5000	400	100	5	cube 25
4	SP	10000	400	100	5	cube 25
5	SP	15000	400	100	5	cube 25
6	SP	5000	800	100	5	cube 25
7	SP	5000	1600	100	5	cube 25
8	SP	5000	400	500	5	cube 25
9	SP	5000	400	1000	5	cube 25
10	SP	5000	400	100	10	cube 25
11	SP	5000	400	100	20	cube 25
12	SP	5000	400	100	40	cube 25
13	SP	5000	400	100	80	cube 25
14	SP	5000	400	100	160	cube 25
15	SP	5000	400	100	5	cube 25
16	SP	5000	400	100	5	cube 20
17	SP	5000	400	100	5	cube 15
18	SP	5000	400	100	5	LABB + 20
19	SP	5000	400	100	5	LABB + 15
20	SP	5000	400	100	5	LABB + 10

Таблиця E5. (Продовження)

	Час докінгу (с)						
№ рядка	**Середнє**	**Медіана**	**RMSD<2A**	**PB-Valid**	**lDDT-PLI>0.5**	**RMSD медіана**	**lDDT-PLI медіана**
1	12	12	0.35	0.94	0.45	3.2	0.45
2	138	93	0.51	0.95	0.58	2.0	0.65
3	26	18	0.49	0.94	0.54	2.1	0.55
4	28	19	0.48	0.93	0.53	2.1	0.55
5	29	20	0.49	0.94	0.53	2.0	0.55
6	27	19	0.48	0.95	0.55	2.2	0.59
7	29	21	0.48	0.94	0.56	2.2	0.59
8	27	19	0.49	0.94	0.54	2.1	0.55
9	27	19	0.49	0.94	0.54	2.1	0.55
10	27	19	0.51	0.93	0.55	1.9	0.60
11	27	19	0.52	0.94	0.56	1.8	0.63
12	28	20	0.52	0.95	0.57	1.8	0.64
13	30	21	0.54	0.94	0.59	1.8	0.66
14	32	23	0.55	0.94	0.59	1.8	0.66
15	26	18	0.49	0.94	0.54	2.1	0.55
16	18	15	0.50	0.95	0.58	2.0	0.62
17	10	10	0.50	0.97	0.57	1.9	0.59
18	33	20	0.49	0.95	0.54	2.1	0.55
19	25	17	0.51	0.92	0.54	1.9	0.56
20	16	13	0.40	0.89	0.48	2.7	0.49

Таблиця E6. Частка передбачень, що відповідають критеріям якості.

Метод	внутрішня енергія	стеричні зіткнення з білком	стеричні зіткнення з органічними кофакторами	стеричні зіткнення з неорганічними кофакторами	стеричні зіткнення з молекулами води
AutoDock-CPU	0.95	0.98	1.00	0.98	0.99
AutoDock-GPU	0.95	0.99	1.00	0.98	0.99
Vina	0.96	0.99	1.00	1.00	1.00
Glide	0.99	0.95	1.00	0.99	0.99
ArtiDock	0.90	0.81	0.99	1.00	0.96
ArtiDock + Vina	0.94	0.90	0.99	1.00	0.97
ArtiDock + UFF	0.98	0.98	1.00	1.00	0.99

Таблиця E7. Кореляція між критеріями точності передбачених положень лігандів та Cα RMSD між накладеними голо- та апо/передбаченими сайтами зв’язування білка.

Кореляція	Метод	RMSD	lDDT-PLI
Спірмена	**AutoDock-CPU**	0.16	-0.159
Спірмена	**AutoDock-GPU**	0.131	-0.119
Спірмена	**Vina**	0.153	-0.124
Спірмена	**Glide**	0.174	-0.185
Спірмена	**ArtiDock**	0.193	-0.212
Спірмена	**ArtiDock + Vina**	0.199	-0.208
Спірмена	**ArtiDock + UFF**	0.185	-0.198
Пірсона	**AutoDock-CPU**	0.17	-0.166
Пірсона	**AutoDock-GPU**	0.136	-0.133
Пірсона	**Vina**	0.162	-0.142
Пірсона	**Glide**	0.222	-0.205
Пірсона	**ArtiDock**	0.265	-0.264
Пірсона	**ArtiDock + Vina**	0.305	-0.274
Пірсона	**ArtiDock + UFF**	0.267	-0.255