

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 пропускну 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 (IDDT-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 (IDDT-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](https://doi.org/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](https://doi.org/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](https://doi.org/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. Колеев, 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. Yesylevskyu. **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. Yesylevskyu. **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.