

Голові разової спеціалізованої ради
Інституту фізики НАН України
доктору фіз.-мат. наук Габовичу О. М.

ВІДГУК

офіційного опонента, доктора фізико-математичних наук, доцента
кафедри молекулярної фізики фізичного факультету Київського національного
університету імені Тараса Шевченка **НІКОЛАЄНКА Тимофія Юрійовича**
на дисертаційну роботу **ВОЙЦІЦЬКОГО Тараса Васильовича**
*«Визначення параметрів нековалентних взаємодій малих органічних молекул із
протеїнами на основі моделей машинного навчання»*,
подану на здобуття ступеня доктора філософії з галузі знань
10 – «Природничі науки» за спеціальністю 104 – «Фізика та астрономія»

Актуальність теми дисертації та її фізичний зміст

Дослідження фізико-хімічних закономірностей нековалентного міжмолекулярного розпізнавання у біополімерних системах є одним із актуальних і практично важливих завдань статистичної фізики конденсованого стану та молекулярної біофізики. Утворення стійких комплексів між білковими макромолекулами та низькомолекулярними органічними лігандами, яке має істотний вплив на біофізичні процеси у живих системах, термодинамічно є проявом складного балансу між, з одного боку, ентальпійним вирашем від формування водневих зв'язків, π -електронних та іонних контактів, дисперсійних взаємодій Ван-дер-Ваальса, та, з іншого — ентропійними факторами, зумовленими десольватацією контактних поверхонь та обмеженням просторових ступенів вільності системи.

У практичній площині комп'ютерного моделювання та дизайну біологічно активних сполук (*in silico drug discovery*) сучасні підходи зіткнулися з вираженою дилемою, адже попри високу точність та фізичну обґрунтованість, атомістичне моделювання відповідних систем методом молекулярної динаміки разом із розрахунками вільних енергій на основі термодинамічного інтегрування є надто повільним для відбору (скринінгу) потрібних сполук із-поміж значної кількості існуючих. Натомість, класичний докінг, який спирається на спрощені моделі міжатомної взаємодії (силові поля), практично вичерпав можливості покращення своєї прогностичної сили. Через такий дисбаланс обчислювальної складності і точності традиційні підходи до *in silico* скринінгу практично досягли межі своєї ефективності. Дисертаційна робота Т.В. Войціцького присвячена розвитку архітектур глибинного навчання (Deep Learning) як можливого шляху розв'язання вказаної проблеми.

Актуальність і наукова цінність роботи зумовлена тим, що розроблені у ній методи машинного навчання не є результатом розв'язання суто інженерної задачі аналізу даних або підбору гіперпараметрів, а спираються на фізичні моделі процесів міжмолекулярної взаємодії у досліджуваних системах. Зокрема, апроксимація багатовимірної (гіпер)поверхні вільної енергії як основа для передбачення спорідненості (афінності), геометрії комплексів «білок-ліганд» та розрізнення типів атомів розглядається подібно до параметризації силових полів у класичній молекулярній фізиці. Застосовується аналогія з ієрархією функцій розподілу в статистичній фізиці, в рамках

якої визначається, який саме порядок статистик — від одновимірних розподілів геометричних параметрів контактів до парних міжмолекулярних кореляцій у вигляді матриць відстаней та вибірок із багатовимірних конформаційних просторів білків — забезпечує генерацію реалістичних моделей досліджуваних комплексів. Фізичний підхід дозволив водночас розв'язати і критично важливі прикладні завдання — подолати дефіцит і структурну однорідність навчальних даних у базі PDB, нейтралізувати перетікання інформації (data leakage) між вибірками молекулярних структур, ефективно моделювати тривимірну структуру молекулярних комплексів.

Робота безпосередньо пов'язана з планами відділу фізики біологічних систем Інституту фізики НАН України та виконувалася в межах держбюджетної теми «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., державний реєстраційний номер 0123U100990) та гранту НФДУ «Низькорозмірні графеноподібні дихалькогеніди перехідних металів з контрольованими полярними та електронними властивостями для перспективних застосувань у нанoeлектроніці та біомедицині» (2020–2021, 2023 рр., № 2020.02/0027).

Наукова новизна одержаних результатів

Наукову новизну роботи визначають такі ключові фундаментальні та прикладні результати, які отримані вперше:

- 1) запропоновано та реалізовано модель 3DProtDTA як новий підхід до прогнозування спорідненості зв'язування малих молекул із макромолекулами на основі графового представлення білкової структури на рівні амінокислот із залученням просторових координат, одержаних в рамках моделей AlphaFold, які інтерпретуються як вибірка з багатовимірного конформаційного розподілу;
- 2) за допомогою комп'ютерного експерименту обґрунтовано достатність використання одновимірних статистичних розподілів геометричних характеристик нековалентних взаємодій для конструювання реалістичних сайтів зв'язування лігандів;
- 3) створено метод молекулярного докінгу ArtiDock, у якому поєднано обчислювально ефективне прогнозування матриць міжмолекулярних контактів еквіваріантними нейромережами із подальшим фізично вмотивованим двоетапним алгоритмом реконструкції валідного просторового положення ліганду.

Новизну додатково підтверджують результати, отримані завдяки:

- 1) удосконаленню дифузійної генеративної нейромережевої моделі PocketCFDM для прогнозування просторових положень лігандів завдяки створеному оригінальному алгоритму аугментації навчальних вибірок синтетичними кишнями, що забезпечило кардинальне зниження частки стеричних колізій порівняно з базовою моделлю з архітектурою DiffDock;
- 2) подальшому з'ясуванню меж деформації остова апо-структур ($\text{RMSD}_{\text{Ca}} > 0.5 \text{ \AA}$), яка впливає на фізичну коректність розрахованих положень ліганда, змінюючи кількість стеричних колізій в моделі негнучкого докінгу.

Наукове та практичне значення результатів дисертації

Наукова цінність роботи полягає у розширенні підходів статистичної фізики до апроксимації складних енергетичних поверхонь комплексів біологічних макромолекул за допомогою спеціалізованих нейромережевих моделей графових та генеративних архітектур.

Практична цінність одержаних результатів полягає у створенні збалансованих за швидкістю та точністю програмних інструментів для ранніх стадій розробки лікарських засобів. Зокрема:

- метод ArtiDock забезпечує час передбачення просторового положення 0.2–2 с на один ліганд, знижуючи розрахункову собівартість скринінгу сполук і роблячи її суттєво дешевшою у порівнянні із традиційними інструментами (Glide, AutoDock, Vina) при вищій точності;
- створене програмне забезпечення інтегровано в розрахункову платформу біотехнологічної компанії Receptor.AI, у якій набуло застосування у розв'язанні практичних задач масивного віртуального скринінгу;
- програмні реалізації, протоколи оцінювання, набори даних та підготовлені у роботі тестові вибірки даних розміщено у загальнодоступних репозиторіях GitHub і Zenodo.

Обґрунтованість і достовірність наукових положень, висновків і результатів дисертаційного дослідження

Наукові висновки й положення дисертації відповідають темі дисертації і мають високий ступінь достовірності, який забезпечується коректним застосуванням апарату статистичної фізики, методів машинного навчання та математичної оптимізації, а також використанням верифікованого програмного забезпечення (PyTorch, PyTorch Geometric, RDKit, Biopython, Pteros) із відкритим вихідним кодом.

Їх обґрунтованість підтверджується:

- використанням репрезентативних, стандартизованих масивів експериментальних даних (PDB, PDBbind, Davis, KIBA, ZINC20);
- залученням сучасних схем розбиття даних, строгим контролем перетікання інформації шляхом структурної кластеризації профілів взаємодій (PLINDER) та хронологічного розбиття (PoseX) задля уникнення штучного завищення метрик точності моделей;
- детальним налаштуванням та калібруванням робочих параметрів порівнюваних алгоритмів (зокрема, у випадках Glide, Vina, AutoDock);
- незалежною верифікацією фізичної та хімічної валідності отриманих структур комплексом тестів PoseBusters;
- проходженням результатів через процедуру незалежного рецензування у міжнародних фахових наукових виданнях.

Загальна характеристика змісту дисертаційної роботи

Дисертація має загальний обсяг 197 сторінок (з них основного тексту — 165 с.) та містить вступ, 4 розділи, висновки, список використаних джерел (185 позицій) та 6 додатків. Її матеріал включає 25 рисунків та 16 таблиць в основному тексті.

У Вступі висвітлено стан проблеми, обґрунтовано вибір об'єкта і предмета дослідження, сформульовано мету, завдання, наукову новизну та практичну цінність результатів, а також наведено дані щодо їх апробації й особистого внеску.

У Розділі 1 проаналізовано термодинаміку молекулярного розпізнавання (моделі індукованої відповідності та конформаційного відбору) та обмеження сучасних методів оцінки спорідненості. Розглянуто термодинаміку асоціації та ефекти ентальпійно-ентропійної компенсації. Проаналізовано еволюцію поглядів на молекулярне розпізнавання. Зіставлено експериментальні методи із обчислювальними підходами класичного докінгу, молекулярної динаміки, PBSA, FEP та новітніми нейромережевими архітектурами (графові мережі, трансформери, дифузійні моделі).

У Розділі 2 викладено створення графової моделі 3DProtDTA для оцінки

характеристик взаємодії білок-ліганд. Розкрито підхід до графового опису структури макромолекули на рівні амінокислотних залишків, топологія якого формується за просторовими координатами моделей AlphaFold, дозволяючи зняти обмеження щодо наявності кристалографічних даних. Доведено перевагу архітектури над аналогами (GraphDTA, DeepDTA тощо) та досліджено специфіку узагальнення на невідомих класах кіназ.

У Розділі 3 розв'язано проблему вираженого дисбалансу та дефіциту просторових комплексів у базі PDBbind в контексті прогнозування взаємодій білків з лігандами. За результатами детального статистичного аналізу значної кількості нековалентних контактів сформовано правила генерації штучних білкових «кишень» із використанням дипептидних блоків. Працездатність підходу валідовано на генеративній моделі RocketCFDM і з'ясовано, що включення синтетичних комплексів забезпечило приріст частки просторових положень високої точності ($\text{RMSD} < 2 \text{ \AA}$) до 10% та істотно мінімізувало стеричні конфлікти.

У Розділі 4 розроблено метод високопродуктивного докінгу ArtiDock та його програмну реалізацію. Метод використовує SE(3)-еквіваріантні графові нейронні мережі у поєднанні з механізмом уваги трансформерів для передбачення матриці міжатомних відстаней. Така матриця перетворюється у валідну просторову конфігурацію за двоетапною процедурою, яка включає градієнтний спуск та диференціальну еволюцію конформерів з локальною мінімізацією потенціальної енергії структур. Показано, що на перевіірочних вибірках PLINDER та PoseX ArtiDock суттєво перевершує за точністю класичні пакети (Glide, Vina, AutoDock-GPU) за значно менших витрат обчислювального часу.

Висновки містять узагальнення одержаних наукових результатів, а Додатки А–Е деталізують процедури попередньої фільтрації, кластеризації даних, підбору параметрів і оцінки обчислювальної вартості.

Оформлення роботи відповідає встановленим вимогам.

Відсутність порушень академічної доброчесності

Порушень принципів академічної доброчесності, фактів академічного плагіату, фабрикації чи фальсифікації даних у дисертаційній роботі та публікаціях здобувача не виявлено. Використані теоретичні концепції та експериментальні бази даних супроводжуються вичерпними бібліографічними посиланнями з чітким розмежуванням авторського напрацювання.

Програмний код для оцінювання результатів та валідаційні вибірки знаходяться у вільному доступі. Збереження закритими знайдених вагових коефіцієнтів моделі ArtiDock зумовлене її комерційним використанням, тоді як усі математичні викладки, структури функцій втрат і числові гіперпараметри детально розкриті в дисертації, що забезпечує наукову прозорість роботи та можливість незалежного відтворення її результатів.

Повнота викладу матеріалів дисертації в наукових публікаціях та апробація

Результати дисертаційної праці повно висвітлено у 3 статтях, опублікованих у міжнародних фахових виданнях, що індексуються у базах Scopus та Web of Science Core Collection і належать до квартилів Q1/Q2 (відповідно до предметної області):

1. *Voitsitskyi T., Stratiichuk R., Koleiev I., Popryho L., Ostrovsky Z., Henitsoi P., Khropachov I., Vozniak V., Zhytar R., Nechepurenko D., Yesylevskyi S., Nafiiiev A., Starosyla S.* 3DProtDTA: a deep learning model for drug-target affinity prediction based on residue-level

protein graphs. RSC Advances, 2023, Vol. 13, No. 15, pp. 10261–10272. DOI: 10.1039/D3RA00281K..

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

3. *Voitsitskiy T.*, Koleiev I., Stratiichuk R., Kot O., Kyrylenko R., Savchenko I., Husak V., Yesylevskyy S., Starosyla S., Nafiiiev A. ArtiDock: Accurate Machine Learning Approach to Protein–Ligand Docking Optimized for High-Throughput Virtual Screening. Journal of Chemical Information and Modeling, 2026, Vol. 66, No. 4, pp. 2117–2128. DOI: 10.1021/acs.jcim.5c02777.

Кількість та науковий рівень публікацій повністю відповідають п. 8 Порядку присудження ступеня доктора філософії (Постанова КМУ № 44). У всіх зазначених роботах здобувач є першим автором із визначальним особистим внеском.

Апробація результатів здійснювалася на 5 міжнародних та всеукраїнських наукових конференціях: ISSSMC-2024 (Воянув, Польща, 2024 р.), ПНК-2024 Інституту фізики НАН України (Київ, 2025 р.), NBP-2025 (Харків, 2025 р.), APS Satellite Workshop (Київ, 2026 р.), Biophysics Week 2026 (Київ, 2026 р.).

Оволодіння здобувачем методологією наукової діяльності та рівень виконання завдань

Здобувач на високому рівні володіє методологією наукового дослідження в галузях фізики, біофізики та машинного навчання. Робота підтверджує здатність автора самостійно вирішувати складні міждисциплінарні проблеми, розробляти математичні моделі та алгоритми, реалізовувати високоефективний програмний код, коректно організовувати багатоетапні обчислювальні експерименти на сучасних апаратних платформах і здійснювати критичну фізичну інтерпретацію отриманих закономірностей. Усі сформульовані в роботі задачі виконано повністю.

Дискусійні положення, зауваження та рекомендації до дисертаційної роботи

Позитивно оцінюючи високий рівень дисертації, вважаю за доцільне виокремити низку зауважень і дискусійних питань:

1. Відсутнє пряме порівняння точності моделі ArtiDock (Розділ 4) при її навчанні з використанням розробленої автором синтетичної аугментації даних (Розділ 3) та без неї. Так, у Розділі 3 переконливо доведено користь синтетичної аугментації для покращення моделі PocketCFDM в умовах дефіциту даних у PDBbind, однак модель ArtiDock (Розділ 4) тренувалася суто на реальних комплексах з PLINDER без використання згенерованих штучних «кишень». Відсутність тестування ArtiDock із синтетичною аугментацією залишає відкритим питання про сумісність і практичну синергію між цими двома розробками.

2. В алгоритмі синтезу штучних сайтів зв'язування (Розділ 3) застосована спрощена формалізація гідрофобного ефекту на основі геометричного критерію. Хоча у теоретичній частині роботи детально обґрунтовано, що рушійною силою гідрофобних контактів є ентропійний виграш від витіснення молекул води, проте в алгоритмі синтезу «кишень» гідрофобний контакт задається суто геометричним критерієм (за відстанню), без урахування розчинника. Це призвело до виникнення значного штучного надлишку гідрофобних взаємодій у згенерованих сайтах (453 тис., проти 343 тис. у природних), що

певною мірою зміщує статистичні властивості утвореної вибірки молекулярних структур.

3. Хоча задача мінімізації функції між- та внутрішньомолекулярних нев'язок на першій стадії післяобробки ArtIDock (Розділ 4.3.3), яка полягає у реконструкції просторової «хмари» точок для конкретної сполуки, є строго детермінованою, для її розв'язання було обрано метод оптимізації Adam, який початково був розроблений як спеціалізований для задач стохастичної оптимізації (що цілком виправдовує його застосування у таких задачах, згаданих, наприклад, у Розділі 2.2.4). Натомість, застосування класичних детермінованих підходів (наприклад, квазіньютонівського L-BFGS, із точним або наближеним лінійним пошуком) виглядало би більш природним, і не позбавленим тих переваг, які у Розділі 4.3.3 згадуються стосовно методу Adam. Відповідно, без явного виділення необхідних (або визначальних) вимог, які в контексті цієї задачі мав би задовольняти метод оптимізації, переваги саме обрання методу Adam у ній залишаються дискусійними.

4. Зважаючи на те, що на тестовому масиві PLINDER MLSB із апо- та змодельованими структурами точність ArtIDock закономірно падає (медіана RMSD зростає з 2.0 Å до 3.4 Å; частка валідних передбачень без додаткової силової релаксації знижується до 53% за зміщення білкового остова, що перевищує $\text{RMSD}_{\text{Ca}}^{\text{holo-apo}} > 0.5 \text{ Å}$), — незрозуміло, чи означає це, що «намагаючись» зберегти істинні міжатомні відстані, модель призводить до стеричних колізій із незсунутими бічними ланцюгами білка. Відповідно, якою мірою це може вказувати на обмеження методу для систем із вираженим ефектом індукованої відповідності (induced fit)?

5. Оскільки модель ArtIDock оптимізована на генерацію єдиної просторової конфігурації ліганду і не здійснює розрахунку термодинамічної вільної енергії ΔG , незрозуміло, чи має ця модель (у її нинішньому або модифікованому вигляді) здатність пріоритетувати хімічні сполуки за спорідненістю, чи вимагає вона додаткового залучення зовнішніх скорингових систем або поетапного використання моделі 3DProtDTA. Зазначена на с. 161 властивість методу знаходити «положення ... оптимальні для даного ліганду та даної конформації сайту зв'язування» залишає відкритим питання про те, оптимуму якої фізичної величини відповідає знайдене положення.

6. У разі коли за допомогою нейромережі AlphaFold було згенеровано (і включено до використаних у роботі вибірок молекул) єдину квазістатичну усереднену структуру для даного білка, — якою мірою вона може в контексті завдань роботи відображати статистичний ансамбль конформерів білка у водному оточенні, його динаміку («гнучкість» структури)?

7. Зіставлення часових характеристик ArtIDock із даними методів ко-фолдингу на наборі PoseX (Розділ 4.4.4) спирається на опубліковані сторонніми авторами показники, отримані на різних апаратних конфігураціях обчислювального обладнання, призводячи до суто орієнтовного характеру оцінки переваги моделей за швидкістю роботи.

8. У тексті присутні незначні одруківки або неточності. Зокрема, на с. 5 та 127 дисертації формулювання щодо підвищення частки «до 10%» (що є неоднозначним, адже згідно із табл. 3.3 (40 згенерованих положень, найкраще з топ 5), частка просторових положень ліганда із $\text{RMSD} < 2 \text{ Å}$ зросла з 15.1% до 25.1%, тобто відбулося зростання на 10 відсоткових пунктів); нашарування двох різних конструкцій (або пропущене слово) на с. 53 у «DynamicBind здатен адаптувати конформацію дозволяє ідентифікувати конформації білків...»; неузгодженість слів у «Центроїди сайті були розташовані в центрі...» на с. 139.

Наведені зауваження мають рекомендаційний і дискусійний характер, стосуються перспектив подальшого розвитку наряду і не знижують загальної високої оцінки виконаного дослідження.

Загальний висновок

Дисертаційна робота Войціцького Тараса Васильовича «Визначення параметрів нековалентних взаємодій малих органічних молекул із протеїнами на основі моделей машинного навчання» є завершеним, самостійним науковим дослідженням, у якому розв'язано актуальну задачу обчислювальної біофізики — підвищення точності та швидкості моделювання нековалентних взаємодій «білок–ліганд» шляхом створення й оптимізації обчислювальних моделей машинного навчання з урахуванням фізико-хімічних закономірностей нековалентного зв'язування.

За своєю актуальністю, рівнем, науковою новизною, теоретичною та практичною цінністю, обґрунтованістю висновків і повнотою викладу результатів в рецензованих публікаціях дисертація відповідає встановленим вимогам «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого постановою КМУ від 12 січня 2022 р. № 44 (зі змінами), та Наказу МОН України № 40 від 12.01.2017 р. (зі змінами).

Здобувач, Войціцький Тарас Васильович, заслуговує на присудження наукового ступеня доктора філософії з галузі знань 10 – «Природничі науки» за спеціальністю 104 – «Фізика та астрономія».

Офіційний опонент:

Доктор фізико-математичних наук,
доцент кафедри молекулярної
фізики фізичного факультету
Київського національного університету
імені Тараса Шевченка

Тимофій НІКОЛАЄНКО