

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

Відгук
офіційного опонента – Степаньяна Степана Григоровича
доктора фізико-математичних наук, старшого наукового співробітника,
завідувача відділу молекулярної біофізики Фізико-технічного інституту
низьких температур ім. Б.І. Веркіна НАН України
на дисертаційну роботу
ВОЙЦІЦЬКОГО ТАРАСА ВАСИЛЬОВИЧА

«Визначення параметрів нековалентних взаємодій малих органічних молекул
із протеїнами на основі моделей машинного навчання»
подану до захисту на здобуття наукового ступеня доктора філософії
з галузі знань 10 – Природничі науки
за спеціальністю 104 – Фізика та астрономія

**Актуальність дисертаційної роботи та її зв'язок з науковими
напрямами, планами, темами.**

Дисертаційна робота «Визначення параметрів нековалентних взаємодій малих органічних молекул із протеїнами на основі моделей машинного навчання» Т.В.Войціцького присвячена актуальній проблемі сучасної фізики біологічних систем, яка полягає у поєднанні фізичного розуміння природи міжмолекулярних взаємодій із сучасними алгоритмами глибинного навчання. Дослідження взаємодії біополімерів, насамперед протеїнів та нуклеїнових кислот, з малими органічними молекулами є надзвичайно важливими для сучасної фармакології. Останнім часом розробка нових ліків починається з проведення масового скринінгу місць можливого зв'язування мільйонів органічних молекул із біополімерами. Визначення параметрів нековалентних взаємодій малих органічних молекул із протеїнами є важливим завданням. Такі взаємодії лежать в основі процесів молекулярного розпізнавання, регуляції активності білків, дії лікарських засобів і формування молекулярних комплексів. Тому точне прогнозування характеристик взаємодії між малими молекулами та протеїнами має практичне значення для пошуку й оптимізації потенційних біологічно активних сполук.

Традиційне експериментальне визначення параметрів зв'язування є достатньо складним, ресурсоємним і потребує значних часових витрат. Обчислювальні методи, зокрема молекулярне докування та молекулярна динаміка, дають змогу частково вирішити цю проблему, однак також характеризуються значною обчислювальною складністю та залежністю від якості використовуваних моделей. У зв'язку з цим особливої актуальності

набуває застосування методів машинного навчання, здатних виявляти складні нелінійні закономірності у великих масивах експериментальних і структурних даних.

Сучасний розвиток машинного навчання та штучного інтелекту створює можливості для прогнозування афінності зв'язування, енергії взаємодії, ймовірності утворення стабільного комплексу та інших параметрів нековалентних взаємодій на основі молекулярних дескрипторів, структурних характеристик протеїнів і властивостей їхніх комплексів. Використання таких підходів може суттєво прискорити первинний відбір перспективних сполук і зменшити обсяг експериментальних досліджень.

Отже, **актуальність дослідження зумовлена** необхідністю розроблення точних і ефективних розрахункових підходів до оцінювання нековалентних взаємодій малих органічних молекул із протеїнами. Поєднання структурних даних, молекулярних дескрипторів і моделей машинного навчання є перспективним напрямом для підвищення ефективності прогнозування молекулярного зв'язування та може бути використане в задачах раціонального дизайну лікарських засобів і пошуку нових біологічно активних сполук.

Дисертаційна робота виконана у відділі фізики біологічних систем Інституту фізики НАН України та відповідає новітнім напрямам досліджень сучасної біофізики, біоінформатики, структурної біології та фармацевтичної хімії. Робота виконувалась 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.) та НФДУ № 2020.02/0027, «Низькорозмірні графеноподібні дихалькогеніди перехідних металів з контрольованими полярними та електронними властивостями для перспективних застосувань у нанoeлектроніці та біомедицині» (2020-21, 2023 рр).

Ступінь достовірності отриманих результатів і обґрунтованості основних наукових положень і висновків.

Достовірність результатів, наведених у дисертаційній роботі, забезпечується поєднанням методів обчислювальної біофізики для встановлення фізичної природи процесів молекулярного розпізнавання, аналізу нековалентних взаємодій та обґрунтування вибору репрезентацій вхідних даних, а також методів чисельного комп'ютерного моделювання для реалізації розроблених підходів у вигляді алгоритмів та моделей машинного навчання. **Обґрунтованість основних висновків** забезпечена ретельним урахуванням усіх можливих типів міжмолекулярних взаємодій, а також використанням великої бази даних експериментально визначених структур комплексів білок-ліганд, отриманих за допомогою методів рентгеноструктурного аналізу, cгуо-ЕМ або NMR, яка використовувалась при тренуванні обчислюваної моделі. Додатковим підтвердженням достовірності та наукової значущості отриманих наукових результатів є публікації у провідних світових періодичних фахових виданнях, всі з яких віднесені до

першого квартилю (Q1), а також апробація отриманих результатів на п'яти вітчизняних та міжнародних наукових конференціях.

Повнота викладу основних наукових положень в опублікованих працях із зазначенням особистого внеску здобувача.

Основні результати дисертаційної роботи опубліковано у 3 наукових працях у періодичних фахових виданнях RSC Advances та Journal of Chemical Information and Modeling, віднесених до першого квартилю (Q1). Кожен із розділів 2, 3 та 4 дисертації відповідає науковій праці, опублікованій зі співавторами.

У цих наукових працях, виконаних за темою дисертації, матеріали дисертації і **основні наукові положення викладені повним чином.**

Особистий внесок здобувача полягає у наступному: У науковій публікації, результати якої описані в розділі 2, розроблено схему експерименту; розроблено модулі, необхідні для тренування та тестування моделі машинного навчання; проведено підбір гіперпараметрів; створено порівняння з наявними методами; проаналізовано результати; написано текст публікації та сформовано таблиці та зображення. У науковій публікації, результати якої описані в розділі 3, розроблено модулі, необхідні для аналізу та обробки тренувальних даних, тренування та тестування моделі машинного навчання для молекулярного докінгу; розроблено модуль для генерації штучних комплексів протеїн-ліганд; проаналізовано результати; написано текст публікації та сформовано таблиці та зображення. У науковій публікації [3], результати якої описані в розділі 4, розроблено схему експерименту; розроблено модулі, необхідні для тренування та тестування моделі машинного навчання, післяобробки передбачень; проведено підбір гіперпараметрів; створено порівняння з наявними методами; проаналізовано результати; написано текст публікації та сформовано таблиці та зображення.

Наукова новизна отриманих результатів.

У дисертації Т.В.Войціцького «Визначення параметрів нековалентних взаємодій малих органічних молекул із протеїнами на основі моделей машинного навчання» **вперше** проведені **наукові дослідження**, на основі яких отримані такі **нові результати**:

1. Розроблено новий 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. Розроблено 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.

Наукове та практичне значення результатів дисертації.

Наукове значення дисертації Т.В.Войціцького полягає у розвитку сучасних підходів до кількісного опису та прогнозування нековалентних взаємодій між малими органічними молекулами та протеїнами із застосуванням методів машинного навчання. Такі взаємодії визначають специфічність і стабільність утворення молекулярних комплексів та значною мірою зумовлюють біологічну активність низки органічних сполук. Застосування моделей машинного навчання дає змогу встановлювати залежності між структурними й фізико-хімічними характеристиками молекул, властивостями білкових мішеней та параметрами їхньої взаємодії.

Наукова цінність такого підходу полягає не лише у можливості прогнозування параметрів зв'язування, а й у виявленні закономірностей, які складно встановити за допомогою традиційних методів аналізу. Дослідження сприяє вдосконаленню методології комп'ютерного моделювання молекулярних взаємодій, зокрема шляхом порівняння різних алгоритмів машинного навчання, визначення найбільш інформативних молекулярних дескрипторів та оцінювання впливу структурних характеристик на точність прогнозування. Отримані результати можуть розширити наукові уявлення про взаємозв'язок між молекулярною структурою та параметрами нековалентного зв'язування.

Таким чином, наукове значення теми полягає у створенні та обґрунтуванні сучасного обчислювального підходу до дослідження взаємодій органічних молекул з протеїнами, що поєднує методи обчислювальної

біофізики, структурної біоінформатики та машинного навчання. Практичне значення результатів дисертації полягає у можливості використання як теоретичної й методичної основи для подальших досліджень молекулярного розпізнавання та раціонального пошуку біологічно активних сполук.

Оцінка основного змісту, структури і оформлення дисертації

Дисертаційна робота Т.В.Войціцького складається з вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації становить 197 сторінки; робота містить 28 рисунків та 23 таблиці. Список використаних джерел містить 185 найменувань на 13 сторінках. Додатки (19 сторінок) містять 3 рисунки та 7 таблиць.

У *вступі* обґрунтовано актуальність обраної теми, наведено мету та завдання дослідження, зв'язок дисертації з тематикою роботи Інституту фізики НАН України, визначено об'єкт, предмет та методи дослідження, наукову новизну та практичну цінність отриманих результатів, особистий внесок здобувача в опубліковані наукові праці, що висвітлюють результати роботи.

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 входних репрезентацій. Показано, що модель 3DProtDTA перевершує наявні аналоги на 1-20% за загальноприйнятими критеріями. 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 протеїн-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 генеративній дифузійній моделі

PocketCFDM: застосування штучних комплексів для розширення тренувальних даних дозволило підвищити частку правильних передбачень положення лігандів до 10% та знизити частку передбачень зі стеричними зіткненнями на 5-6%.

У **четвертому розділі** представлено ArtIDock - метод молекулярного докінгу на основі глибинного навчання, оптимізований для високопродуктивного віртуального скринінгу. ArtIDock поєднує легку архітектуру штучних нейронних мереж із фізично обґрунтованою післяобробкою передбачень, що дозволяє досягти пропускну здатності на рівні або кращої за найшвидші традиційні інструменти. Показано, що ArtIDock стабільно перевершує класичні методи в різноманітних реалістичних сценаріях, включаючи білкові сайти з кофакторами, іонами та молекулами води. На розділеному за часом тестовому наборі PoseX ArtIDock демонструє конкурентоспроможну точність порівняно з найефективнішими інструментами машинного навчання, досягаючи 73.5% успішних передбачень та найкращих результатів серед усіх порівняних підходів (63.9%) з урахуванням хімічної валідності передбачень.

Розділ **Висновки** містить підсумки та основні висновки роботи.

Дисертація написана українською мовою з застосуванням сучасної наукової термінології, результати досліджень викладено з логічною послідовністю, робота складає враження цілісної та закінченої наукової праці.

Оформлення дисертації відповідає вимогам до кваліфікаційних наукових робіт такого рівня.

Зауваження та рекомендації

1. Було б корисно навести приклад отриманих при моделюванні структур комплексів протеїн-ліганд хоча б для кількох типів лігандів, а також порівняння їх структурних параметрів з експериментальними даними.
2. Також було б доцільним встановити чи впливає машинне навчання на отриману структуру комплексів і, якщо впливає, навести кілька прикладів для різних типів лігандів.
3. У роботі необхідно було провести хоча б короткий аналіз оціночних значень енергій міжмолекулярної взаємодії в комплексах протеїн-ліганд та порівняння цих енергій з результатами, отриманими іншими методами, наприклад, методом молекулярної динаміки.
4. У списку наукових праць, в яких опубліковані основні наукові результати дисертації (анотація, стор. 11), не вказані квартали журналів.

Зроблені зауваження жодним чином не знижують високої наукової цінності дисертаційного дослідження, та можуть бути враховані здобувачем при подальшому розвитку теми та підготовці наукових праць.

Загальний висновок та оцінка дисертації

Враховуючи актуальність, наукову новизну і практичне значення одержаних результатів, вважаю, що дисертаційна робота Т.В.Войціцького «Визначення параметрів нековалентних взаємодій малих органічних молекул

із протейнами на основі моделей машинного навчання», подана на здобуття наукового ступеня доктора філософії, **відповідає вимогам** наказу МОН України № 40 від 12.01.2017 р. «Про затвердження Вимог до оформлення дисертації» зі змінами, внесеними згідно з Наказом Міністерства освіти і науки № 759 від 31.05.2019 та постанові КМ України № 44 від 12.01.2022 р. «Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», з останніми змінами, внесеними постановою КМ № 507 від 03.05.2024 р., а її автор Войціцький Тарас Васильович **заслуговує** на присудження наукового ступеня доктора філософії за спеціальністю 104 – Фізика та астрономія.

Офіційний опонент:

завідувач відділу молекулярної біофізики

Фізико-технічного інституту низьких

температур ім. Б.І. Веркіна НАН України

доктор фіз.-мат. наук, с.н.с.

Степаньян С.Г.