

Рецензія
на дисертаційну роботу
ВОЙЦІЦЬКОГО ТАРАСА ВАСИЛЬОВИЧА

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

Актуальність теми дослідження

Вивчення живої матерії було і залишається одним із найактуальніших напрямів сучасної науки, що фактично сформувало окрему індустрію: про це свідчать щорічні багатомільярдні фінансові вкладення у цю галузь у всьому світі. Одним із ключових напрямів у цій царині є дослідження взаємодії живої матерії (зокрема людського організму) з різними речовинами, включно зі штучно синтезованими, що є необхідною передумовою розробки більш ефективних лікарських засобів.

Процедура розробки нового лікарського засобу відбувається у кілька етапів. Завершальний етап - клінічні дослідження - є не лише довготривалим, але й характеризується малою вибіркою пацієнтів. Попередній етап - експериментальні біологічні дослідження - також тривалий і потребує вартісних процедур синтезу досліджуваних сполук. Отже, існує потреба максимально скоротити кількість хімічних речовин (включно з новоствореними), що відбираються для подальшого аналізу.

Саме цю задачу розв'язує моделювання взаємодії біологічної тканини, яка розглядається на рівні біомолекул, з молекулами досліджуваного препарату. Сучасні методи розрахунку такої взаємодії, зокрема молекулярна динаміка, дають змогу проводити моделювання з високою точністю порівняно з експериментом. Проте далекодійний характер кулонівської взаємодії змушує для досягнення прийнятної точності враховувати порядку десятків тисяч атомів навколишнього середовища, що потребує значних обчислювальних ресурсів і часу розрахунків.

Через це виникає потреба у суттєвому скороченні обсягу наявної інформації на основі її первинного аналізу щодо відповідності певним

критеріям. Оскільки обсяг наявної інформації дуже великий, а кількість критеріїв значна, застосування штучних нейронних мереж у поєднанні з методами машинного навчання для розв'язання цієї проблеми видається цілком доречним і, до певної міри, визначальним, оскільки формує вибірку для подальшого поглибленого аналізу. Класичні обчислювальні підходи, зокрема молекулярний докінг, значною мірою досягли плато своєї практичної точності, що додатково обґрунтовує актуальність розробки нових методів на основі машинного навчання. З огляду на це, тема дисертаційної роботи Т. В. Войціцького, спрямована на розробку та оптимізацію таких методів, є актуальною науковою проблемою фізики біологічних систем, а її актуальність не викликає сумніву.

Зв'язок роботи з науковими програмами, планами, темами

Дисертаційна робота виконана у відділі фізики біологічних систем Інституту фізики Національної академії наук України в межах планових науково-дослідних тем і грантових проєктів:

- 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».

Таким чином, результати дисертаційної роботи безпосередньо пов'язані з виконанням державних науково-дослідних тем і сучасних грантових проєктів, що підтверджує її інтеграцію у пріоритетні напрями фундаментальних і прикладних досліджень.

Рівень виконання поставленого наукового завдання та оволодіння методологією наукової діяльності

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 поставленої мети здобувач сформулював і послідовно розв'язав такі завдання:

- розробку методу передбачення афінності зв'язування малих молекул із протеїнами на основі штучних нейронних мереж із використанням структурної інформації про протеїн і ліганд, який перевершував би наявні аналоги за точністю;
- проведення статистичного аналізу фізико-хімічних параметрів основних типів нековалентних взаємодій у відомих експериментальних комплексах протеїн - ліганд;
- розробку алгоритму генерації модельних систем, які імітують комплекси протеїн - ліганд на рівні сайту зв'язування та відтворюють фізико-хімічні закономірності реальних комплексів, для збільшення обсягу тренувальних даних моделей молекулярного докінгу;
- розробку методу високопродуктивного молекулярного докінгу на основі штучних нейронних мереж, який перевершував би класичні підходи та інші методи машинного навчання за точністю та швидкістю.

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

Методологія дослідження ґрунтується на послідовному поєднанні методів обчислювальної біофізики (для встановлення фізичної природи молекулярного розпізнавання, аналізу нековалентних взаємодій та обґрунтування вибору репрезентацій вхідних даних, архітектур моделей і їхніх параметрів) з методами чисельного комп'ютерного моделювання (для реалізації розроблених теоретичних підходів у вигляді алгоритмів і моделей машинного навчання). Такий підхід свідчить про належне оволодіння здобувачем сучасною методологією наукової діяльності: постановка задач є фізично обґрунтованою,

вибір інструментів - методологічно виваженим, а отримані результати систематично зіставлені з даними, отриманими іншими сучасними методами. Усі поставлені завдання виконано у повному обсязі, а мету роботи досягнуто.

Наукова новизна отриманих результатів

Наукова новизна дисертаційного дослідження полягає в отриманні таких основних результатів:

- уперше 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; розроблену на цій основі графову нейронну мережу 3DProtDTA підтверджено як точнішу за наявні підходи машинного навчання на 1-20 %;
- уперше запропоновано підхід до збільшення обсягу 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 підвищує частку правильних передбачень до 10 %;
- уперше створено 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 сукупними критеріями RMSD і хімічної коректності порівняно з іншими підходами машинного навчання.

Наведені результати є новими, мають самостійний науковий характер і повністю відповідають темі дисертації.

Ступінь обґрунтованості та достовірності наукових положень, висновків і рекомендацій

Обґрунтованість і достовірність отриманих результатів забезпечуються коректною постановкою задач, використанням сучасного та адекватного меті роботи математичного апарату машинного навчання, а також застосуванням загальновизнаних наборів даних і критеріїв оцінювання. Розроблені моделі систематично тестувалися на незалежних наборах даних, а їхня ефективність кількісно порівнювалася з провідними класичними методами та сучасними інструментами машинного навчання.

Наукові положення та висновки сформульовані на підставі коректних порівнянь із сучасними літературними даними та узгоджуються з висновками, зробленими в кожному з розділів дисертації. Це забезпечує належний рівень обґрунтованості результатів, їх відтворюваність і відповідність сучасним науковим уявленням. Достовірність підтверджується також тим, що частина результатів уже впроваджена у практику - вони стали основою програмного забезпечення, що застосовується у проєктах віртуального скринінгу.

Наукове та практичне значення отриманих результатів

Наукове значення результатів полягає у поглибленні розуміння фізичної природи нековалентних взаємодій у системах протеїн - мала органічна молекула та у створенні фізично обґрунтованих методів їх моделювання на основі машинного навчання, що є внеском у розвиток фізики біологічних систем і обчислювальної біофізики.

Практичне значення полягає в тому, що розроблені підходи дають змогу прискорити процес *in silico* пошуку нових лікарських засобів (віртуальний скринінг) завдяки швидкому та точному передбаченню афінності й геометрії зв'язування між терапевтичними мішенями (протеїнами) та потенційними ліками (малими органічними молекулами). Запропонована методологія генерації штучних комплексів дозволяє підвищити точність моделей машинного навчання для нових класів протеїнів і лігандів, для яких відсутні експериментально визначені дані. Розроблені у межах дисертації програмні продукти 3DProtDTA та ArtiDock здатні не лише аналізувати взаємодію між різними комбінаціями молекул, але й прогнозувати найбільш вірогідні комбінації для досягнення поставленої мети, що дозволяє скоротити час і витрати на розробку нових лікарських засобів.

Повнота викладу результатів у публікаціях та особистий внесок здобувача

Основні результати дисертації повно й достатньою мірою висвітлені у наукових публікаціях здобувача. За темою дисертації опубліковано 3 статті у наукових журналах, що індексуються у базах даних Scopus і Web of Science (у всіх здобувач є першим автором), а також оприлюднено результати у 5 доповідях на міжнародних і всеукраїнських наукових конференціях.

Кожен із розділів 2, 3 та 4 дисертації відповідає окремій опублікованій науковій праці. Особистий внесок здобувача є визначальним і полягає у розробці графової нейронної мережі 3DProtDTA (розділ 2), алгоритму генерації штучних комплексів протеїн - ліганд (розділ 3) та методу молекулярного докінгу ArtiDock (розділ 4); у реалізації відповідних програмних модулів попередньої обробки даних, генерації ознак, підбору гіперпараметрів, тренування й тестування моделей; у систематичному порівнянні з наявними підходами; а також у написанні текстів публікацій і підготовці таблиць та рисунків. Таким чином, дисертація відображає значний особистий внесок здобувача на всіх етапах дослідження - від постановки задачі до апробації результатів.

Загальна оцінка дисертації, її структура та оформлення

Дисертація загалом справляє гарне враження. У ній детально описано сучасний стан проблеми, чітко поставлено мету, сформульовано відповідні критерії, послідовно та ґрунтовно викладено шляхи досягнення мети. Отримані результати зіставлено з результатами інших сучасних методів аналізу; поліпшення ефективності на 10-20 % порівняно з наявними методами можна вважати вагомим з огляду на економію обчислювальних ресурсів і коштів у цій царині. Дисертація складається зі вступу, чотирьох розділів, висновків і додатків; повний обсяг становить 197 сторінок, робота містить 28 рисунків, 23 таблиці та список використаних джерел зі 185 найменувань. Структура роботи є логічною і повністю відповідає її меті та завданням.

Зауваження та дискусійні положення

Разом з тим до дисертаційної роботи є низка зауважень, що переважно стосуються тексту та оформлення рукопису:

1. У тексті наявне засилля англомовної термінології (докінг, афінність, стекінг, коефіцієнт конкордації, пулінг та інші), що ускладнює читання

рукопису. Деякі з цих слів мають усталений український відповідник (афінність - спорідненість, конкордація - відповідність), інші є багатозначними, і автор сам наводить їх тлумачення українською (mean pooling - усереднення, max pooling - вибір максимуму, add pooling - сумування, graph pooling - графовий пулінг), проте далі послуговується «англо-вимовним» варіантом. Зокрема, ключове слово «докінг» варто було б замінити на «швартування» - термін хоч і теж іноземного походження, але зрозумілий і достатньо точно описує процес. Це передусім справа мовних спеціалістів, однак без залучення фізиків, які мають стимулювати українізацію фізичного лексикону, унормування термінології не відбудеться.

2. З попереднім тісно пов'язане непослідовне використання скорочень латиницею та кирилицею. Наприклад, доцільно застосовувати стандартне скорочення ЯМР замість NMR та МН замість ML, тим більше, що автор уже використовує МД (молекулярна динаміка) та ІТК (ізотермічна титраційна калориметрія). Виняток становлять скорочення, що застосовуються для індексування параметрів у формулах.
3. У тексті трапляється відсутність потрібних розділових знаків (ком, тире); рукопис потребує редагування синтаксису.

Наведені зауваження мають переважно редакційний характер, не стосуються суті отриманих наукових результатів і не знижують загального рівня дисертаційної роботи. Вони можуть бути враховані автором у подальшій науковій роботі та публікаціях.

Загальний висновок

Дисертаційна робота Войціцького Тараса Васильовича «Визначення параметрів нековалентних взаємодій малих органічних молекул із протеїнами на основі моделей машинного навчання» є завершеною самостійною науковою працею, яка робить вагомий внесок у заявлену тематику. У роботі отримано нові науково обґрунтовані результати, що в сукупності розв'язують актуальне наукове завдання фізики біологічних систем. Отримані результати дають змогу значно скоротити час і витрати на розробку новітніх лікарських засобів, а розроблені програмні продукти дозволяють здійснювати прогнозування з метою створення необхідних речовин для досягнення бажаного ефекту; частина результатів уже впроваджена у практику.

За актуальністю, науковою новизною, обґрунтованістю й достовірністю результатів, а також їх науковим і практичним значенням дисертація відповідає вимогам наказу МОН України № 40 від 12.01.2017 р. «Про затвердження Вимог до оформлення дисертації» та постанови Кабінету Міністрів України від 12.01.2022 р. № 44 «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 постановами КМ № 341 від 21.03.2022 р., № 502 від 19.05.2023 р. та № 507 від 03.05.2024 р.

Автор дисертаційної роботи, Войціцький Тарас Васильович, заслуговує на присудження наукового ступеня доктора філософії з галузі знань 10 - Природничі науки за спеціальністю 104 - Фізика та астрономія.

Головний науковий співробітник
відділу фізики кристалів
Інституту фізики НАН України,
доктор фізико-математичних наук
Войтенко Олександр Іванович