А.Д. ТЕРЕЩЬ,[1,2] Т.Ю. НІКОЛАЄНКО[1]

[1] Київський національний університет імені Тараса Шевченка
(Вул. Володимирська, 64/13, Київ 01601)
[2] Інститут хімії поверхні ім. О.О. Чуйка НАН України
(Вул. Генерала Наумова, 17, Київ 03164)

МОДЕЛЬ ДЛЯ ПАРАМЕТРИЗАЦІЇ ПОТЕНЦІАЛІВ МІЖАТОМНОЇ ВЗАЄМОДІЇ ЗА КВАНТОВО-МЕХАНІЧНИМИ ДЕСКРИПТОРАМИ НА ОСНОВІ ГРАФОВОЇ НЕЙРОННОЇ МЕРЕЖІ

УДК 539.6

На основі графових нейронних мереж розроблено модель машинного навчання для передбачення повної енергії біомолекул за їхньою структурною формулою та квантово-механічними дескрипторами шляхом прогнозування параметрів функцій, що апроксимують міжатомні потенціали. Доведено застосовність створеної моделі для передбачення повної енергії біомолекул, енергії їхньої взаємодії, а також для прогнозування впорядкування конформерів за енергіями. Показано фізичну обґрунтованість отриманих параметрів, що відкриває можливості для подальшого використання моделі в задачах молекулярного моделювання.

Ключові слова: машинне навчання, потенціали міжатомної взаємодії, конформери біомолекул, квантово-механічні дескриптори, нейронні мережі, зв'язування біомолекул.

1. Вступ

Обчислення енергії системи атомів є необхідним етапом у фізичних методах комп'ютерного моделювання структури й динаміки як ізольованих молекул, так і конденсованого середовища. У рамках квантово-хімічного підходу [1], таке обчислення виконується на основі розв'язку квантово-механічної задачі про знаходження стану електронної підсистеми із заданими координатами досліджуваних атомів. Попри високу надійність, такі методи потребують числового розв'язання рівнянь, які описують електронну структуру (наприклад, рівняння Гартрі–Фока чи рівняння теорії функціоналу густини), на кожному кроці основного моделювання, що потребує значних затрат обчислювальних ресурсів. Це накладає практичне обмеження на їхнє використання для моделювання систем, зокрема – біомолекулярних, які мають значну кількість атомів та/або потребують спільного аналізу значної кількості конфігурацій методами статистичної фізики. Зменшення обчислювальної складності можливе шляхом апроксимації енергії системи як функції координат її атомів, зокрема – сумою атом-атомних потенціальних функцій. Такий підхід, відомий як метод силових полів (force fields), є найменш обчислювально затратним і лежить в основі поширених реалізацій методу молекулярної динаміки [2], докінгу [3] методів пошуку структур атомних кластерів [4–6] і моделювання структури білків [7, 8].

Хоча заміна розв'язку складних квантово-механічних рівнянь наближеним, але обчислювально ефективним класичним представленням силових полів, уможливлює моделювання систем на масштабах часу й розмірів, недосяжних для прямих квантово-механічних розрахунків, надійність і ефективність такого моделювання повністю залежить від коректності обраних міжатомних потенціалів. Це передбачає як вибір коректної функціональної форми для їхнього представлення (зокрема, із коректною асимптотикою на великих міжатомних відстанях), так і належне калібрування параметрів обраних функцій.

Цитування: Терещь А.Д., Ніколаєнко Т.Ю. Модель для параметризації потенціалів міжатомної взаємодії за квантово-механічними дескрипторами на основі графової нейронної мережі. *Укр. фіз. журн.* **71**, №7, 577 (2026).

Окрім традиційних підходів [9], коли функціональна форма потенціалу є наперед заданою, міжатомні потенціали можна апроксимувати за допомогою методів машинного навчання (МН). Методи МН, які здатні апроксимувати функціональну залежність на основі набору прикладів аргументів і значень функції, успішно застосовуються в хімії для прогнозування шляхів реакцій [10, 11], енергій збуджених станів [12, 13], енергій утворення [14, 15] і пошуку нових хімічних сполук [16–18]. Останнім часом, було показано [19], що МН-потенціали, в яких методами МН представляються потенціали міжатомної взаємодії, дають змогу передбачати енергію молекул з точністю квантово-механічних розрахунків, витрачаючи набагато менше обчислювальних ресурсів. З обчислювальної точки зору, в основі сучасних МН-моделей, як правило, лежать нейронні мережі. Представниками таких моделей є Grappa [20], TorchANI [21] і TorchMD-Net [22].

Важливо, що підхід до представлення міжатомних потенціалів на основі МН-моделей дає змогу не лише розв'язати задачу апроксимації як таку, але й водночас уникнути необхідності калібрування параметрів у цих потенціалах. Для цього вирішують задачу пошуку МН-моделі, яка дає змогу представити енергію системи атомів як функцію не лише міжмолекулярних відстаней, але й структури самих молекул. Ключовою передумовою для цього є вибір інформативних дескрипторів для представлення структури молекул, який полегшує параметризацію ("тренування") моделей і забезпечує їхню узагальнювальну здатність. Відповідні дескриптори як числові представлення молекулярної структури мають бути інваріантними до обертань і трансляцій молекули, правильно враховувати перестановку однакових атомів і унікально описувати конфігурацію молекули.

Однак, попри універсальність МН-моделей, заснованих на використанні нейронних мереж як засобу апроксимації енергії системи атомів, їхнє застосування в методах моделювання вимагає більшого обсягу обчислень, ніж метод класичних силових полів. Крім того, представлення даних у вигляді абстрактних багатокомпонентних "векторів", яким оперують нейронні мережі в проміжних обчисленнях, є складними для інтерпретації і не дає змогу якісно простежити вплив структури системи на її енергію. Це може призводити, зокрема, до того, що модель вивчає статистичні кореляції замість причинно-наслідкових зв'язків і, як наслідок, – до ефекту "компенсації похибок", коли, наприклад, вплив двох різних функціональних груп на властивості молекули співвідноситься з цими групами довільним чином, якщо такі групи переважно зустрічаються разом у наборі тренувальних даних моделі й моделі фактично достатньо апроксимувати лише суму внесків цих груп.

На відміну від існуючих підходів, у даній роботі аналізується можливість двостадійної побудови МН-моделі силових полів на основі нейронних мереж. На першій стадії застосовуються дескриптори, які одержані з квантово-хімічних розрахунків за допомогою теореми віріала й асоційовані з окремими атомами або парами атомів, а не з молекулою в цілому. Це дає змогу зменшити ефект компенсації похибок, характерний для прямої апроксимації повної енергії, і водночас оцінити ефективність "перенесення знань" (transfer learning) [23] у параметризації міжатомних потенціалів. Крім того, на відміну від численних нейромережевих моделей, що безпосередньо апроксимують повну енергію або потенціал молекули, наш підхід спрямований на передбачення параметрів класичних міжатомних потенціалів, які потім використовуються для обчислення енергії.

2. Теорема віріала і її застосування для побудови квантово-механічних дескрипторів

Одним із основних елементів моделі, що пропонується в даній роботі, є використання квантово-механічних дескрипторів, заснованих на фундаментальному зв'язку між кінетичною енергією електронів молекули і її енергією як розв'язком стаціонарного рівняння Шредингера, який встановлюється теоремою віріала.

Згідно з теоремою віріала, можемо записати середнє значення оператора кінетичної енергії електронів у системі через повну енергію системи [24]:

$$\langle \hat{T}_e \rangle = -E. \tag{1}$$

Представимо це значення через хвильову функцію Ψ:

$$\langle \Psi | \hat{T}_e | \Psi \rangle = \sum_i -\frac{\hbar^2}{2m_e} \langle \Psi | \Delta_{r_i} | \Psi \rangle =$$

$$= -N_e \frac{\hbar^2}{2m_e} \langle \Psi | \Delta_{r_1} | \Psi \rangle, \tag{2}$$

де $\hat{T}_e$ – оператор кінетичної енергії електронів, $\hbar$ – приведена стала Планка, m_e – маса електрона, N_e – кількість електронів у системі, а Δ_{r_i} – оператор Лапласа за координатою r_i. Його квантове середнє з багаточастинковою хвильовою функцією електронної підсистеми Ψ молекули є

$$\langle\Psi|\Delta_{r_1}|\Psi\rangle = \int dr_1\, dr_2 \ldots dr_N \times$$
$$\times\, \Psi^*(r_1, r_2, ..., r_N)\Delta_{r_1}\Psi(r_1, r_2, ..., r_N) \quad (3)$$

(тут і надалі підсумовування за спіновими ступенями вільності суміщено з операцією інтегрування за просторовими координатами задля компактності запису). Воно може бути виражене через приведену одночастинкову матрицю густини

$$\gamma(r, \tilde{r}) = N_e \int dr_2 \ldots dr_N \times$$
$$\times\, \Psi^*(r, r_2, ..., r_N)\Psi(\tilde{r}, r_2, ..., r_N). \quad (4)$$

Тоді, записавши кінетичну енергію через $\gamma(r, \tilde{r})$, отримаємо

$$-E = \langle\hat{T}_e\rangle = -N_e\frac{\hbar^2}{2m_e}\langle\Psi|\Delta_{r_1}|\Psi\rangle =$$
$$= -\frac{\hbar^2}{2m_e}\int (\Delta_{\tilde{r}}\gamma(r, \tilde{r}))_{\tilde{r}=r}\, dr. \quad (5)$$

Представимо тепер матрицю густини розвиненнями за базисними функціями $\chi_\mu(r)$, як це зазвичай робиться при числовому розв'язуванні рівнянь квантово-хімічними методами [25]:

$$\gamma(r, \tilde{r}) = \sum_\mu \sum_\nu D_{\mu\nu}\chi_\mu(r)\chi_\nu(\tilde{r}), \quad (6)$$

де коефіцієнти $D_{\mu\nu}$ утворюють так звану електронну матрицю густини, яка містить необхідну інформацію про електронну структуру системи. Підставляючи це розвинення у вираз для енергії, отримаємо:

$$-E = \langle\hat{T}_e\rangle = -\frac{\hbar^2}{2m_e}\sum_\mu \sum_\nu D_{\mu\nu} \times$$
$$\times \int \chi_\mu(r)\,(\Delta_{\tilde{r}}\chi_\nu(\tilde{r}))_{\tilde{r}=r}\, dr. \quad (7)$$

Для подальшого тут зручно ввести матричний елемент $K_{\mu\nu}$ кінетичної енергії для базисних функцій:

$$K_{\mu\nu} = -\frac{\hbar^2}{2m_e}\int \chi_\mu(r)\,(\Delta_{\tilde{r}}\chi_\nu(\tilde{r}))_{\tilde{r}=r}\, dr. \quad (8)$$

Таким чином, квантове середнє кінетичної енергії можна записати у вигляді:

$$-E = \langle\hat{T}_e\rangle = \sum_\mu \sum_\nu D_{\mu\nu}K_{\mu\nu} = \mathrm{Tr}(D \cdot K), \quad (9)$$

де $\mathrm{tr}(D \cdot K)$ – слід матриці, що є скалярним значенням. За значення кінетичної енергії електронів у цій системі приймалася така енергія у допоміжній системі Кона–Шема, що, строго кажучи, робить твердження теореми віріала для повної енергії наближеним.

Далі розділимо суму (9) за окремими атомами, щоб виокремити внески кожного з них (аналогічно введенню зарядів атомів за Малікеном [25]):

$$-E = \langle\hat{T}_e\rangle = \sum_A \sum_B \left(\sum_{\mu\in A}\sum_{\nu\in B} D_{\mu\nu}K_{\mu\nu}\right), \quad (10)$$

де A і B – індекси атомів.

Наведений зв'язок між матрицями $D_{\mu\nu}$ і $K_{\mu\nu}$ та повною енергією системи дає підстави очікувати, що нейронна мережа, натренована на внесках окремих атомів або пар атомів (зокрема, хімічних зв'язків), представлених у цих матрицях, зможе формувати інформативні вектори ознак. Ці вектори, у свою чергу, можуть бути використані іншими моделями для точнішого прогнозування енергетичних характеристик молекул.

3. Характеристики й підготовка використаних наборів даних

Необхідною передумовою для тренування нейронної мережі є наявність достатньо великого, різноманітного, репрезентативного й збалансованого набору даних, зокрема для задач, пов'язаних із моделюванням властивостей молекул. У даній роботі використано набір QMugs [26], який є підмножиною бази ChEMBL [27] і містить лише біологічно й фармакологічно релевантні молекули. Саме такий підхід дає змогу налаштовувати модель на прогнозування властивостей молекул, подібних до лікарських засобів, що підвищує її практичну цінність.

Оригінальний набір даних QMugs містив близько 665000 молекул. З огляду на обмеження обчислювальних ресурсів, було обрано підмножину з 304331 молекул. Ці дані далі було поділено на три піднабори: 19814 молекул (6,5%), 56904 молекули

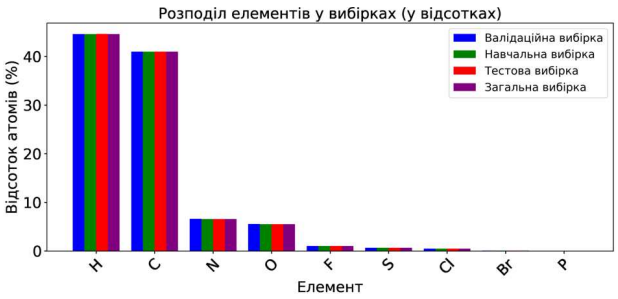


Рис. 1. Розподіл за кількості атомів різних хімічних елементів для молекул тренувальної, валідаційної та тестової вибірок

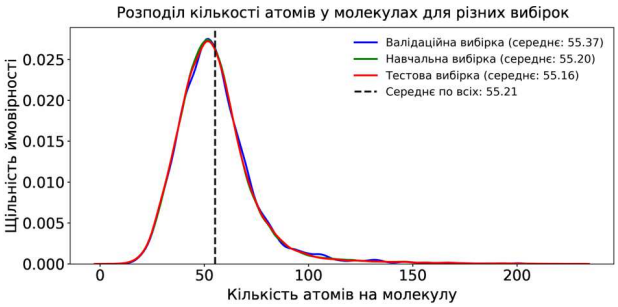


Рис. 2. Розподіл за кількістю атомів у молекулі для тренувальної, валідаційної та тестової вибірок

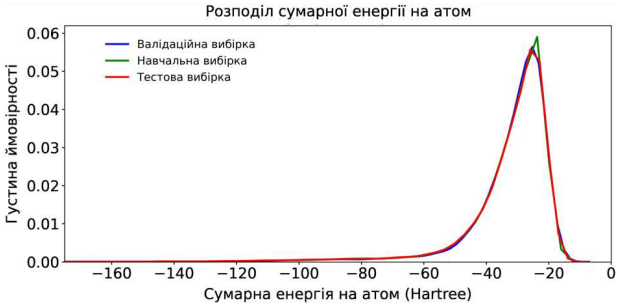


Рис. 3. Розподіл за повною енергією, яка припадає на один атом, для молекул із тренувальної, валідаційної та тестової вибірок

(18,7%) і 227613 молекул (74,8%), які використовувалися, відповідно, як валідаційна, тестова й тренувальна вибірки.

У рамках попереднього аналізу проведено статистичне дослідження ключових характеристик набору даних. Було проаналізовано розподіл молекул за хімічними елементами – H, C, N, O, F, S, Cl, Br та P (рис. 1), за кількістю атомів у молекулах (рис. 2) та за енергією на атом (рис. 3). Подібність одержаних розподілів для валідаційної, тестової та тренувальної вибірок підтверджує коректність виконаного розбиття, що є необхідною умовою для побудови надійних і узагальнювальних МН-моделей.

Для кожної молекули нами за допомогою програми Psi4 [28] було пораховано матричні елементи кінетичної енергії за формулою (10), використовуючи той самий набір базисних функцій Def2-SVP, що і при створенні початкового набору даних QMugs.

Оскільки кінцевою метою є побудова відображення структури молекули на параметри силових полів, необхідно представити молекулярну структуру у вигляді, придатному для методів машинного навчання. У таких задачах молекулу часто представляють у вигляді графа [29], що дає змогу ефективно враховувати їхню структуру й властивості при розв'язанні задач машинного навчання. У цьому підході молекула описується графом, де вузли відповідають атомам, а ребра – хімічним зв'язкам чи нековалентним взаємодіям між атомами. У створеній у цій роботі МН-моделі, кожен вузол графа (атом) містить інформацію про хімічний елемент атома, представлену як вектор із єдиним ненульовим компонентом, номер якого і визначає хімічний елемент атома (так зване one-hot кодування).

Одним із етапів формування графа є визначення хімічних зв'язків між атомами. На основі даних про структуру молекули визначається кратність кожного зв'язку (одинарний, подвійний тощо). Додатково враховуються просторові "взаємодії": обчислюється відстань між кожною парою атомів, і якщо вона перевищує встановлений поріг, взаємодія виключається. Для атомів, що перебувають у межах цього порогу, зв'язок класифікується як нековалентний.

Ребра графа описуються за допомогою набору параметрів. Серед них – координати атомів, тип зв'язку – ковалентний (з урахуванням кратності зв'язку) або нековалентний. Завдяки цьому граф зберігає інформацію як про ковалентні, так і нековалентні взаємодії в молекулі.

На основі цих даних граф створюється за допомогою бібліотеки DGL [30] (Deep Graph Library). Він зберігає як структурну, так і фізико-хімічну інформацію про молекулу, що робить його зручним для подальшого аналізу методами МН.

4. Архітектури створених нейромереж

Створена МН-модель містила в собі дві нейронні мережі, які відповідали запропонованому двоетапному підходу до передбачення енергії системи атомів як функції їхніх координат.

Метою першої з них було кодування конфігурації хімічного оточення кожного із атомів і зв'язків у вигляді векторів заданої розмірності – так званих ембедингів. Для її реалізації було використано підхід графових нейронних мереж (Graph Neural Network, GNN), відповідно до обраного способу представлення структури молекул. З-поміж існуючих різновидів GNN, було обрано архітектуру Message Passing Neural Network (MPNN) [31]. Докладніше архітектура й особливості тренування цієї мережі описані далі (див. 4.1). Для реалізації усіх моделей МН у цій роботі був використаний фреймворк PyTorch [32].

Після аналізу структурної формули молекули за допомогою створеної MPNN, для атомів і пар атомів було створено окремі повнозв'язні нейронні мережі для передбачення параметрів міжатомних потенціалів. Оскільки інформація про міжатомні відстані не використовувалася згаданими мережами, то параметри таких потенціалів у запропонованому підході передбачаються на основі лише структурних формул молекул. Натомість, міжатомні відстані впливають на одержувані енергії молекул лише через вирази для міжатомних потенціалів. Ці вирази докладніше описані далі в розділі 4.2.

4.1. Графова нейромережа для створення ембедингів

Для побудови ембедингів як векторних представлень локального оточення атомів і зв'язків у молекулі було розроблено графову нейронну мережу архітектури MPNN, яка складалася з 4 кроків передачі повідомлень і прихованого шару розміром 350 нейронів. Для тренування цієї мережі її було застосовано як передбачувач внесків елементів матриць кінетичної енергії та густини електронів (див. (10)), асоційованих із окремими атомами

$$\mathrm{KD}_i^{\mathrm{atom}} = \sum_{\mu,\nu\in i} K_{\mu\nu} D_{\mu\nu}, \tag{11}$$

і хімічними зв'язками

$$\mathrm{KD}_{ij}^{\mathrm{edge}} = \sum_{\mu\in i}\sum_{\nu\in j} K_{\mu\nu} D_{\mu\nu}, \tag{12}$$

молекули до загальної кінетичної енергії її електронної підсистеми. На етапі тренування MPNN використовувалися допоміжні доповнення до її архітектури у вигляді двох окремих повнозв'язних нейромереж – однієї для атомів ($\mathrm{MLP}_{\mathrm{atom}}$) та іншої для зв'язків ($\mathrm{MLP}_{\mathrm{edge}}$). А саме: на кожній ітерації процесу тренування спочатку для кожного атома i за допомогою MPNN обчислювався ембединг h_i як

$$h_i = \mathrm{MPNN}\big(x_i, \{(x_j, e_{ij}) \mid j \in \mathcal{N}(i)\}\big), \tag{13}$$

де x_i – вектор вхідних ознак атома, e_{ij} – вектор ознак зв'язку між атомами i та j, а $\mathcal{N}(i)$ – множина сусідніх атомів (з'єднаних із даним хімічними зв'язками). Після цього допоміжна повнозв'язна мережа $\mathrm{MLP}_{\mathrm{atom}}$ приймає на вхід отриманий ембединг h_i і обчислює відповідний внесок для атомів

$$\widehat{\mathrm{KD}}_i^{\mathrm{atom}} = \mathrm{MLP}_{\mathrm{atom}}(h_i). \tag{14}$$

Мережа для атомів складалася з трьох лінійних шарів з активаціями LeakyReLU з шириною прихованого шару 160 нейронів.

Мережа $\mathrm{MLP}_{\mathrm{edge}}$ для зв'язків приймає на вхід суму ембедингів двох атомів, які утворюють ребро графа молекули, сконкатеновану із вектором ознак зв'язку e_{ij}:

$$\widehat{\mathrm{KD}}_{ij}^{\mathrm{edge}} = \mathrm{MLP}_{\mathrm{edge}}\left((h_i + h_j) \parallel e_{ij}\right), \tag{15}$$

де $\parallel$ означає операцію конкатенації векторів. Архітектура $\mathrm{MLP}_{\mathrm{edge}}$ складалася з чотирьох лінійних шарів з активаціями LeakyReLU між ними з шириною прихованого шару 199 нейронів. У тренуванні MPNN враховувалися лише ковалентні зв'язки, а внеском нековалентних взаємодій нехтували.

Після того, як мережі передбачали відповідні внески $\widehat{\mathrm{KD}}_i^{\mathrm{atom}}$ на рівні атомів і зв'язків $\widehat{\mathrm{KD}}_{ij}^{\mathrm{edge}}$, їхні суми порівнювалися з еталонними значеннями, розрахованими квантово-хімічним методом. Таким чином, функція втрат задавалася як середня абсолютна похибка (MAE) як

$$\mathrm{MAE}_{\mathrm{MPNN}} = \frac{1}{n}\sum_{M=1}^{n}\left|\sum_{i\in M}\widehat{\mathrm{KD}}_i^{\mathrm{atom}} + \sum_{(i,j)\in M}\widehat{\mathrm{KD}}_{ij}^{\mathrm{edge}} - KD_M^{\mathrm{true}}\right|, \tag{16}$$

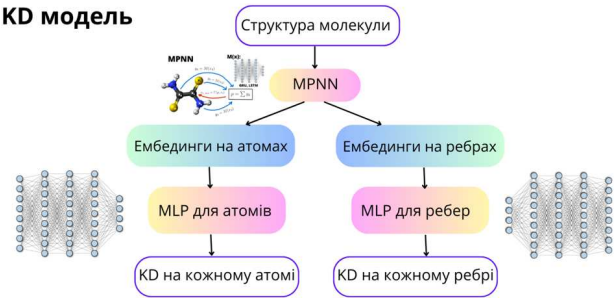


Рис. 4. Архітектура моделі для створення ембедингів

де KD_M^{true} – істинне значення, розраховане квантово-хімічними методами. Після цього виконувалося стандартне зворотне поширення похибки для одночасного оновлення ваг MPNN, MLP_{atom} та MLP_{edge} з метою мінімізації цієї похибки. Тренування моделі для створення ембедингів здійснювалося на відеокарті Nvidia GeForce GTX 1080, використовуючи алгоритм оптимізації Adam [33] і розмір батчу 64.

4.2. Повнозв'язна мережа для передбачення повної енергії молекули

Як результат застосування мережі MPNN, натренованої в описаний вище спосіб, для кожного атома обраних молекул було одержано ембединги h_i, які кодують хімічний "контекст" атома й визначаються структурною формулою молекули, але не залежать від її просторової структури. Наступним кроком було створення двох окремих повнозв'язних мереж, вхідними даними для яких були такі ембединги й вектори ознак ребер.

Так, одна із таких мереж (позначимо її як `charge_net`) приймає на вхід ембединги h_i й передбачає часткові заряди атомів:

$$Q_i = \texttt{charge\_net}(h_i). \tag{17}$$

Інша (позначимо її як `feats_net`) працює на рівні пар атомів. Для кожного ребра графа молекули вона приймає на вхід конкатенацію ембедингів атомів i і j

$$x_{ij} = [h_i \parallel h_j \parallel e_{ij}], \tag{18}$$

де символ $\parallel$ означає операцію послідовного об'єднання (конкатенації) елементів векторів, а також – вектор ознак ребра e_{ij}. За цим вектором, `feats_net` передбачає для кожної пари атомів у молекулі набір параметрів для силового поля:

$$C = \texttt{feats\_net}(x_{ij}) = [C_0, C_1, C_2, ..., C_{13}]. \tag{19}$$

Задаючи далі відстань між атомами як

$$r_{ij} = |R_i - R_j|, \tag{20}$$

далекодійна складова міжатомної взаємодії в запропонованому підході обчислюється на основі передбачених параметрів як

$$E_{ij}^{\text{LR}} = \frac{Q_i Q_j}{r_{ij}} - \frac{|C_0|}{r_{ij}^6} + \frac{|C_1|}{r_{ij}^{12}}, \tag{21}$$

а короткодійна складова – як

$$E_{ij}^{\text{SR}} = \sum_{k=1}^{6} C_{2k} \times \\ \times \exp\left(-\frac{\left(r_{ij} - (1.2 + 0.6(k-1))\right)^2}{C_{2k+1}^2 + \varepsilon}\right), \tag{22}$$

де $\varepsilon = 10^{-5}$ – доданок, введений для уникнення ділення на нуль у процесі тренування. Водночас (21) застосовується лише для пар атомів, між якими відсутній ковалентний зв'язок, тоді як для з'єднаних ним атомів приймається $E_{ij}^{\text{LR}} = 0$. Нарешті, загальна енергія взаємодії атомів i та j знаходиться як:

$$E_{ij} = E_{ij}^{\text{SR}} + E_{ij}^{\text{LR}}. \tag{23}$$

Передбачені внески повної енергії зчитуються з ребер і вузлів графа, для отримання повної енергії. Вона порівнюється з енергією E^{QC}, порахованою квантово-хімічним методом для цієї молекули, за метрикою MAE (середня абсолютна похибка). Щоб полегшити тренування нейронної мережі, її використовували для передбачення не безпосередньо повної енергії, а поправки до оцінки E^{lin} енергії, отриманої за допомогою лінійної регресії. Вхідними даними для такої регресії був вектор із кількостей атомів кожного хімічного елемента у молекулі:

$$E^{\text{lin}} = w^\top N + b = \sum_{\alpha=1}^{m} w_\alpha N_\alpha + b, \tag{24}$$

де $N = (N_1, N_2, ..., N_m)$ – вектор кількостей атомів кожного з m хімічних елементів, w – вектор коефіцієнтів лінійної регресії, b – її вільний член.

Таким чином, функція втрат при тренуванні й тестуванні мереж обчислювалася за формулою

$$\mathrm{MAE} = \frac{1}{n_{\mathrm{mols}}} \sum_{M=1}^{n_{\mathrm{mols}}} \left| \sum_{(i,j)\in M} E_{ij} + E_M^{\mathrm{lin}} - E_M^{QC} \right|,$$

де індекс M нумерує молекули вибірки, а запис $(i,j) \in M$ відповідає врахуванню в сумі внесків від усіх пар атомів M-ої молекули.

Задля уникнення явища перетренування нейронних мереж і щоб підвищити їхню стійкість до невеликих змін у просторових положеннях атомів, під час тренування мереж до координат атомів додавалися нормально розподілені випадкові величини ("гаусів шум") із середнім 0 і стандартним відхиленням $\sigma = 0{,}005$ Å.

При тренуванні моделей `charge_net` і `feats_net` для передбачення повної енергії молекули використовували алгоритм Adam із початковою швидкістю тренування 0,001 і планувальником зменшення швидкості тренування, який кожні 20 ітерацій ("епох") знижував її в 0,8 раза. Тренування моделей здійснювалося з використанням відеокарти Nvidia A100 й розміром батчу 400. Такий підхід дає змогу моделі ефективніше наближатися до мінімуму функції втрат на пізніших етапах тренування. Загалом тренування тривало 400 ітерацій.

Варто відзначити, що в процесі тренування моделей використовувалися лише окремі молекули, але не їхні комплекси. Це робить процес тренування швидшим, однак, загалом, переводить із режиму інтерполяції в режим екстраполяції застосування параметрів потенціалів міжатомної взаємодії (21) і (22), отриманих за (17) і (19), до найбільш практично важливого випадку міжмолекулярної взаємодії. Відтак, наведені нижче результати валідації моделей у цьому випадку показують їхню узагальнювальну здатність.

Окрім моделей `charge_net` та `feats_net` для передбачення повної енергії молекули, було створено також їхні аналоги для передбачення кінетичної енергії електронів молекули. Їх нееквівалентність моделям для передбачення повної енергії молекули пов'язана з наближеним характером теореми віріала в разі прийняття за кінетичну енергію електронів відповідного значення для допоміжної системи Кона–Шема у методі DFT. Разом із тим, оскільки за теоремою віріала середнє значення оператора кінетичної енергії з точністю до знака дорівнює повній енергії системи (1), для тренування на передбачення кінетичної енергії була використана аналогічна параметризація в архітектурі нейронної мережі зі зміною знака на протилежний перед сумою доданків E_{ij}^{LR} та $E_{ij}^{\mathrm{SR}} = 0$.

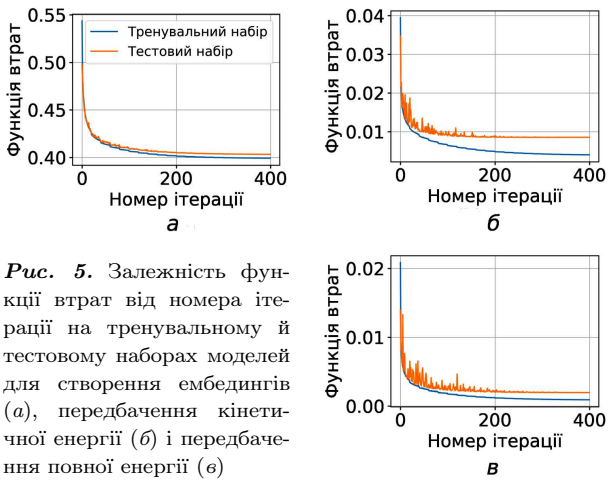


Рис. 5. Залежність функції втрат від номера ітерації на тренувальному й тестовому наборах моделей для створення ембедингів (*а*), передбачення кінетичної енергії (*б*) і передбачення повної енергії (*в*)

5. Результати та обговорення

5.1. Результати тренування й валідації моделей для передбачення повної енергії молекул

Динаміку зниження функції втрат при тренуванні моделі для створення ембедингів наведено на рис. 5, *а*, а моделей для передбачення кінетичної та повної енергії – на рис. 5, *б* і 5, *в* відповідно.

На тестовому наборі середнє значення похибки при передбаченні кінетичної енергії електронів і повної енергії молекул становило 6,1 ккал/моль і 2,2 ккал/моль, як показано на рис. 6 і 7 відповідно, що для випадку повної енергії є близьким до загальноприйнятого порогу хімічної точності (1 ккал/моль). Такий рівень точності вважається прийнятним для більшості завдань у обчислювальній хімії, зокрема для якісного відтворення термодинамічних характеристик і адекватного моделювання хімічних процесів.

З огляду на невелику похибку створених моделей при їхньому застосуванні до окремих молекул (у режимі інтерполяції), було перевірено їхню роботу в режимі екстраполяції – для типів задач, відмінних від тих, що ставилися при тренуванні

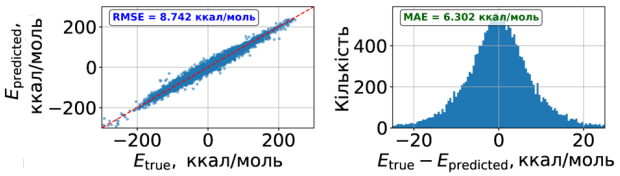


Рис. 6. Оцінка передбачень моделі для кінетичної енергії на валідаційному наборі даних: зіставлення істинних і передбачених значень (ккал/моль) (*а*); розподіл похибки у визначенні кінетичної енергії (ккал/моль), де E_{true} – референтні значення кінетичної енергії (ккал/моль), $E_{\text{predicted}}$ – передбачення моделі (*б*)

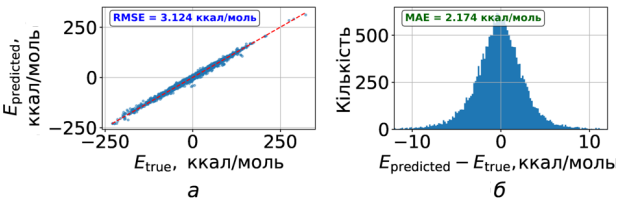


Рис. 7. Оцінка передбачень моделі для повної енергії на валідаційному наборі даних: зіставлення істинних і передбачених значень (ккал/моль) (*а*); розподіл похибки у визначенні повної енергії (ккал/моль), де E_{true} – референтні значення повної енергії (ккал/моль), $E_{\text{predicted}}$ – передбачення моделі (*б*)

Результати обчислення енергії міжмолекулярної взаємодії (ккал/моль) для пар основ AT і GC за допомогою створеної MH-моделі й інших методів

Метод	Пара AT	Пара GC
Референтне значення [34]	12,4	25,4
Створена модель	14,8	30,5
ANI	10,7	17,2
GFN2-xTB	16,1	29,2

моделей. Зокрема, було розглянуто передбачення енергії міжмолекулярної взаємодії – величини, яка не була включена до тренувальних даних. Таке застосування також відповідає більш практичному випадку, коли ключову роль відіграють не абсолютні значення енергії системи, а різниці енергій при зміні конфігурації системи.

5.2. Застосування моделей до оцінки енергій взаємодії в парах нуклеотидних основ ДНК

Енергія взаємодії між парами основ дезоксирибонуклеїнової кислоти (ДНК), – аденіну (A) й тиміну (T), гуаніну (G) й цитозину (C), – є важливою характеристикою, що впливає на міцність спарювання між її полімерними ланцюгами. Ці енергетичні характеристики є важливими для розуміння термодинамічних властивостей макромолекули ДНК, її температури плавлення, здатності до утворення стабільних вторинних структур, а також механізмів розриву або пошкодження ланцюгів в умовах різних фізичних і хімічних впливів. Крім того, оцінка енергії взаємодії як приклад застосування певної моделі потенціалу міжатомної взаємодії є важливою для побудови точних моделей молекулярної динаміки для біотехнологічних застосувань.

У парі аденін–тимін формуються два сильні водневі зв'язки, тоді як у парі гуанін–цитозин –три, що робить пару G–C загалом стабільнішою за A–T пару. Для пари A–T енергія взаємодії в газовій фазі зазвичай оцінюється на рівні приблизно 12,4 ккал/моль, тоді як для пари G–C вона складає близько 25,4 ккал/моль [34]. Значення енергії взаємодії можуть варіюватися залежно від обраного методу квантово-хімічних обчислень і конкретних умов моделювання.

Енергію взаємодії двох молекул із використанням розроблених моделей знаходили як

$$E_{\text{int}} = (E_{\text{mol 1}} + E_{\text{mol 2}}) - E_{\text{complex}}, \qquad (25)$$

де E_{int} – енергія міжмолекулярної взаємодії, E_{complex} – енергія повної системи (комплексу) із двох молекул, $E_{\text{mol 1}}$ і $E_{\text{mol 2}}$ – енергії ізольованих молекул з рівноважною геометрією. Оскільки для тренування MH-моделей використовувався набір молекул QMugs, рівноважні геометрії яких були знайдені напівемпіричним квантово-хімічним методом GFN2-xTB [35], цей самий метод було використано для знаходження геометрій молекулярного комплексу і його складових, з якими в рамках створених MH-моделей обчислювалися енергії E_{complex}, $E_{\text{mol 1}}$ і $E_{\text{mol 2}}$ у (25).

Для порівняння з іншими методами моделювання міжмолекулярної взаємодії, енергії E_{int} для пар A–T і G–C також були знайдені за допомогою нейронної мережі TorchANI і напівемпіричного методу GFN2-xTB. Отримані результати наведено у таблиці.

Отримані значення показують, що створена MH-модель демонструє для енергії взаємодії результати, співставні з іншими методами й досить близькі

до референтних значень: у випадку пар нуклеотидних основ ДНК аденін–тимін і гуанін–цитозин відхилення від відомих значень становлять відповідно 2,4 та 5,1 ккал/моль.

5.3. Прогнозування енергії взаємодії молекул у нековалентно-зв'язаних комплексах

Для повнішої оцінки точності створеної МН-моделі доцільно протестувати її на ширшому наборі молекул і отримати статистично більш обґрунтовану її характеристику для передбачення енергій взаємодії молекул. З цією метою було використано набір даних GMTKN55 [36], а саме – його піднабір S66 [37], який містить 66 хімічно репрезентативних димерів, утворених нековалентно зв'язаними молекулами. Енергії взаємодії в цьому наборі варіюються від 2,82 до 19,49 ккал/моль із середнім значенням 5,47 ккал/моль.

У цій роботі з набору S66 було використано лише геометрії димерів, які для подальшого аналізу було дооптимізовано методом GFN2-xTB. Після цього, енергії взаємодії молекул у комплексах обчислювалися методом DFT із використанням обмінно-кореляційного функціонала ωB97X-D [38] і базисного набору def2-SVP з урахуванням ефекту перекриття базисних функцій (Basis Set Superposition Error, BSSE [39]), який оцінювався та компенсувався за допомогою процедури *counterpoise correction*, реалізованої в програмі `Psi4`.

Хоча в оригінальній роботі [37] енергії взаємодії були отримані високоточним методом CCSD(T)/CBS [40], який вважається еталонним у квантово-хімічних розрахунках, застосований різновид ωB97X-D/def2-SVP методу DFT дав можливість отримати результати, близькі до референтних. Відповідне порівняння енергій представлено на рис. 8. Важливо, що цей самий різновид методу DFT використовувався і для обчислення повної енергії молекул при створенні тренувального набору для нейронних мереж. Така відповідність свідчить про доцільність і коректність вибору даного методу для нашого дослідження.

Одержані під час застосування розробленої МН-моделі результати для енергії міжмолекулярної взаємодії в розглянутих комплексах наведено на рис. 9. Із них можна зробити висновок, що створена модель здатна прогнозувати енергії взаємодії, що відповідають референтним значенням із

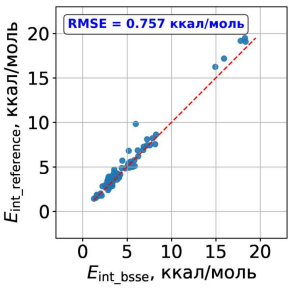
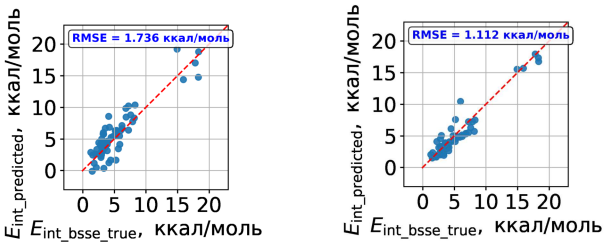


Рис. 8. Порівняння енергій взаємодії молекул E_{int_bsse}, знайденої методом DFT (ωB97X-D/def2-SVP), зі значеннями $E_{int\_reference}$ з оригінальної роботи [37]



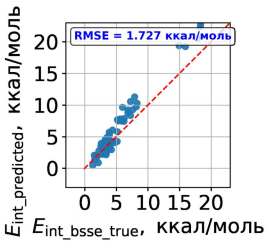


Рис. 9. Порівняння референтних значень $E_{int\_bsse\_true}$ енергій взаємодії молекул у комплексах набору S66 з відповідними енергіями $E_{int\_predicted}$, передбаченими створеною моделлю (*а*), методами GFN2-xTB (*б*) та ANI (*в*)

середньоквадратичною похибкою 1,7 ккал/моль. Тим самим, для задачі визначення енергії взаємодії молекул створена модель демонструє коректні результати незважаючи на те, що енергія взаємодії не була основною цільовою метрикою під час її тренування.

5.4. Прогнозування відносних енергій конформерів

Варіативність просторового розташування атомів у біомолекулах, таких як білки, нуклеїнові кислоти або ліганди, з незмінним графом хімічних зв'язків визначається конформацією молекул. Завдяки "гнучкості" своєї структури, біомолекули можуть набувати різних конформацій через обертання груп атомів навколо одинарних σ-зв'язків. Конформаційна різноманітність суттєво впливає на функціональні властивості молекул: можливість взаємодії з іншими молекулами й каталітичну активність [41]. Зокрема, альфа-спіраль і

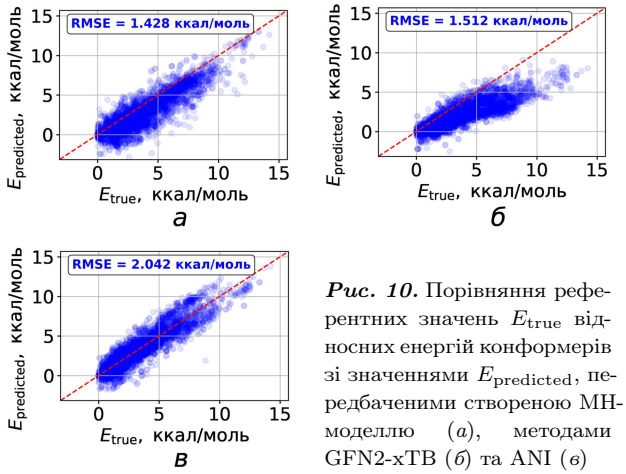


Рис. 10. Порівняння референтних значень E_{true} відносних енергій конформерів зі значеннями $E_{\text{predicted}}$, передбаченими створеною МН-моделлю (*а*), методами GFN2-xTB (*б*) та ANI (*в*)

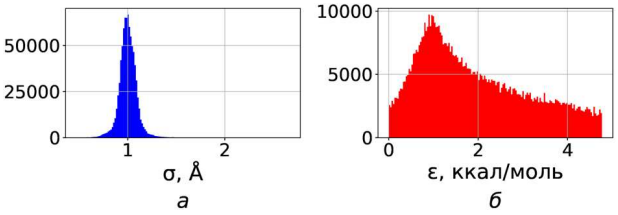


Рис. 11. Розподіли параметрів міжатомного потенціалу Леннард-Джонса, оцінених на основі передбачень створеної МН-моделі: характерної відстані мінімуму σ (*а*); глибини потенціальної ями ε (*б*)

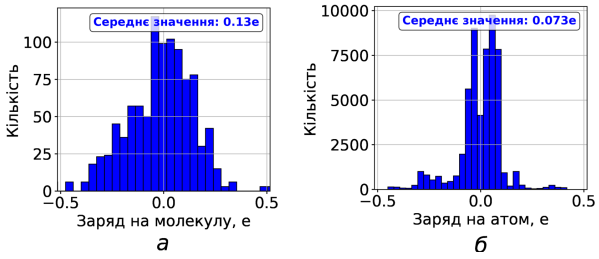


Рис. 12. Розподіл передбачених зарядів для молекул як цілого (*а*) і зарядів для окремих атомів (*б*)

бета-листи білків є різними способами згортання одного й того самого поліпептидного ланцюга. Аналізуючи експериментальні дані оберальної спектроскопії [42], а також спектри поглинання та комбінаційного розсіяння [43], важливо враховувати існування багатьох конформерів у складі зразка, адже під час вимірювання в газовій фазі внесок у загальний спектр даватимуть усі присутні конформери. Водночас, інтенсивність внеску різних конформерів буде пропорційною ймовірності їхнього існування в конкретних термодинамічних умовах. Ця ймовірність визначається відносною енергією кожного конформера згідно з розподілом Гіббса. Із цим пов'язана практична важливість методів оцінювання відносних енергій конформерів молекул.

Для аналізу застосовності розробленої МН-моделі для такої задачі, з набору даних PC9 [44] було відібрано підвибірку молекул, що містять 7–9 атомів (не рахуючи атомів H). Загалом було взято 117 молекул, для кожної з яких за допомогою програми CREST [45] було виконано пошук можливих конформацій. У сумі отримано 4253 конформерів. Далі геометрії молекул були дооптимізовані за допомогою методу GFN2-xTB. Після дооптимізації методом DFT ωB97X-D/def2-SVP за допомогою програми Psi4 були розраховані відносні енергії конформерів як різниця між енергією цієї конформації молекули та енергією її конформера із найменшою енергією.

Для порівняльного аналізу коректності передбачень створеної МН-моделі знайдено відносні енергії, отримані методом GFN2-xTB, а також за допомогою нейронної мережі ANI. Із наведених на рис. 10 результатів такого аналізу випливає, що створена модель показує хорошу кореляцію в задачі визначення відносних енергій конформерів із RMSE = 1,4 ккал/моль. Оскільки всі конформери однієї молекули мають однаковий граф і, відповідно, ті самі параметри силового поля, то такий тест додатково перевіряє коректність самих аналітичних формул, які поєднують міжатомні відстані з параметрами силового поля для обчислення енергії системи.

5.5. Аналіз фізичного змісту параметрів створеного міжатомного потенціалу

У створеному потенціалі (21)–(23) функціональна форма доданків (21), що моделюють далекодійні взаємодії, відповідає електростатичній взаємодії та потенціалу Леннард-Джонса $V(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$. Однак, її вибір автоматично не гарантує відповідність цих доданків вказаним фізичним механізмам міжатомної взаємодії, адже в процесі тренування МН-моделей можливі ефекти компенсації похибок між різними доданками. Щоб перевірити фізичну коректність отриманих параметрів у цих доданках окремо, було проведено їхній додатковий аналіз.

Потенціал Леннард-Джонса визначається двома величинами – σ, що відповідає відстані, на якій потенціал набуває мінімального значення, та ε, що характеризує глибину потенціальної ями. Ці параметри можна обчислити на основі отриманих коефіцієнтів моделі (23) за такими формулами:

$$\sigma = \frac{C_0}{C_1}, \quad \varepsilon = \frac{C_0^2}{4C_1}. \tag{26}$$

Побудовані розподіли одержаних у такий спосіб значень σ і ε показано на рис. 11.

Із одержаних розподілів видно, що за порядком величини, типові значення σ і ε є зіставними з відповідними параметрами, використовуваними в класичних силових полях, зокрема GROMOS [46], у якому типові значення σ становлять близько 3 Å, а ε – приблизно 0,13 ккал/моль. Це підтверджує фізичну обґрунтованість запропонованого підходу, використаного для тренування мережі `feats_net`.

На рис. 12 для вибірки, яка нараховує 1000 електронейтральних молекул з валідаційного набору даних наведено розподіл їхніх загальних зарядів, знайдених підсумовуванням Q_i із (17), передбачених мережею `charge_net` як зарядів окремих атомів у кожній з молекул. Хоча процес тренування мережі не накладав на Q_i жодних спеціальних обмежень, окрім як застосування цих величин у доданку, що відповідає закону Кулона, одержані результати свідчать, що значення загального заряду молекул містяться переважно в межах $\pm 0{,}25e$, що є допустимим для електронейтральних молекул. Таким чином, передбачені значення зарядів атомів не є довільними, а узгоджені між собою відповідно до структурної формули молекули. Сумарний заряд молекули $0{,}13e$ (де e – заряд електрона), рис. 12, *а*, а атома $0{,}073e$, рис. 12, *б*, співпадає за порядком величини з дробовими зарядами, які використовують для атомів в інших міжатомних потенціалах.

6. Висновки

Таким чином, запропоновано й реалізовано двоетапний підхід до тренування нейронних мереж для відтворення залежності енергії біомолекул від їхньої просторової структури й квантово-механічних дескрипторів. Продемонстровано, що за структурною формулою та складовими кінетичної енергії електронів молекули, знайденими методами теорії функціонала густини й локалізованими на атомах і парах атомів, можливо знайти вектори ознак, які є аналогічними до типів атомів у класичних силових полях методу молекулярної динаміки. Для цього створено графову нейромережеву модель для обчислення таких векторів ознак для усіх атомів довільної електронейтральної молекули, що складається з хімічних елементів H, C, N, O, F, S, Cl, Br і P. Запропоновано функціональну форму для потенціалу міжатомної взаємодії, параметри якого є функціями векторів ознак атомів, знайдених графовою нейромережевою моделлю. Для апроксимації такої функції створено модель на основі повнозв'язної нейронної мережі. Показано, що з її допомогою запропоновані потенціали дають змогу в наближенні попарної взаємодії визначати повну енергію молекул і кінетичну енергію її електронів із середнім абсолютним відхиленням 2,2 ккал/моль і 6,1 ккал/моль відповідно. Підтверджено можливість узагальнення створених моделей на молекули, які не містилися в тренувальній вибірці, а також – узагальнення на задачі, які виходять за межі прямого тренування. Зокрема, підтверджено застосовність потенціалів міжатомної взаємодії, знайдених створеними моделями на основі структурної формули молекул, для прогнозування енергії їхньої міжмолекулярної взаємодії. Для набору з 66 молекулярних комплексів S66 середньоквадратична похибка визначення такої енергії становить 1,7 ккал/моль. Показано, що створена двостадійна нейромережева модель дає змогу визначати відносні енергії конформерів біомолекул з набору PC9 із середньоквадратичною похибкою 1,4 ккал/моль. Виявлено, що параметри міжатомних потенціалів типу Леннард-Джонса, визначені створеними нейромережевими моделями, за порядком величини узгоджуються з аналогічними параметрами класичного силового поля GROMOS.

Автори щиро вдячні Єнському університету імені Фрідріха Шиллера за наданий доступ до ресурсів Університетського обчислювального центру, що суттєво прискорило наше дослідження.

1. R. Parr, Y. Weitao. *Density-Functional Theory of Atoms and Molecules*, *International Series of Monographs on Chemistry* (Oxford University Press, 1994).
2. D. Frenkel, B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*, *Computational science* (Academic Press, 2001).

3. D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Rev. Drug Disc.* **3**, 935 (2004).
4. R. Johnston. *Atomic and Molecular Clusters* (Taylor and Francis, 2002).
5. Y. Xiang, D.Y. Sun, X.G. Gong. Generalized simulated annealing studies on structures and properties of Nin (n = = 2–55) clusters. *J. Phys. Chem. A* **104**, 2746 (2000).
6. P.N. Day, R. Pachter, M.S. Gordon, G.N. Merrill. A study of water clusters using the effective fragment potential and Monte Carlo simulated annealing. *J. Chem. Phys.* **112**, 2063 (2000).
7. K.A. Dill, J.L. MacCallum. The protein-folding problem, 50 years on. *Science* **338**, 1042 (2012).
8. S. Mayewski. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. *Proteins Struct. Func. Bioinform.* **59**, 152 (2005).
9. M. Allen, D. Tildesley. *Computer Simulation of Liquids*, *Computer Simulation of Liquids* (Clarendon Press, 1989).
10. Q. Zhao, D.M. Anstine, O. Isayev, B.M. Savoie. Δ^2 machine learning for reaction property prediction. *Chem. Sci.* **14**, 13392 (2023).
11. U.V. Ucak, I. Ashyrmamatov, J. Ko, J. Lee. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nature Commun.* **13**, 1186 (2022).
12. R. Souza, J. Duarte, R. Goldschmidt, I. Borges, Jr. Machine learning prediction of electronic molecular excited state properties. *J. Braz. Chem. Soc.* **36**, 1 (2025).
13. Š. Sršeň, O.A. von Lilienfeld, P. Slavíček. Fast and accurate excited states predictions: Machine learning and diabatization. *Phys. Chem. Chem. Phys.* **26**, 4306 (2024).
14. D. Zhang, Q. Chu, D. Chen. Predicting the enthalpy of formation of energetic molecules via conventional machine learning and GNN. *Phys. Chem. Chem. Phys.* **26**, 7029 (2024).
15. M.R. Dobbelaere, I. Lengyel, C.V. Stevens, K.M. Van Geem. Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *J. Cheminform.* **16**, 99 (2024).
16. U.K. Ghosh, F. Al Abir, N. Rifaat, S.M. Shovan, A. Sayeed, M.A.M. Hasan. Most dominant metabolomic biomarkers identification for lung cancer. *Inform. Med. Unlock.* **28**, 100824 (2022).
17. X. Zhang, I. Jonassen, A. Goksøyr. *Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data* (Exon Publications, 2021).
18. Z. Zhang, Z.-P. Liu. *Intelligent Computing Theories and Application: Proceedings of the 15th International Conference ICIC 2019, Nanchang, China, August 3–6, 2019* (Springer 2019), Part II, p. 517 [ISBN: 978-3-030-26968-5, 978-3-030-26969-2].
19. J. Behler. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
20. L. Seute, E. Hartmann, J. Stühmer, F. Gräter. Grappa – a machine learned molecular mechanics force field. *Chem. Sci.* **16**, 2907 (2025).
21. X. Gao, F. Ramezanghorbani, O. Isayev, J.S. Smith, A.E. Roitberg. TorchANI: A free and open source pytorch-based deep learning implementation of the ANI neural network potentials. *J. Chem. Inform. Model.* **60**, 3408 (2020).
22. S. Doerr, M. Majewski, A. Pérez, A. Krämer, C. Clementi, F. Noe, T. Giorgino, G. De Fabritiis. TorchMD: A deep learning framework for molecular simulations. *J. Chem. Theor. Comput.* **17**, 2355 (2021).
23. S.J. Pan, Q. Yang. A Survey on Transfer Learning. *IEEE Trans. Knowledg. Data Eng.* **22**, 1345 (2010).
24. E. Weislinger, G. Olivier. The classical and quantum mechanical virial theorem. *Int. J. Quant. Chem.* **8**, 389 (1974).
25. F. Jensen. *Introduction to Computational Chemistry* (Wiley, 2017).
26. C. Isert, K. Atz, J. Jiménez-Luna, G. Schneider. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273 (2022).
27. D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M. Magariños, J. Mosquera, P. Mutowo, M. Nowotka *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucl. Acids Res.* **47**, D930 (2018).
28. D.G.A. Smith, L.A. Burns, A.C. Simmonett, R.M. Parrish, M.C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan *et al.* Psi4 1.4: Open-source software for high-throughput quantum chemistr. *J. Chem. Phys.* **152**, 184108 (2020).
29. Z. Guo, K. Guo, B. Nan, Y. Tian, R.G. Iyer, Y. Ma, O. Wiest, X. Zhang, W. Wang, C. Zhang, N.V. Chawla. Graph-based molecular representation learning. In: *Proc. of the 32nd International Joint Conference on Artificial Intelligence IJCAI '23* (Publisher, 2023), p. XXX.
30. M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, Z. Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv:1909.01315.
31. J. Gilmer, S.S. Schoenholz, P.F. Riley. Neural Message Passing for Quantum Chemistry. In: *Proc. of the 34th International Conference on Machine Learning.* Edited by D. Precup, Y.W. Teh (PMLR, 2017), p. 1263.
32. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Publisher, 2019), p. 8024.
33. D.P. Kingma, J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980.

34. Y. Mo. Probing the nature of hydrogen bonds in DNA base pairs. *J. Mol. Model.* **12**, 665 (2006).
35. C. Bannwarth, S. Ehlert, S. Grimme. GFN2-xTB – an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theor. Comput.* **15**, 1652 (2019).
36. L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. Grimme. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **19**, 32184 (2017).
37. J. Řezáč, K.E. Riley, P. Hobza. S66: A Well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theor. Comput.* **7**, 2427 (2011).
38. J.-D. Chai, M. Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615 (2008).
39. S. Boys, F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **19**, 553 (1970).
40. K. Raghavachari, G.W. Trucks, J.A. Pople, M. Head-Gordon. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **157**, 479 (1989).
41. H.N. Motlagh, J.O. Wrabl, J. Li, V.J. Hilser. The ensemble nature of allostery. *Nature* **508**, 331 (2014).
42. A. Kovács, A.Y. Ivanov. Vibrational analysis of α-D-glucose trapped in Ar matrix. *J. Phys. Chem. B* **113**, 2151 (2009).
43. I. Peña, E.J. Cocinero, C. Cabezas, A. Lesarri, S. Mata, P. Écija, A.M. Daly, Á. Cimas, C. Bermúdez, F.J. Basterretxea, S. Blanco, J.A. Fernández, J.C. López, F. Castaño, J.L. Alonso. Six pyranoside forms of free 2-deoxy-D-ribose. *Angew. Chem. Internat. Edit.* **52**, 11840 (2013).
44. M. Nakata, T. Shimazaki. PubChemQC Project: A large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model* **57**, 1300 (2017).
45. P. Pracht, F. Bohle, S. Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **22**, 7169 (2020).
46. N. Schmid, A.P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A.E. Mark, W.F. van Gunsteren. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **40**, 843 (2011).

Одержано 08.09.25

A.D. Terets', T.Yu. Nikolayenko

MODEL FOR PARAMETERIZATION
OF INTERATOMIC INTERACTION POTENTIALS
BY QUANTUM-MECHANICAL DESCRIPTORS
ON THE BASIS OF A GRAPH NEURAL NETWORK

Based on graph neural networks, a machine learning model has been developed to predict the total energy of biomolecules from their structural formulas and quantum-mechanical descriptors by predicting the parameters of the functions that approximate interatomic potentials. The applicability of the created model for predicting the total energy of biomolecules, their interaction energy, and the ordering of conformers by energies has been proven. The physical validity of the obtained parameter values was demonstrated, which opens opportunities for further application of the model in molecular modeling problems.

Keywords: machine learning, interatomic interaction potentials, biomolecule conformers, quantum-mechanical descriptors, neural networks, biomolecular binding.