


Відгук отримано 19.12.2021 року
Голова спеціалізованої вченої ради ДФ 64.051.041

1

Відгук офіційного опонента на дисертацію

Бердника Михайла Ігоровича

 Олександр
Курченко

**« Метод L_1 -регуляризації для опису фізико-хімічних властивостей
молекул »**

яка подана на здобуття наукового ступеня
доктор філософії за спеціальністю "102 Хімія"

Дослідження зв'язку "структура-властивість/активність" є однією із ключових, дуже загальних завдань сучасної хімії. Широкий спектр теоретичних методів включає не тільки методи квантової хімії і молекулярного моделювання але й різноманітні підходи, що ґрунтуються на статистичних (хемометричних) методах. Загалом такі методи – методи QSAR/QSPR (Quantitative-Structure Activity/Property Relationships) включають ряд підходів, що призначені описати різні аспекти в побудові прогностичних моделей. Незважаючи на досить довгу історію (перші моделі QSPR з'явилися у роботах Вінера ще у 1947-1948 роках) ця методологія продовжує інтенсивно розвиватись. В останні роки методи Data mining, Machine learning, Big Data Analysis стали необхідною частиною багатьох досліджень хімії. Однак, при реалізації моделей QSAR, виникає загальна проблема відбору найбільш важливих дескрипторів (предикторів) які здатні не тільки адекватно описати властивості первинної – навчальної вибірки, але й надати якісний прогноз для нових, ще не синтезованих систем. Слід зауважити, що така проблема відбору виникає і в інших хімічних дисциплінах. Так, методи квантової хімії, які оперують розкладами багаточастинкової хвильової функції, також потребують ефективних процедур відбору найбільш важливих компонент (конфігурацій), що гарантують якісний, опис системи за помірними витратами розрахункових ресурсів.

На сьогоднішній день, в царині QSAR, використовують кілька варіантів для такого відбору серед яких наприклад факторний аналіз. Однак, вочевидь, однозначного розв'язку цієї проблеми не існує і вивчення інших, альтернативних підходів є актуальним завданням.

Отже дисертаційна робота Михайла Бердника є **актуальною** оскільки вона пов'язана із дослідженням можливостей використання в завданнях QSAR одного із ефективних статистичних методів звуження предикторного простору – методу L_1 -регуляризації. Хоч сам метод і було запропоновано більше двадцяти років тому, інтерес до нього значно зріс тільки у останні роки. До моменту початку роботи над дисертацією L_1 -регуляризацію ще не було використано в хімічній науці.

В зв'язку з вказаним аспірантом реалізовано унікальний комплекс програм що відтворює не тільки різноманітні регуляризовані підходи, а й інші статистичні

методи як-то: PCR, PLS, метод нейронних мереж, метод логістичної регресії. На додаток до стандартного методу найменших квадратів програмно реалізовано методи найменших модулів та метод ортогональних відстаней. Розроблено комплекс програмних засобів для тестування отриманих статистичних моделей.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій

В роботі наведено наукові результати та висновки з них, які базуються на коректному використанні статистичних та хемометричних методів. Також значну частину наукових результатів складають коректні використання методів квантової хімії. Порівняння результатів роботи із відомими стандартними методами статистики і квантової хімії також доводять коректність і обґрунтованість отриманих результатів та висновків. Наприклад результати наближених розрахунків запропонованими автором методами (L_1 -CCD, L_1 -CCSD) знаходяться у доброму і логічному узгодженні із необмеженим за конфігураційним розкладом методом CCSD. Розрахунки лінійної регресії альтернативними методами знаходяться у логічному узгодженні із стандартним методом найменших квадратів. Отримані наукові результати аргументовано інтерпретовано на основі сучасних уявлень.

Стислий аналіз змісту дисертації

Дисертаційна робота складається зі вступу, 5 розділів, загальних висновків, списку використаних джерел та 5 додатків. Обсяг загального тексту дисертації складає 205 сторінок. Основний текст включає 142 сторінки. В роботі 19 таблиць та 47 рисунків. Список використаних джерел містить 218 найменувань.

У вступі передано основний зміст дисертації. Вказано актуальність, мету і завдання дослідження. Перераховано наукову новизну та практичне значення одержаних результатів. Показано основний внесок здобувача в дисертацію, представлено зв'язок теми дисертаційної роботи з тематикою науково-дослідницької роботи кафедри хімічного матеріалознавства хімічного факультету Харківського національного університету імені В.Н. Каразіна.

У першому розділі дисертації зроблено літературний огляд що стосується методів L_2 - та L_1 -регуляризації. Розглянуто зокрема методи LASSO та LARS-LASSO. Певну увагу було приділено альтернативним методам побудови лінійної регресії. Серед них розглянуто метод найменших модулів, метод ортогональних відстаней та метод найменших модулів ортогональних відстаней. Певну увагу було приділено методам побудови регресійних моделей, що ґрунтуються на вивченні

факторної будові даних (PCR,PLS). Представлено також стислий аналіз методу логістичної регресії та методів валідації отриманих регресійних рівнянь.

Вказано, що для усіх, описаних в огляді, методів у роботі розроблено комп'ютерні програми.

У другому розділі дисертації описано використані алгоритми L1-регуляризації. Крім найпростішого прикладу регуляризації середнього значення описано програмно реалізовані дисертантом алгоритми LARS-LASSO. Крім того детально описано алгоритми методу найменших модулів, методу ортогональних відстаней та методу LADOD. Також реалізовано розрахунки методами PCR, PLS та методом нейронних мереж. Описані алгоритми було використано в ряді завдань. Першим і найпростішим описаним завданням є проблема рКа карбонових кислот. Це завдання було тестовою проблемою в якій дано методологічні основи використаних методів. Показано зокрема принципові відмінності методу L2-регуляризації (метод Тихонова) від методу L1-регуляризації. Проведене скорочення дескрипторного набору дозволило отримати кілька простих регресійних рівнянь.

Розроблені методичні основи регуляризованих розрахунків, далі, були використані в завданнях що стосуються неемпіричних оцінок рКа фенолів, та регресійних моделей констант іонізації органічних сполук різної природи. Встановлено, що навіть для дуже різних за будовою систем можливо отримати досить прості регресійні рівняння. Описано використання L1-регуляризації у комбінації із методом нейронних мереж.

Ще кілька прикладів використання дослідженого підходу стосується опису температури кипіння органічних сульфідів, температури кипіння флуороалканів, та дослідженню можливостей побудови кількісної моделі "В'язкість рідин – тиск насиченого пару".

Третій розділ роботи присвячено теоретичному дослідженню валідаційних характеристик лінійних регресійних моделей. Було обрано достатньо просте лінійне рівняння із введеними за нормальним розподілом похибками. Досліджено залежності (зв'язки) між різними параметрами внутрішньої та зовнішньої валідації. Встановлено умови адекватності прогностичних оцінок якості лінійних регресій на основі тестових вибірок. Встановлено зокрема необхідність раціонального розбиття на тестову і навчаючу вибірки. Для адекватної оцінки необхідне дослідження поведінки параметрів внутрішньої й зовнішньої валідації в області значної густини точок розбиття.

В цьому розділі представлено результати розрахунків лінійної регресії з використанням альтернативних підходів. Представлено ситуації, коли альтернативні методи виявляють кращі оцінки ніж стандартний метод найменших квадратів.

В четвертому розділі описано використання L1-регуляризованої класифікаційної функції на основі логістичної регресії. Представлено дані щодо основності органічних сполук до катіонів літію. Розглянуто проблему знаходження потенційних лігандів рецепторів естрогену. Встановлено ряд простих логістичних функцій які дають якісний прогноз на рівні найкращих підходів.

П'ятий розділ дисертації присвячено використанню процедури L1-регуляризації в квантовій хімії, зокрема в теорії зв'язаних кластерів. Для оптимізації та прискорення розрахунків, в п'ятому розділі досліджено різні багатокрокові методи першого порядку для розв'язку робочих рівнянь теорії зв'язаних кластерів. Встановлено, що достатньо простий алгоритм (метод "важкої кульки") гарантує добру швидкість отримання розв'язків нелінійних рівнянь теорії.

В розділі описано принципово новий підхід до формування впорядкованого набору електронно-збуджених конфігурацій. На його основі виникає можливість створення прогресивної системи наближень в теорії зв'язаних кластерів. Показано, що такі наближення дають систематичні розв'язки квантово-хімічного завдання від найпростішого та найдешевшого до більш точного.

Представлено дані щодо напівемпіричних π -електронних розрахунків в рамках нових розроблених L1-регуляризованих підходів (L1-MP2, L1-CCD, L1-CCSD). Показано, що навіть короткі розклади хвильової функції здатні достатньо точно описати енергетичні характеристики молекул.

Представлені дані щодо побудови кривих дисоціації малих молекул демонструють значну ступінь паралельності наближених до точних кривих.

Наукова новизна отриманих результатів.

- 1) Показано, що за допомогою процедури L1-регуляризації можливо створити впорядкований список молекулярних дескрипторів, завдяки чому можлива побудова простих малопараметричних регресійних рівнянь.
- 2) Показано, що прості малопараметричні регресійні рівняння, які отримані за допомогою алгоритму LARS-LASSO можуть мати значно

кращі прогностичні характеристики, ніж ті результати, що отримані методами PCR, PLS.

- 3) Побудовані прості регресійні рівняння для опису констант іонізації органічних кислот та основ різної будови.
- 4) Продемонстровано важливість коректного розділення вхідних даних на тестову й тренувальну вибірки для отримання адекватних оцінок обраної статистичної моделі.
- 5) Показано, що алгоритм LARS-LASSO може бути з успіхом використано для побудови простих класифікаційних функцій на основі логістичної регресії.
- 6) Побудовано класифікаційну функцію, що описує основність органічних сполук різної природи до катіонів літію.
- 7) Побудовано класифікаційну функцію, що дозволяє описати активність органічних сполук стероїдної та нестероїдної природи по відношенню до рецепторів естрогену.
- 8) Вперше показано, що за допомогою методу L1-регуляризації можливо сгенерувати впорядкований список електронно-збуджених конфігурацій, на основі якого може бути створено прогресивну систему наближень квантовохімічного методу.

Зауваження

- 1) На стор. 30 (Розділ 1) дисертант вказує, що, зокрема, метод Random Forest схильний до перенавчання. Дійсно, навчальна вибірка апроксимується цим методом з високим коефіцієнтом детермінації, але застосування процедури “out-of-bag” дозволяє надійно прогнозувати нові сполуки. Часто такі прогнози більш надійні ніж при використанні лінійних регресійних моделей.
- 2) В дисертації показано, що лінійні прогностичні регресійні моделі може бути реалізовано чотирма способами (OLS, LAD, ODR, LADOD). Але не сформульовано загальних критеріїв для вибору тієї чи іншої моделі як найкращої для даного випадку.
- 3) Чим все ж таки краще метод L1-регуляризації ніж метод генетичних алгоритмів ?
- 4) В задачах побудови класифікаційних функцій, в якості методу порівняння, обрано метод випадкових лісів. Чим обумовлено такий вибір ? Адже на теперішній час існують і інші методи, як-то: SVM (метод опорних векторів).

- 5) Опонент категорично не згоден з твердженням дисертанта на стор. 113: « ... для молекул, властивості яких прогнозуються методами SVM, RF та ANN, неможливо визначити, які саме структурні особливості чи молекулярні параметри відповідають наявності або відсутності певного рівня активності. Таким чином, ці методи не можуть дати загальне розуміння, чому певні молекули мають відповідний рівень активності / властивості.». Впевнено вважаю, що не існує QSAR/QSPR моделей які не можливо інтерпретувати. Просто, в деяких випадках, зокрема для методів «чорної скриньки» необхідно застосовувати спеціальні підходи (див. наприклад огляд: A. Cherkasov etc. QSAR modeling: where have you been? Where are you going to? //Journal of Medicinal Chemistry. – 2014. – vol.57. – P. 4977-5010).
- 6) В українському правописі замість слова «задача» коректніше використовувати слово «завдання».

Загалом слід відзначити, що результати дисертації **достатньо повно викладені** в наукових публікаціях. Також аналіз дисертації довів **відсутність ознак** порушення академічної доброчесності.

Висновок

Зроблені зауваження не впливають на високу оцінку дисертаційної роботи в цілому. Одержані результати відповідають сформульованій меті та поставленим завданням.

Таким чином, я вважаю, що за своєю актуальністю, науковим рівнем, практичною значущістю, достовірністю отриманих результатів і ступенем обґрунтованості висновків, робота «Метод L_1 -регуляризації для опису фізико-хімічних властивостей молекул» відповідає вимогам до дисертаційних робіт на здобуття наукового ступеня доктор філософії, а її автор Бердник Михайло Ігорович заслуговує присудження наукового ступеня доктор філософії зі спеціальності "102 Хімія".

Член-кореспондент НАН України,
професор, доктор хімічних наук,
директор фізико-хімічного інституту
ім. О. В. Богатського НАН України

Віктор КУЗЬМІН

Підпис чл.-кор. НАН України Віктора Кузьміна засвідчую.
Учений секретар ФХІ НАН України

К.Х.Н., С.Н.С.

Євген ШАБАНОВ

