

Відгук отримано 10.12.2021 р.-кз.

Голова спеціалізованої вченої ради ДФ 64.051.041

Відгук

офіційного опонента на дисертацію

Бердника Михайла Ігоровича

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

подану на здобуття наукового ступеня доктора філософії

за спеціальністю 102 «Хімія»

галузь знань 10 «Природничі науки»

Олександр Керченко

Актуальність теми дисертації

Передбачення фізико-хімічних властивостей молекул, виходячи з структурних параметрів, є класичною проблемою теоретичної хімії. Існує щонайменше два основних підходи для вирішення поставленої задачі. Перший – фундаментальне дослідження конкретної властивості й побудова теоретичної моделі виходячи з суто фізико-хімічних міркувань механізму реалізації властивості молекули (зазвичай такий підхід є індивідуальним для кожної окремої досліджуваної характеристики). Другий, суто емпіричний підхід, – полягає в встановленні взаємозв'язку між структурою молекули та відповідного їй експериментального значення шуканої властивості чи активності.

Представлена робота присвячена дослідженню і розвитку нових методів й алгоритмів вказаного другого підходу, який відомий в літературі як QSAR/QSPR, *Quantitative Structure-Activity / Property Relationship*. Реалізація QSAR містить певну серію послідовних етапів – починаючи з розробки нових дескрипторів та закінчуючи валідацією та оцінкою передбачувальної здатності отриманих теоретичних моделей. На сьогоднішній день існує ряд типових проблем у використанні методики QSAR. Наведемо декілька з них:

- відбір компактного набору предикторів з десятків тисяч існуючих молекулярних дескрипторів;
- побудова моделей взаємозв'язку цих предикторів із значеннями шуканої властивості за наявності надлишкових або “пошкоджених” даних;
- проблема перенавчання моделей;
- нетривіальність інтерпретації отриманих результатів;
- відсутність надійної процедури валідації отриманих моделей.

Кожна з перерахованих вище проблем сама по собі включає значну кількість деталей і нюансів, які необхідно враховувати при використанні методу QSAR.

У зв'язку з цим, пропонований у дисертаційній роботі підхід, а саме використання методу L_1 -регуляризації (на відміну від більш відомої та добре дослідженої L_2 -регуляризації, що зазвичай використовується для запобігання перенавчання моделей), є вкрай привабливим з огляду на те, що вказаний метод дозволяє одночасно вирішити цілий комплекс перерахованих вище проблем. Крім того, як наочно показано в розділі 5 дисертації, потенціальна сфера застосування L_1 -регуляризації в теоретичній хімії не обмежується рамками QSAR, а виявляється значно ширше, оскільки може бути застосована до проблем підвищення ефективності квантово-хімічних підходів.

Ступінь обґрунтованості наукових положень, висновків, рекомендацій

Сформульовані в роботі наукові положення, висновки та рекомендації є цілком обґрунтованими, оскільки вони є узагальненням великого обсягу теоретичного та розрахункового матеріалу, отриманого з використанням сучасних методів хемоінформатики, хеометрики та квантової хімії. Для більшості реалізованих автором алгоритмів наведено порівняння з літературними даними та результатами аналогічних підходів. Наукові та практично значущі факти аргументовані та інтерпретовані на підставі сучасних уявлень. У зв'язку з цим, ступінь обґрунтованості наукових положень, висновків та рекомендацій дисертації не викликає сумнівів.

Наукова новизна одержаних результатів

Серед низки нових результатів, винесених на захист, слід відмітити наступне:

- Показано, що метод L_1 -регуляризації в комбінації з методом LARS (Least-Angle Regression Stagewise) дозволяє отримувати компактні регресійні рівняння, що легко інтерпретуються, як при вирішенні проблем передбачення фізико-хімічних властивостей, так і для бінарної класифікації молекулярних систем.
- Наочно доведено ефективність L_1 -регуляризації у боротьбі з проблемою «перенавчання» шляхом цілеспрямованого виключення малозначних дескрипторів із усієї сукупності дескрипторного простору.
- Продемонстровані області застосування та точність альтернативних способів побудови регресійних рівнянь у порівнянні зі стандартним для QSAR – методом найменших квадратів.
- Завдяки використанню досліджуваного у роботі підходу, отримано набори компактних регресійних моделей для опису експериментальних даних констант іонізації, температур кипіння, тиску насиченої пари та в'язкості для різних класів органічних сполук.
- Сформульовані основні вимоги до процедур валідації, оцінки адекватності та передбачуваної здатності регресійних рівнянь у рамках методу QSAR.
- Показано перспективи використання регуляризаційного методу для спрощення процедури розв'язку рівнянь у рамках квантово-хімічних підходів, що враховують електронну кореляцію.

Практичне значення одержаних результатів

Одним з важливих результатів даної роботи є демонстрація потенціалу методу L_1 -регуляризації стосовно різних проблем теоретичної хімії. Зокрема це стосується спрощення ключових етапів QSAR підходу, а саме строго-детермінованого та надійного способу скорочення дескрипторного набору та запобігання перенавчанню регресійних моделей. Вивчено альтернативні підходи до побудови регресійних рівнянь на основі незаслужено недооцінених у хімії підходів, таких як метод найменших модулів, метод ортогональних відстаней та нещодавно розробленого методу найменших абсолютних відхилень ортогональних відстаней. Проведено дослідження та порівняння ефективності та надійності різних

валідаційних метрик. Крім того, продемонстровано ефективність L_1 -регуляризації у застосуванні до проблеми класифікації на прикладі логістичної регресії, а також перспективи даного підходу для спрощення схеми проведення квантово-хімічних розрахунків у рамках теорії зв'язаних кластерів. Що не менш важливо, автором не лише проведено теоретичні дослідження щодо застосовності та ефективності даного підходу, а й програмно реалізовано та тестовано відповідні алгоритми на великому наборі експериментальних даних.

Повнота викладення матеріалів дисертації в публікаціях

Основні результати роботи викладені в 13 публікаціях: 5 статей (з них 3 у фахових журналах України, 2 статті входять до міжнародних наукометричних баз) та тези 8 доповідей на наукових конференціях різного рівня. Публікації з належною повнотою передають зміст дисертації.

Загальні дані щодо структури роботи

Дисертація складається зі вступу, п'яти розділів, висновків, п'яти додатків і списку використаних літературних джерел із 218 бібліографічних найменувань. Робота викладена на 205 сторінках, з них 142 сторінки основного друкованого тексту. Вона вміщує 19 таблиць та 47 рисунків.

Стислий аналіз змісту дисертації

У **вступі** передано суть дисертації: обґрунтовано її актуальність, сформульовано мету і задачі дослідження, показано наукову новизну та практичне значення одержаних результатів, апробації, показано основний внесок здобувача, висвітлений зв'язок теми дисертаційної роботи з плановою тематикою науково-дослідницьких тем кафедри хімічного матеріалознавства хімічного факультету Харківського національного університету імені В. Н. Каразіна.

Перший розділ присвячений літературному огляду, в якому стисло розглянуті існуючі методи побудови регресійних моделей, математичні викладки та геометричні ілюстрації, необхідні для розуміння суті пропонованого підходу. Викладено сучасні дані щодо проблемі валідації моделей QSAR, існуючі алгоритми та метрики.

У **другому розділі** описано основи застосування регуляризаційних підходів у комбінації з різними методами побудови регресійних рівнянь. На простому прикладі, обчислення середнього значення нестандартним способом, продемонстрована особливість реалізації введення L_1 -регуляційної поправки в функцію, що мінімізується. Викладено подробиці та деталі реалізації алгоритмів побудови регресійних моделей методів найменших модулів, ортогональних відстаней та найменших абсолютних відхилень ортогональних відстаней. Отримано рівняння залежності фізико-хімічних властивостей (константи іонізації, температури кипіння, в'язкості та тиску насиченої пари) від структурних дескрипторів для широкого набору органічних молекул.

Третій розділ містить детальне дослідження валідаційних характеристик регресійних рівнянь QSAR. На модельному прикладі продемонстровано вплив розміру тестових вибірок, способу їх побудови на значення валідаційних метрик та їхній взаємозв'язок із істинними значеннями коефіцієнтів рівнянь у різних методах.

Наочно сформульована послідовність проведення аналізу надійності та передбачувальної здатності одержаних моделей.

У четвертому розділі наведено реалізацію L_1 -регуляризаційного підходу до проблеми класифікації молекулярних систем. Представлено алгоритм логістичної регресії (LR) в рамках методу LARS-LASSO. Реалізацію підходу тестовано на прикладі спорідненості органічних систем різної природи до іону Li^+ (на основі констант рівноваги відповідної реакції). Результати демонструють ефективність методу LR-LARS-LASSO на структурно різномірних вибірках.

П'ятий розділ присвячений вивченню можливості застосування L_1 -регуляризації для підвищення ефективності розрахунків методами квантової хімії. Перша частина описує способи прискорення ітераційної процедури розв'язку рівнянь теорії зв'язаних кластерів за рахунок модифікації градієнту, що представляє собою певну аналогію із процедурою регуляризації. Проведено порівняння із загальноприйнятим методом DIIS (Direct Inverse in Iteration Subspace). Показано, що використання досить простого та дешевого (в обчислювальному плані) підходу, дозволяє отримати результат, який можна порівняти за ефективністю з «дорогими» багатокроковими методами.

Далі було досліджено використання регуляризації для зниження обчислювальної складності методів урахування електронної кореляції, яке досягається шляхом скорочення конфігураційного простору. В результаті такого скорочення «виживають» найбільш значимі електронні конфігурації які дають максимальний внесок в точну хвильову функцію. Тестування даного підходу проводиться як на рівні напівемпіричних, так і неемпіричних (*ab initio*) методів. На прикладі деяких спряжених вуглеводнів показано зміну конфігураційного складу хвильової функції зі зміною величини регуляризуючої поправки.

В рамках *ab initio* підходу, досліджено дисоціативні криві молекул LiH , BH та води.

Достовірність отриманих результатів та зроблених висновків не викликає сумніву оскільки вони спираються на коректно проведені статистичні, хемометричні та квантово-хімічні розрахунки.

Зауваження

1. У роботі в декларативній формі сказано, що альтернативні способи відбору дескрипторів (зокрема, так звані евристичні підходи, такі як методи генетичних алгоритмів, мурашиної колонії тощо) можуть демонструвати менш надійні результати, ніж запропоновані автором, проте чисельних результатів не надано.

2. В третьому розділі наведено гістограми розподілу величин валідації R_{test}^2 та R_{train}^2 . Зрозуміло, що отримані гістограми мають якусь форму. Оскільки при цьому вивчалася тестова задача виникає питання: що саме може впливати на форму таких розподілів? Чи можна зробити якісь висновки з аналізу безпосередньо гістограм о якості вхідних даних?

3. Застосування крос-валідаційної процедури LOO (Leave-One-Out) у додатку до штучних нейронних мереж у тому вигляді, в якому воно описане у пункті 2.6 є нетиповим у загальновідомій практиці. Як правило, подібні підходи застосовуються для встановлення оптимальних метопараметрів нейронної мережі, а

ніяк не для визначення дескрипторного набору моделі. Причина використання LOO для цілей потребує детального пояснення.

4. У розділі 3, присвяченому валідації, недостатньо чітко описані критерії вибору параметру k -Nearest-Neighbors (кількості найближчих сусідів) і особливо вплив цієї величини на значення валідаційних характеристик отриманих з урахуванням Applicability Domain.

5. У дисертаційній роботі недостатньо відображена ефективність (реальний вигреш у обчислювальній складності) застосування регуляризації в методах теорії пов'язаних кластерів. І не сказано яким чином може бути досягнуто вигреш в розрахунковій складності методу.

Висновок

Зроблені зауваження не впливають на високу оцінку дисертаційної роботи в цілому, яку можна розглядати як непросте, завершене, багатопланове, оригінальне дослідження, яке виконане на високому науковому рівні із залученням сучасних методів хімічної інформатики, хеометрії та квантової хімії.

Одержані результати відповідають сформульованій меті та поставленим задачам. Також аналіз дисертації довів **відсутність ознак порушення академічної доброчесності**.

Таким чином, за своєю актуальністю, науковим рівнем, практичною значущістю, достовірністю отриманих результатів і ступенем обґрунтованості висновків та рекомендацій дисертаційна робота «Метод L_1 -регуляризації для опису фізико-хімічних властивостей молекул» відповідає вимогам до дисертацій на звання доктор філософії, а її автор **Бердник Михайло Ігорович** заслуговує на присудження наукового ступеня доктор філософії за спеціальністю 102 «Хімія» галузь знань 10 «Природничі науки».

Доктор фізико-математичних наук,
старший науковий співробітник,
провідний науковий співробітник
Фізико-технічного інституту
низьких температур
ім. Б. І. Веркіна НАН України



Степан СТЕПАНЬЯН

