

Харківський національний університет імені В. Н. Каразіна

Міністерство освіти і науки України

Кваліфікаційна наукова  
праця на правах рукопису

**Бердник Михайло Ігорович**

УДК 544.169+544.15+519.237.5

## **ДИСЕРТАЦІЯ**


**Метод  $L_1$ -регуляризації для опису фізико-хімічних властивостей  
молекул**

Спеціальність 102 - «Хімія»


(Галузь знань 10 - Природничі науки)

Подається на здобуття ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

 М.І. Бердник

Науковий керівник: Іванов Володимир Венедиктович, доктор хімічних наук, професор.

Усі примірники дисертації  
ідентичні за змістом.  
Голова спеціалізованої вченої  
ради ДФ 64.051.041  
 Олександр Куріченко

Харків – 2021

## АНОТАЦІЯ

Бердник М. І. Метод  $L_1$ -регуляризації для опису фізико-хімічних властивостей молекул. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття ступеня доктора філософії за спеціальністю 102 - Хімія (Галузь знань 10 – Природничі науки). – Харківський національний університет імені В. Н. Каразіна Міністерства освіти і науки України, Харків, 2021.

Роботу присвячено дослідженню можливостей використання  $L_1$ -регуляризації в побудові хеометричних моделей «структура-активність» і квантовохімічних розрахунках. Для виконання завдань дисертації розроблено оригінальний комплекс програм, що реалізують різні статистичні (хеометричні) підходи до побудови регресійних моделей й аналізу їх прогностичної здатності. Також створено комплекс квантовохімічних програм, у яких  $L_1$ -регуляризація використовується для побудови хвильових функцій методів, що ураховують електронну кореляцію.

Зокрема, у дисертаційній роботі розглядалося використання  $L_1$ -регуляризації для побудови лінійних емпіричних моделей опису різних фізико-хімічних параметрів молекул. Серед таких параметрів розглянуто  $pK_a$  та температури кипіння органічних сполук різної природи, які включають карбонові кислоти, феноли, сульфіді, флуороалкани. Розглядалися також кореляції в'язкості рідин та тиску насиченого пару різних органічних сполук. Спираючись на досліджені вибірки молекул, було показано, що з використанням  $L_1$ -регуляризації завжди можна сформулювати послідовний (упорядкований) набір дескрипторів. Систематично додаючи дескриптори з цього набору до моделей лінійної регресії або штучних нейронних мереж, можна отримати рівняння (або відповідно нейронні мережі) з послідовно зростаючими величинами критеріїв валідації. Оскільки після ранжування дескрипторного набору обрані предиктори можуть використовуватися в різних підходах до побудови лінійної регресії, нами було проведено відповідне дослідження якості цих альтернативних моделей. При цьому розглядалися:

метод найменших квадратів (*Ordinary Least squares*, OLS), метод найменших модулів (*Least Absolute Deviation*, LAD), метод ортогональних відстаней (*Orthogonal Distances Regression*, ODR), а також запропонований нещодавно метод найменших абсолютних відхилень ортогональних відстаней (*Least Absolute Deviation of Orthogonal Distances*, LADOD). Було показано, що той чи інший метод може мати кращі прогностичні властивості відповідно до критеріїв зовнішньої або внутрішньої валідації. Показано, що методом штучних нейронних мереж з використанням впорядкованого дескрипторного набору, який був отриманий методом  $L_1$ -регуляризації, також може бути зроблено якісні прогнози властивостей речовини. Також було проведено співставлення отриманих рівнянь лінійної регресії з альтернативними підходами, що працюють із нескороченими (неоптимізованими) дескрипторними наборами. А саме: з методом PCR (*Principal Component Regression*), а також методом PLS (*Partial Least Squares* або *Projection on Latent Structure*). Слід зазначити, що хоча з використанням цих методів для деяких задач і були отримані досить надійні прогностичні моделі, але такі моделі не надають ясної інформації стосовно природи отриманих рівнянь і не відповідають на питання: які структурно-хімічні особливості або молекулярні дескриптори, призводять до змін у відгуку (активності). У вивчених прикладах  $L_1$ -регуляризація дозволила сформулювати компактні одно-, двух- або трьох- параметричні моделі, які здатні задовільно описати набір даних. Відповідно до вивчених прикладів, моделі отримані з попереднім відбором із використанням LARS-LASSO виявились кращими, ніж результати розрахунків PLS та PCR.

Певну увагу в дисертації приділено методам валідації й оцінкам якості регресійних рівнянь. З цією метою було використано модельну задачу, у яку вносилися похибки як в залежну, так і в незалежну змінні. Для полегшення аналізу, а також, щоб вивчити валідаційні характеристики рівнянь в усіх досліджених методах лінійної регресії, розглядався найпростіший, але далеко нетривіальний випадок – регресія з однією незалежною змінною. Така постановка задачі дала можливість оцінювати рівняння відповідно до

близькості коефіцієнтів регресійних рівнянь до «ідеальних» теоретичних значень. З використанням модельної задачі було досліджено вплив раціонального розбиття вибірки на тренувальну та тестову на якість отриманих регресійних рівнянь. Було продемонстровано, що випадкове одиничне розбиття вибірки не є інформативним, оскільки в залежності від систем, що опинилися в тестовій вибірці, валідаційні характеристики для початкової (повної) вибірки можуть бути як дуже погані, що веде до недооцінки, так і дуже добрі, що веде до переоцінки якості рівняння. Отже, показано, що для адекватної оцінки регресійного рівняння, а також дослідження якості вхідних даних у цілому, необхідно створювати та вивчати якомога більше розбивань на тренувальну й тестову вибірку. Також було досліджено вплив урахування границь застосовності моделі (*Applicability Domain*, AD) на валідаційні характеристики регресійних рівнянь. Встановлено, що при випадкових розбиваннях вибірки на тренувальну та тестову, максимумами розподілу внутрішніх та зовнішніх характеристик, зазвичай, співпадають. На відміну від цього, коли розбиття на тренувальну й тестову вибірки здійснюються таким чином, що тестова вибірка знаходиться в AD моделі, отриманої виходячи з тренувальної вибірки, відповідний розподіл зовнішніх критеріїв валідації зміщується відносно внутрішніх у сторону збільшення. При цьому найбільш інформативними є розбиття, що попадають близько до максимуму густини точок, оскільки саме з аналізу цих областей можна отримати найбільш повне адекватне розуміння якості моделі. Також було досліджено відомі, запропоновані на сьогодні, критерії валідації. Виходячи з модельної задачі, було зроблено висновок, що деякі з критеріїв валідації надто сильно корельовані один з одним, що робить їх одночасне використання малоінформативним. Серед таких параметрів пари: ( $R_{test}^2$  – CCC та  $Q_{F3}^2$  – RMSEP<sub>test</sub>). Встановлено, що для даних із вираженим розкидом типовою картиною є зворотна (суттєво нелінійна) залежність  $R_{train}^2$  –  $R_{test}^2$ . При цьому покращення (збільшення) коефіцієнтів внутрішньої валідації ( $R_{train}^2$ ,  $Q_{LOO}^2$ ), взагалі кажучи, не є свідомством покращення прогностичної



властивості моделі, оскільки для лінійної регресії за достатньо великої кількості точок залежність між цими двома критеріями була близька до лінійної. Проте, критерій  $Q_{Loo}^2$  може бути успішно використано для малих вибірок. Спираючись на розрахункові дані, показано, що для більшості випадків метод OLS давав найкращі результати. Однак, для великих вибірок із похибкою як в залежній, так і в незалежних змінних, у методі ODR (та LADOD) можна отримати найкращі рівняння.

Інша тісно пов'язана із побудовою статистичних моделей проблема - це побудова класифікаційної функції. З цією метою в роботі використано  $L_1$ -регуляризований розрахунок логістичної регресії. Розглянуто дві задачі. У першій проведено класифікацію молекул на сильні та слабкі основи відносно іону літію. У другій задачі органічні системи були класифіковані на активні або неактивні відповідно до спорідненості зв'язування молекул до рецепторів естрогену. Показано, що з використанням  $L_1$ -регуляризованої логістичної регресії можна досягнути таких результатів класифікації, які є конкурентно-спроможними до результатів, отриманих з використанням інших, більш складних у розрахунковому сенсі, методів. Використання спеціального  $L_1$ -регуляризованого алгоритму (його позначено як LR-LARS-LASSO) дало можливість отримати досить прості класифікаційні рівняння, які є інтерпретуємими (на відміну від результатів, отриманих в інших популярних методах класифікації, таких як: метод опорних векторів, метод випадкових лісів, метод штучних нейронних мереж). Також отримані рівняння логістичної регресії є однозначними й відтворюваними.

Показано, що метод  $L_1$ -регуляризації може бути використаний і в квантовій хімії. За допомогою процедури  $L_1$ -регуляризації можливо створення впорядкованого (ранжованого) набору електронно-збуджених відносно Гартрі-Фоківського стану конфігурацій. Включаючи різну кількість конфігурацій з створеного набору, можливо отримати прогресивний набір наближених розв'язків до точних даних методу. Метод реалізовано в рамках теорії збурень Меллера-Плессета другого порядку (MP2) та різних рівнів теорії зв'язаних

кластерів. Продemonстровано, що такі наближені розв'язки дають доволі точні значення енергетичних характеристик молекул, при цьому кількість конфігурацій у розрахунках може бути значно нижчою, ніж у розрахунках з використанням повного конфігураційного набору точного методу. Для ефективного розв'язку відповідних рівнянь теорії зв'язаних кластерів, реалізовано низку розрахункових алгоритмів з використанням багатокрокових методів першого порядку.

**Ключові слова:**  $L_1$ -регуляризація, QSAR/QSPR,  $pK_a$  органічних сполук, температури кипіння органічних сполук, в'язкість рідини, тиск насиченого пару, ліганди рецептору естрогену, основність до катіону літію, лінійна регресія, метод найменших квадратів, метод найменших модулів, метод ортогональних відстаней, штучні нейронні мережі, валідація, логістична регресія, теорія збурень Меллера-Плессета, теорія зв'язаних кластерів.

## ABSTRACT

Berdnyk M. I.  $L_1$ -regularization method for the description of the physical and chemical properties of molecules. Qualification scholarly paper: a manuscript.

Thesis submitted for obtaining the Doctor of Philosophy degree in Natural Sciences, Speciality 102 – Chemistry. – V. N. Karazin Kharkiv National University, Ministry of Education and Science of Ukraine, Kharkiv, 2021

This thesis focuses on the study of the possibilities of  $L_1$ -regularization application in the construction of "structure-activity" chemometric models and quantum chemical calculations. To perform the tasks of the thesis, an original set of programs has been developed that implement various statistical (chemometric) approaches to the construction of regression models and analysis of their prognostic properties. A set of quantum chemical programs has also been created, in which  $L_1$ -regularization is used to construct wave functions of methods that take into account electronic correlation.

In particular, in the thesis we consider application of  $L_1$ -regularization to obtain linear empirical models for describing various physicochemical parameters of molecules. Among the such parameters the pKa and boiling points for different nature organic compounds including carbonic acids, phenols, sulfides, fluoroalkanes were investigated. Also the correlations between viscosity of media and vapor pressure for different organic compounds have been described.

Based on the studied samples of molecules, it was shown that with the use of  $L_1$ -regularization it is always possible to form a sequential (ordered) set of descriptors. By systematically adding descriptors from this set to linear regression models or artificial neural networks, it is possible to obtain equations (or neural networks, respectively) with successively increasing values of validation criteria. Due to the fact that after ranking of the descriptors set, the selected predictors can be used in different approaches to construct linear regression models, we conducted a corresponding study of the quality of these alternative models. We considered: the *Ordinary Least Squares* method (OLS), the *Least Absolute Deviation* method (LAD), and the *Orthogonal Distances Regression* method (ODR), as well as the recently

proposed method of the *Least Absolute Deviation of Orthogonal Distances* (LADOD). It has been shown that depending on the set of data the different methods can have better prognostic abilities according to the criteria of external or internal validation. It is shown that with the use of artificial neural networks, based on the preliminary ordered by the method of  $L_1$ -regularization descriptor set, high-quality predictions of the properties of matter can also be made. The obtained linear regression equations were also compared with alternative approaches that work with non-shrunked (non-optimized) descriptor sets, namely: with the PCR method (Principal Component Regression), as well as with the PLS method (Partial Least Squares or Projection on Latent Structure). It should be noted that although with the use of these methods for some problems we obtained fairly reliable predictive models, however, such models do not provide clear information about the nature of the obtained equations and do not answer the question of what structural and chemical features, or molecular descriptors, lead to changes in response (activity). In the studied examples, we used  $L_1$ -regularization to formulate compact one-, two- or three-parametric models that are able to satisfactorily describe the data set. According to the studied examples, the models obtained with pre-selection, using LARS-LASSO, turned out to be better than the results of PLS and PCR calculations.

In the proposed PhD thesis some attention is paid to validation methods and quality of regression equations estimates. For this purpose, a model problem was used in which errors were introduced in both the dependent and independent variables. To simplify the analysis, as well as to study the validation characteristics of the equations obtained in all the studied methods of linear regression, we considered the simplest, but not the trivial case - regression with one independent variable. Such formulation of the problem made it possible to estimate the equations in accordance with the proximity of the coefficients of the regression equations to the "ideal" theoretical values. With the use of the mentioned model problem, the influence of rational sampling on the training and test sets on the quality of the obtained regression equations was investigated. It has been shown that random single sampling is not informative because, depending on the molecules in the test sample, the

validation characteristics for the initial (complete) sample can be both very poor, leading to underestimation, and very good, leading to overestimation of the equation quality. Therefore, it is shown that in order to adequately estimate the quality of the regression equation, as well as to study the quality of the input data in general, it is necessary to create and study as many samplings into a training and test sample as possible. The influence of taking into account the limitations of Applicability Domain (AD) of the model on the validation characteristics of regression equations was also investigated. It is established that, at random divisions of the sample into training and test maxima of distribution of internal and external characteristics, as a rule, coincide. In contrast, when the breakdowns into training and test samples are performed in such a way that the test sample is in the AD of model obtained from the training sample, the corresponding distribution of external validation criteria is shifted relative to the internal in the direction of increase. The most informative are the partitions, which fall close to the maximum density of points. It is from the analysis of these areas one gets the most complete adequate understanding of the quality of the model. The known validation criteria proposed to date were also investigated. Based on the model problem, it was concluded that some of the validation criteria are too highly correlated with each other, which makes their simultaneous use uninformative. Among the following parameters are the pairs: ( $R_{test}^2$  - CCC and  $Q_{F3}^2$  - RMSEP<sub>test</sub>). It is established that for data with substantial scatter the typical picture is the inverse (essentially nonlinear) dependence  $R_{train}^2 - R_{test}^2$ . In this case, the improvement (increase) of the internal validation coefficients ( $R_{train}^2$ ,  $Q_{LOO}^2$ ) is generally not the evidence of an improvement in the oracle properties of the model, because for linear regression at a sufficiently large number of points the relationship between these two criteria was always close to linear. However, the criterion  $Q_{LOO}^2$  can be successfully used for small samples. Based on the calculated data, it is shown that for most cases, the OLS method gave the best results. However, for large samples, with errors in both the dependent and independent variables, in the ODR (and LADOD) method, the best equations can be obtained.

Another problem that is closely related to the construction of statistical models is the construction of the classification function. For this purpose, the  $L_1$ -regularized calculation of logistic regression was performed in this work. Two problems are considered. In the first classification of molecules on strong and weak bases according to their binding affinity towards  $\text{Li}^+$  ion in the gas phase is carried out. In the second problem, organic molecules were classified as active or inactive according to the estrogen receptor relative binding affinity. It is shown that with the use of  $L_1$ -regularized logistic regression it is possible to achieve classification results that are competitive with those obtained using other, more complex in the computational sense, methods. The use of a special  $L_1$ -regularized algorithm (denoted as LR-LARS-LASSO) made it possible to obtain fairly simple classification equations that are interpretable (in contrast to the results obtained in other popular classification methods such as: Support-Vector Machines, Random Forest, Artificial Neural Networks). Also, the obtained logistic regression equations are unambiguous and reproducible.

It is shown that the  $L_1$ -regularization method can be used in quantum chemistry. Using the  $L_1$ -regularization procedure, it is possible to create an ordered (ranked) set of electronically excited configurations relative to the Gartree-Fock state. By including a different number of configurations from the created set, it is possible to obtain a progressive set of approximations to the exact calculations of the methods. The method is implemented in the framework of Meller-Plessett's theory of second-order perturbations (MP2) and different levels of the coupled clusters theory. It has been shown that such approximate solutions give fairly accurate values of the energy characteristics of molecules, and the number of configurations in the calculations can be much lower than in calculations using a complete configuration set of the exact method. A number of computational algorithms using first-order multi-step methods have been implemented to effectively solve the corresponding equations of the coupled clusters theory.

**Key words:**  $L_1$ -regularization, QSAR/QSPR,  $\text{pK}_a$  of organic compounds, boiling point of organic compounds, viscosity of liquids, vapor pressure, Estrogen Receptor

Ligands, Lithium cation basicity, Linear regression, Ordinary Least Squares, Least Absolute Deviations, Orthogonal Distances Regression, Least Absolute Deviation of Orthogonal Distances, Artificial Neural Networks, validation, Logistic Regression, Moller-Plessett's perturbation theory, Coupled Clusters theory.

## СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

**Наукові праці у наукових фахових виданнях України, що входять до міжнародної наукометричної бази Scopus:**

- [1] Berdnyk, M. I.; Zakharov, A. B.; Ivanov, V. V. Application Of  $L_1$ -Regularization Approach In QSAR Problem. Linear Regression And Artificial Neural Networks. *Methods Objects Chem. Anal.* **2019**, *14* (2), 79–90. <https://doi.org/10.17721/moca.2019.79-90>.

(Особистий внесок здобувача: програмна реалізація застосованих методів регресії а також штучних нейронних мереж, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, участь у обговоренні результатів, написання публікації).

**Наукові праці в наукових фахових виданнях України**

- [2] Бердник, М. И.; Иванов, В. В. Многошаговые Методы Первого Порядка в Решении Уравнений Теории Связанных Кластеров. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія Хімія* **2015**, № 25, 39-45.

(Особистий внесок здобувача: програмна реалізація методу зв'язаних кластерів, а також методів оптимізації до впроваджених ітеративних схем, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання публікації).

- [3] Бердник, М. И.; Иванов, В. В.  $L_1$ -Регуляризация в Квантовой Химии.  $\pi$ -Электронная Теория Связанных Кластеров с Учетом Двукратных Возбуждений. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія Хімія* **2016**, № 26, 58–64.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованого методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованого методу, участь у обговоренні результатів, написання публікації).

- [4] Berdnyk, M. I.; Onizhuk, M. O.; Ivanov, V. V. Methods for Building Linear Regression Equations in the “Structure-Property” Problems. *Kharkov Univ. Bull. Chem. Ser.* **2018**, № 30, 6–17. <https://doi.org/10.26565/2220-637x-2018-30-01>.



(Особистий внесок здобувача: програмна реалізація застосованих методів регресії, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання публікації).

**Наукові праці, в яких опубліковані основні наукові результати дисертації у періодичних наукових виданнях закордонних держав, що входять до ОЕСР, і реферуються у міжнародній наукометричній базі Scopus**

[5] Ivanov, V. V.; Berdnyk, M. I.; Adamowicz, L.  $L_1$ -Regularisation of the Coupled-Cluster Solutions. *Mol. Phys.* **2017**, *115* (21–22), 2892–2902. <https://doi.org/10.1080/00268976.2017.1359345>.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованого методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованого методу, участь у обговоренні результатів, написання публікації).

**Наукові праці, які засвідчують апробацію матеріалів дисертації**

[6] Бердник, М. И.; Иванов, В. В.  $L_1$ -регуляризация. от статистики до квантовой химии, *Хімічні Каразінські читання - 2016* : тези доп. VIII всеукр. наук. конф. студентів та аспірантів, Харків, Україна, квітень 18–20, 2016; ХНУ імені В. Н. Каразіна: Харків, 2016; С. 132-133.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованих статистичних методів і методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[7] Бердник, М. И.; Иванов, В. В. Применение  $l_1$ -регуляризации в неэмпирических и полуэмпирических расчетах квантовой химии, *XII Всеукраїнська конференція молодих вчених та студентів з актуальних питань хімії* : збірка праць всеукр. наук. конф., Харків, Україна, травень 11-13, 2016; ДНУ НТК ІМК НАНУ: Харків, 2016; С. 32.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованого методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[8] Бердник, М.И.; Дяченко, А.В.; Иванов, В. В; Регрессионные модели QSAR, *Збірник тез доповідей, Хімічні Проблеми Сьогодення (ХПС-2018)*, Вінниця, Україна, березень 27-29, 2018; Донецький національний університет імені Василя Стуса: Вінниця, 2018; С. 177.

(Особистий внесок здобувача: програмна реалізація методу LARS-LASSO а також методів регресії, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[9] Berdnyk, M.; Ivanov, V.; Zakharov, A;  $L_1$ -Regularization In Different Applications Of Chemical Modeling, *Molecular Engineering And Computational Modelling For Nano- And Biotechnology: From Nanoelectronics To Biopolymers* : Book of Abstracts International Scientific Conference, Cherkasy, Ukraine, September 25–26, 2018; Bohdan Khmelnytsky Cherkasy National University: Cherkasy, 2018; P. 30-33.

(Особистий внесок здобувача: програмна реалізація методів регресії, регуляризованих квантовохімічних методів, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[10] Бердник, М.І.;  $L_1$ -регуляційний підхід у розрахунках фізикохімічних властивостей молекул, *Сучасні Проблеми Хімії* : тези доповідей XX Міжнародної конференції студентів та аспірантів, Київ, Україна, травень 15–17, 2019; Київський національний університет імені Тараса Шевченка: Київ, 2019; С. 140.

(Особистий внесок здобувача: програмна реалізація використаних методів, розрахунки фізико-хімічних властивостей молекул, участь у обговоренні результатів, написання тез, доповідь на конференції).

[11] Berdnyk, M. I.; Denysenko, K. A.; Zakharov, A. B.; Ivanov, V. V.; Validation Of Regression Equations In QSAR Problem, *Сучасні Тенденції 2020* : Тези доповідей Київської Конференції з аналітичної хімії, Київ, Україна,

жовтень 21-23, 2020; Київський національний університет імені Тараса Шевченка: Київ, 2020; С.79-80.

(Особистий внесок здобувача: програмна реалізація методів валідації, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[12] Денисенко, К. А.; Бердник, М. И.; Захаров, А. Б.; Метод валидации уравнений линейной регрессии, *Хімічні Каразінські читання - 2021* : тези доп. XIII всеукр. наук. конф. студентів та аспірантів, Харків, Україна, квітень 20–21, 2021; ХНУ імені В. Н. Каразіна: Харків, 2021; С. 122-123.

(Особистий внесок здобувача: програмна реалізація методів валідації, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, участь у написанні тез).

[13] Berdnyk, M.; Ivanov, V.; Application Of Lasso Logistic Regression To Classification Problems In Chemistry, *Modern Chemistry Problems* : Book of abstracts XXII International Conference for Students, PhD Students and Young Scientists, Київ, Україна, травень 19–21, 2021; Київський національний університет імені Тараса Шевченка: Київ, 2021; С. 9.

(Особистий внесок здобувача: програмна реалізація методів класифікації, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

## ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ.....	19
Вступ.....	24
РОЗДІЛ 1. $L_1$ -регуляризація і сучасні статистичні підходи до побудови моделей "структура-активність" (Літературний огляд).....	30
1.1. Метод найменших квадратів.....	30
1.2. Метод найменших квадратів із регуляризацією.....	32
1.3. Знаходження $L_1$ -регуляризованих розв'язків методу найменших квадратів.....	35
1.4. Альтернативні методи побудови лінійної регресії.....	39
1.4.1. Метод найменших модулів.....	40
1.4.2. Метод ортогональної регресії та метод абсолютних відхилень ортогональних відстаней.....	42
1.4.3. Регресійні моделі основані на аналізі головних компонент (методи PCR та PLS).....	43
1.5. Логістична регресія.....	46
1.6. Проблема валідації регресійних QSAR/QSPR рівнянь.....	48
1.6.1. Внутрішня валідація.....	49
1.6.2. Y-Рандомізація.....	51
1.6.3. Зовнішня валідація.....	52
1.6.4. Рациональне розбиття вибірки на навчаючу та тестову.....	53
Висновки до розділу 1.....	55
РОЗДІЛ 2. Лінійні $L_1$ -регуляризаційні моделі в описі фізико-хімічних параметрів молекул.....	57
2.1. $L_1$ -регуляризація середнього значення ( <i>a toy example</i> ).....	60
2.2. Алгоритм LARS-LASSO.....	62
2.3. Робочі формули для розрахунку регресійних рівнянь.....	63
2.3.1. Побудова регресійного рівняння методом LAD.....	64
2.3.2. Однопараметрична регресія методом ODR.....	66
2.3.3. Однопараметрична регресія методом LADOD.....	67

2.4. Константи іонізації карбонових кислот.....	68
2.5. Оцінки якості неемпіричних розрахунків $pK_a$ фенолів.....	72
2.6. QSAR моделі опису констант іонізації органічних сполук різної природи.....	75
2.7. Температура кипіння органічних сульфідів.....	82
2.8. Температура кипіння флуороалканів.....	85
2.9. В'язкість рідин та тиск насиченого пару органічних сполук.....	91
Висновки до розділу 2.....	92
РОЗДІЛ 3. Тестові дослідження валідаційних характеристик регресійних QSAR/QSPR рівнянь.....	94
3.1. Вплив розміру тестової вибірки на критерії валідації.....	95
3.2. Великі вибірки даних ( $N=100$ , $sd(y)=20$ , $sd(x)=0$ ).....	97
3.3. Великі вибірки даних ( $N=100$ , $sd(y)=20$ , $sd(x)=10$ ).....	100
3.4. Середні за розміром вибірки ( $N=40$ , $sd(y) = 10$ , $sd(x)=0$ ).....	102
3.5. Малі за розміром вибірки ( $N=20$ , $sd(y)=5$ , $sd(x)=2.5$ ).....	107
Висновки до розділу 3.....	109
РОЗДІЛ 4. $L_1$ регуляризація в побудові класифікаційних моделей.....	111
4.1. Загальна проблема побудови класифікаційних функцій.....	111
4.2. Алгоритм розрахунку LR-LARS-LASSO.....	114
4.3. Тестові набори даних.....	115
4.4. Результати LR-LARS-LASSO розрахунків.....	119
Висновки до розділу 4.....	123
РОЗДІЛ 5. Можливості використання $L_1$ -регуляризації в квантові хімії.....	125
5.1. Градієнтні алгоритми розв'язку нелінійних рівнянь теорії CC.....	126
5.1.1. Тестові оцінки ефективності різних алгоритмів.....	129
5.2. Теорія зв'язаних кластерів і регуляризація.....	135
5.3. Теоретичні засади $L_1$ -CC розрахунків.....	137
5.4. Напівемпіричні розрахунки $L_1$ -CC.....	139
5.5. Неемпіричні розрахунки $L_1$ -CC.....	145
5.5.1. $L_1$ -CCD та $L_1$ -CCSD розрахунки дисоціації молекули LiH.....	145

5.5.2. $L_1$ -CCSD розрахунки молекули ВН.....	150
5.5.3. Симетрична дисоціація молекули води в теорії $L_1$ -CCSD.....	153
Висновки до розділу 5.....	154
ЗАГАЛЬНІ ВИСНОВКИ.....	156
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	158
ДОДАТОК А СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ.....	181
ДОДАТОК Б Простий приклад регуляризації: розрахунок $L_1$ -регуляризованого середнього значення .....	185
ДОДАТОК В Константи іонізації органічних сполук різної природи.....	187
ДОДАТОК Г Температури кипіння флуороалканів.....	196
ДОДАТОК Д Розбиття на кластери органічних сполук основних до катіону Li .....	197

### ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

Скорочення	Розшифровка скорочення
AD	<i>Applicability domain</i> (Границі застосовності моделі)
ANN	<i>Artificial Neural Networks</i> (Штучна нейронна мережа)
AUC	<i>Area Under Curve</i> (Площа під кривою)
CC	<i>Coupled Cluster method</i> (теорія зв'язаних кластерів)
CCC	<i>Concordance correlation coefficient</i> (Узгоджений кореляційний коефіцієнт)
CCD	<i>Coupled Cluster Doubles</i> (Теорія зв'язаних кластерів з урахуванням двократних збуджень)
CCSD	<i>Coupled Cluster Singles and Doubles</i> (Теорія зв'язаних кластерів з урахуванням однократних і двократних збуджень)
DIIS	<i>Direct Inverse in Iterative Subspace</i> (Пряме обернення в ітеративному підпросторі)
EIV	<i>Errors in variables</i> (Похибка в усіх змінних)
FS	<i>Forward Stepwise</i> (Покроковий, "уперед" метод)
FPR	<i>False positive rate</i> (Частка об'єктів що помилково віднесені до активних, відносно загальної кількості неактивних молекул)
GA	<i>Genetic Algorithms</i> (Генетичний алгоритм)
HF	<i>Hartree-Fock</i>

	(Метод Гартрі-Фока)
НВ	<i>Heavy Ball</i> (Ітераційний метод «важкої кульки»)
НОМО	<i>Highest Occupied Molecular Orbital</i> (Найвища зайнята молекулярна орбіталь)
ISTA	<i>Iterative Shrinkage-Thresholding Algorithms</i> (Ітеративний алгоритм бар'єрного скорочення)
ІС	<i>Index of Ideality of Correlation</i> (Індекс ідеальності кореляції)
KNN	<i>k-nearest neighbors</i> (Метод k-найближчих сусідів)
LASSO	<i>Least Absolute Selection and Shrinkage Operator</i> (Найменший абсолютний вибір із оператором скорочення)
LARS	<i>Least-Angle Regression Stagewise</i> (Покрокова регресія: метод найменших кутів)
LAD	<i>Least Absolute Deviation</i> (Метод найменших модулів)
LADOD	<i>Least Absolute Deviation of Orthogonal Distances</i> (Метод найменших модулів ортогональних відстаней)
LOO або LOO-CV	<i>Leave-one-out</i> або <i>Leave-one-out Cross Validation</i> (Процедура перехресного оцінювання з вилученням точок по одній)
$L_2$ -OLS	<i>L<sub>2</sub>- Ordinary Least Squares (ridge regression)</i> (Метод найменших квадратів із регуляризацією за Тихоновим, інша назва – гребенева регресія)
$L_1$ -CCD	<i>L<sub>1</sub> - Coupled Cluster Doubles</i> ( $L_1$ регуляризований метод CCD)
$L_1$ -CCSD	<i>L<sub>1</sub> - Coupled Cluster Singles and Doubles</i> ( $L_1$ регуляризований метод CCSD)



$L_1$ -CCSDT	<i>L<sub>1</sub> – Coupled Cluster Singles, Doubles and Triples</i> ( $L_1$ регуляризований метод CCSDT)
LR	<i>Logistic Regression</i> (Логістична регресія)
LR-LARS-LASSO	<i>LARS-LASSO Logistic Regression</i> (Логістична регресія з відбором предикторів за допомогою методу LARS-LASSO)
LUMO	<i>Lowest Unoccupied Molecular Orbital</i> (Найнижча вакантна молекулярна орбіталь)
MP	<i>Møller–Plesset perturbation theory</i> (Багаточастинкова теорія збурень Меллера-Плессет)
MO	<i>Molecular Orbital</i> (Молекулярна орбіталь)
MAE	<i>Mean absolute Error</i> (Середня абсолютна помилка)
MCSCF	<i>Many Configurational Self Consistent field method</i> (Багатоконфігураційний метод самоузгодженого поля)
NIPALS	<i>Nonlinear Iterative Partial Least Squares</i> (Нелінійний ітеративний неповний метод найменших квадратів)
NPE	<i>Nonparallelity error</i> (Похибка непаралельності потенціальної кривої відносно точної функції)
OLS (або МНК)	<i>Ordinary Least Squares</i> (Стандартний метод найменших квадратів)
ODR	<i>Orthogonal Distances Regression</i> (Ортогональна регресія)
PCA	<i>Principal component analysis</i> (Аналіз головних компонент)
PCR	<i>Principal Component Regression</i>

	(Регресія на головних компонентах)
PLS	<i>Partial Least Squares or Projection on Latent Structures</i> (Неповний метод найменших квадратів або проекція на латентні структури)
PPP (або ППП)	<i>Pariser-Parr-Pople approach</i> ( $\pi$ -електронне наближення – Метод Парізера-Парра-Попла для розрахунку $\pi$ -електронних систем)
QSAR	<i>Quantitative Structure-Activity Relationship</i> (Кількісний зв'язок структура-активність)
QSPR	<i>Quantitative Structure-Property Relationship</i> (Кількісний зв'язок структура-властивість)
RMSE	<i>Residual Mean Square Error</i> (Остаточна середньоквадратична похибка)
RMSEP	<i>Root Mean Square Error of Prediction</i> (Середньоквадратична помилка прогнозу)
RS	<i>Rational selection</i> (Метод раціонального розбиття на тестову і навчальну вибірки)
RF	<i>Random Forest</i> (Випадковий ліс)
ROC	<i>Receiver Operating Characteristic</i> (Робоча характеристика приймача – характеристика точності класифікаційної функції)
SVM	<i>Support Vector Machine</i> (Метод опорних векторів)
SVD	<i>Singular Values Decomposition</i> (Метод сингулярного розкладу довільної матриці)
ST	<i>Soft Threshold</i> (М'який поріг)
SGA	<i>Standard Gradient Approach</i>

	(Стандартний градієнтний метод)
TLS	<i>Total Least Squares</i> (Узагальнений, тотальний, метод найменших квадратів)
TPR	<i>True positive rate</i> (Частка між кількістю правильно визначених активних молекул та загальною кількістю активних молекул)
TIC1	<i>Total informational content index</i> (Інформаційний топологічний індекс першого порядку)

## Вступ

Побудова статистичних (хемометричних) моделей за типом "структура-активність" (в англomовній літературі QSAR) є необхідним етапом багатьох досліджень у різних галузях хімії. Зокрема, такі моделі затребувані у фізичній хімії: в описі температур кипіння, плавлення, ліпофільності, критичних параметрів й інших характеристик речовини, термодинамічних параметрів систем і процесів тощо. Інша галузь, що потребує такі моделі, є дослідження й прогноз біологічної активності органічних сполук (лікарські ефекти, токсичність). Для побудови моделей QSAR хемометрія та хемоінформатика на основі статистики пропонує широкий набір різних підходів. Серед них, зокрема, регресійний аналіз, методи класифікації й метод нейронних мереж. Однак на шляху реалізації вказаних підходів постає **актуальна й дуже загальна проблема** – проблема відбору найбільш важливих теоретичних параметрів молекул (незалежних змінних – дескрипторів), яких було б достатньо для адекватного опису шуканої активності чи властивості.

Серед існуючих підходів для такого відбору, звертає на себе увагу метод  $L_1$ -регуляризації (R. Tibshirani, 1996). Цей метод успішно використовується у вирішенні певних технічних завдань, але наразі він ще не використовувався в хімії. Ряд характеристик цього методу може зробити його корисним у розв'язку різноманітних задач QSAR. Серед них – побудова компактних (малопараметричних) регресійних та класифікаційних моделей. Також цієї проблеми безпосередньо стосується задача побудови регресійного рівняння для даних зі значним розкидом, що є типовим для досліджень QSAR. У зв'язку зі сказаним, **актуальною задачею** є вивчення можливостей  $L_1$ -регуляризації, створення відповідних алгоритмів і спеціалізованих комп'ютерних програм для потреб QSAR, що реалізують: а) підходи до відбору дескрипторів; б) альтернативні методи побудови лінійних рівнянь і бінарних класифікаційних функцій; с) методи валідації (тестування) прогностичної здатності отриманих моделей. Також **актуальною є задача** аналізу даних щодо різних фізико-хімічних характеристик молекулярних систем, зокрема тих, які мають значення

як лікарські сполуки (константи іонізації  $pK_a$ , температури кипіння, в'язкість та ін.). Такий аналіз було проведено із використанням розроблених програм.

Іншим аспектом дисертаційної роботи, що стосується  $L_1$ -регуляризації, є можливість її використання в квантовій хімії. Незважаючи на досить різні основи статистичної науки і квантової механіки молекул, необхідність ефективного скорочення системи незалежних параметрів є загальною проблемою. Уже давно було усвідомлено, що для адекватного розрахунку енергетичних параметрів молекулярних систем, необхідне урахування ефектів електронної кореляції. У цій царині, серед класичних наближень квантової теорії, відомі такі методи як-от: багаточастинкова теорія збурень та теорія зв'язаних кластерів. Реалізація цих підходів потребує явного включення у хвильову функцію електронно-збуджених конфігурацій різної кратності. Але сучасні розширені базиси атомних орбіталей ведуть до надто широкого конфігураційного простору, робота з яким пов'язана зі значними витратами комп'ютерних ресурсів. Отже, постає проблема стиснення (скорочення) конфігураційного складу хвильової функції. Ми вважаємо, що процедура  $L_1$ -регуляризації може бути корисним й ефективним підходом до такої задачі.

**Мета і завдання дослідження**, у зв'язку із вищесказаним, полягали у вивченні можливостей  $L_1$ -регуляризації в задачах, що стосуються побудови прогностичних QSAR-моделей фізико-хімічних властивостей і квантовій хімії молекул. Для цього необхідно:

- розробити основні розрахункові алгоритми  $L_1$ -регуляризації, реалізувати їх у комп'ютерних програмах та провести тестові розрахунки;
- програмно реалізувати й дослідити різні альтернативні способи побудови лінійної регресії для систем зі значним розкидом даних;
- провести конкретні розрахунки регресійних моделей констант іонізації ( $pK_a$ ) різних органічних сполук, порівняти результати, що отримані в різних регресійних моделях, із прогнозами штучних нейронних мереж;

- дослідити регресійні моделі опису температур кипіння (ВР) органічних сполук різної будови;
- програмно реалізувати й дослідити методи валідації отриманих регресійних рівнянь, провести модельні розрахунки валідаційних параметрів;
- розробити, програмно реалізувати й тестувати метод побудови класифікаційних функцій на основі  $L_1$ -регуляризації;
- побудувати бінарну класифікаційну функцію для опису основності органічних сполук до катіону літію;
- побудувати загальну класифікаційну функцію для опису стероїдних та не стероїдних лігандів рецептору естрогену;
- дослідити можливість використання  $L_1$ -регуляризації в проблемі створення наближених моделей хвильової функції квантовохімічних методів урахування електронної кореляції на основі теорії збурень і теорії зв'язаних кластерів;
- оптимізувати ітераційну процедуру пошуку розв'язків теорії зв'язаних кластерів;
- провести тестування розробленого підходу на прикладі розрахунків  $\pi$ -спряжених систем (полієни, ароматика, каліцени) методами теорії зв'язаних кластерів;
- провести розрахунки кривих дисоціації малих молекул.

**Предмет дослідження:** регресійні й класифікаційні моделі QSAR, отримані за допомогою  $L_1$ -регуляризації; альтернативні методи побудови лінійної регресії; порівняльний аналіз різних регресійних підходів;  $pK_a$  органічних сполук; температури кипіння органічних сполук; ліганди естрогену;  $L_1$ -регуляризовані багаточастинкові квантовохімічні методи (MP2, CCSD);

**Об'єкт дослідження:** фізико-хімічні характеристики молекулярних систем, як-то: константи іонізації органічних сполук, температури кипіння,

в'язкість рідин; хвильові функції теорії зв'язаних кластерів побудовані для  $\pi$ -спряжених систем та для двохатомних молекул; криві дисоціації двохатомних молекул.

**Методи дослідження:** лінійний регресійний аналіз як метод побудови моделей, що здатні прогнозувати величини активностей/властивостей; логістичний регресійний аналіз для побудови класифікаційних функцій; багаточастинкові методи квантової хімії, як-от: MP2, CCD, CCSD, які здатні адекватно описати енергетичні характеристики молекул. Квантовохімічні методи оптимізації геометрії органічних молекул.

### **Наукова новизна отриманих результатів.**

- Показано, що  $L_1$ -регуляризація здатна створити ранжований список молекулярних дескрипторів, завдяки чому можлива побудова простих малопараметричних регресійних рівнянь на основі методу найменших квадратів, методу найменших модулів, методу ортогональних відстаней (OLS, LAD, ODR).
- Показано, що отримані на основі алгоритму LARS-LASSO малопараметричні регресійні рівняння (OLS, LAD) можуть мати значно кращі прогностичні характеристики, ніж стандартні методи, у яких не робиться відбір дескрипторів (наприклад, неповний метод найменших квадратів, PLS).
- Отримано регресійні рівняння для опису констант іонізації ( $pK_a$ ) органічних кислот та основ різної будови.
- Отримано регресійні рівняння для опису температур кипіння ( $BP$ ) різних класів органічних сполук у залежності від будови молекули.
- Представлено умови, за яких розділення вхідних даних на тестову й тренувальну вибірки гарантують адекватну оцінку обраної моделі.
- Показано, що алгоритм LARS-LASSO може бути з успіхом використано в побудові компактного рівняння логістичної регресії для бінарної класифікації молекул за активністю.

- Побудовано загальну класифікаційну функцію, що описує основність органічних сполук різної природи до катіонів літію.
- Отримано класифікаційну функцію, що дозволяє описати активність органічних сполук стероїдної та нестероїдної природи по відношенню до рецепторів естрогену.
- Вперше показано, що за допомогою методу  $L_1$ -регуляризації можливе створення ранжованого списку електронно-збуджених конфігурацій, який може бути використаний для створення прогресивної системи наближень квантовохімічного методу (на прикладі MP2, CCD, CCSD).

**Особистий внесок здобувача** полягає в аналізі літературних даних за темою дисертації. Особисто здобувачем створено ряд нових квантовохімічних і хемометричних комп'ютерних програм. Також особисто здобувачем зроблено розрахунки різних регресійних моделей, навчання нейронних мереж та квантовохімічні розрахунки молекул. Формулювання напряму досліджень, інтерпретація результатів розрахунків та написання статей зроблено спільно з науковим керівником проф. В. В. Івановим. Співавтори публікацій – проф. Л. Адамович (університет Арізони, м. Тусон), доц. А. Б. Захаров та М. Оніжук (Харківський національний університет імені В. Н. Каразіна) приймали участь в обговоренні результатів та написанні статей.

**Апробація матеріалів дисертації.** Основні результати роботи були представлені на VIII Всеукраїнській науковій конференції студентів та аспірантів «Хімічні Каразінські читання – 2016» (Харків, 2016), XII Всеукраїнській конференції молодих вчених та студентів з актуальних питань хімії (Харків, 2016), I Міжнародній (XI Української) науковій конференції студентів, аспірантів і молодих вчених «Хімічні проблеми сьогодення» (Вінниця, 2018), міжнародній науковій конференції: «Молекулярна інженерія та комп'ютерне моделювання для нано- і біотехнологій: від наноелектроніки до біополімерів» (Черкаси, 2018), XX Міжнародній



конференції студентів та аспірантів «Сучасні Проблеми Хімії» (Київ, 2019), Київській Конференції з аналітичної хімії «Сучасні Тенденції 2020» (Київ, 2020), XIII Всеукраїнській науковій конференції студентів та аспірантів «Хімічні Каразінські читання - 2021» (Харків, 2021) та XXII Міжнародній конференції студентів, аспірантів та молодих вчених «Сучасні Проблеми Хімії» (Київ, 2021).

**Структура та обсяг дисертації.** Дисертаційна робота складається зі вступу, 5 розділів, загальних висновків, списку використаних джерел та 5 додатків. Обсяг загального тексту дисертації складає 205 сторінок, з них основного тексту 142 сторінки. Робота ілюстрована 19 таблицями та 47 рисунками. Список використаних джерел містить 218 найменувань.

**Зв'язок роботи з науковими програмами, планами, темами.** Протягом виконання дисертаційної роботи, результати досліджень були використані в НДР кафедри хімічної матеріалознавства Харківського національного університету імені В. Н. Каразіна: "Органічні модифікатори та іон-молекулярні системи й нові матеріали на їх основі для аналітичного та електрохімічного застосування" № держреєстрації 0118U002025.

**Практичне значення отриманих результатів.** Одним із важливих практичних результатів роботи є створений здобувачем комплекс статистичних програм, за допомогою яких можлива побудова компактного набору дескрипторів і далі альтернативних регресійних, або класифікаційних моделей різноманітних характеристик органічних молекул, а також тестування прогностичної здатності цих моделей. Здобувачем створено ряд моделей (регресійних і нейромережевих) для опису констант іонізації органічних систем різної природи. Ця інформація є принципово важливою у дослідженні й розробці сполук із біологічною активністю.

# РОЗДІЛ 1

## **$L_1$ -РЕГУЛЯРИЗАЦІЯ І СУЧАСНІ СТАТИСТИЧНІ ПІДХОДИ ДО ПОБУДОВИ МОДЕЛЕЙ "СТРУКТУРА-АКТИВНІСТЬ" (ЛІТЕРАТУРНИЙ ОГЛЯД)**

У представленому розділі наведено літературний огляд сучасних статистичних методів, які є об'єктами дисертаційного дослідження. Дано короткий опис підходів, що ґрунтуються на прийомі регуляризації і, в особливості,  $L_1$ -регуляризації. Ми розглядаємо  $L_1$ -регуляризацію як один з ефективних прийомів скорочення набору незалежних змінних, можливості використання якого в хімії майже не досліджені. У розділі також надано опис методів побудови лінійних регресійних моделей, які також недостатньо висвітлені в сучасній науковій літературі. Крім стандартного методу найменших квадратів, дано короткий опис альтернативних підходів, які можуть бути корисними при обробці даних із достатньо великим розкидом (розсіюванням).

Втім, у цьому розділі ми **не розглядаємо** методи, що ґрунтуються на стохастичних стратегіях перебору предикторів. Крім, власне, повного перебору, до них ми відносимо також метод генетичних алгоритмів (див. наприклад (1,2)), метод "мурашиної ферми" (*ant colony*)<sup>3,4</sup> та метод "випадковий ліс" (*random forest*)<sup>5,6</sup>. Вказані підходи ідеологічно відрізняються від  $L_1$ -регуляризаційної техніки, яка є предметом нашого дослідження. Зауважимо, до речі, що на відміну від регуляризаційних підходів, вказані методи перебору схильні до перенавчання<sup>7</sup>.

Певну увагу в огляді також приділено способам оцінки якості отриманих регресійних рівнянь (проблеми валідації).

### **1.1 Метод найменших квадратів**

Метод найменших квадратів, який ми вважаємо стандартним методом побудови регресійних рівнянь (*Ordinary Least Squares*, OLS<sup>1</sup>), широко

---

<sup>1</sup> У зв'язку з тим, що предметом дисертації є низка сучасних статистичних методів, для яких вже встановлено типові англійські абревіатури, то, надалі, ми і будемо їх використовувати.

висвітлено в науковій літературі (див. наприклад (8-10)). Тому, у цьому підрозділі ми опишемо лише загальні моменти, які є концептуально важливими для подальшого викладення.

Отже, вважаємо, що молекулярні дескриптори (незалежні змінні) сконцентровано у векторах-стовпчиках  $x_j$ . Тоді рівняння регресії для залежної змінної  $y$  (властивість, активність, "лінійний відгук") може бути представлено наступним чином:

$$y_i = \beta_0 + \sum_{j=1}^m \beta_j x_{ij} + e_i, \quad (1.1)$$

тут індекс "i" відповідає номеру системи (молекули), а індекс "j" – є номером дескриптору.  $\beta$  – коефіцієнти регресії,  $m$  – кількість дескрипторів, а  $e_i$  – випадкова похибка. Розв'язок задачі OLS (1.1) може бути отримано мінімізацією функції

$$\min_{\beta_0, \beta} \left[ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^m x_{ij} \beta_j)^2 \right]. \quad (1.2)$$

Тут  $n$  – кількість молекул (об'єктів). Мінімізаційна задача (1.2) в компактному матричному запису має вигляд:

$$\beta_{OLS} = \arg \min_{\beta} \|Y - X\beta\|_2^2, \quad (1.3)$$

де

$$Y = \{y_i\}; \quad X = \{1, x_{i,1}, x_{i,2}, x_{i,3}, \dots, x_{i,m}\}; \quad \beta = \{\beta_0, \beta_1, \dots, \beta_m\} \quad i = 1, \dots, n, \quad (1.4)$$

а символ  $\|\dots\|_2^2$  відповідає квадрату евклідової норми вектору ( $L_2$  – норма). Для довільного вектору  $Z$ :  $\|Z\|_2^2 = Z^+ Z = \sum_i z_i^2$ . Верхній індекс "плюс" ( $Z^+$ ) означає операцію транспонування.

Якщо матриця  $X^+ X$  є невинродженою і відповідна обернена  $(X^+ X)^{-1}$  існує, то розв'язок OLS може бути отримано за відомою матричною формулою<sup>8,9</sup>:

$$\beta = (X^+ X)^{-1} X^+ Y. \quad (1.5)$$

Тут і надалі, ми будемо вважати, якщо не сказано інакше, що матриця  $X$  має перший одиничний стовпчик, що дозволяє урахувати "вільний" член  $\beta_0$ .

## 1.2. Метод найменших квадратів із регуляризацією

Наразі було запропоновано і досліджено багато  $L_q$  норм (зокрема норм матриць і векторів), при  $q \geq 0$ <sup>11-13</sup> в якості регуляризуючих параметрів. Найбільш вивченою з практичної точки зору є  $L_2$ -регуляризація, або регуляризація некоректних задач за Тихоновим<sup>14,15</sup>. В англomовній літературі з статистики  $L_2$ -регуляризація OLS позначається як гребенева регресія (*ridge regression*)<sup>16,17</sup>. Цей метод дозволяє надати наближений розв'язок OLS (1.1, 1.2) у ситуації, коли задача (1.5) є погано обумовленою. Тобто матриця  $X^+X$  може бути **квазівиродженою** із великим числом обумовленості. Результати розрахунків з такою матрицею стають надто чутливими до варіацій вхідних даних. Спеціальним випадком використання  $L_2$ -регуляризації є ситуація, коли матриця  $X^+X$  є виродженою і взагалі не може бути обернена. Зазвичай, для таких проблем в англomовній літературі використовується термін *ill-posed tasks*. Такі ситуації реалізуються, наприклад, у випадках, коли дескриптори навчаючої вибірки лінійно зв'язані. Це може трапитись у задачах, де кількість дескрипторів (**m**) значно перевищує кількість зразків/молекул/об'єктів (**n**). У такому разі розв'язок усе ще може бути отримано завдяки введенню  $L_2$  норми<sup>14,15</sup>:

$$\min_{\beta_0, \beta} \left[ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^m x_{ij} \beta_j)^2 \right], \quad (1.6)$$

за умови:  $\|\beta\|_2 \equiv \sum_{j=1}^m \beta_j^2 \leq t^2$ .

Отже, особливістю  $L_2$ -регуляризованого методу OLS ( $L_2$ -OLS, *ridge regression*) є оптимізація з обмеженням за евклідовою нормою на величини регресійних коефіцієнтів.

Інший підхід до розв'язку погано-зумовлених задач регресії, що тісно пов'язаний із (1.6), є метод псевдорозв'язку, у якому за методом Мура-Пенроуза знаходять псевдообернену матрицю  $X^+X$ <sup>18,19</sup>

Існує безліч прикладів успіху використання  $L_2$ -OLS у різних галузях науки (зокрема хімії) і техніки (див. наприклад (20,21)).

Інший спосіб регуляризації пов'язано з введенням обмеження на  $L_1$ -норму шуканого набору регресійних коефіцієнтів (метод LASSO – *Least Absolute Selection and Shrinkage Operator*)<sup>11,22</sup>:

$$\min_{\beta_0, \beta} \left[ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^m x_{ij} \beta_j)^2 \right], \quad (1.7)$$

за умови  $\|\beta\|_1 \equiv \sum_{j=1}^m |\beta_j| \leq t$ .

Задачі (1.6) та (1.7), хоча візуально й схожі, але відрізняються алгоритмічною складністю, а також дають суттєво різні розв'язки OLS задачі. Згідно Tibshirani *et al*<sup>11,22,23</sup> ми наведемо якісну графічну ілюстрацію різниці між  $L_1$  та  $L_2$  регуляризацією (рис. 1.1).

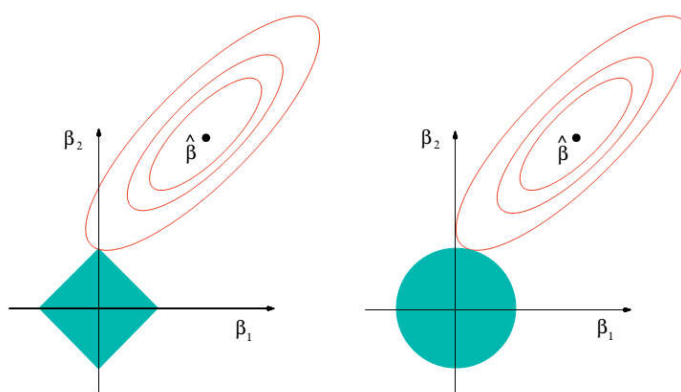


Рис. 1.1  $L_1$ - (зліва) і  $L_2$ - (справа) регуляризація в двовимірному просторі дескрипторів

Тут можна бачити, що за певної величини норми  $\|\beta\|_1$  двовимірний простір дескрипторів, при  $L_1$ -регуляризації (на відміну від  $L_2$ -регуляризації), скорочується до одновимірного ( $\beta_1 = 0$ ). Отже  $L_1$ -регуляризовані рішення можуть давати лінійні залежності, що є більш компактними.

Рис. 1.2 ілюструє практичне використання таких властивостей  $L_1$ -регуляризованих рішень задачі OLS. Тут розглядається задача, яка включає в лінійну модель вісім дескрипторів. Можна бачити, що при зменшенні граничного значення норми  $t$ , коефіцієнти регресії  $\beta$  зменшуються. Але, якщо у

випадку  $L_2$ -регуляризації  $\beta$  зменшуються пропорційно й рівномірно, то для  $L_1$ -регуляризації для кожного дескриптору існує таке граничне значення норми  $\|\beta\|_1$ , коли відповідний коефіцієнт  $\beta$  дорівнюватиме нулю і, отже, відповідний дескриптор не дає внесок у регресійне рівняння. Ця властивість характерна тільки для  $L_q$ -регуляризацій з  $q \leq 1$  (22-23). Зауважимо при цьому, що для  $q < 1$ , задача регуляризації не є опуклою<sup>23</sup>, що значно ускладнює розв'язок оптимізаційної проблеми.

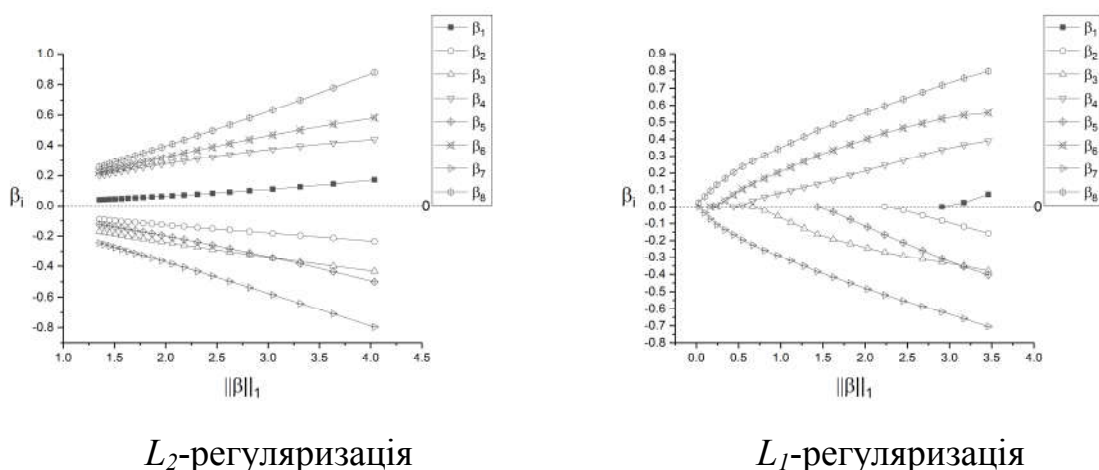


Рис. 1.2 Залежність коефіцієнтів регресії від  $L_1$ -норми вектору  $\beta$  для двох методів регуляризації

Детальне дослідження  $L_1$ -регуляризованих рішень показало, що вони мають кілька особливостей у порівнянні зі стандартним розв'язком задачі OLS, а також з іншими альтернативними методами побудови регресійних моделей<sup>11,22-24</sup>. А саме:

- 1)  $L_1$ -регуляризоване рішення OLS може бути отримано навіть якщо кількість дескрипторів значно більша за кількість молекул/зразків/об'єктів;
- 2)  $L_1$ -регуляризовані рішення можуть бути досить компактними й дозволяють виділити "найбільш важливі" дескриптори для опису досліджуваної властивості;
- 3) можна довести, що хоч методом повного перебору доступних дескрипторів можна виділити дескрипторні набори, що формально описують тренувальну (навчальну) вибірку більш точно, ніж  $L_1$ -рішення, але далеко не

завжди ці дескриптори надають більш точний опис тестової вибірки. А саме: згідно до (24), у ситуації, коли відносна похибка у вхідних даних велика, властивості прогнозу  $L_1$ -регуляризованих рівнянь можуть бути кращими за рівняння, що отримано повним перебором;

4) у порівнянні з альтернативними рівняннями, отриманими з використанням жадібного (*greedy*) алгоритму *Forward Stepwise*,  $L_1$ -регуляризовані рівняння майже завжди є кращими<sup>24</sup>;

5) похідна  $L_1$ -регуляризованої задачі – кусочно (локально) лінійна, і тому розв'язки задачі можуть бути отримані класичними методами, які маніпулюють з похідними. Задача повного перебору належить до так званого класу пр-повних задач, для яких, взагалі кажучи, не існує оптимальних алгоритмів<sup>24,25</sup>;

6) на відміну від  $L_1$ -OLS (LASSO), особливістю підходів, що базуються на так званих генетичних алгоритмах<sup>1,2</sup> (їх можна інтерпретувати як варіант стратегії перебору), є відсутність відтворюваності розв'язків. Тобто набори знайдених найважливіших дескрипторів є випадковими і, при наступному "прогоні" розрахунків, не відтворюються.

### 1.3. Знаходження $L_1$ -регуляризованих розв'язків методу найменших квадратів

Для знаходження  $L_1$ -регуляризованих розв'язків OLS, як правило, не використовують алгоритмічно складну задачу оптимізації з обмеженнями (1.7). Замість цього її переформулюють у так звану форму Лагранжа, яка має наступний матричний вигляд<sup>11,26</sup>:

$$\beta(\lambda)_{\text{LASSO}} = \arg \min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad (1.8)$$

де  $L_1$ -норма має вигляд:

$$\|\beta\|_1 = \sum_{i=1}^n |\beta_i| \quad (1.9)$$

У рівнянні (1.8)  $\lambda \geq 0$  – це коефіцієнт, що регулює внесок  $L_1$ -норми в загальний функціонал. Для довільного коефіцієнту  $\lambda$  завжди може бути

знайдено таке значення  $t$ , для якого розв'язок рівняння (1.7) буде еквівалентним до розв'язку рівняння (1.8)<sup>11,27</sup>. Також слід зазначити, що до рівняння (1.8) не входить  $\beta_0$ . Зазвичай, його позбуваються в рівняннях там, де це можливо, за рахунок процедури центрування дескрипторів/предикторів та властивості  $y$  на середнє значення:

$$Y^{\text{new}} = \{y_i - \bar{y}\}; \quad X = \{x_{i,1} - \bar{x}_1, x_{i,2} - \bar{x}_2, x_{i,3} - \bar{x}_3, \dots, x_{i,m} - \bar{x}_m\}; \quad i = 1, \dots, n, \quad (1.10)$$

де

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}; \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (1.11)$$

Слід зазначити, що така процедура дозволяє позбутися вільного члену тільки в рівняннях лінійної регресії. У рівняннях, наприклад, логістичної регресії (дивись підрозділ 1.5 і розділ 4) таким чином позбутися вільного члена неможливо.

Також, зазвичай, проводять нормалізацію на середнє-квадратичне:

$$Y^{\text{new}} = \{y_i / \hat{y}\}; \quad X = \left\{ \frac{x_{i,1}}{\hat{x}_1}, \frac{x_{i,2}}{\hat{x}_2}, \frac{x_{i,3}}{\hat{x}_3}, \dots, \frac{x_{i,m}}{\hat{x}_m} \right\}; \quad i = 1, \dots, n, \quad (1.12)$$

де

$$\hat{x}_j = \sqrt{\sum_{i=1}^n x_{ij}^2}, \quad \hat{y} = \sqrt{\sum_{i=1}^n y_i^2}. \quad (1.13)$$

Слід зазначити, що подібних нормалізацій існує декілька (див. наприклад (28)). Їх основна ідея – позбутися впливу одиниць вимірювання, а також зробити дескриптори "відносно однаковими". Вище наведено тільки ту нормалізацію, яка використовувалася в наших розрахунках. Якщо ж дескриптори є величинами в однакових одиницях вимірювання, то подібну нормалізацію, зазвичай, не роблять, щоб не втратити корисну інформацію<sup>11</sup>.

Як було вже сказано раніше, для отримання розв'язків LASSO задачі можна використовувати стандартні методи для роботи з похідними. Однак для того, щоб знайти похідну від функції (1.8), необхідно спочатку позбутися розриву в точці  $\|\beta\|_1 = 0$ .

Для цього, наприклад, можна представити  $\|\beta\|_1$  у наступному вигляді:



$$\|\beta\|_1 = \sum_i \frac{\beta_i \beta_i}{|\beta_i|} = \beta^+ V^{-1} \beta, \quad (1.14)$$

де  $V^{-1}$  – діагональна матриця<sup>29</sup>:

$$V^{-1} = \begin{pmatrix} 1/|\beta_1| & 0 & 0 & 0 \\ 0 & 1/|\beta_2| & 0 & 0 \\ 0 & 0 & 1/|\beta_3| & 0 \\ 0 & 0 & 0 & \dots \end{pmatrix}. \quad (1.15)$$

Слід зазначити, що у формі (1.15) діагональні елементи матриці є невизначеними коли  $\alpha_x \rightarrow 0$ , що критично, оскільки розв'язок задачі LASSO може мати, як було сказано вище, компакту форму з великою кількістю нульових регресійних коефіцієнтів. У такому разі, відповідний член прирівнюється до нуля:  $V_{ii}^{-1} = 0$ . Зауважимо, що така матриця може бути інтерпретована як псевдообернена. Такий підхід, при наближенні до розв'язків з великою кількістю нульових регресійних коефіцієнтів, призводить до того, що подальша робота алгоритму значно сповільнюється. Крім того, слід мати на увазі, що при використанні наближення (1.14-1.15) можливе отримання багатьох локальних мінімумів. Однією з причин цього є той факт, що при наближенні регресійного коефіцієнта до 0, на певному етапі ітераційної процедури його буде виключено з рівняння без можливості подальшого змінення<sup>30</sup>. Втім, у нашій практиці, а також у практиці багатьох авторів<sup>29,30</sup>, така ситуація не зустрічалась.

Зауважимо, що вираз (1.14) не є єдино можливим. Наприклад, може бути і таке представлення:  $\|\beta\|_1 = \sum_i \beta_i \text{sign}(\beta_i)$  воно, вочевидь, веде до іншої форми робочих рівнянь.

Використовуючи формулу (1.14), та беручи похідну від функції, що мінімізується (1.8), отримуємо вектор градієнту:

$$\frac{\partial}{\partial \beta^+} \left( \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right) = X^+ X \beta - X^+ Y + \lambda V^{-1} \beta. \quad (1.16)$$

Причому вважається, що матриця  $V$  є сталою на кожній ітерації. Прирівнюючи похідну до нуля отримуємо рівняння для регресійних коефіцієнтів:

$$\beta = (X^+X + \lambda V^{-1})^{-1} X^+Y. \quad (1.17)$$

Очевидно, що це рівняння розв'язується ітеративно, оскільки  $V$  (1.15), залежить від  $\beta$ .

Раніше було запропоновано також й інші процедури розв'язку рівняння (1.8) (див. наприклад (29)).

Наші дослідження показали, що зручними в задачах QSAR є група методів заснованих на ISTA (*Iterative Shrinkage-Thresholding Algorithms*)<sup>31</sup>. Це методи, які засновані на застосуванні оператору "стиснення" (оператору м'якого порогу):

$$T_\lambda(x_i) = (|x_i| - \lambda)_+ \text{sign}(x_i), \quad (1.18)$$

де операції  $(\dots)_+$  відповідає така умова:

$$(c)_+ = \begin{cases} c, & \text{if } c > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1.19)$$

Крок в алгоритмах ISTA має наступний вигляд:

$$\beta_{k+1} = T_\lambda(\beta_k - \mu X^+(X\beta_k - Y)), \quad (1.20)$$

де  $\mu$  – параметр величини шагу.

У наших дослідженнях найбільш ефективним методом виявився метод LARS (*Least-Angle Regression Stagewise*), за допомогою якого з використанням модифікації запропонованої у (32,33), можна отримати всі розв'язки LASSO (надалі цей алгоритм будемо позначати LARS-LASSO).

В алгоритмі LARS на кожному кроці до регресійної моделі додається лише один дескриптор (тут і надалі ми будемо позначати множину обраних дескрипторів –  $\epsilon$ ). На першій ітерації до "пустої" регресійної моделі ( $y = \beta_0$  або  $y = 0$ , якщо проведено операцію автомасштабування (1.10-1.13)) додається найбільш скорельований з передбачуваною властивістю дескриптор. Далі в

напрямку вектору дескриптору ( $X_\ell$ ) робиться максимально допустимий крок доки інший дескриптор ( $k$ ) не стане еквівалентно скорельованим із залишком:

$$Y_{\text{new}} = Y - X\beta_{\text{LASSO}}, \quad (1.21)$$

де вектор коефіцієнтів ( $\beta_{\text{LASSO}}$ ) після першої ітерації містить тільки один не нульовий елемент. Відповідну кореляцію ( $C$ ) розраховують за формулою:

$$C = X^T (Y - X\beta_{\text{LASSO}}). \quad (1.22)$$

Таким чином, до моделі додається дескриптор  $k$  і рух продовжується далі в напрямку цих двох дескрипторів таким чином, щоб кореляція дескрипторів множини  $\epsilon$  із залишком (1.22) була однаковою.

У методі LARS додавання дескрипторів до моделі може відбуватися допоки всі дескриптори не опиняться в множині  $\epsilon$ . Тоді відповідний розв'язок буде еквівалентним стандартному OLS методу без відбору дескрипторів.

У LARS-LASSO модифікації кожен раз, коли додається дескриптор до множини  $\epsilon$ , береться до уваги знак коефіцієнту при дескрипторі (більш детально алгоритм буде наведено в розділі 2.2). При цьому, коли протягом руху в еквікорельованому напрямку знак коефіцієнта одного з дескрипторів множини  $\epsilon$  змінюється до того як інший дескриптор виявиться однаково скорельованим з остатком (1.22), то рух у цьому напрямку зупиняється, а дескриптор має бути виключеним із моделі. Таким чином, на відміну від стандартної *stepwise* регресії, на кожній ітерації LARS-LASSO дескриптор може бути як додано до моделі, так і виключено з неї! Ця особливість означає, що LARS-LASSO нетривіально урахує факторну структуру даних.

#### 1.4. Альтернативні методи побудови лінійної регресії

$L_1$ -регуляризований розрахунок методу найменших квадратів дозволяє виділити найбільш важливі дескриптори (або, строго кажучи, послідовність дескрипторів) для опису молекулярної властивості. Далі цей виділений набір дескрипторів може бути використано деінде. Наприклад, у побудові QSAR/QSPR моделей різноманітних фізико-хімічних або біологічних

властивостей. При цьому зауважимо, що такі рівняння, в умовах певного розкиду вхідних даних, можуть бути отримані кількома альтернативними способами. Окрім OLS, лінійні рівняння можуть бути розраховані методом найменших модулів (*Least Absolute Deviation*, LAD)<sup>34,35</sup>, методом ортогональних відстаней (*Orthogonal Distances Regression*, ODR)<sup>36,37</sup>, а також запропонованим та вперше дослідженим нами методом найменших абсолютних відхилень ортогональних відстаней (*Least Absolute Deviation of Orthogonal Distances*, LADOD)<sup>38,39</sup>.

Слід зауважити, що перераховані методи (окрім OLS) не тільки не знайшли розповсюдження в хімічній науці, але, і це дуже дивно, взагалі не були достатньо досліджені. Тому однією з цілей представленої дисертації було якісне порівняння результатів регресійного аналізу реалізованого за допомогою різних (альтернативних) підходів.

#### 1.4.1. Метод найменших модулів

В методі LAD мінімізується наступна функція:

$$\beta_{\text{LAD}} = \arg \min_{\beta} \|Y - X\beta\|_1. \quad (1.23)$$

На відміну від виразу (1.3), тут використовується  $L_1$ -норма (1.9).

Цікаво, що метод найменших модулів було запропоновано ще у 1757 році. За 50 років до того, як у роботах Гауса й Лежандра з'явився метод OLS<sup>40</sup>! Не зважаючи на це, метод LAD і на сьогоднішній день майже не використовується. Зокрема, він не використовується і в хімії, і в дослідженнях QSAR. Це пов'язано з кількома обставинами:

- 1) апріорно вважається, що для моделі (вхідних даних) виконуються умови теореми Гауса-Маркова<sup>23</sup>. Звідси оцінки методом OLS є оптимальними;
- 2) розрахункова складність LAD значно вища ніж у OLS, оскільки похідна від функції, що мінімізується (1.23), не є безперервною функцією й розв'язок задачі (1.23) може бути реалізовано ітеративно.

Незважаючи на складність розрахункових алгоритмів LAD, він має певні переваги над OLS. Насамперед LAD – робастний підхід, тобто це підхід, який

може адекватно описати дані з "викидами" (англ. *outliers*). Робастність LAD пов'язана з тим, що форма (1.23) може бути реалізована як різновид зваженого методу OLS (дивись нижче). Вагові фактори LAD автоматично налаштовуються для відповідних точок даних. Отже LAD є аналогом зваженого методу OLS, хоч і не використовує апріорну інформацію стосовно похибок даних.

Існує кілька алгоритмів розв'язку проблеми LAD<sup>34,35</sup>. У нашій роботі було реалізовано алгоритм, що відповідає так званому "варіаційно-зваженому" методу Мудрова та Кушко<sup>41</sup>. У цьому методі вираз (1.23) трансформують у проблему зваженого методу OLS:

$$H_{LAD}(\beta) = \|Y - X\beta\|_1 = \sum_i |y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_m x_{im}| = \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_m x_{im})^2 / w_i. \quad (1.24)$$

Тут  $w_i$  – вагові множники:

$$w_i = |\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots - y_i|. \quad (1.25)$$

Використання такого підходу веде до системи лінійних рівнянь:

$$\begin{cases} \beta_0 \sum_i \frac{1}{w_i} + \beta_1 \sum_i \frac{x_{i1}}{w_i} + \beta_2 \sum_i \frac{x_{i2}}{w_i} + \dots = \sum_i \frac{y_i}{w_i} \\ \beta_0 \sum_i \frac{x_{i1}}{w_i} + \beta_1 \sum_i \frac{x_{i1}^2}{w_i} + \beta_2 \sum_i \frac{x_{i1} x_{i2}}{w_i} + \dots = \sum_i \frac{y_i x_{i1}}{w_i} \\ \dots \dots \dots \end{cases} \quad (1.26)$$

Рівняння (1.26) має бути вирішено ітеративно за процедурою самоузгодження коефіцієнтів  $\beta$ .

Варто також зазначити, що у випадку, коли матриця дескрипторів складається лише з одного дескриптору, методу LAD відповідає пряма, що проходить через дві точки вхідних даних на площині XY. Ця обставина також може бути використана для отримання розв'язків методу LAD.

### 1.4.2 Метод ортогональної регресії та метод абсолютних відхилень ортогональних відстаней

У методах лінійної регресії, розглянутих в попередніх підрозділах, вважається, що матриця дескрипторів  $X$  не містить помилок / похибок. Така ситуація можлива, коли в якості дескрипторів використовуються деякі теоретичні індекси, які не містять похибки за визначенням. Коли ж у якості незалежної, і залежних змінних використовуються експериментальні результати визначені з ненульовою похибкою, то має сенс використовувати спеціальні підходи.

До зазначених підходів можна віднести повний (узагальнений) метод найменших квадратів (*Total Least Squares*, TLS), окремим випадком якого є метод ортогональних відстаней (*Orthogonal Distance Regression*, ODR). У методі ODR для того, щоб знайти шукане рівняння, необхідно мінімізувати суму евклідових відстаней від точок вхідних даних до гіперплощини, що відповідає рівнянню регресії (рис. 1.3).

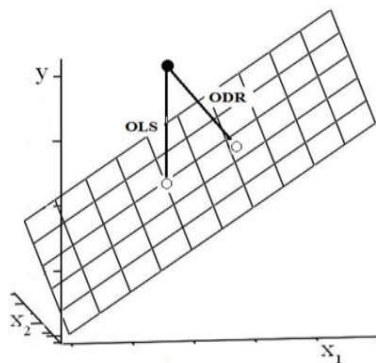


Рис. 1.3 Геометрична інтерпретація різниці методів OLS та ODR на прикладі рівняння з двома незалежними змінними  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Виходячи з геометричної відстані<sup>42</sup>, можна записати функцію ODR, з мінімізацією якої отримаємо розв'язок методу:

$$\beta_{\text{ODR}} = \arg \min_{\beta} \left\{ \|Y - X\beta\|_2^2 / (1 + \|\beta\|_2^2) \right\}. \quad (1.27)$$

Якщо ж, за аналогією до методу LAD, шукати відстань як модуль (*Least Absolute Deviation of Orthogonal Distance*, LADOD) то:

$$\beta_{\text{LADOD}} = \arg \min_{\beta} \left\{ \|Y - X\beta\|_1 / \sqrt{1 + \|\beta\|_2^2} \right\}. \quad (1.28)$$

Метод LADOD вперше було запропоновано на кафедрі хімічного матеріалознавства університету імені В. Н. Каразіна<sup>38,39</sup>.

Робочі рівняння для методів LAD, ODR та LADOD будуть описані нами в розділі 2.

### 1.4.3 Регресійні моделі основані на аналізі головних компонент (методи PCR та PLS)

До сих пір ми обговорювали лише підходи у яких або проводиться ефективне скорочення дескрипторного набору, або такий набір вже є скороченим настільки, що можливо використання якогось з вищезгаданих методів. Зрозуміло, що при цьому втрачається частина інформації, що міститься в дескрипторах, які не було обрано для опису властивості. Але існує й інший підхід до роботи з такими даними – перетворення вхідних даних таким чином, щоб уся інформація, що утримується в дескрипторах, певним чином увійшла до лінійної моделі.

Для цього в методі PCA (*Principal Component Analysis*), як першої частини методу PCR (*Principal Component Regression*)<sup>43,44</sup>, із вхідної матриці  $X$  формуються головні компоненти як суперпозиція вхідних дескрипторів  $X$  таким чином, щоб:

1) зміна властивостей молекул у напрямку головних компонентів була найбільшою. При цьому зрозуміло: для того, щоб зміна властивостей у напрямку головної компоненти була найбільшою, необхідно формувати фактори як суперпозицію дескрипторів, які є найбільш чутливими для молекул тренувальної вибірки.

2) головні компоненти мають бути ортогональні одні до одного, що гарантується тим, що вони є власними значеннями матриці кореляцій.

3) використання невеликої кількості головних компонент у якості регресійних параметрів дозволяє суттєво скоротити розмірність проблеми побудови регресійного рівняння.

Для реалізації цього алгоритму вхідна матриця  $X$  рангу  $r$  розбивається на  $r$  матриць рангу 1:

$$X = M_1 + M_2 + \dots + M_r, \quad (1.29)$$

де кожна матриця  $M_i$  може бути представлена як добуток векторів рахунків (*scores*)  $t_i$  та навантажень (*loadings*)  $p_i^+$ , або в матричному виді:

$$X = TP^+. \quad (1.30)$$

Таким чином, пошук розв'язків PCA зводиться до пошуку матриць  $T$  та  $P$ , після чого в методі PCR властивість  $u$  описується з використанням матриці рахунків  $T$  у методі OLS як нової матриці, що є суперпозицією дескрипторів. Причому матриця  $T$  складається з ортогональних стовпчиків і тому матриця  $T^+T$  не є сингулярною і може бути обернена. Розмірність матриці  $T$  також буде нижче, ніж розмірність матриці  $X$ , якщо в матриці  $X$  були присутні лінійно залежні дескриптори.

Для знаходження розв'язків методу PCA нами було використано алгоритм NIPALS (*Nonlinear Iterative Partial Least Squares*)<sup>45</sup>, у якому головні компоненти розраховувалися послідовно один за одним. При цьому кожний раз з  $X$  вилучається  $i$ -тий добуток  $t_i p_i^+$  і отримують матрицю-залишок  $X'$ , яку використовують для подальших розрахунків. Розрахунок головної компоненти (скажімо  $j$ -тої компоненти) у методі NIPALS має наступний вигляд:

$$\begin{aligned} 1) & t_j = x_j \\ 2) & p_j^+ = t_j^+ X / (t_j^+ t_j) \\ 3) & p_j = p_j / \|p_j\|_2 \\ 4) & t_j^{\text{new}} = X p_j \\ 5) & \text{if } \|t_j^{\text{new}} - t_j\|_2 < \text{accuracy} \quad \text{then : exit;} \\ & \quad \text{else : goto step 2} \end{aligned} \quad (1.31)$$

Зазвичай, у методі PCR використовують тільки невелику кількість перших компонент, які вважаються "змістовними"  $j \leq r$ . Чим більшу кількість



добутків  $t_i p_i^+$  було виключено з  $X$ , тим менший інформаційний зміст має залишок, який, зазвичай, з кожним "вилученням" вже розрахованої компоненти складається все більше й більше з шумів<sup>44</sup>. Надалі, говорячи про використання PCR розмірності  $k$ , будемо мати на увазі, що використовується  $k$ -перших головних компонентів.

У представлений роботі метод PCR було реалізовано нами на мові програмування FORTRAN.

Як можна бачити з наведеного алгоритму, у методі PCR не використовується інформація про передбачувану властивість для отримання головних компонентів. Отримані головні компоненти – це напрямки, на яких найсильніше змінюється дескрипторний набір. Таким чином, отримані величини  $t_i$  не завжди добре корелюють із властивістю  $y$  (45).

Для виправлення цієї ситуації пошук головних компонентів може бути зроблено одночасно як для дескрипторів  $X$ , так і для властивості  $y$ . Регресійна модель, що основана на таких засадах, має назву PLS (*Partial Least Squares* або *Projection on Latent Structure*). Існує кілька підходів розв'язання задачі PLS. Порівняння ефективності алгоритмів, а також чисельної стабільності PLS з однією незалежною змінною  $y$  можна знайти в (46). У даній роботі було використано алгоритм NIPALS запропонований Волдом<sup>47</sup> для методу PLS.

for  $k = 1$ , кількість латентних змінних :

$$w^{(k)} = X^{(k)*} y / \|X^{(k)*} y\|_2$$

$$t^{(k)} = X^{(k)} w^{(k)} / \left( t^{(k-1)*} t^{(k-1)} \right)$$

$$p^{(k)} = X^{(k)T} t^{(k)} \quad (1.32)$$

$$q^{(k)} = y^T t^{(k)}$$

$$\text{if } q^{(k)} = 0: \text{ break}$$

$$X^{(k+1)} = X^{(k)} - t^{(k)} p^{(k)*}$$

end for

Отримавши необхідні компоненти  $(w^{(k)}, p^{(k)})$  маємо змогу розрахувати коефіцієнти регресії, що визначають внески відповідних факторів (латентних змінних).

$$\beta = W(P^+W)^T y. \quad (1.33)$$

У формулі (1.33) вважається, що  $W$ :  $w^{(0)}, w^{(1)}, \dots$  та  $P$ :  $p^{(0)}, p^{(1)}, \dots$  – матриці-стовпчики. Кожен з цих стовпчиків відповідає "своїй" головній компоненті.

### 1.5. Логістична регресія

Логістична регресія (*Logistic Regression*, LR) відрізняється від розглянутих вище методів. Перш за все, LR не призначена для оцінок величин властивостей (активностей). Замість цього LR дозволяє класифікувати молекули на декілька типів відносно величини активності. Найпоширенішою є класифікація на два класи (бінарна класифікація). Наприклад, «активний» – «неактивний». Тут і надалі під активністю ми будемо мати на увазі значну наявність або відсутність (чи слабку вираженість) шуканої властивості.

Класифікаційна задача бінарної LR може бути записана наступним чином<sup>23</sup>:

$$p_i = 1 / (1 + \exp(-f_i)), \quad f_i = \sum_j \beta_j x_{ij}, \quad (1.34)$$

де  $p_i$  – це розрахована вірогідність того, що молекула виявиться активною.  $\beta_j$  – параметри регресії моделі. Значення відгуку  $p_i$  вказує на скільки вірогідним є те, що молекула має ту чи іншу властивість. Отже, чим ближча величина  $p_i$  до одиниці, тим вірогідніше, що молекула є активною, виходячи з результатів навчальної моделі. І навпаки: чим ближче значення  $p_i$  до нуля, тим більша вірогідність того, що молекула не є активною або не має шуканої властивості. Зазвичай, і в більшості випадків даної роботи, за граничне значення приймають  $p_i = 0.5$ , але для розрахунку якості моделі ця границя може бути змінена, наприклад, для отримання ROC-кривих (буде розглянуто далі)<sup>48</sup>. Загалом функція логістичної регресії має наступний графічний вигляд (рис. 1.4).

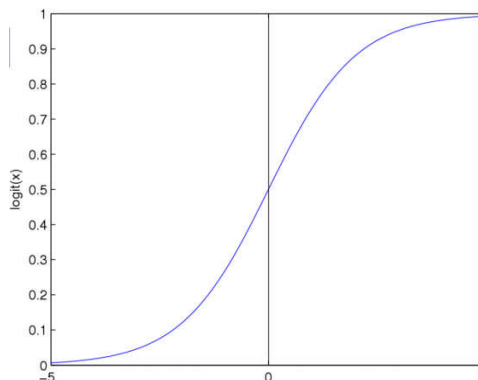


Рис.1.4 Загальний вигляд логістичної кривої для бінарної класифікації

Для того, щоб знайти значення коефіцієнтів  $\beta_j$  LR, знаходять максимум log-вірогідності, яка для бінарної класифікації має наступний вигляд:

$$\ell(\beta) = \sum_{i=1}^n \{y_i \ln p(x_i; \beta) + (1 - y_i) \ln (1 - p(x_i; \beta))\}. \quad (1.35)$$

Для бінарної класифікації вважаємо, що  $y_i = 0$  або 1 для неактивних і активних молекул відповідно. Підставляючи вираз (1.34) у (1.35), отримуємо:

$$\ell(\beta) = \sum_{i=1}^n \{y_i f_i - \ln(1 + e^{f_i})\}. \quad (1.36)$$

Для знаходження максимуму виразу (1.36) розрахуємо частинну похідну виразу і прирівняємо її до нуля:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^n x_i (y_i - p(x_i; \beta)) = X^+ (y - p) = 0, \quad (1.37)$$

Будемо розв'язувати рівняння (1.37) із застосуванням алгоритму Ньютона-Рафсона<sup>49</sup>. Для цього вирахуємо другу похідну:

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^n x_i x_i^+ p(x_i; \beta) (1 - p(x_i; \beta)) = -X^+ W X, \quad (1.38)$$

тут  $W$  – діагональна матриця з  $i$ -тим елементом:

$$W_{ii} = p(x_i; \beta) (1 - p(x_i; \beta)). \quad (1.39)$$

Крок ітераційного алгоритму Ньютона-Рафсона тоді може бути представлено наступним чином:

$$\begin{aligned}
\beta^{\text{new}} &= \beta^{\text{old}} - \left( \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta} = \beta^{\text{old}} + (X^+ W X)^{-1} X^+ (y - p) \\
&= (X^+ W X)^{-1} X^+ W (X \beta^{\text{old}} + W^{-1} (y - p)) \\
&= (X^+ W X)^{-1} X^+ W z
\end{aligned} \tag{1.40}$$

Тут для зручності введено величину:

$$z = X \beta^{\text{old}} + W^{-1} (y - p). \tag{1.41}$$

Таким чином, ітераційна процедура знаходження рішення логістичної регресії може бути сформульована як ітеративний пошук розв'язку рівнянь зваженого методу найменших квадратів з "відгуком"  $z$  (1.41) та вагами  $W$ , доки відповідний критерій зупинки (1.37) не буде досягнуто.

### 1.6. Проблема валідації регресійних QSAR/QSPR рівнянь

На важливість проблеми валідації регресійних рівнянь QSPR/QSAR було вказано вже доволі давно<sup>50</sup>. Але лише останнім часом опубліковано кілька робіт у яких пропонується ряд коефіцієнтів, що характеризують якість отриманих рівнянь<sup>51-54</sup>. З проблемою валідації також тісно пов'язано питання про оптимальне розбиття наборів даних на тренувальну (навчальну) та тестову вибірки. Існує певна кількість публікацій, у яких обговорюються підходи до такого розбиття<sup>55,56</sup>. У цих роботах, зокрема, стверджується, що "гарне" розбиття вибірок покращує якість моделей у порівнянні з розбиттям випадковим чином. У той же час, можна зустріти роботи, у яких стверджується, що реальна передбачувальна здатність таких моделей, навпаки, може й погіршуватись<sup>57</sup>.

На жаль, для великої кількості запропонованих регресійних підходів до опису різноманітних фізико-хімічних (та біохімічних) властивостей наводиться лише стандартний набір параметрів, що включає F-критерій, коефіцієнт кореляції (за Пірсоном) та стандартне відхилення. Отже, на цей час не існує загальноприйнятих підходів для адекватної характеристики точності отриманих рівнянь.

Таким чином, питання валідації є надзвичайно актуальним. Особливе значення воно набуває в задачах вибору регресійної моделі серед кількох альтернатив. Для детального дослідження різних підходів у представленій дисертації ми використовуємо найпростішу регресійну модель з одним дескриптором. Такі моделі зустрічаються в хімії доволі часто. Однак, не зважаючи на достатню простоту задачі, її прогностичні властивості ще не були вивчені належним чином.

Наразі відомо, що після отримання моделі ефективність останньої повинна бути підтверджена набором кроків валідації, які визначають, до яких задач може бути використана отримана модель, наскільки надійні розрахунки з використанням цієї моделі<sup>50</sup>.

Історично першими валідаційними критеріями, з використанням яких оцінювали ефективність моделей, стала низка індексів, які наразі відомі як підходи внутрішньої валідації<sup>50</sup>. Ці методи роблять висновки про якість моделі виходячи з характеристик тренувальної вибірки, тобто використовують лише інформацію, що була необхідна і достатня для побудування моделі.

### 1.6.1. Внутрішня валідація

У якості внутрішньої валідації нами було розглянуто наступні (добре відомі) коефіцієнти:

$$R^2_{train} = 1 - \frac{\sum_{i=1}^{n_{train}} (y_i^{exp} - y_i^{calc})^2}{\sum_{i=1}^{n_{train}} (y_i^{exp} - \bar{y}^{train})^2}, \quad (1.42)$$

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n_{train}} (y_i^{exp} - y_{i/i}^{pred})^2}{\sum_{i=1}^{n_{train}} (y_i^{exp} - \bar{y}^{train})^2}, \quad (1.43)$$

тут  $n_{train}$  – це кількість зразків/молекул у тренувальній вибірці (*train*),  $y_i^{exp}$  – експериментально визначені значення властивості,  $y_i^{calc}$  – розраховані значення властивості для тренувальної вибірки,  $\bar{y}^{train}$  – середнє значення  $y_i^{exp}$  тренувальної вибірки,  $y_{i/i}^{pred}$  – значення властивості розраховане для кожного зразка з використанням моделі, в якій всі точки були використані як точки тренувальної

вибірки, окрім тої, для якої властивість передбачується (процедура *Leave-one-out*, LOO).

Критерій  $Q_{LOO}^2$ , отриманий з використанням процедури LOO, довгий час розглядався як найважливіший коефіцієнт, який спроможний оцінювати якість моделі. Однак, наразі відомо<sup>58,59</sup>, що цей критерій може погано корелювати з критеріями зовнішньої валідації. Проте різниця між цим критерієм та  $R_{train}^2$  все ще вважається важливим критерієм **перенавчання**<sup>II</sup> моделі. Якщо різниця між цими двома коефіцієнтами достатньо велика ( $> 0.3$  (50)), то вважається, що модель може перенавчатися і в результаті не може гарантувати адекватне передбачення властивостей для систем, які суттєво відрізняються від тих, що не увійшли до тренувальної (навчальної) вибірки.

У дослідженнях, що були представлені в цій дисертації, було використано також запропонований нещодавно критерій ІС (*Index of ideality of correlation*) як коефіцієнт внутрішньої валідації<sup>51,52</sup>. В основі ІС лежать дві наступні величини, що характеризують відхилення даних:

$$MAE_{train}^{-} = \frac{1}{N^{-}} \sum_{k=1}^{N^{-}} |\Delta_k| \text{ where } \Delta_k < 0 \text{ for } k \in [1, N^{-}], \quad (1.44)$$

$$MAE_{train}^{+} = \frac{1}{N^{+}} \sum_{k=1}^{N^{+}} |\Delta_k| \text{ where } \Delta_k > 0 \text{ for } k \in [1, N^{+}], \quad (1.45)$$

де

$$\Delta_k = y_k^{exp} - y_k^{calc}, \quad (1.46)$$

$N^{+}$  – кількість значень (1.46) з позитивним значенням, а  $N^{-}$  відповідно кількість значень (1.46) з негативним значенням. Тоді ІС визначається наступним чином:

$$IIC = r_{train} \frac{\min(MAE_{train}^{-}, MAE_{train}^{+})}{\max(MAE_{train}^{-}, MAE_{train}^{+})}, \quad (1.47)$$

---

<sup>II</sup> Термін "перенавчання" характеризує ситуацію, коли модель **надто добре** описує тренувальну (навчальну) вибірку. Це може вести до опису скоріше випадкової похибки ніж власне тієї інформації для опису якої будується регресія. Див. (60)

де  $r_{train}$  – відомий коефіцієнт кореляції Пірсона<sup>61</sup> для рівняння теорія-експеримент:

$$r_{train} = \frac{\sum_{i=1}^{n_{train}} (y_i^{exp} - \bar{y}^{train})(y_i^{calc} - \bar{y}^{calc})}{\sqrt{\sum_{i=1}^{n_{train}} (y_i^{exp} - \bar{y}^{train})^2 \times \sum_{i=1}^{n_{train}} (y_i^{calc} - \bar{y}^{calc})^2}} \quad (1.48)$$

### 1.6.2. Y-Рандомізація

Існує й інший підхід до внутрішньої валідації рівнянь – рандомізація значень залежної змінної  $y$ . У багатьох роботах описано так звану випадкову кореляцію (*casual correlation*)<sup>62,63</sup>. У такому випадку вважається, що модель описує похибки визначених властивостей, а не реальні лінійні зв'язки залежних та незалежних змінних. Для того, щоб впевнитися, що в наших розрахунках не реалізується *casual correlation*, ми розраховували  $R_{shuffled}^2$  (63).

Нещодавно було запропоновано підхід, у якому робиться спроба апроксимувати поведінку середніх абсолютних відхилень для тестових даних на основі аналізу розподілу рандомізованих даних для тренувальної вибірки<sup>62</sup>. У цьому підході тренувальний набір даних випадковим чином перемішується так, що вихідним векторам  $X$  більш не відповідають початкові вхідні значення залежної змінної (властивості, активності). Для отриманого невпорядкованого набору даних розраховують рівняння регресії. Розраховують  $MAE_{yRAND}$  (див. рівняння 1.44,1.45) для кожної такої моделі й для відповідної перемішаної тренувальної вибірки. Після чого розраховують  $MAE_{CCE}$ :

$$MAE_{CCE} = MAE_{train} + (MAE_{mean} - MAE_{yRAND}), \quad (1.49)$$

де  $MAE_{train}$  – це MAE розраховане для  $y$ -неперемішаної тренувальної вибірки,  $MAE_{mean}$  відповідає MAE розрахованому для тренувальної вибірки з розрахованим  $y_i^{calc} \equiv \bar{y}^{train}; \forall i \in [1; n^{train}]$ .

Ця процедура повторюється багато разів ( $N_{rand}$ ) для отримання розподілу  $MAE_{CCE}$ . Згідно з роботою<sup>62</sup>, значення  $MAE_{test}$  для тестових вибірок з

точками/молекулами, що містяться в границях застосовності моделі (*Applicability Domain*, AD) повинні опинитися в діапазоні  $[0; MAX_{N_{rand}}(MAE_{CCE})]$ .

Довгий час вважалося, що внутрішньої валідації достатньо для характеристики якості моделі. Однак, на теперішній час, спеціалісти QSAR дійшли до висновку, що для оцінки передбачувальної здатності моделі внутрішньої валідації не достатньо і що для валідації моделі необхідно використовувати зовнішній (тестовий) набір даних.

### 1.6.3. Зовнішня валідація

Для зовнішньої валідації, у більшості випадків, наразі найчастіше використовується критерій  $R^2_{test}$  (64) або еквівалентно відомий у літературі  $Q^2_{F2}$  (65,66).

$$R^2_{test} \equiv Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i^{exp} - y_i^{pred})^2}{\sum_{i=1}^{n_{test}} (y_i^{exp} - \bar{y}^{test})^2}, \quad (1.50)$$

тут  $n_{test}$  – кількість точок у тестовій вибірці,  $y_i^{pred}$  – передбачені значення властивості з відповідними експериментальними властивостями  $y_i^{exp}$ ,  $\bar{y}^{test}$  – середнє значення експериментального значення властивості для тестового набору. Надалі в цій роботі ми будемо використовувати позначення  $R^2_{test}$ .

Згідно з роботою (64), критерій (1.50) є необхідним, але недостатнім для зовнішньої валідації моделі. Також повинна досліджуватися остаточно середня-квадратична помилка (*Residual mean square error*, RMSE). Для цілей нашої роботи ми використовували замість цього залежний коефіцієнт RMSEP:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{test}} (y_i^{exp} - y_i^{pred})^2}{n^{test}}}. \quad (1.51)$$

Ми також досліджували у якості альтернативи до цього коефіцієнту середню абсолютну помилку (*Mean Absolute Error*, MAE):

$$MAE_{test} = \frac{\sum_{i=1}^{n_{test}} |y_i^{exp} - y_i^{pred}|}{n^{test}}. \quad (1.52)$$



У літературі наразі запропоновано також кілька альтернатив коефіцієнту

$R_{test}^2$  :

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i^{exp} - y_i^{pred})^2}{\sum_{i=1}^{n_{test}} (y_i^{exp} - \bar{y}^{train})^2}, \quad (1.53)$$

$$Q_{F3}^2 = 1 - \frac{n_{train} \sum_{i=1}^{n_{test}} (y_i^{exp} - y_i^{pred})^2}{n_{test} \sum_{i=1}^{n_{train}} (y_i^{exp} - \bar{y}^{train})^2}. \quad (1.54)$$

Узгоджений кореляційний коефіцієнт (*concordance correlation coefficient*, CCC) має наступний вигляд:

$$CCC = \frac{2 \sum_{i=1}^{n_{test}} (y_i^{exp} - \bar{y}^{test})(y_i^{pred} - \bar{y}^{pred})}{\sum_{i=1}^{n_{test}} (y_i^{exp} - \bar{y}^{test})^2 + \sum_{i=1}^{n_{test}} (y_i^{pred} - \bar{y}^{pred})^2 + n_{test} \times (\bar{y}^{test} - \bar{y}^{pred})^2}, \quad (1.55)$$

тут  $\bar{y}^{pred}$  – середнє значення  $y_i^{pred}$ .

Серед запропонованих коефіцієнтів відомо, що коефіцієнт  $Q_{F1}^2$  (67) у багатьох випадках<sup>65</sup> дає занадто оптимістичні результати й тому вважається гіршим ніж  $R_{test}^2$ , який було впроваджено, щоб компенсувати цю проблему. У той же час, коефіцієнти  $Q_{F3}^2$  (66,68) та CCC (54,69) було впроваджено, щоб компенсувати відомі проблеми коефіцієнту  $R_{test}^2$ . Автори критерію  $Q_{F3}^2$  також зазначили, що цей критерій добре узгоджується з коефіцієнтом RMSE.

Детальний опис використаних коефіцієнтів може бути знайдено в огляді (70).

#### 1.6.4. Рациональне розбиття вибірки на навчаючу та тестову

Зрозуміло, що QSAR/QSPR моделі повинні використовуватися для передбачення властивостей таких систем (молекул, зразків), що мають схожі властивості до систем тренувальної вибірки<sup>56,71</sup>. Так, наприклад, модель, що тренувалась на алканах, не може бути використана для передбачення властивостей ароматичних сполук. Така область параметрів (властивостей/дескрипторів), у якій можуть знаходитися лише такі системи, що

мають параметри, схожі до систем тренувальної вибірки, називають границями застосовності моделі (*Applicability Domain*, AD). Існує кілька підходів формування AD. Деякі з них використовують інформацію щодо властивості молекули зазвичай її активності<sup>72</sup>. Деякі дослідники використовують структурні дескриптори молекули так, що тестова молекула, що належить AD моделі, повинна мати схожі структурні дескриптори як і молекули тренувальної вибірки<sup>56</sup>.

Таким чином, виникає проблема раціонального розбиття (*rational selection*, RS) вхідної вибірки на тестову і тренувальну. При цьому, тестова вибірка повинна бути представницькою, тобто описувати активності (властивості) за інтенсивністю, бо інакше характеристики зовнішньої валідації можуть виявитися завищеними.

Існує кілька робіт, у яких порівнюється ефективність різних підходів на основі RS (55,57). У представленій роботі ми воліли бути впевненими, що точки тестової вибірки якісно схожі (у термінах дескрипторів) з точками тренувального набору. Для досягнення цієї мети нами було використано декілька методів раціонального розбиття.

У розділі 3 нами був використаний метод *k*-найближчих сусідів (*k-nearest neighbors*, KNN)<sup>73</sup>. Цей метод базується на розрахунку узагальнених відстаней між точками в дескрипторному просторі. У розрахунках цього розділу використовувався лише один дескриптор, втім, оскільки досліджувана властивість також є певним дескриптором, ми, таким чином, використовували двовимірний простір для розрахунку відстаней і далі – розподілу точок між тестовою й тренувальною вибірками. Очевидно, що такий підхід не дозволяє побудувати AD для моделі, у якій ще не визначено величину властивості (активності). Вичерпний опис алгоритму методу KNN-AD, що був використаний у даній роботі, може бути знайдено в (73).

У розділі 4, для кластеризації сполук на набори молекул з подібними властивостями, ми використали метод *k*-середніх<sup>74</sup>. Методи *k*-середніх та *k*-найближчих сусідів є ідеологічно схожими, оскільки обидва розраховують

відстані в дескрипторному просторі, але, на відміну від методу  $k$ -найближчих сусідів, метод  $k$ -середніх розраховує відстань від деякого середнього значення (центру кластеру), що належить до кластеру, до всіх молекул, що відносить метод до цього кластеру. При цьому центр кластеру обирається таким чином, щоб мінімізувати відстань молекул у кожному кластері до цього центру. На відміну від розділу 3, ми не використовували властивість для класифікації – замість цього використовувався ортонормований набір дескрипторів, що містив порядку 1000 розрахункових дескрипторів.

### Висновки до розділу 1

1. Останнім часом досягнуто значний прогрес у розробці й використанні набору регуляризаційних технік у загальних проблемах розпізнавання образів. Завдяки  $L_1$ -регуляризації створено, зокрема, ефективні підходи до аналізу зображень. Незважаючи на досить складні в чисельному аспекті задачі негладкої, а часом і не опуклої оптимізації задач  $L_x$ -регуляризації, наразі вже створено досить ефективні алгоритми, які можуть бути використані для широкого класу хімічних проблем.

2. Аналіз літератури показав, що  $L_1$ -регуляризація практично не використовувалась у хімічних дисциплінах і, зокрема, у проблемах QSAR/QSPR та в квантовій хімії. Разом з тим, надійний (і що важливо – відтворюваний) підхід до систематичного скорочення набору незалежних змінних (дескрипторів, параметрів моделі) може бути надзвичайно корисним при побудові прогностичних моделей.

3. Одним з важливих аспектів хімічної інформатики й хеометрії (зокрема QSAR/QSPR) є проблема побудови регресійних рівнянь у випадку певного (іноді значного) розкиду вхідних даних. Для таких розрахунків, крім стандартного методу найменших квадратів (OLS), можна використовувати ще кілька альтернативних підходів, які ведуть до різних лінійних залежностей. На диво, на сьогоднішній день, такі альтернативи ще й досі є мало дослідженими.

4. Значною проблемою на сьогоднішній день є проблема тестування (валідації) отриманих QSAR/QSPR моделей. Уже давно було визнано, що отримати регресійне рівняння набагато легше, ніж довести його надійність. Лише останнім часом було запропоновано ряд індексів (характеристик), що призначені оцінити точність моделей, але надто мало інформації щодо можливостей їх використання. Більше того, невідомо наскільки запропоновані характеристики узгоджуються одна з одною!

5. Для усіх, описаних в огляді, методів у роботі розроблено комп'ютерні програми з використанням алгоритмічних мов FORTRAN і Python3.

## РОЗДІЛ 2

### ЛІНІЙНІ $L_1$ -РЕГУЛЯРИЗАЦІЙНІ МОДЕЛІ В ОПИСІ ФІЗИКО-ХІМІЧНИХ ПАРАМЕТРІВ МОЛЕКУЛ

Пошук моделей для опису кількісних співвідношень структура-активність/властивість (QSAR/QSPR) є важливим кроком у великій кількості наукових та технологічних задачах. У застосуванні до хімічних проблем QSAR/QSPR моделі використовуються в найрізноманітніших галузях. Зокрема, у медичній хімії моделі QSAR спрямовані на розробку сполук з шуканими лікарськими властивостями, у (еко)токсикології – оцінку токсичності, канцерогенності, тератогенності<sup>III</sup>, мутагенності й факторів біоконцентрації хімічних сполук. Загальні огляди див. у (75-77). Значну частину QSPR складає дослідження фізико-хімічних властивостей: ліпофільності, розчинності, октанові числа, температури кипіння і плавлення органічних систем тощо<sup>78,79</sup>. Крім вищезгаданих звичайних фізико-хімічних характеристик, варто відзначити також важливість QSAR у дослідженнях хроматографічних параметрів (фактори утримання)<sup>80</sup>, критичних властивостей речовини<sup>81</sup>, детонаційних характеристики молекул<sup>IV</sup> (82), основності сполук по відношенню до катіонів лужних металів<sup>83</sup> та багато *ін.*

Незважаючи на поширення QSAR моделей у наукових галузях, набір математичних інструментів, що використовується для побудови рівнянь залишається дещо обмеженим. Найчастіше використовується побудова регресійних моделей. Багато з них, стосовно біоактивності, засновано на теорії Хенча<sup>84-87</sup>, яка базується на оцінках ліпофільності й електронних та стеричних емпіричних дескрипторах. Відомо, що такі моделі інколи досить непогано передбачають біоактивність молекул. Однак у загальних випадках теорія Хенча стає непридатною.

Неможливість простих лінійних, або параболічних чи білінійних<sup>88</sup> за ліпофільності залежностей, описати широкий спектр молекулярних

<sup>III</sup> фактор, що веде до порушень ембріонального розвитку.

<sup>IV</sup> у англійській літературі вони відомі як *impact sensitivity*.

властивостей призвела до лавиноподібного зростання кількості молекулярних дескрипторів. На сьогоднішній день, таких індексів тільки серед тих, що реалізовані у відомих програмах, налічується більше 7000. Вони включають параметри, що описують різноманітні аспекти молекулярної структури. Зазвичай їх класифікують згідно відповідної "мірності": 0D, 1D, 2D, 3D. Серед них топологічні, електричні, геометричні та інші дескриптори. Достатньо повний їх опис може бути знайдено в (89,90).

Отже, для загальних лінійних моделей QSAR характерною є ситуація, коли кількість дескрипторів значно перевищує кількість спостережень<sup>47,91</sup>, що веде до мультиколінеарності регресійної (лінійної) проблеми.

У таких випадках стандартний метод OLS не може бути використано безпосередньо. Щоб розв'язати проблему мультиколінеарності реалізовані інші підходи, які умовно можна розбити на дві групи. **Перша група** підходів – це методи, у яких не робиться нетривіальний відбір дескрипторів (але можливе попереднє відсіювання константних, або сильно корельованих між собою дескрипторів). Це методи, які працюють з факторною будовою задачі. До них відносяться PCR та PLS (див. підрозділ 1.4.3).

Але нас, перш за все, буде цікавити **друга група** методів, яка базується на скороченні набору дескрипторів для того, щоб зробити можливим роботу з такими методами як OLS, LAD, ODR та LADOD (див. підрозділи 1.4.1-1.4.2). Також після такого скорочення стає можливим використати методи бінарної класифікації як то дескримінаційний аналіз та логістична регресія, так і звичайний (не конволюційний) метод нейронних мереж.

Найпростішим підходом, що дозволяє скоротити набір дескрипторів, є покроковий метод – *Forward Stepwise* (FS) регресія<sup>23,24</sup>. Цей метод можна розуміти як OLS, який доповнено процедурою ранжування (впорядкування) дескрипторів відповідно до їх кореляції із властивістю / активністю. У методі FS до регресійного рівняння послідовно додаються найбільш важливі дескриптори. Звичайно така процедура не бере до уваги можливу кореляцію між дескрипторами, а також факторну будову матриці дескрипторів і тому не

може розглядатися як універсальний підхід. Не зважаючи на це, існує доволі багато робіт, у яких FS регресія або аналогічні методи успішно використовуються<sup>92,93</sup>.

Інший підхід до відбору дескрипторів базується на ідеології перебору, різновидом якого є генетичні алгоритми (*Genetic Algorithms*, GA)<sup>1,2</sup>. Сфера застосування GA досить широка. Такі методи використовують, зокрема, при реалізації різноманітних багатопараметричних оптимізаційних задач (у хімії – оптимізація геометрії великих систем, докінг тощо)<sup>94-96</sup>. У контексті проблем QSAR/QSPR можна відзначити кілька важливих недоліків GA: 1) GA розв'язки, строго кажучи, не є відтворюваними; 2) GA розв'язки схильні до перенавчання; 3) розрахункові витрати GA можуть бути надто великими при наявності значного набору дескрипторів та об'єктів.

Радикально інший підхід до проблеми відбору дескрипторів ми вбачаємо в методах, що базуються на ідеї  $L_1$ -регуляризації. Після публікацій піонерських робіт R. Tibshirani<sup>22,97</sup>  $L_1$ -регуляризація знайшла використання в ряді технічних проблем<sup>11,98</sup>. Представлена робота пов'язана з дослідженням можливостей  $L_1$ -регуляризації в хімії й QSAR/QSPR зокрема.

У даному розділі ми розглядали аналіз регресійних моделей побудованих методом  $L_1$ -регуляризації. При цьому нас цікавила скоріш суто прагматична (хеометрична) точка зору на проблему відбору дескрипторів, ніж статистична. Крім того, ми зацікавлені в порівнянні результатів, що отримані різними підходами для побудови регресійних рівнянь (а саме з результатами методів OLS, LAD, ODR і LADOD). Ми використовували набір даних щодо активності "як є" і намагалися отримати найкращі регресійні рівняння виходячи з цих даних.

У роботі особлива увага приділялася методу LASSO. Завдяки використанню методики LARS-LASSO (див. розділ 1.3.1.) розв'язки можливо отримувати з високою розрахунковою ефективністю навіть для задач з тисячами дескрипторів.

У якості тестових систем розглядалися експериментальні вибірки різних фізико-хімічних властивостей. Геометрія молекул у цьому розділі оптимізувалася з використанням напівемпіричного методу AM1 доступного в GAMESS<sup>99</sup>. Після чого розраховувалися дескриптори з використанням програми E-Dragon 1.0 (100,101) для флуороалканів та PaDEL-Descriptor<sup>102</sup> для всіх інших задач. Для проведення розрахунків було розроблено відповідний комплекс програм на мовах програмування FORTRAN та Python.

### 2.1. $L_1$ -регуляризація середнього значення (*a toy example*)

У сучасних теоретичних дослідженнях доволі часто стали зустрічаються так звані "іграшкові приклади" (англ. *toy examples*). Вони призначені для викладення методичних основ того підходу, що використовується в подальшому. У цьому підрозділі ми пропонуємо досить просту, майже іграшкову, задачу  $L_1$ -регуляризації середнього значення вибірки. Очевидно, що вона може бути сформульована як задача мінімізації наступної функції:

$$F(a) = \|x - a\|_2^2 + \lambda |a|, \quad (2.1)$$

де  $x$  є вибірка даних  $\{x_1, x_2, \dots, x_N\}$  розмірності  $N$ ,  $\bar{x}_\lambda = \arg \min_a (\|x - a\|_2^2 + \lambda |a|)$  – шукане середнє (регуляризоване) значення,  $\lambda$  – параметр регуляризації. Виходячи лише з першої складової, легко знайти звичайне середнє значення  $\bar{x}$ . Для цього, у дусі методу найменших квадратів, вираховуємо похідну:

$$d = \frac{\partial}{\partial a} \sum_i (x_i - a)^2 = 2 \sum_i (x_i - a). \quad (2.2)$$

Прирівнюючи її до нуля ( $d = 0$ ), відразу отримуємо

$$\bar{x} = a = \frac{\sum_i x_i}{N}. \quad (2.3)$$

Для повної функції (2.1), з метою реалізації градієнтного алгоритму розрахунку  $a = \bar{x}_\lambda$ , формально, можемо використати похідну від (2.1):

$$\partial F = d + \lambda \cdot \text{sign}(a), \quad (2.4)$$



де  $\text{sign}(a)$  – знак числа  $a$ , який може бути представлений, наприклад, таким чином:  $\text{sign}(a) = \frac{a}{|a|}$ , а величина  $d$  розраховується згідно (2.2).

Однак в околиці точки  $a \sim 0$  функція (2.1) є розривною, що веде до невизначеності (2.4) і далі – алгоритму розрахунку. У таких випадках один з методів оптимізації пов'язаний із використанням так званого “м'якого порогу” (*soft threshold*, ST)<sup>11,103</sup>. Метод полягає в тому, що в точках  $|a| > 0$  градієнт розраховується згідно (2.4). В іншому випадку ( $a \sim 0$ ) використовується стратегія ST, у якій субградієнт визначається наступним чином<sup>11,103</sup>:

$$\partial F = \begin{cases} d + \lambda, & \text{якщо } d < -\lambda \\ d - \lambda, & \text{якщо } d > \lambda \\ 0, & \text{якщо } -\lambda \leq d \leq \lambda \end{cases} \quad (2.5)$$

Рис. 2.1 ілюструє типову поведінку  $\partial F$  як функції від  $d$ .

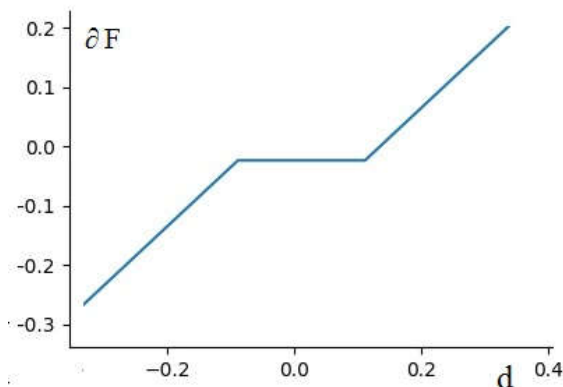


Рис. 2.1 Залежність субградієнту  $\partial F$  від параметру  $d$  (2.2)

Сформувавши субградієнт,  $k+1$ шій крок ітераційного методу мінімізації (2.1) може бути виконаний, наприклад, так:

$$a^{(k+1)} = a^{(k)} - \xi \partial F^{(k)}, \quad (2.6)$$

де  $\xi$  параметр методу, який може бути або постійним, або оптимізуватися на кожному кроці. Звісно, представлена задача є досить тривіальною і при зростанні  $\lambda \rightarrow \infty$ ,  $\bar{x}_\lambda \rightarrow 0$ .

У якості прикладу наведемо результати розрахунків для вибірки  $x = [0.62, 0.6, 0.75, 0.65, 0.62, 1.1]$ . При  $\lambda = 0$  маємо звичайне середнє значення

$a = \bar{x} = 0.7233$ . Використовуючи постійне значення параметру збіжності  $\xi \sim 0.1$  для розв'язку задачі з  $\lambda = 1$ , з точністю  $|\partial F| \sim 10^{-6}$ , за 9 ітерацій, отримуємо  $\bar{x}_{\lambda=1} = 0.640$ .

Зауважимо, що  $\lambda = 1$  відповідає спеціальному випадку, коли дві компоненти входять у (2.1) з рівними вагами. Цікаво, що медіана вибірки ( $M = \arg \min_a \sum |x_i - a|$ ) дає близьке до  $\bar{x}_{\lambda=1}$  значення  $M = 0.635$ . Отже, наш  $L_1$ -регуляризований розрахунок, при належному виборі  $\lambda$ , дає реалістичні оцінки середнього значення вибірки при наявності викидів. Програму для розрахунку  $\bar{x}_\lambda$  наведено в додатку Б.

## 2.2. Алгоритм LARS-LASSO

Найбільш ефективним методом для вирішення LASSO проблеми є метод LARS-LASSO. Детальний опис LARS-LASSO алгоритму, що був використаний у представлений роботі, може бути знайдено в (33). У розділі 1.3 було наведено базові концепції методу. У цьому підрозділі ми опишемо основні розрахункові кроки.

Почати з ітерації  $k = 0$ , порожньої множини  $\varepsilon$  та порожньої множини знаків дескрипторів  $s$  і параметру регуляризації  $\lambda_0 = +\infty$ .

Ітерації проводять доки  $\lambda_k > 0$  (або поки не буде набрана необхідна кількість дескрипторів) наступним чином:

1. Розрахувати  $\beta_\varepsilon^{\text{LARS}}(\lambda_k)$  за формулою:

$$\beta_\varepsilon^{\text{LARS}}(\lambda_k) = c - \lambda_k d. \quad (2.7)$$

Тут  $c = (X_\varepsilon)^+ Y$  та  $d = (X_\varepsilon X_\varepsilon^T)^+ s$ ,  $s$  – вектор знаків коефіцієнтів дескрипторів множини  $\varepsilon$ .

2. Розрахувати крок, коли наступний дескриптор буде рівно корельованим згідно з (1.22):

$$t_i^{\text{join}} = \frac{X_i^T (y - X_\varepsilon c)}{\pm 1 - X_i^T X_\varepsilon d}, \quad (2.8)$$

із двох отриманих  $t_i^{\text{join}}$  обирати таку величину, що належить інтервалу  $[0, \lambda_k]$ .

Знайти дескриптор, який виявиться рівнокорельованим із залишком.  
Зафіксувати величину  $\max_{i \in \mathcal{E}} t_i^{\text{join}}$ :

$$i_{k+1}^{\text{join}} = \arg \max_{i \in \mathcal{E}} t_i^{\text{join}}, \quad (2.9)$$

$$\lambda_{k+1}^{\text{join}} = \max_{i \in \mathcal{E}} t_i^{\text{join}}. \quad (2.10)$$

Розрахувати відповідний знак при дескрипторі:

$$s_{k+1}^{\text{join}} = \text{sign} \left( X_{i_{k+1}^{\text{join}}}^T \{Y - X\beta^{\text{LARS}}(\lambda_{k+1}^{\text{join}})\} \right). \quad (2.11)$$

3. Знайти умову зміни знаку при дескрипторах доданих до множини  $\mathcal{E}$ .

$$t_i^{\text{cross}} = \frac{c_i}{d_i}. \quad (2.12)$$

При цьому розглядають тільки такі  $t_i^{\text{cross}} \leq \lambda_k$ . Таким чином, "найближча" зміна знаку:

$$i_{k+1}^{\text{cross}} = \arg \max_{i \in \mathcal{E}} t_i^{\text{cross}} \quad (2.13)$$

$$\lambda_{k+1}^{\text{cross}} = \max_{i \in \mathcal{E}} t_i^{\text{cross}} \quad (2.14)$$

4. Обрати  $\lambda_{k+1}$ :

$$\lambda_{k+1} = \max(\lambda_{k+1}^{\text{join}}; \lambda_{k+1}^{\text{cross}}) \quad (2.15)$$

Якщо  $\lambda_{k+1}^{\text{cross}} \geq \lambda_{k+1}^{\text{join}}$ , то дескриптор  $i_{k+1}^{\text{cross}}$ , а також його знак  $s_i$ , виключають з множини  $\mathcal{E}$ . Якщо  $\lambda_{k+1}^{\text{join}} > \lambda_{k+1}^{\text{cross}}$ , то відповідний дескриптор  $i_{k+1}^{\text{join}}$ , а також знак коефіцієнта  $s_i$ , що йому відповідає, додають до множини  $\mathcal{E}$ .

### 2.3. Робочі формули для розрахунку регресійних рівнянь

У представленій дисертації проведено ряд розрахунків однопараметричних та багатопараметричних регресій за допомогою різних альтернативних методів. Тому в даному підрозділі ми наводимо явні розрахункові формули для методів LAD, ODR та LADOD. Метод OLS добре відомий, отже відповідні формули для нього ми не будемо описувати.

### 2.3.1. Побудова регресійного рівняння методом LAD

Спочатку розглянемо однопараметричне лінійне рівняння:

$$y = \beta_0 + \beta_1 x. \quad (2.16)$$

Для знаходження коефіцієнтів  $\beta_0$  та  $\beta_1$  маємо мінімізувати функцію, що відповідає абсолютному відхиленню вхідних даних ( $y_i$ ) і розрахованих за формулою (2.16).

$$H_{LAD}(\beta_0, \beta_1) = \sum_i |y_i - \beta_0 - \beta_1 x_i| = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 / w_i, \quad (2.17)$$

де вагова функція має вигляд:

$$w_i = |y_i - \beta_0 - \beta_1 x_i| + \eta. \quad (2.18)$$

Тут параметр  $\eta$  є достатньо малою величиною, що гарантує уникнення сингулярності при проведенні ітераційної процедури. У представленні  $H_{LAD}(\beta_0, \beta_1)$ , за допомогою вагової функції  $w_i$ , ми реалізуємо так званий "варіаційно-зважений" метод<sup>41</sup>. Зауважимо, що такий підхід не є єдино можливим (дивись наприклад огляд (34,35)).

Отже, спершу вираховуємо похідну по  $\beta_0$  і прирівнюємо її до нуля:

$$\frac{\partial H_{LAD}}{\partial \beta_0} = -2 \left( \sum_i \frac{y_i}{w_i} - \beta_1 \sum_i \frac{x_i}{w_i} - \beta_0 \sum_i \frac{1}{w_i} \right) = 0. \quad (2.19)$$

При розрахунку цієї похідної ми, згідно варіаційно-зваженому методу, вважаємо, що ваги  $w_i$  є константами, але далі, при реалізації ітераційного кроку, ваги змінюються відповідно змінам коефіцієнтів регресії.

Виходячи з (2.19), маємо можливість виразити  $\beta_0$  через узагальнені (зважені) середні  $\tilde{y}$  та  $\tilde{x}$ :

$$\beta_0 = \tilde{y} - \beta_1 \tilde{x}. \quad (2.20)$$

Зважені середні мають вигляд:

$$\tilde{y} = \sum_i \frac{y_i}{w_i} / \sum_i \frac{1}{w_i}, \quad (2.21)$$

$$\tilde{x} = \sum_i \frac{x_i}{w_i} / \sum_i \frac{1}{w_i}. \quad (2.22)$$

Далі вираховуємо похідну від (2.17) по  $\beta_1$  (також вважаємо  $w_i$  константами):

$$\frac{\partial H_{\text{LAD}}}{\partial \beta_1} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) \frac{x_i}{w_i} = 0. \quad (2.23)$$

Після елементарної алгебри отримуємо компактку формулу для  $\beta_1$

$$\beta_1 = \frac{\tilde{s}_{xy}}{\tilde{s}_{xx}}, \quad (2.24)$$

де

$$\tilde{s}_{xy} = \sum_i (y_i - \tilde{y}) x_i / w_i = \sum_i (y_i - \tilde{y})(x_i - \tilde{x}) / w_i, \quad (2.25)$$

$$\tilde{s}_{xx} = \sum_i (x_i - \tilde{x}) x_i / w_i = \sum_i (x_i - \tilde{x})^2 / w_i. \quad (2.26)$$

Оскільки (2.20), (2.24), (2.25) та (2.26) залежать від вагових факторів (2.18), які, у свою чергу, залежать від шуканих параметрів  $\beta_0$  та  $\beta_1$ , то задача LAD розв'язується методом ітерацій. У якості початкового наближення для  $\beta_0$  та  $\beta_1$  ми обирали OLS рішення. Ітераційна процедура LAD завершується при умові, що зміна  $\beta_0$  та  $\beta_1$  на сусідніх ітераціях менша ніж наперед задана достатньо мала величина **eps**. Блок-схема ітераційного розрахунку методу LAD представлена на рис. 2.2.

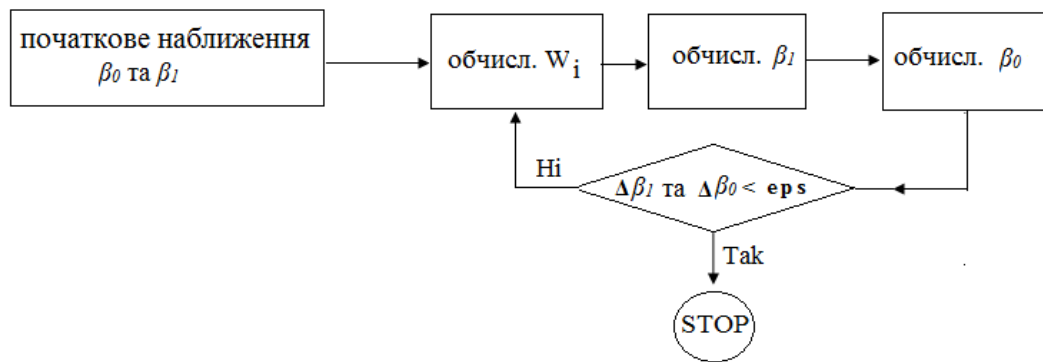


Рис. 2.2 Блок-схема ітераційного розрахунку коефіцієнтів однопараметричного регресійного рівняння методом LAD

У загальному випадку, коли справа іде про полілінійне (багатопараметричне) рівняння вигляду

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m, \quad (2.27)$$

зручно використати матричне формулювання LAD. Функціонал (2.17) тепер буде мати наступний вигляд:

$$H_{LAD}(\beta) = \|Y - X\beta\|_1 = (Y^+ - \beta^+ X^+) W(\beta)^{-1} (Y - X\beta). \quad (2.28)$$

У цьому виразі  $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)$  – вектор-стовпчик, у якому зібрані регресійні коефіцієнти (2.27),  $W(\beta)$  – діагональна матриця, елементами якої є величини (2.18), а  $X$  – матриця предикторів (дескрипторів). При цьому обернена матриця  $W^{-1}$ , завдяки параметру  $\eta$ , існує завжди. Або в іншому варіанті розрахунку можливо покласти  $\eta = 0$  і розраховувати псевдообернену матрицею.

Після певних матричних маніпуляцій, виходячи з (2.28) та виразу для  $\partial H_{LAD} / \partial \beta^+ = 0$ , отримуємо матричне рівняння:

$$\beta = (X^+ W(\beta)^{-1} X)^{-1} X^+ W(\beta)^{-1} Y. \quad (2.29)$$

Оскільки й тут вираз для  $\beta$  залежить від  $W(\beta)$ , рівняння (2.29) розв'язується методом ітерацій.

#### **Алгоритм LAD:**

1) Методом OLS обчислити початкове наближення  $\beta^{old} = (\beta_0, \beta_1, \beta_2, \dots)$ ;

*Ітераційний крок:*

2)  $\beta = \beta^{old}$ . Обчислити  $W(\beta)^{-1}$ . Якщо  $w_{ii} < \epsilon ps$  то  $w_{ii}^{-1} = 0$ ;

3) Обчислити  $\beta^{new} = (X^+ W(\beta)^{-1} X)^{-1} X^+ W(\beta)^{-1} Y$ ;

4) Якщо  $|\beta^{old} - \beta^{new}| > \epsilon ps$ , покласти  $\beta^{old} = \beta^{new}$  і перейти до кроку 2); інакше STOP;

Згідно нашому досвіду представлений алгоритм гарантує надійний розв'язок задачі LAD за помірну кількість кроків – 10-20 ітерацій.

#### **2.3.2. Однопараметрична регресія методом ODR**

Для повноти картини ми наведемо тут також і рівняння для ODR, хоч воно раніше й було описано в літературі<sup>36,37</sup>. Функцію, що відповідає методу ODR, можна отримати за допомогою добре відомої в аналітичній геометрії формули для відстані від точки до прямої (див. наприклад (42)). Отже, для суми квадратів відстаней від точок вибірки  $(x_i, y_i)$  до лінії (2.16) з параметрами  $\beta_0$  та  $\beta_1$  маємо функцію:

$$H_{ODR}(\beta_0, \beta_1) = \frac{1}{1 + \beta_1^2} \sum_i (\beta_0 + \beta_1 x_i - y_i)^2. \quad (2.30)$$

Прирівнюючи похідну до нуля:

$$\frac{\partial H_{\text{ODR}}}{\partial \beta_0} = \frac{2}{1 + \beta_1^2} \sum_i (\beta_0 + \beta_1 x_i - y_i) = 0. \quad (2.31)$$

Отримуємо вираз для  $\beta_0$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}, \quad (2.32)$$

де  $\bar{y}$  та  $\bar{x}$  – звичайні середні значення.

Похідна від (2.30) по  $\beta_1$ , з урахуванням (2.32) має вигляд:

$$\frac{\partial H_{\text{ODR}}}{\partial \beta_1} = \frac{2}{1 + \beta_1^2} (\beta_1 s_{xx} - s_{xy}) - \frac{2\beta_1}{(1 + \beta_1^2)^2} (\beta_1^2 s_{xx} - 2\beta_1 s_{xy} + s_{yy}), \quad (2.33)$$

де використано позначення для коваріаційних матриць

$$s_{xy} = \sum_i (y_i - \bar{y}) x_i = \sum_i (y_i - \bar{y})(x_i - \bar{x}) \quad (2.34)$$

$$s_{xx} = \sum_i (x_i - \bar{x}) x_i = \sum_i (x_i - \bar{x})^2, \quad s_{yy} = \sum_i (y_i - \bar{y}) y_i = \sum_i (y_i - \bar{y})^2 \quad (2.35)$$

Прирівнюючи похідну (2.33) до нуля, після певних алгебраїчних маніпуляцій отримуємо квадратне рівняння для  $\beta_1$ :

$$\beta_1^2 s_{xy} + \beta_1 (s_{xx} - s_{yy}) - s_{xy} = 0. \quad (2.36)$$

Розв'язок цього рівняння і дає шукане (зауважимо – не ітераційне) рішення задачі ODR.

### 2.3.3. Однопараметрична регресія методом LADOD

Функція для нещодавно запропонованого методу найменших модулів ортогональних відстаней (LADOD)<sup>38,39</sup> може бути отримана аналогічно до (2.17) та (2.30):

$$H_{\text{LADOD}}(\beta_0, \beta_1) = \frac{1}{\sqrt{1 + \beta_1^2}} \sum_i |\beta_0 + \beta_1 x_i - y_i| = \frac{1}{\sqrt{1 + \beta_1^2}} \sum_i (\beta_0 + \beta_1 x_i - y_i)^2 / w_i. \quad (2.37)$$

Тут також введено вагову функцію  $w_i$  вигляду (2.18). Слідуючи схемам, що були викладені в підрозділах (2.2.1) та (2.2.2), знаходимо похідні й прирівнюємо їх до нуля.

$$\frac{\partial H_{\text{LADOD}}}{\partial \beta_0} = \frac{2}{\sqrt{1 + \beta_1^2}} \sum_i (\beta_0 + \beta_1 x_i - y_i) / w_i = 0. \quad (2.38)$$

Звідки

$$\beta_0 = \tilde{y} - \beta_1 \tilde{x} . \quad (2.39)$$

Узагальнені (зважені) середні  $\tilde{y}$  та  $\tilde{x}$  мають такий же вигляд як і в методі LAD (2.21, 2.22).

Похідна по другому параметру  $\beta_1$  має вигляд:

$$\frac{\partial H_{\text{LADOD}}}{\partial \beta_1} = \frac{1}{1 + \beta_1^2} \left( \sqrt{1 + \beta_1^2} (2\beta_1 \tilde{s}_{xx} - 2\tilde{s}_{xy}) - \frac{\beta_1}{\sqrt{1 + \beta_1^2}} (\beta_1^2 \tilde{s}_{xx} - 2\beta_1 \tilde{s}_{xy} + \tilde{s}_{yy}) \right) . \quad (2.40)$$

Де введено допоміжні зважені величини аналогічні (2.25, 2.26):

$$\tilde{s}_{xy} = \sum_i (y_i - \tilde{y}) x_i / w_i = \sum_i (y_i - \tilde{y})(x_i - \tilde{x}) / w_i , \quad (2.41)$$

$$\tilde{s}_{xx} = \sum_i (x_i - \tilde{x}) x_i / w_i = \sum_i (x_i - \tilde{x})^2 / w_i , \quad (2.42)$$

$$\tilde{s}_{yy} = \sum_i (y_i - \tilde{y}) y_i / w_i = \sum_i (y_i - \tilde{y})^2 / w_i . \quad (2.43)$$

І знову, виходячи з  $\partial H_{\text{LADOD}} / \partial \beta_1 = 0$ , знаходимо рівняння (тепер вже кубічне!) для  $\beta_1$ :

$$\beta_1^3 \tilde{s}_{xx} + \beta_1 (2\tilde{s}_{xx} - \tilde{s}_{yy}) - 2\tilde{s}_{xy} = 0 . \quad (2.44)$$

Отримане рівняння вже записано в канонічній формі й може бути розв'язано, наприклад, за допомогою формули Кардано. Серед трьох можливих рішень обираємо дійсне, яке відповідає мінімуму (2.37).

Звернемо увагу на те, що метод LADOD, як і LAD, реалізовано за допомогою ітеративної схеми, яка показана на рис. 2.2. Але, на відміну від LAD, на кожній ітерації LADOD потрібен розв'язок кубічного рівняння (2.44).

## 2.4. Константи іонізації карбонових кислот

Для демонстрації методичних основ регуляризованих розрахунків ми розглядаємо побудову QSPR моделі іонізації карбонових кислот. Було обрано експериментальні значення  $pK_a$  для 15 насичених карбонових кислот<sup>104</sup> за 25°C:

НСООН	CH <sub>3</sub> COOH	C <sub>2</sub> H <sub>5</sub> COOH	C <sub>3</sub> H <sub>7</sub> COOH	(CH <sub>3</sub> ) <sub>2</sub> CHCOOH
CH <sub>3</sub> (CH <sub>2</sub> ) <sub>3</sub> COOH	(CH <sub>3</sub> ) <sub>2</sub> CHCH <sub>2</sub> COOH	(CH <sub>3</sub> ) <sub>3</sub> CCOOH	CH <sub>2</sub> FCOOH	CH <sub>2</sub> ClCOOH
CH <sub>2</sub> BrCOOH	CH <sub>2</sub> ICOOH	CHCl <sub>2</sub> COOH	CCl <sub>3</sub> COOH	CF <sub>3</sub> COOH



Досліджувалася залежність  $pK_a$  від дев'яти молекулярних дескрипторів: заряд на кисні карбонільної групи  $C=O$  ( $x_1$ , ат. од.), заряд на кисні гідроксильної -ОН групи ( $x_2$ , ат. од.), заряд на водні гідроксильної групи ( $x_3$ , ат. од.), площа поверхні молекули ( $x_4$ , Å<sup>2</sup>), її об'єм ( $x_5$ , Å<sup>3</sup>), молекулярна рефракцію ( $x_6$ , Å<sup>3</sup>), поляризованість ( $x_7$ , Å<sup>3</sup>), індекс Рендіча ( $x_8$ ), а також інформаційний індекс шляхів у молекулярному графі ( $x_9 = -\sum_i \frac{P_i}{n} \log_2 \frac{P_i}{n}$ ,  $n = \sum_i P_i$ ). Тут величина  $P_i$  – кількість маршрутів довжини «і» у молекулярному графі. При цьому в графі для простоти розглядалися лише атоми карбону. Таким чином, повне можливе рівняння для  $pK_a$  має наступний вигляд:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_9 x_9. \quad (2.45)$$

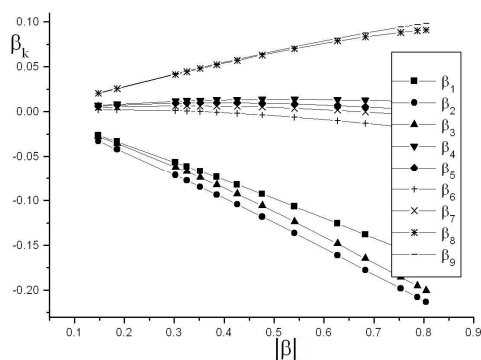
Звісно, вибір необхідних дескрипторів з цих дев'яти може бути зроблено вручну із структурно-хімічних міркувань або суто комбінаторно. Зауважимо, що комбінаторне перерахування навіть для такої відносно простої функції потребує аж  $2^9 = 512$  розрахунків і досліджень можливих регресій (включно тривіальний випадок, коли  $y = \beta_0$ )!

Тут нам було цікаво як поведуть себе регресійні коефіцієнти при розрахунках з  $L_2$ - та  $L_1$ -регуляризацією. Для цього розв'язувались регуляризовані рівняння ( $L_2$ -OLS та LASSO) в Лагранжевій формі:

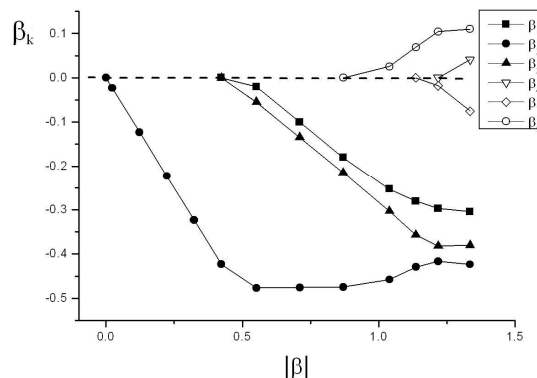
$$\beta(\lambda)_{L_2-OLS} = \arg \min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}, \quad (2.46)$$

$$\beta(\lambda)_{LASSO} = \arg \min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}. \quad (2.47)$$

За різних значень параметру  $\lambda$ , який регулює внесок відповідної норми в (2.46) та (2.47), були отримані залежності коефіцієнтів рівняння (2.45) від  $L_1$ -норми коефіцієнтів  $|\beta| = \sum_i \varpi_i$  (рис. 2.3). При цьому за віссю абсцис відкладали  $L_1$ -норму як для  $L_2$ -OLS, так і для LASSO коефіцієнтів для того, щоб мати можливість зіставити результати розрахунків.



(а)



(б)

Рис. 2.3 Залежності коефіцієнтів регресії для  $pK_a$  карбонових кислот від  $L_1$ -норми вектору коефіцієнтів регресії. (а)  $L_2$ -регуляризація, (б)  $L_1$ -регуляризація

З наведеної залежності (рис. 2.3) можна побачити, що за доволі строгих обмежень ( $\|\beta\|_1 < 0.7$ ) у методі LASSO, у рівнянні залишаються лише три з дев'яти дескрипторів ( $x_1, x_2, x_3$ ). Тут  $\beta_1 \approx \beta_3$  а  $|\beta_2| \gg |\beta_3|$ .

Подальше збільшення параметру  $\lambda$ , відповідно до задачі (2.47), у регресії LASSO призводить до того, що залишається лише один дескриптор –  $x_2$  (заряд на кисні гідроксильної групи). На відміну від LASSO, у методі  $L_2$ -OLS значення всіх коефіцієнтів  $\beta_k$  монотонно змінюються. Очевидно, що така поведінка дескрипторів у  $L_2$ -OLS, на відміну від  $L_1$ -аналогу (LASSO), не дозволяє робити висновки про важливість того чи іншого дескриптору.

Таким чином, відповідно до поведінки коефіцієнтів, отриманих у методі LASSO рис. 2.3 (б), найбільш важливим є дескриптор  $x_2$  (переріз  $\|\beta\|_1 < 0.5$ ). Відповідне рівняння з цим коефіцієнтом в методі OLS має наступний вигляд:

$$pK_a = -24.44 - 91.35x_2, R^2 = 0.852, s = 0.27, Q^2 = 0.805, \theta \approx 0.05, \quad (2.48)$$

а в методі LAD:

$$pK_a = -20.53 - 79.06x_2, R^2 = 0.839, s = 0.28, Q^2 = 0.798, \theta \approx 0.04. \quad (2.49)$$

Тут  $R^2 = R^2_{\text{train}}$  (див. формулу 1.50) – коефіцієнт детермінації отриманий для навчаючої (*train*) вибірки,  $Q^2 = Q^2_{\text{LOO}}$  – коефіцієнт детермінації отриманий за

процедурою *Leave-one-Out* (LOO), див. формулу (1.51),  $\theta = R^2 - Q^2$ , а  $s$  – стандартне відхилення. Рівняння (2.48) та (2.49) можуть розглядатися як задовільні й відповідні одне до одного. Невелика величина  $\theta$  є свідомством відсутності перенавчання. Також ці два рівняння мають близькі стандартні відхилення  $s$ .

Тепер перевіримо рівняння, що мають три дескриптори обрані LASSO  $\|\beta\|_1 \approx 0.7$  (див. рис. 2.3 (б)). Отримуємо для OLS:

$$pK_a = -1.08 - 19.93x_1 - 46.80x_2 - 67.61x_3, R^2 = 0.971, s = 0.62, Q^2 = 0.746. \quad (2.50)$$

Для LAD:

$$pK_a = -4.28 - 17.22x_1 - 55.12x_2 - 61.17x_3, R^2 = 0.969, s = 0.67, Q^2 = 0.201. \quad (2.51)$$

Для методу OLS можна бачити (2.50), що в порівнянні з (2.48) значення коефіцієнту  $R^2$  виявилось кращим, але передбачувальна здатність відповідно до критерію  $Q^2$  є значно гіршою із різницею  $\theta \approx 0.23$ . У той же час рівняння, що отримано в методі LAD ( $Q^2 = 0.201$ ), є й зовсім неадекватним. Таке падіння передбачувальної здатності в трьохпараметричному рівнянні потребує подальшого вивчення.

Зазвичай із збільшенням кількості дескрипторів у рівнянні, параметр  $R^2 = R^2_{\text{train}}$  збільшується. Але це далеко не завжди пов'язано із покращеннями передбачувальних властивостей моделі. Так, у цій тестовій задачі рішення з одним дескриптором (2.48, 2.49) виявилось кращим ніж рішення з трьома.

Ідеологія PCR не вбачає попереднього відбору дескрипторів. У методі PCR варіюється кількість сингулярних чисел ( $n_s$ ), що включають у розклад (див. підрозділ 1.4.3). Результати PCR розрахунку з різними значеннями  $n_s$  наведено на рис. 2.4.

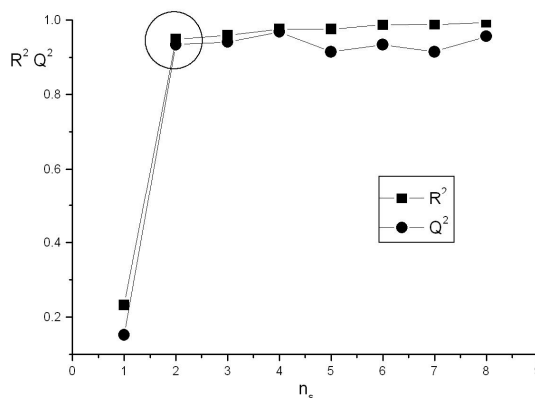


Рис. 2.4 Залежність величин  $R^2$  та  $Q^2$  від кількості сингулярних чисел у методі PCR для розрахунку  $pK_a$  карбонових кислот

З наведеної залежності можна бачити, що за  $n_s=1$  PCR не дозволяє отримати надійне рівняння ( $R^2 \approx 0.2$ ,  $Q^2 \approx 0.1$ ). При збільшенні кількості сингулярних чисел до двох передбачувальна властивість методу значно покращується. Подальше збільшення  $n_s$  не призводить до значного покращення значення  $Q^2$ . Таким чином, метод PCR з лише двома сингулярними числами  $n_s=2$  дає найкраще регресійне рівняння. Ми не наводимо тут громіздкого рівняння, що включає в себе усі дев'ять дескрипторів. Але даний приклад демонструє очевидну перевагу OLS (2.48) та LAD (2.49) рішень – вони є компактними і з гарною прогностичністю.

## 2.5. Оцінки якості неемпіричних розрахунків $pK_a$ фенолів

Стандартний підхід для демонстрації якості теоретичної моделі – це побудова графіка "теорії-експеримент" та розрахунок відповідних лінійних рівнянь. У цьому підрозділі ми дослідили кореляцію між неемпіричним розрахунком  $pK_a$  фенолів та експериментальними даними. Теоретичні, а також експериментальні дані були взяті з (105). Ми обрали два методи Гартрі-Фоківського (*Hartree-Fock*, HF) розрахунку  $pK_a$  з базисними функціями, що включають (b) і не включають (a) дифузні функції на важких атомах

нейтральної молекули. Урахування середовища було виконано в рамках поляризаційно континуальної моделі (CPCM).

Таблиця 2.1

**Два варіанти квантовохімічного розрахунку  $pK_a$  фенолів**

Варіант	Нейтральна молекула	Аніон
a	CPCM/HF/6-31G(d)	CPCM/HF/6-31+G(d)
b	CPCM/HF/6-31+G(d)	CPCM/HF/6-31+G(d)

Результати побудови регресійних моделей наведено в табл. 2.2, 2.3 та рис. 2.5.

Таблиця 2.2

**Регресійні коефіцієнти, а також критерії внутрішньої валідації для залежностей “теорія-експеримент” для  $pK_a$  (варіант а)**

Метод	Коефіцієнти регресії		$R^2$	$Q^2$	$\theta$
<b>OLS</b>	$\beta_0$	0.312	0.860	0.816	0.044
	$\beta_1$	0.970			
<b>LAD</b>	$\beta_0$	1.001	0.855	0.842	0.013
	$\beta_1$	0.898			
<b>ODR</b>	$\beta_0$	-0.423	0.855	0.800	0.055
	$\beta_1$	1.049			
<b>LADOD</b>	$\beta_0$	0.050	0.859	0.859	0.000
	$\beta_1$	1.000			

Очевидно, що найкраще рівняння “теорія-експеримент” відповідає рівнянню з вільним членом, який дорівнює нулю, і тангенсом куту нахилу рівним одиниці:

$$y^{(\text{theor})} = y^{(\text{exp})}. \quad (2.52)$$

Ненульові значення вільного члену говорять про наявність систематичної похибки, а відхилення тангенсу куту нахилу від одиниці характеризує невідповідність якості  $pK_a$  розрахунків для різних молекул. У наших розрахунках LADOD демонструє велику розрахункову точність у розрахунках

$pK_a$  різних молекул у порівнянні з іншими лінійними моделями. LADOD має найменше значення вільного члену за модулем, а також найбільш близьке значення тангенсу куту нахилу до одиниці  $\beta_1 = 1$ , найбільше значення  $Q^2$  і  $\theta \approx 0$ . Інші методи демонструють значно гіршу прогностичну здатність *ab initio* розрахунків  $pK_a$ .

Таблиця 2.3

Регресійні коефіцієнти, а також критерії внутрішньої валідації для залежностей "теорія-експеримент" для  $pK_a$  (варіант b)

Метод	Коефіцієнти регресії		$R^2$	$Q^2$	$\theta$
<b>OLS</b>	$\beta_0$	0.318	0.877	0.833	0.043
	$\beta_1$	0.987			
<b>LAD</b>	$\beta_0$	0.504	0.867	0.864	0.003
	$\beta_1$	0.977			
<b>ODR</b>	$\beta_0$	-0.339	0.872	0.821	0.051
	$\beta_1$	1.058			
<b>LADOD</b>	$\beta_0$	0.290	0.868	0.867	0.000
	$\beta_1$	1.000			

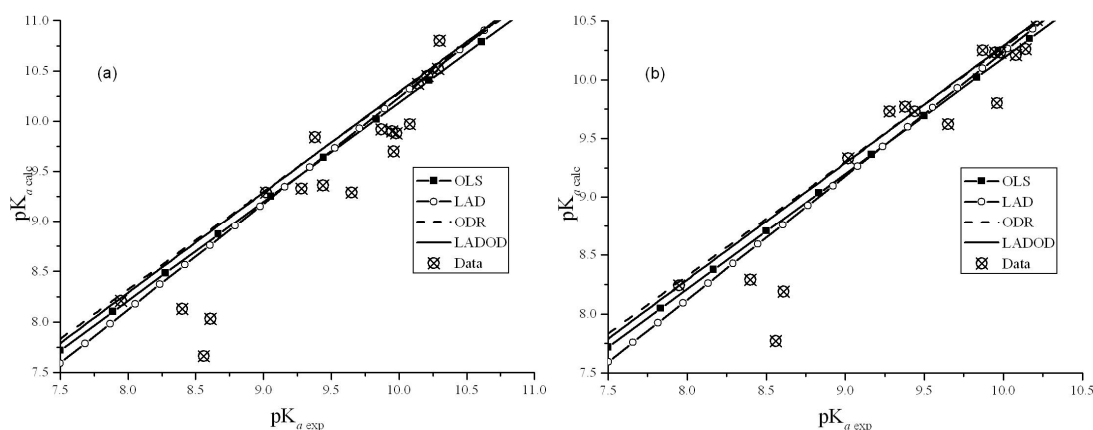


Рис. 2.5 Залежності "теорія-експеримент"  $pK_a$  для двох варіантів розрахунків ("a"- зліва, та "b"- справа)

У кінці цього підрозділу зауважимо, що зіставлення результатів, які отримані різними методами розрахунку регресійних рівнянь, є корисними не тільки тому, що дають можливість побудувати альтернативні моделі

"структура-властивість", але й оцінити ту інформацію, яку в принципі ми здатні вилучити з представлених даних. Такий підхід ми трактуємо як "прагматичний" підхід QSAR.

## 2.6. QSAR моделі опису констант іонізації органічних сполук різної природи

Для цього підрозділу нами було використано дані про  $pK_a$  43 молекул органічних кислот та основ з (106). Структури молекул дивись у додатку В. При цьому, у кислотно-основній рівновазі приймали участь різні функціональні групи. Таким чином, у цьому підрозділі набір молекул був доволі різноманітним у структурному сенсі. Це призводило до того, що для деяких молекул  $pK_a$  моделювалася добре одним набором дескрипторів, у той час як для інших молекул даний набір дескрипторів був не застосовним. Такі обставини повинні призводити до погіршення якості моделей, отриманих у методах, що використовують повний набір дескрипторів (PCR та PLS). У цьому розділі ми не виділяли тестову вибірку, оскільки молекул було доволі мало й були вони надто різноманітними.

Для знаходження відповідного набору рівнянь було розраховано 1083 різноманітних дескрипторів. Далі, за допомогою методу LARS-LASSO, ми отримали ранжовану послідовність дескрипторів, які є найбільш важливими для опису  $pK_a$  органічних кислот (позначення дескрипторів згідно PaDEL-Descriptor<sup>102</sup>):

$$AATS4e > AATSC5e > AATS1s > C2SP3 > AATSC5s > GATS5s > \max Hother > JGI3. \quad (2.53)$$

Відповідний профіль змін коефіцієнтів у LASSO розрахунку наведено на графіку (Рис. 2.6).

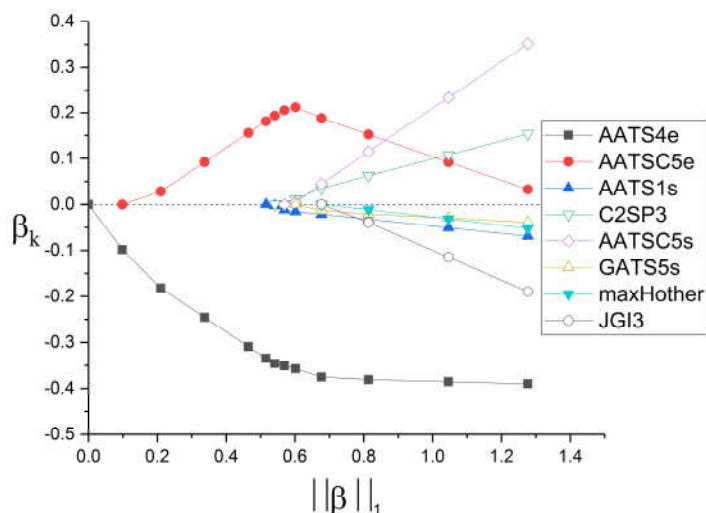


Рис. 2.6 Профілі зміни дескрипторів у методі LASSO у дослідженні  $pK_a$  органічних кислот

З наведеного рис. 2.6 можна бачити, що найбільш важливі дескриптори – це AATS4e та AATSC5e. Внесок усіх інших дескрипторів зникає з моделі дуже швидко. Таким чином, ми можемо припустити, що включення більш ніж двох дескрипторів у модель призведе до малих загальних покращень у моделях. Зауважимо також, що поведінка профілів LASSO є нетривіальною. Так, внесок індексу AATSC5e при великих значеннях  $\|\beta\|_1$  є досить малим. Але зі зменшенням  $\|\beta\|_1$  його значимість суттєво зростає.

Ефективність відбору за допомогою процедури LASSO підтверджує рис. 2.7, який демонструє порівняння якості моделей, що були отримані в методах PCR, PLS та OLS і LAD. Як і очікувалось, якість рівнянь, отриманих у методах PCR та PLS, виявилася доволі низькою. Хоча критерій  $R^2$  у PLS розрахунку є високими, передбачувальна здатність відповідно до критерію  $Q^2$  у таких моделей дуже низька.



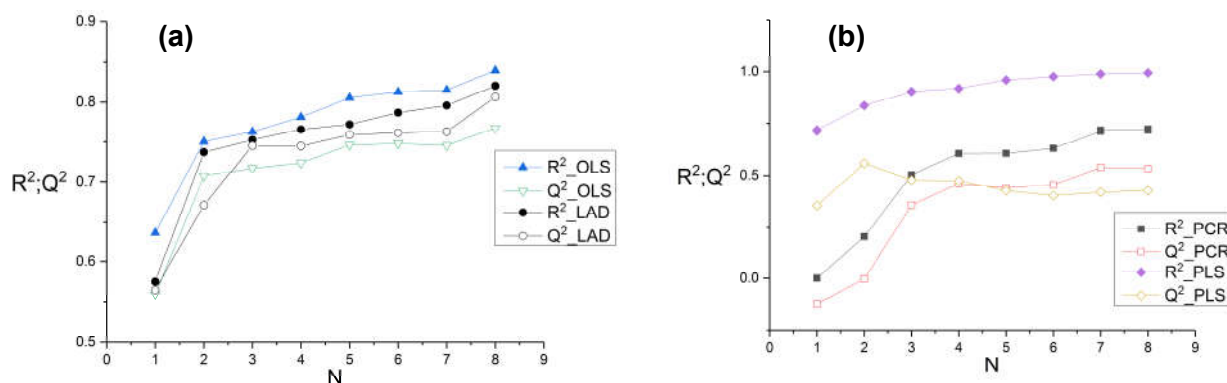


Рис. 2.7 Порівняння коефіцієнтів внутрішньої валідації, що були отримані для рівнянь у методах (a) OLS та LAD, (b) PCR та PLS для  $pK_a$  органічних сполук.

Для графіку (a) N – кількість дескрипторів, для графіку (b) N – кількість урахованих сингулярних чисел

Двопараметричні рівняння, що отримані в методах OLS та LAD для  $pK_a$ , мають наступний вигляд:

$$\text{OLS: } pK_a = 42.189 - 4.478AATS4e + 112.5AATSC5e, R^2 = 0.750, Q^2 = 0.707, \quad (2.54)$$

$$\text{LAD: } pK_a = 48.096 - 5.282AATS4e + 112.0AATSC5e, R^2 = 0.737, Q^2 = 0.672, \quad (2.55)$$

тут AATS4e – *Average Broto-Moreau autocorrelation – lag 4/weighted by Sanderson electronegativities* а AATSC5e – *Average centered Broto-Moreau autocorrelation - lag 5/weighted by Sanderson electronegativities* (детальні пояснення щодо цих дескрипторів див. (90)).

Графіки залежності теорія-експеримент у методах OLS та LAD мають наступний вигляд:

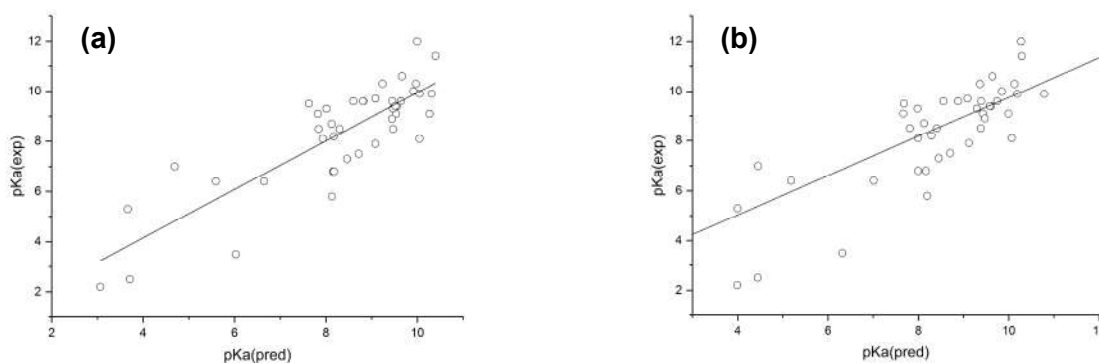


Рис. 2.8 Залежність експериментальних значень  $pK_a$  органічних сполук (а) у методі OLS і (б) у методі LAD від передбачених за процедурою LOO-CV для регресійних рівнянь із двома дескрипторами

Для методу OLS було отримано наступне рівняння теорія-експеримент:

$$pK_a(\text{exp}) = 0.966pK_a(\text{pred}) + 0.287, \quad (2.56)$$

а для LAD:

$$pK_a(\text{exp}) = 0.787pK_a(\text{pred}) + 1.889. \quad (2.57)$$

Висока різноманітність молекул у вибірці не дозволила в цьому підрозділі вилучити навіть невелику кількість молекул для цілей тестування. Виключення з тренувального набору навіть однієї молекули (процедура LOO-CV) призводило до значного погіршення передбачувальних властивостей моделей для цієї молекули в методі PLS (див. рис. 2.7 (b)). Такий висновок, а також висновок про погану поведінку отриманих моделей, може бути зроблено з аналізу критеріїв внутрішньої валідації. Наприклад, велика різниця між критеріями  $R^2$  та  $Q^2$  для моделей, отриманих методом PLS, пов'язана з перенавчанням, що говорить про те, що моделі отримані цим методом для цієї проблеми є не надійними. Виключення з навчаючої вибірки навіть однієї молекули призводить до того, що значення коефіцієнту  $Q^2$  опиняються дуже низькими, у той час як значення  $R^2$  доволі велике. Метод PCR, на відміну від OLS, виявився не спроможним надати прийнятні результати навіть для тренувальної вибірки відповідно до значень  $R^2$  та  $Q^2$  і тому не може бути визнаний як адекватний. Регресія OLS та LAD дала в цьому випадку найбільш

надійні рівняння, що менш чутливі до зниження кількості молекул у тренувальній вибірці. Такий висновок було зроблено, оскільки різниці між критеріями  $R^2$  та  $Q^2$  доволі мала. Навіть із зменшеною кількістю молекул у тренувальній вибірці ці методи можуть давати задовільні результати для  $pK_a$  органічних сполук, хоча з не дуже високими значеннями критеріїв внутрішньої валідації.

Для того, щоб отримати моделі з кращими критеріями валідації, ми припустили нелінійну структуру задачі, що може бути промодельована з використанням методу штучних нейронних мереж (*Artificial Neural Networks*, ANN). Для цього нами було використано програмний пакет **NeuPy**<sup>107</sup>.

Ми використали багат шарову нейронну мережу прямого розповсюдження (*feedforward neural network*). Використана штучна нейронна мережа мала три шари: вхідний та вихідний шар з лінійною функцією активації, а також одним прихованим шаром з функцією активації у вигляді гіперболічного тангенсу.

Прихований шар мав чотири нейрони. Моделі з такою кількістю нейронів давали найкращі критерії внутрішньої валідації. Структура нейронної мережі, що була використана нами, наведено на рис. 2.9.

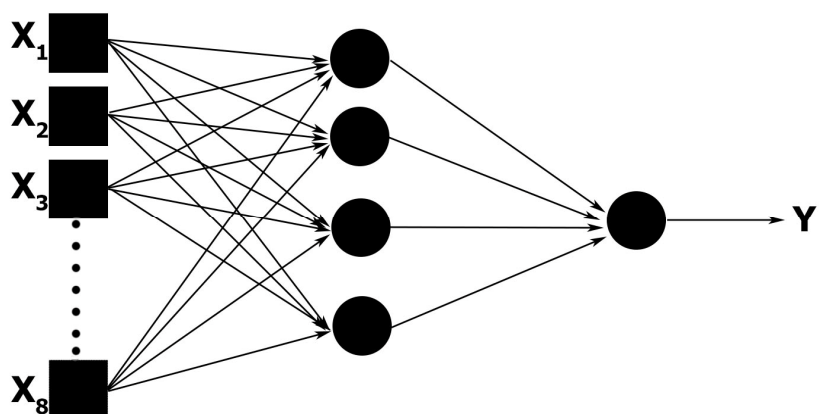


Рис. 2.9 Архітектура штучної нейронної мережі що була використана в розрахунках (вісім дескрипторів). Тут квадрати позначають вхідні дескриптори, а кола – нейрони

Параметри ініціалізації нейронної мережі були отримані з нормального розподілу з середнім значенням, що дорівнює нулю та стандартним відхиленням = 0.01. Для тренування ANN було використано модифікований метод спряжених градієнтів.

Для того, щоб порівняти отримані ANN моделі з отриманими в інших методах, ми також використовували процедуру LOO-CV. Ця процедура була реалізована наступним чином: ANN була натренована з використанням повного набору даних, після чого параметри ANN були використані в якості початкового наближення для кожного етапу процедури LOO-CV. Критерій  $Q^2$  був достатньо високим лише для таких ANN, що практично не змінювались в ході процедури LOO-CV (ANN натренована для всієї вибірки була також оптимальною для всіх наборів даних, отриманих виключенням однієї молекули з вхідного набору).

У табл. 2.4 наведено результати критеріїв внутрішньої валідації для різної кількості дескрипторів використаних у побудуванні ANN. Дескриптори включалися в рівняння виходячи з послідовності 2.53 таким же чином як і в методах OLS та LAD.

Таблиця 2.4

**Коефіцієнти внутрішньої валідації отримані для ANN з різною кількістю дескрипторів**

Кількість дескрипторів	$R^2$	$Q^2$
1	0.680	0.529
2	0.833	0.787
3	0.854	0.679
4	0.932	0.709
5	0.967	0.865
6	0.980	0.882
7	0.994	0.943
8	0.994	0.935

Як можна бачити з табл. 2.4 ANN з лише двома дескрипторами дає значно кращі критерії валідації ніж моделі отримані в інших методах, що вивчалися. Моделі з більшою кількістю дескрипторів, згідно критеріїв внутрішньої валідації, також не втрачають у передбачувальній здатності. Найкращі результати для ANN були отримані із сьома дескрипторами з  $Q^2 = 0.943$ . Відповідна залежність "теорія-експеримент" для ANN з двома дескрипторами наведена на рис. 2.10.

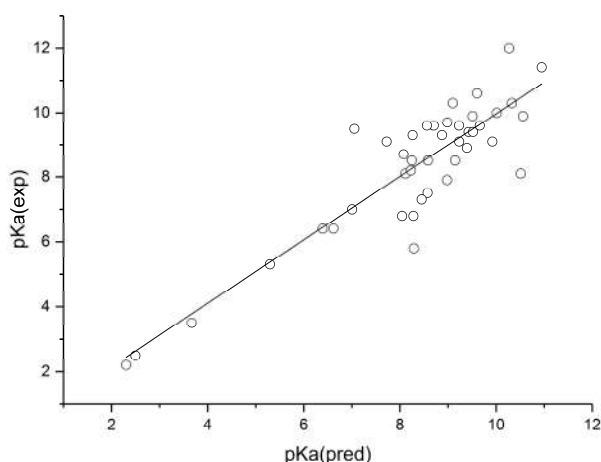


Рис. 2.10 Залежність експериментальних значень  $pK_a$  органічних сполук від передбачених значень за процедурою LOO-CV для ANN з двома дескрипторами

Відповідне рівняння для залежності "теорія-експеримент" (рис. 2.10) має вигляд:

$$pK_a(\text{exp}) = 0.977pK_a(\text{pred}) + 0.196; \quad Q^2 = 0.787 \quad (2.58)$$

У кінці параграфа зауважимо, що одночасне використання багатьох регресійних моделей дає можливість:

1) обрати найкращу; 2) співставлення результатів для різних моделей дозволяє оцінити дійсну якість дескрипторного набору.

## 2.7. Температура кипіння органічних сульфідів

Для цих розрахунків нами використовувався набір даних з 43 молекулами органічних сульфідів<sup>108</sup>. Оскільки ця вибірка містить у собі лише молекули одного виду з різними аліфатичними залишками, можна припустити, що для опису температур кипіння (*Boiling Point*, BP) цих сполук буде достатньо використання лише 1D та 2D дескрипторів.

Такі дескриптори описують порядок зв'язку атомів у молекулі – "молекулярну топологію". Тут і надалі ми видаляли дескриптори, що мають константні значення для всіх молекул вибірки. Після цього вибірка з 501 дескриптора використовувалася в подальших розрахунках.  $L_1$ -регуляризовані методи (LASSO, LARS) дозволили нам отримати найбільш важливі дескриптори. Після чого саме ці дескриптори використовувалися в подальших розрахунках з використанням методів OLS та LAD. Наші розрахунки показали, що використання одного дескриптора, а саме MLFER\_L (*Solute gas-hexadecane partition coefficient*)<sup>109</sup>, достатньо для опису BP. Відповідні рівняння мають наступний вигляд:

у методі OLS:

$$BP(^{\circ}C) = -52.04 + 48.58MLFER\_L, R^2 = 0.982, Q^2 = 0.979, \quad (2.59)$$

у методі LAD:

$$BP(^{\circ}C) = -49.31 + 47.75MLFER\_L, R^2 = 0.981, Q^2 = 0.981. \quad (2.60)$$

Можна побачити, що рівняння практично однакові. Значення критеріїв валідації  $Q^2$  та  $R^2-Q^2$  у методі LAD трохи краще.

Метод PCR для цього завдання потребував декілька сингулярних чисел для досягнення близьких значень до тих, що були отримані в методі OLS (2.59) або LAD (2.60), табл. 2.5. При цьому сам метод PCR генерує рівняння, що містить 501 терм!

Таблиця 2.5

**Коефіцієнти внутрішньої валідації як функція від сингулярних чисел  
( $n_s$ ) в методі PCR**

$n_s$	$R^2$	$Q^2$
1	0.919	0.926
2	0.979	0.978
3	0.981	0.976
4	0.981	0.976
5	0.989	0.985

Зрозуміло, що однопараметричні рівняння типу (2.59, 2.60) не є єдино можливими. Для того, щоб знайти альтернативні розв'язки, ми виключили "найкращий" дескриптор (MLFER\_L) і повторили розрахунок. На рис. 2.11 зображені профілі змін коефіцієнтів  $\beta_i$  отримані методом LARS-LASSO.

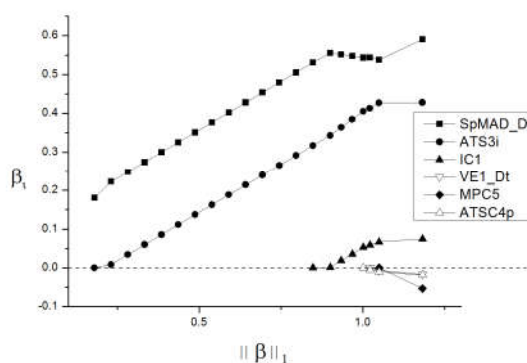


Рис. 2.11 ВР тіоетерів. Профілі коефіцієнтів отримано в методі LARS-LASSO.

З наведеної залежності рис. 2.11 можна бачити, що найбільш важливим дескриптором є SpMAD\_Dt (*Spectral mean absolute deviation from detour matrix*), ATS3i (*Broto-Moreau autocorrelation - lag 3 / weighted by first ionization potential*), а також добре відомий індекс IC<sub>1</sub> (*First-order Informational Contents Index*). Більш детальну інформацію з цих параметрів можна знайти в PaDEL-Descriptor мануалі<sup>102</sup>, а також у книзі (89). Порівняльну характеристику рівнянь (OLS та LAD) наведено в табл. 2.6. Можна бачити, що рівняння мають дуже непогані значення критеріїв валідації.

Таблиця 2.6

**Коефіцієнти внутрішньої валідації для альтернативних рівнянь, отриманих для ВР тіоефірів. Методи OLS/LAD, m – кількість дескрипторів в рівнянні**

m	$\beta_0$	SpMAD_Dt	ATS3i	IC <sub>1</sub>	R <sup>2</sup>	Q <sup>2</sup>
1	-13.02/-17.54	33.83/34.48	–	–	0.961/0.958	0.956/0.955
2	1.23/2.01	20.98/21.99	$6.12 \cdot 10^{-3} / 5.27 \cdot 10^{-3}$	–	0.978/0.977	0.974/0.975
3	-35.26/-35.71	18.47/19.14	$6.93 \cdot 10^{-3} / 6.77 \cdot 10^{-3}$	26.64/26.07	0.985/0.984	0.981/0.981

Графічну репрезентацію залежності "теорія LOO-CV–експеримент" для регресійного рівняння, отриманого в методі LAD (m = 3), наведено на рис. 2.12. Тут ми не приводимо відповідну залежність для методу OLS, оскільки вона майже повністю співпадає із залежністю для LAD.

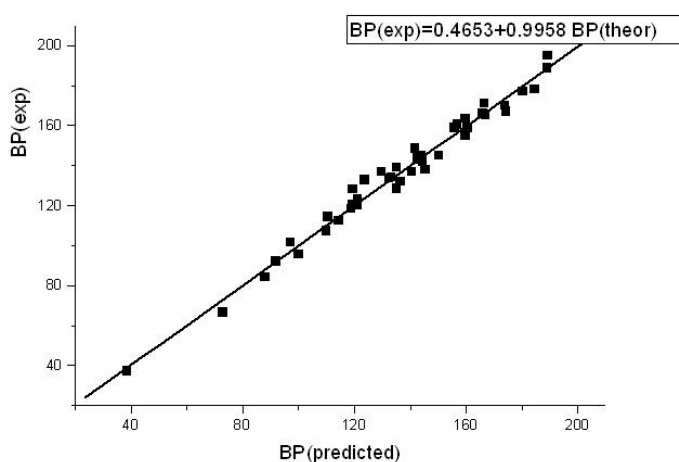


Рис. 2.12 Залежність експериментальних значень (LAD) від теоретичних значень (LOO-CV) для ВР тіоефірів

Для валідації результатів, отриманих у розрахунку без використання дескриптору MLER\_L, ми вибрали тестову вибірку, що складалася з 10 молекул. Отже  $43 - 10 = 33$  молекули були використані як тренувальний (навчальний) набір для отримання моделей в методах PCR, OLS, та LAD.

У якості критерію зовнішньої валідації ми використовували критерій  $R^2_{\text{test}}$ . Виявилося, що для найгіршого випадку лише з однією латентною змінною,



метод PCR дає параметр  $R^2_{\text{test}} = 0.875$ . При збільшенні кількості латентних змінних критерій зовнішньої валідації збільшувався до значень  $R^2_{\text{test}} \approx 0.97$ . У методах OLS та LAD з використанням лише одного дескриптору коефіцієнт  $R^2_{\text{test}} \approx 0.9$ . Подальше збільшення кількості дескрипторів у OLS та LAD також призводило до покращення критерію  $R^2_{\text{test}}$ . Коефіцієнти внутрішньої валідації практично не змінилися при зменшенні кількості молекул в тренувальній вибірці з 43 до 33.

Загалом відзначимо, що, згідно нашого "прагматичного підходу", близькість результатів, які були отримані в рамках різних регресійних підходів, є свідомством адекватності обраної моделі QSAR/QSPR.

## 2.8. Температура кипіння флуороалканів

У цьому розділі ми використовували дані щодо температури кипіння (BP) вибірки флуороалканів ( $^{\circ}\text{C}$ ), яка складалася з 82 молекул<sup>110</sup>. Структури молекул дивись у додатку Г). Було розраховано 919 дескрипторів з використанням програми E-Dragon 1.0 (100,101). 24 молекули з 82 (30%) було використано в якості тестової вибірки для валідації моделей, отриманих з використанням 68 молекул, що не увійшли до тестового набору, у якості тренувальної (навчальної) вибірки.

З використанням LARS-LASSO підходу, ми отримали послідовний набір дескрипторів, що, відповідно до статистичної точки зору, є найбільш важливими для передбачення температур кипіння флуороалканів:

$$\text{IAC} > \text{TIC2} > \text{X0A} > \text{RDF020v} > \text{Mor08m} > \text{HATS2v} > \text{Mor30u} > \text{HATS2e}. \quad (2.61)$$

Інформацію з використаних дескрипторів можна знайти документації DRAGON<sup>111,89</sup>.

Для отримання регресійних рівнянь з найкращою передбачувальною здатністю, необхідно послідовно додавати дескриптори з 2.61 до рівняння. Профіль зміни коефіцієнтів дескрипторів в методі LASSO зображено на рис. 2.13.

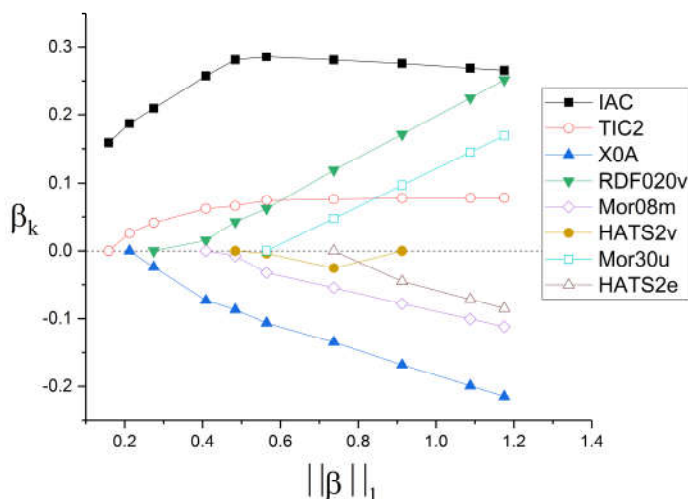


Рис. 2.13 Профіль зміни коефіцієнтів при дескрипторах у методі LASSO у дослідженні температур кипіння флуороалканів

З рис. 2.13 можна зробити висновок, що практично всі дескриптори зменшуються до нуля із зменшенням  $\|\beta\|_1$ , але існують і винятки. Дескриптор HATS2V опиняється в рівнянні одразу після зникнення дескриптора HATS2e. Ці дескриптори доволі сильно скорельовані ( $R = 0.986$ ). Таким чином, заміна дескриптору HATS2e дескриптором HATS2V практично не впливає на якість регресійного рівняння. Метод LASSO зазвичай не залишає сильно корельовані дескриптори в рівнянні. Це важлива властивість методу, оскільки в отриманих таким чином рівняннях не виникає проблем з колінеарністю дескрипторів.

Для цієї вибірки ми отримали регресійні рівняння, послідовно добавляючи дескриптори в рівняння, виходячи з набору (2.61). У табл. 2.7 наведено порівняння коефіцієнтів валідації отриманих моделей з використанням методів OLS, LAD, а також PCR та PLS. У цій таблиці  $N$  – це кількість дескрипторів (для OLS та LAD) та латентних змінних (для PCR та PLS).

Таблиця 2.7

**Порівняння якості моделей, отриманих у методах OLS, LAD, PCR та PLS, для ВР флуороалканів**

N	OLS			LAD			PCR			PLS		
	$R^2$	$Q^2$	$R^2_{\text{test}}$	$R^2$	$Q^2$	$R^2_{\text{test}}$	$R^2$	$Q^2$	$R^2_{\text{test}}$	$R^2$	$Q^2$	$R^2_{\text{test}}$
1	0.74	0.72	0.75	0.74	0.74	0.75	0.15	0.08	0.19	0.66	0.59	0.68
2	0.77	0.75	0.73	0.77	0.77	0.75	0.67	0.63	0.74	0.82	0.78	0.85
3	0.86	0.83	0.77	0.82	0.81	<b>0.58</b>	0.68	0.61	0.76	0.94	0.90	0.91
4	0.92	0.89	0.92	0.90	0.85	0.89	0.84	0.81	0.80	0.96	0.92	0.94
5	0.94	0.92	0.91	0.94	0.93	0.92	0.86	0.82	0.82	0.97	0.93	0.96
6	0.95	0.93	0.94	0.95	0.93	0.94	0.87	0.83	0.83	0.98	0.94	0.96

З табл. 2.7. можна побачити, що, незважаючи на те, що в методі PCR було використано всі дескриптори, відповідна якість моделей, отриманих методом, виявилася значно гіршою, ніж якість рівнянь, отриманих у OLS. Таким чином, рівняння, отримані в двоступеневій процедурі LARS-LASSO-OLS, є більш точними, а також їх значно легше інтерпретувати, ніж відповідні рівняння, отримані в PCR. З іншого боку, якість рівнянь, отриманих в PLS, близька до рівнянь, отриманих в процедурі LARS-LASSO-OLS.

Як і раніше, перевага методів OLS та LAD, що використовують скорочений набір дескрипторів, полягає в простоті отриманих рівнянь, у порівнянні з формально багатопараметричними методами!

Найпростіші регресійні рівняння для ВР флуороалканів, отримані в методах OLS та LAD, мають наступний вигляд. Для OLS:

$$\begin{aligned} \text{BP}(^{\circ}\text{C}) &= 516.881 + 4.126\text{IAC} + 0.718\text{TIC2} - 692.497\text{X0A}, \\ R^2 &= 0.861, \quad Q^2 = 0.828, \quad R^2_{\text{test}} = 0.774. \end{aligned} \quad (2.62)$$

Для LAD:

$$\begin{aligned} \text{BP}(^{\circ}\text{C}) &= 884.796 + 3.128\text{IAC} + 0.752\text{TIC2} - 1107.716\text{X0A}, \\ R^2 &= 0.821, \quad Q^2 = 0.811, \quad R^2_{\text{test}} = 0.583. \end{aligned} \quad (2.63)$$

На наступних рисунках (рис. 2.14, 2.15, 2.16, 2.17) наведено залежності "теорія-експеримент" (для процедури LOO-CV, а також для тестового набору),

розраховані в методах OLS, LAD, PCR та PLS відповідно. Для побудови цих залежностей у рівняннях регресії використовували 3 дескриптори, а також 3 латентні змінні.

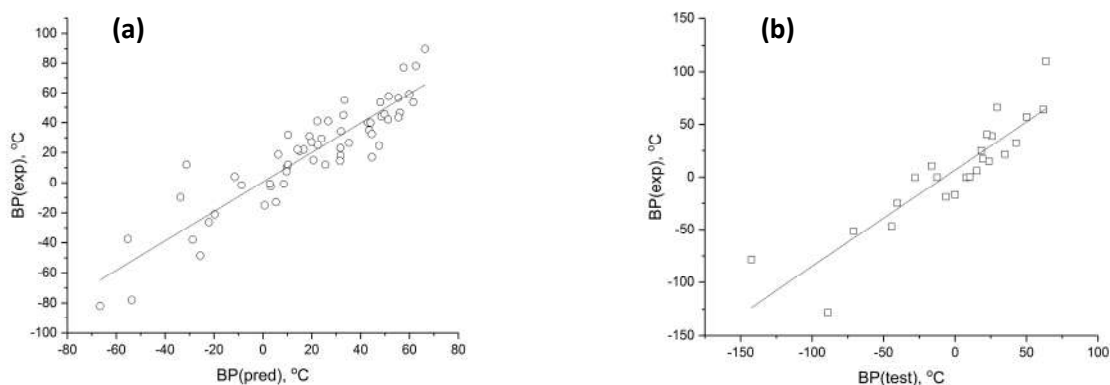


Рис. 2.14 Залежність експериментальних значень ВР від передбачених значень ВР для флуороалканів за процедурою LOO-CV (а) та передбачених для тестового набору (b) (результати було отримано в методі OLS з трьома дескрипторами у рівнянні)

Відповідні рівняння для залежності  $BP(exp) - BP(pred/test)$  для залежностей на рис. 2.14 (a, b), отримані в методі OLS, мають наступний вигляд:

$$\begin{aligned} BP(exp) &= 0.979BP(pred) + 0.471, \\ BP(exp) &= 0.875BP(test) + 6.613. \end{aligned} \quad (2.64)$$

Тут і надалі ми будемо вважати, що  $BP(exp)$  – експериментальні значення для властивості,  $BP(pred)$  – значення отримані в LOO-CV процедурі, а  $BP(test)$  – значення розраховані для тестової вибірки.

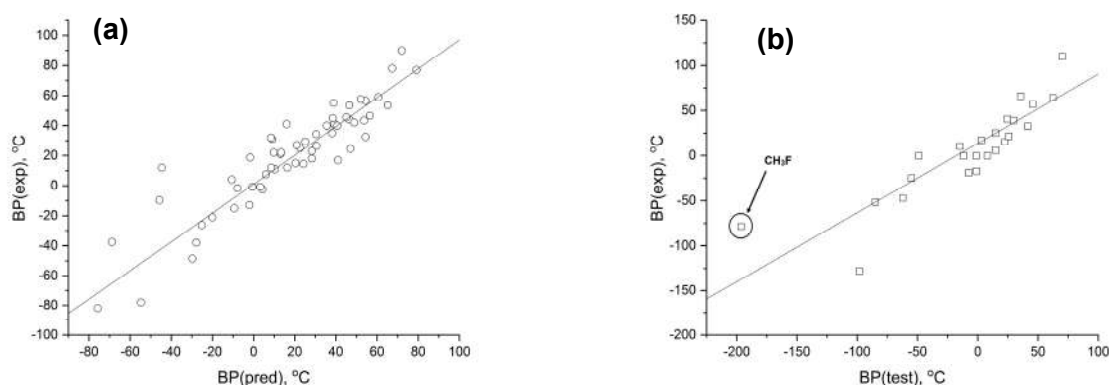


Рис. 2.15 Залежність експериментальних значень ВР від передбачених значень ВР (метод LAD з трьома дескрипторами в рівнянні) для флуороалканів за процедурою LOO-CV (а) та передбачених ВР для тестового набору (б)

Відповідні рівняння для залежності  $BP(exp)$  –  $BP(pred/test)$  (рис. 2.15 (а,б)), що отримані в методі LAD, мають наступний вигляд:

$$\begin{aligned} BP(exp) &= 0.960BP(pred) + 0.994, \\ BP(exp) &= 0.768BP(test) + 13.579. \end{aligned} \quad (2.65)$$

Найвіддаленіша точка тестової вибірки від ідеальної лінії теорія-експеримент із координатами  $BP(exp) = -78.5^\circ C$ ;  $BP(test) = -196.3^\circ C$  (див. рис. 2.15 (б)) відповідає молекулі  $CH_3F$ , і, вірогідно, вона спричинила погіршення оцінок регресійної моделі LAD. Такий викид пов'язано з простотою молекули, що зробило її "дуже відмінною" від інших молекул тестової вибірки. Це призвело до того, що "робастний" метод LAD не брав до уваги цю точку в рівнянні з трьома параметрами. У рівняннях PCR та PLS проблем з цією точкою не було.

Для методу PCR ми отримали наступні рівняння:

$$\begin{aligned} BP(exp) &= 0.980BP(pred) - 0.074, \quad R^2 = 0.615, \\ BP(exp) &= 1.203BP(test) + 2.016, \quad R^2 = 0.789. \end{aligned} \quad (2.66)$$

А для PLS:

$$\begin{aligned} BP(exp) &= 1.019BP(pred) - 0.550, \quad R^2 = 0.900, \\ BP(exp) &= 1.004BP(test) + 3.643, \quad R^2 = 0.912. \end{aligned} \quad (2.67)$$

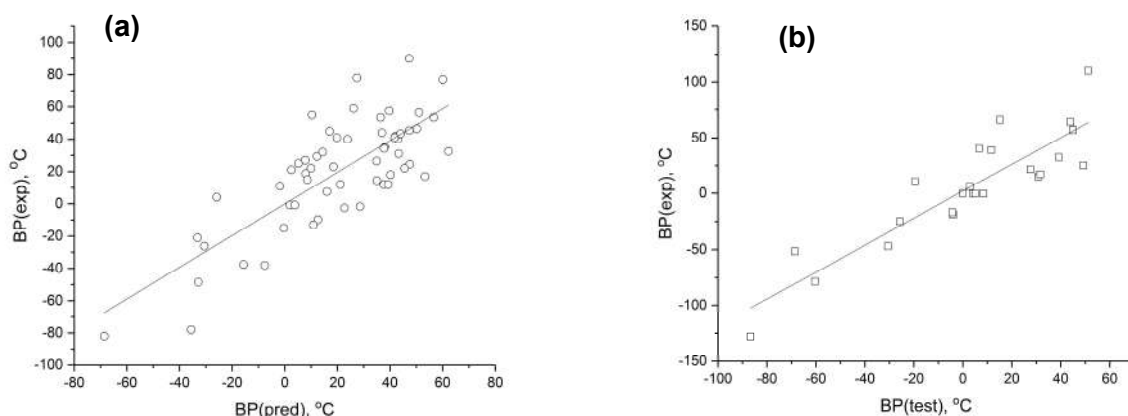


Рис. 2.16 Залежність експериментальних значень ВР від передбачених за процедурою LOO-CV (a) та передбачених ВР для тестового набору (b).

Результати було отримано в методі PCR з трьома латентними змінними у рівнянні

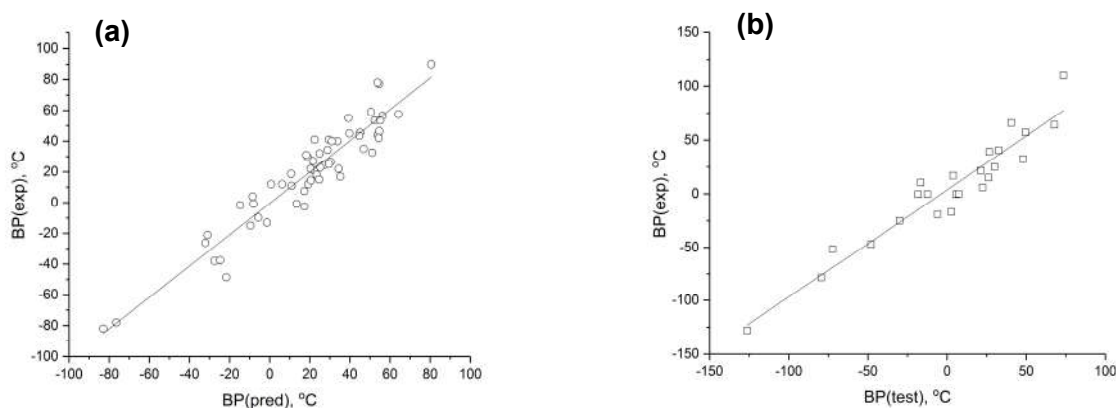


Рис. 2.17 Залежність експериментальних значень ВР флуороалканів від передбачених за процедурою LOO-CV (a) та передбачених для тестового набору (b). Результати було отримано в методі PLS з трьома латентними змінними у рівнянні

Слід зазначити, що структури молекул у цьому підрозділі були дуже схожі. Саме тому  $Q^2$ , розрахований в усіх методах, виявився дуже схожим на  $R^2$  для тренувальної вибірки.

## 2.9. В'язкість рідин та тиск насиченого пару органічних сполук

У цій частині ми демонструємо застосування методів ODR та LADOD. При цьому досліджується кореляція між двома експериментальними параметрами: в'язкістю ( $\log \eta$ ) та тиском насиченої пари органічних сполук за температури 20 °C. Експериментальні дані для 116 структурно-різних сполук було взято з (112). Набір структур включав насичені та ненасичені вуглеводні, ароматичні системи, спирти, етери, естери, азотовмісні та галогеновмісні системи, кетони та кислоти.

Попередній аналіз продемонстрував наявність деякого ступеню кореляції між цими двома параметрами, а найбільше відхилення від лінійної залежності з'являється лише для систем з великою в'язкістю відповідних рідин. Зрозуміло, що такі прості залежності не можуть бути використані для опису рідин з сильними міжмолекулярними взаємодіями (наприклад такими, що мають сильні водневі зв'язки). Не зважаючи на це, обрані дані демонструють слабку лінійну залежність між обраними експериментальними параметрами (табл. 2.8. та рис. 2.18).

Таблиця 2.8

**Регресійні коефіцієнти, а також критерії внутрішньої валідації для залежності  $\log \eta$  (мПа·сек) від  $\log P$  (кПа) за температури  $T = 20^\circ\text{C}$**

Метод	Коефіцієнти регресії		$R^2$	$Q^2$	$\theta$
<b>OLS</b>	$\beta_0$	-0.0043	0.677	0.663	0.014
	$\beta_1$	-0.300			
<b>LAD</b>	$\beta_0$	-0.09	0.629	0.611	0.019
	$\beta_1$	-0.257			
<b>ODR</b>	$\beta_0$	0.0004	0.676	0.661	0.015
	$\beta_1$	-0.312			
<b>LADOD</b>	$\beta_0$	-0.0870	0.634	0.634	0.000
	$\beta_1$	-0.262			

Рівняння, отримані в методах LAD та LADOD, виявились практично ідентичними й сильно відрізняються від отриманих у методах OLS та ODR

(рис. 2.18). Такі відмінності пов'язані з робастністю методів LADOD та LAD. Слід також зазначити, що метод LADOD навіть за погано скорельованого набору даних демонструє стабільність у процедурі LOO-CV ( $\theta \approx 0$ ).

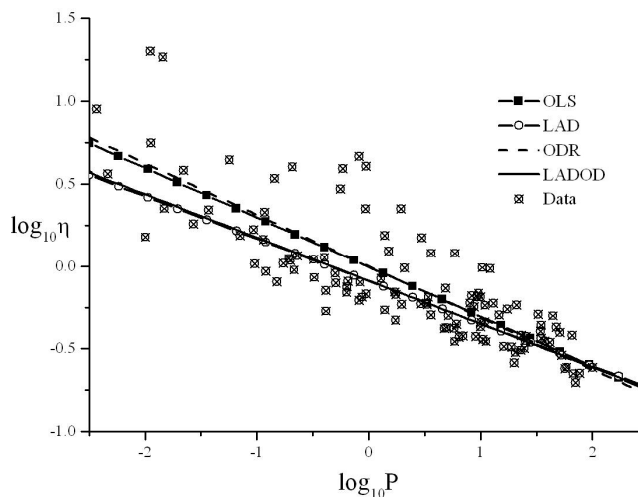


Рис. 2.18 Залежність  $\log \eta$  (мПа·с) від  $\log P$  (кПа) за температури  $T = 20^\circ\text{C}$

Відзначимо, що загальне узгодження результатів у рамках представлених альтернатив (згідно нашого "прагматичної підходу") дає змогу припустити, що наявні моделі є найкращими з можливих для запропонованої гіпотези про зв'язок  $\log \eta - \log P$ .

## Висновки до розділу 2

Теоретичне співставлення результатів, які отримані різними регресійними підходами у даному розділі дисертації, дозволяє сформулювати кілька важливих висновків. А саме:

1.  $L_1$ -регуляризація навіть у "важких випадках" здатна сформулювати послідовний (ранжований) набір дескрипторів, завдяки якому можливо сформулювати достатньо прості регресійні OLS- чи LAD-рівняння.

2. У методі PCR або PLS можна отримати надійні, з прогностичної точки зору, моделі, але з хімічної точки зору, на відміну від OLS чи LAD, такі моделі не надають ясної інформації стосовно природи отриманих рівнянь і не



відповідають на питання: які структурно-хімічні особливості молекули призводять до змін у відгуку (активності).

3. У прикладах, що були наведені у даному розділу,  $L_1$ -регуляризація дозволила сформулювати компактні одно- двох- або трьох- параметричні моделі, які здатні задовільно описати набір даних. Відповідно до вивчених прикладів, моделі, отримані з попереднім відбором з використанням LARS-LASSO, можуть виявитись кращими, ніж результати розрахунків PLS та PCR.

4. Метод нейронних мереж, що використовує дескриптори, обрані з використанням LASSO (або LARS-LASSO), може бути використаний у якості альтернативи лінійним методам регресії для отримання моделей з кращими критеріями валідації.

5. Серед представлених у дисертації методів побудови лінійної регресії розглянуто чотири альтернативних підходи (OLS, LAD, ODR та LADOD). Той чи інший підхід може виявитись кращим з точки зору прогнозу властивостей систем. Для оцінки працездатності методів (чи моделей) ми пропонуємо так званий "прагматичний підхід", оснований на зіставленні результатів розрахунків різними методами. При цьому значні розбіжності в отриманих моделях та відповідних прогнозах можуть бути свідомством недостатньо глибокого аналізу даних та необхідності пошуку іншого, більш вдалого, дескрипторного набору або регресійної моделі.

6. Використаний набір математичних інструментів для побудови регресійних моделей дозволив сформулювати адекватні рівняння для ряду важливих фізико-хімічних параметрів, серед яких:  $pK_a$ , температури кипіння, в'язкість. Важливо, що отримані рівняння були побудовані для "важких" ситуацій, коли навчаючі вибірки включали структурно-різноманітні системи.

Основні положення цього розділу викладено в публікаціях автора (39,113-116).

### РОЗДІЛ 3

## ТЕСТОВІ ДОСЛІДЖЕННЯ ВАЛІДАЦІЙНИХ ХАРАКТЕРИСТИК РЕГРЕСІЙНИХ QSAR/QSPR РІВНЯНЬ

Результати розрахунків QSAR/QSPR зазвичай характеризуються помітним (часом – значним) розкидом даних. Такий розкид веде до певної невизначеності отриманих теоретичних величин та тих оцінок і прогнозів, які робляться на основі розроблених моделей. У зв'язку з цим, постає важливе питання щодо валідації QSAR/QSPR моделей (зокрема регресійних рівнянь). Як уже було сказано в літературному огляді (див. підрозділ 1.6), на сьогоднішній день уже запропоновано певну кількість параметрів, які призначені охарактеризувати прогностичну здатність регресійних рівнянь QSAR/QSPR. Однак досі не було проведено детального дослідження і їх порівняння.

У зв'язку з цим, у представленому розділі дисертації нами проведено ряд чисельних експериментів – розрахунків валідаційних параметрів для модельної функції. Ми вивчали штучні вибірки з похибками, з кількістю точок:  $N = 20, 40$  та  $100$ , що відповідало малим, середнім та великим за розміром вибіркам, які виникають у завданнях QSAR/QSPR. Розглядалися наступні питання, які є дуже загальними для хемоінформатики як наукової дисципліни:

- 1) Як у загальному випадку впливає розмір вхідних даних на поведінку критерію валідації?
- 2) Як збільшення стандартного відхилення розкиду вхідних даних впливає на прогностичні властивості моделі?
- 3) Як змінюються критерії валідації при зміні розміру тестової вибірки?
- 4) Вхідні дані можуть включати похибки вимірювань як в залежній, так і незалежній змінній. Тому ми розглядаємо більш загальний випадок знаходження регресійного рівняння (метод ортогональних відстаней).

Модельні завдання було отримано, виходячи з наступного рівняння:

$$y = 1 + 2x. \quad (3.1)$$

Величина  $x$  лежала в інтервалі від 1 до  $N$  з кроком 1. Де  $N$  – загальна кількість точок. У  $y$  вносилися похибки згідно з нормальним розподілом (функція Гаусу)

із стандартним відхиленням  $sd(y)$ . У кількох розрахункових експериментах, що описано нижче, похибка вносилося також і в незалежну змінну згідно стандартного відхилення, яке позначалось  $sd(x)$ . Для генерації випадкових чисел використовувалась бібліотека **python random**.

### 3.1. Вплив розміру тестової вибірки на критерії валідації

Для цього завдання до вибірки з  $N = 40$  точок, генерованої з рівняння (3.1), до значень властивостей  $y$  вносилися випадкові величини (похибки) за функцією Гауса із стандартним відхиленням  $sd(y) = 10$ . Після цього початкова вибірка з похибками розбивалася на тренувальну (*train*) та тестову (*test*) таким чином, щоб точки тестової вибірки входили до k-NN AD (*Applicability Domain*, див. підрозділ 1.6.4) з кількістю сусідів  $k = 1$ . Кількість таких розбивань досягала  $1.2 \cdot 10^4$  пар вибірок (*train-test*). Моделі в цьому підрозділі розглядались тільки для OLS методу. Вважалося, що результати розрахунків величин коефіцієнтів регресії

$$y = \beta_0 + \beta_1 x \quad (3.2)$$

є кращими, якщо вони ближче до початкових, умовно кажучи – "ідеальних", (3.1). Для цього проводилося дослідження порівняльної статистики. Для кожного інтервалу значень критерію валідації ми отримували середню величину коефіцієнтів регресійного рівняння (3.2) і наводили його на графіку (рис. 3.1). Параметр "part" відповідає частині точок, що було включено в тестовий набір. При цьому для першого інтервалу бралися усі точки із значеннями критерію валідації менше наведеної величини, а для останнього інтервалу – усі точки із значенням критерію більше наведеної величини. Також розраховувалася і наводилися у вигляді інтервалу стандартні відхилення середнього значення параметрів регресії на кожному інтервалі.

У цьому розділі ми використали лише критерії  $R_{\text{test}}^2$  та  $Q_{F3}^2$  (див. підрозділ 1.6). Більш детальне порівняння критеріїв буде наведено в наступних підрозділах.

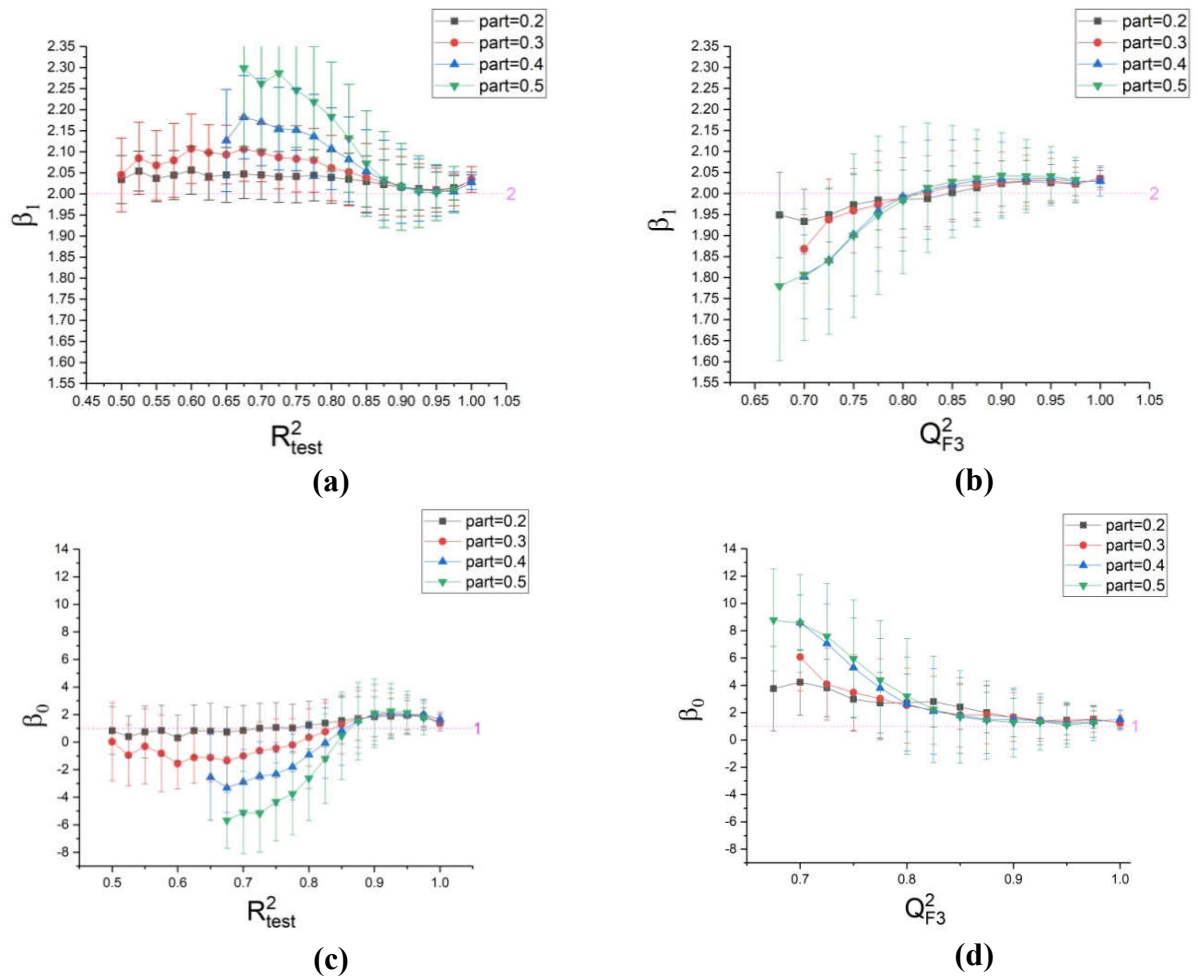


Рис. 3.1 Залежність середніх значень коефіцієнтів регресії  $\beta_1$  та  $\beta_0$  від значень критеріїв зовнішньої валідації при різних долях тестової вибірки (part), ( $N = 40$ ,  $\text{sd}(y) = 10$ ,  $k(\text{NN-AD}) = 1$ )

З представлених на рис. 3.1 залежностей, перш за все, можна бачити, що в цілому для даної задачі обидва критерії валідації ( $R^2_{\text{test}}$  та  $Q^2_{F3}$ ) дійсно покращуються відповідно з покращенням моделі. Так, більші значення критеріїв відповідають ближчим до теоретичних значень коефіцієнтів  $\beta_1$  та  $\beta_0$ . При цьому стандартне відхилення, що відповідає розкиду величин  $\beta_1$  та  $\beta_0$ , також систематично зменшується. Кількість точок тестової вибірки, однак, впливає на інформативність валідаційного критерію. Так, якщо тестова вибірка відносно мала (part = 0.2), хоча середнє значення коефіцієнтів є близьким до теоретичного, виявляється, що критерії валідації практично не спроможні відрізнити добру модель від поганої ( $R^2_{\text{test}} \sim 0.5-1.0$ )! Разом з тим, при надто

великій кількості точок у тестовій вибірці, кількість точок у тренувальній вибірці виявляється не достатньою, щоб отримати якісну модель. Тоді у середньому, значення коефіцієнтів  $\beta_1$  та  $\beta_0$  опиняється досить далекими від теоретичних. Тому надалі в цій роботі ми обирали помірно розбиття з кількістю точок тестової вибірки, що дорівнювало 30% від загальної кількості точок.

### 3.2. Великі вибірки даних ( $N=100$ , $sd(y)=20$ , $sd(x)=0$ )

Надалі ми досліджували розбиття вибірки на тренувальну та тестову двома способами:

- 1) Вхідні тестові точки повинні знаходитися в AD, що визначений виходячи з тренувальної вибірки.
- 2) Розбиття на тестову і тренувальну вибірку проводилось випадковим чином (без урахування AD).

У цьому розділі ми вивчали поведінку критеріїв валідації на основі вибірки з 100 точок, зі стандартним відхиленням тільки в залежній змінній  $y$   $sd(y)=20$ . Кількість розбивань –  $10^5$ . Кількість сусідів з урахуванням AD дорівнювала  $k=1$ .

Аналіз коефіцієнтів валідації ми почнемо з критеріїв внутрішньої валідації. Величина  $Q_{LOO}^2$  досить тривалий час вважалася одним із найбільш важливих критеріїв валідації. Дослідивши залежності цього критерію від  $R_{train}^2$  та  $R_{test}^2$  рис. 3.2, ми дійшли до висновку, що, на жаль, збільшення цього критерію не тільки не гарантує покращення моделі відповідно до критерію  $R_{test}^2$ , а навпаки – може бути в зворотній залежності. Крім того, з рис. 3.2 можна побачити, що для великих вибірок залежність критерію  $Q_{LOO}^2$  від  $R_{train}^2$  опиняється лінійною. Виходячи з цього, можна було б зробити висновок, що чим краще модель описує тренувальну вибірку, тим краще вона повинна описувати й тестову. Але, на жаль, це не так. Із аналізу залежності  $R_{test}^2$  від  $Q_{LOO}^2$  бачимо, що все навпаки – чим краще модель описує тренувальну вибірку, тим гірше вона описує тестову!

Тут і надалі залежності, що зображені чорними точками, відповідають моделям, у яких AD не враховувалася, а червоними – там, де тестові точки входили до AD ( $k=1$ ) отриманого з точок тренувальної вибірки.

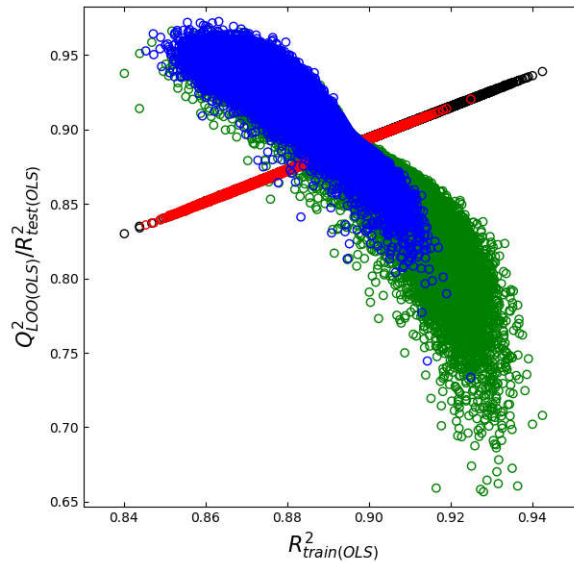


Рис. 3.2 Залежність  $Q^2_{LOO}$   $R^2_{test}$  від  $R^2_{train}$  Чорним точкам відповідають розрахунки без AD, а червоним – розрахунки з AD. Синіми й зеленими точками зображено залежність  $R^2_{test}$  від  $R^2_{train}$  з та без урахування AD відповідно. Усі залежності було отримано в методі OLS ( $N=100$ ,  $sd(y) = 20$ ,  $k(NN-AD) = 1$ )

Також з рис. 3.2 можна бачити, що введення AD призводить до того, що коефіцієнти внутрішньої валідації зосереджуються в області менших значень, а параметри зовнішньої валідації розташовуються в області більш високих значень. Отже AD не дозволяє точкам з великим значенням похибки, що значно відрізняються від інших точок, опинитися в тестувальній вибірці. Таким чином, ці точки обов'язково попадуть у тренувальний набір. Зауважимо, що при урахуванні AD зміни в критеріях внутрішньої валідації є незначними  $\sim 0.1$ . У той же час, критерій зовнішньої валідації змінюється значно сильніше  $\sim 0.3$ .

Для того, щоб ілюструвати розподіл точок серед вибірок, були побудовані відповідні гістограми (рис. 3.3).

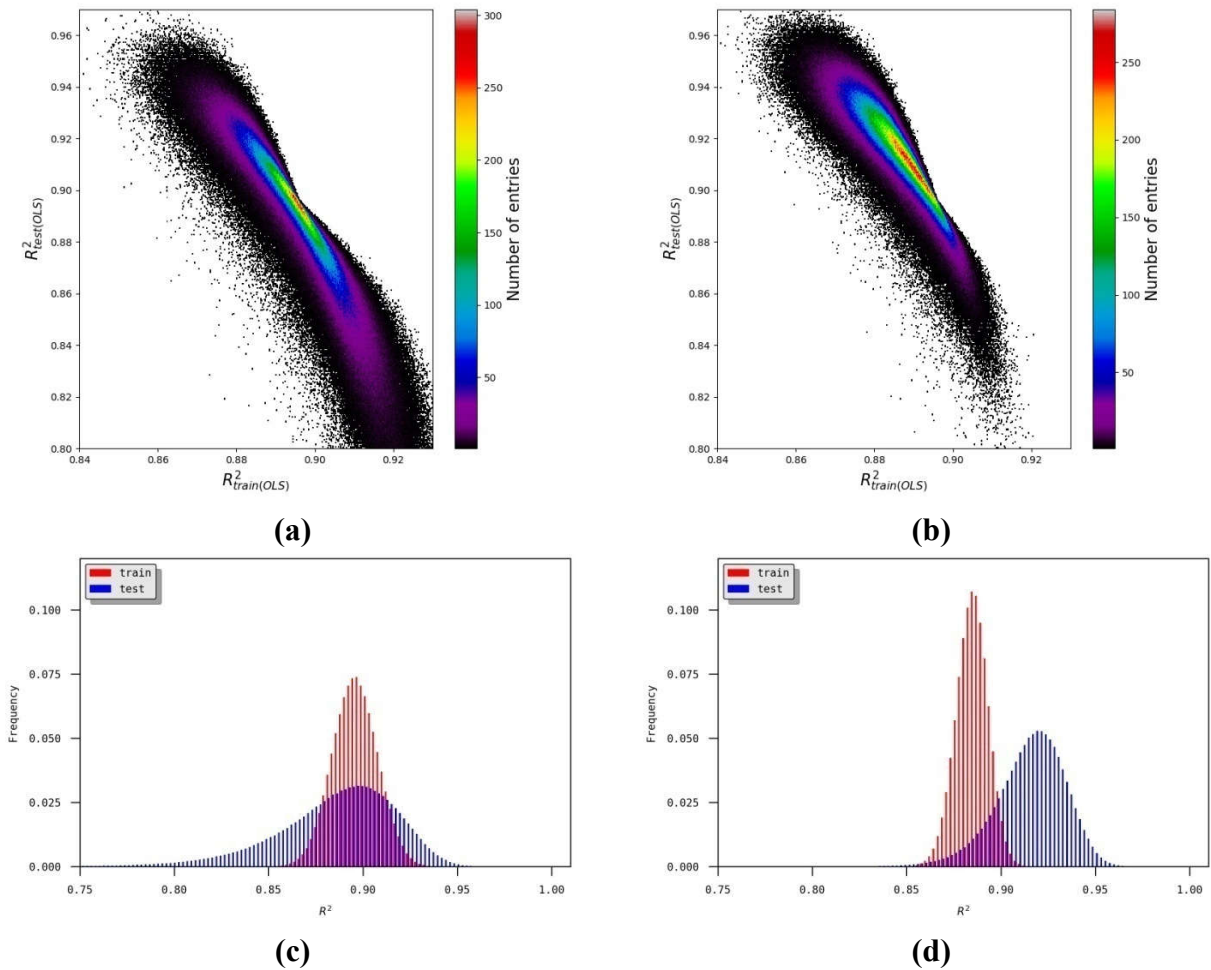


Рис. 3.3 Розподіл  $R^2_{\text{test}}$  від  $R^2_{\text{train}}$  ( $N=100$ ,  $\text{sd}(y)=20$ ,  $\text{sd}(x)=0$ ). (b) та (d) відповідають  $k(\text{NN-AD}) = 1$ , а для (a) та (c) AD не контролювався. "Теплові карти" (a) та (b) ілюструють густину розподілу точок. Розбиття на тестову й тренувальну вибірки проводилося  $10^6$  разів

З наведених діаграм розподілу можна побачити, що основну масу точок сконцентровано у відносно малому діапазоні. Положення цього діапазону залежить від врахування AD. Причому AD "зсуває" максимум у сторону кращих значень критеріїв зовнішньої валідації.

Слід зазначити, що максимумами критеріїв валідації, що співпадають у ситуації, коли не ураховуються AD, зсуваються у різні сторони при урахування AD (рис. 3.3 (c, d)). При цьому значення зовнішнього критерію в обох випадках має значно більший діапазон, ніж критерій внутрішньої валідації. Дещо оптимістичним виявився той факт, що область вищої густини точок

$R^2_{test(OLS)} \approx 0.86-0.92$  відповідає  $R^2_{train(OLS)} \approx 0.88-0.91$ . Таким чином, **якщо ми знаходимося "достатньо близько" до центру густини точок** (червона-жовта-зелена область у "теплових" картах рис. 3.3 (а, b)), внутрішня валідація може давати адекватні оцінки зовнішньої. Згідно наших розрахунків, така область, що гарантує якісну відповідність  $R^2_{test}$  і  $R^2_{train}$ , складає до 30% розбиттів для розрахунків без AD. У розрахунках з AD область високої густини включає 50% точок розбиття *test-train*.

Отже, за допомогою лише критеріїв внутрішньої валідації не можливо зробити висновки про якість прогностичної здатності моделі (не виправляє ситуацію й нещодавно запропонований критерій ПС, що також опинився в наших розрахунках не скорельованим з критеріями зовнішньої валідації). Для цього необхідно використовувати критерії зовнішньої валідації. Втім, не зважаючи на те, що запропоновано доволі таких критеріїв, виявилось, що більшість з них значно корелюють одне з одним (рис. 3.4).

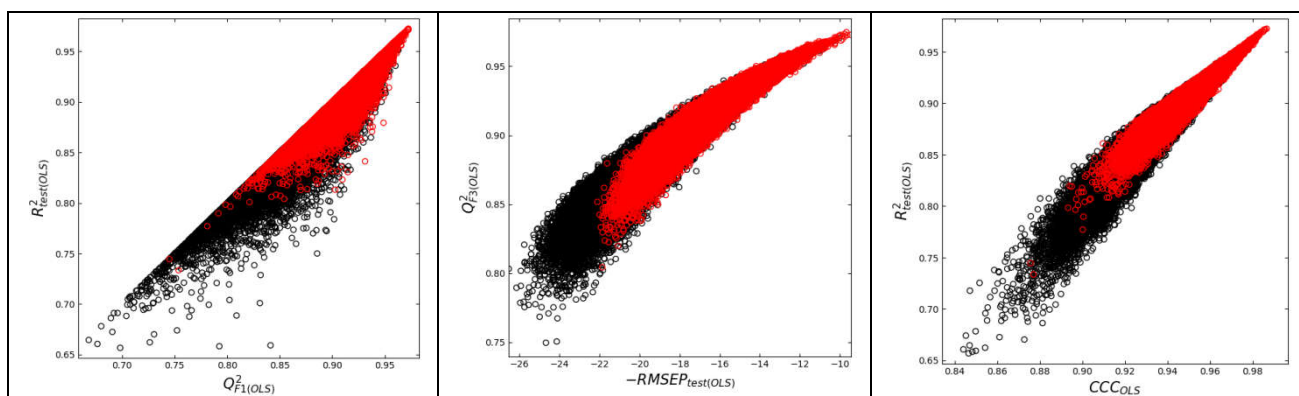


Рис. 3.4 Кореляція критеріїв зовнішньої валідації ( $N=100$ ,  $sd(y)=20$ ,  $sd(x)=0$ ,  $k(\text{NN-AD}) = 1$ )

### 3.3. Великі вибірки даних ( $N=100$ , $sd(y)=20$ , $sd(x)=10$ )

У цьому підрозділі ми ставили за мету з'ясувати як змінюються отримані регресійні рівняння при наявності похибок і в залежній, і в незалежній змінних. Таким чином, у набір даних вносились похибки із стандартними відхиленнями  $sd(y) = 20$ ,  $sd(x) = 10$  за Гаусом. У зв'язку з таким способом внесення похибок, для побудови регресійних рівнянь ми використовували як метод OLS, так і



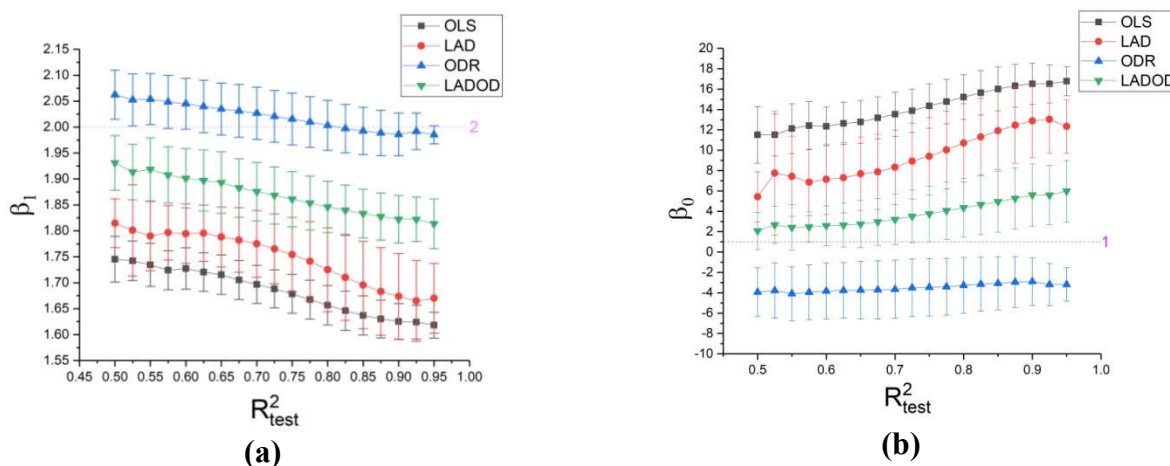
нестандартні методи LAD, ODR, LADOD (див. РОЗДІЛ 1 і РОЗДІЛ 2). Останні два методи відносяться до групи EIV (*errors in variables*), яка призначена для оцінок з узагальненими розподілами похибок між залежними й незалежними змінними.

Введення похибки в  $x$  принципово не змінило форми залежностей, що були отримані в попередньому підрозділі. Порівняння залежностей представлено рис. 3.5.

З наведених діаграм можна бачити, що метод ODR дає найкраще узгодження із теоретичними значеннями  $\beta_1 = 2$  та  $\beta_0 = 0$ . Метод LADOD дає значно ближчі до "ідеальних" величини регресійних коефіцієнтів ніж LAD і далі OLS.

Подивимось тепер як пов'язані розраховані регресійні коефіцієнти з параметрами зовнішньої валідації. Для методу ODR маємо дещо оптимістичну картину. А саме: по мірі зростання  $R_{test}^2$ , величина  $\beta_1$  наближується до точного значення 2 (рис. 3.5, а). Дещо гірша, хоч і якісно вірна, поведінка величини  $\beta_0$  (рис. 3.5, б). Аналогічним чином веде себе й величина  $Q_{F_3}^2$  (рис. 3.5, с, d). Однак величина CCC ніяк не пов'язана із точністю опису параметрів регресії (рис. 3.5, е, f).

Таким чином, серед усіх застосованих методів, найбільш точно відтворив теоретичні коефіцієнти регресії метод ODR, при цьому коефіцієнти зовнішньої валідації проявили помітну узгодженість із середніми значеннями  $\beta$ .



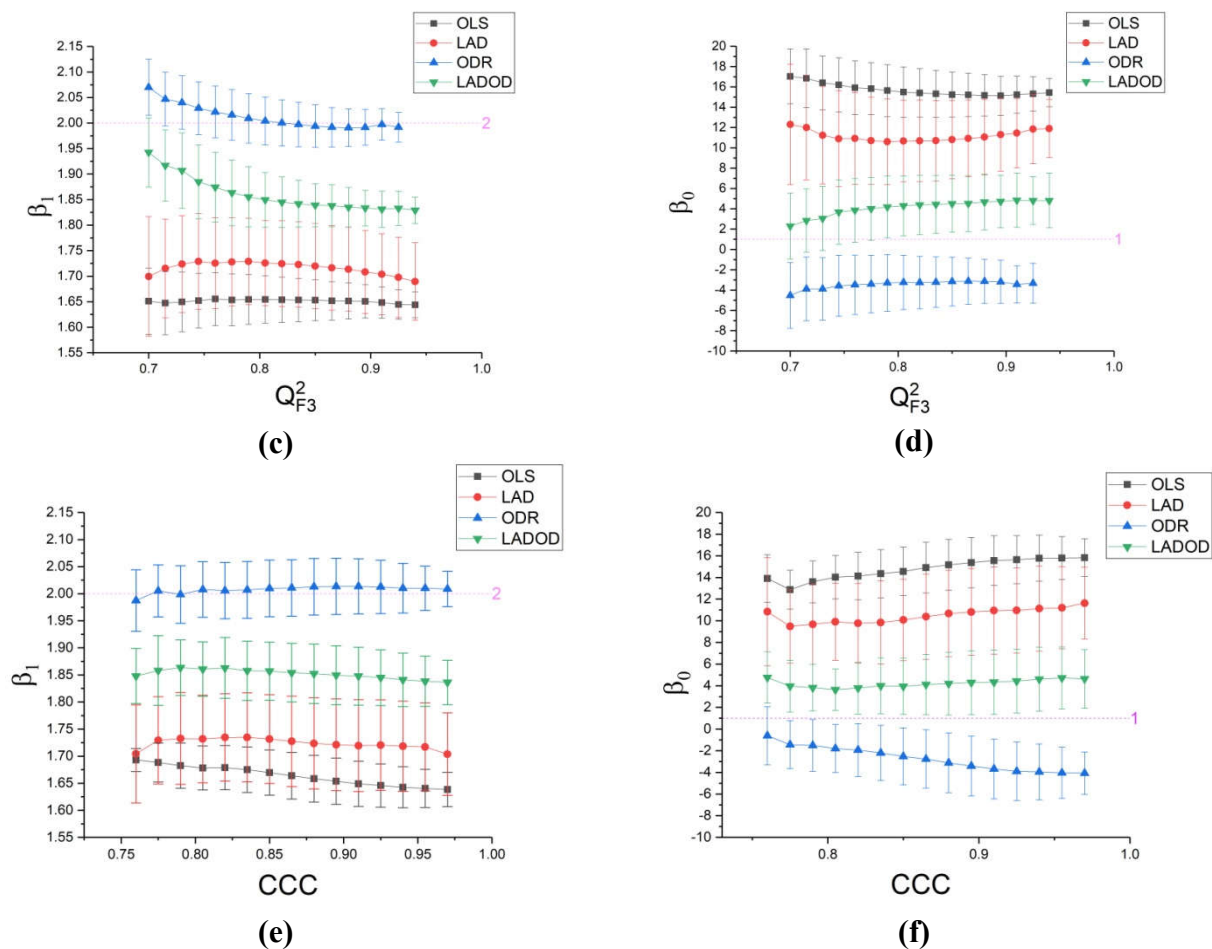


Рис. 3.5 Залежність середніх значень коефіцієнтів рівняння  $\beta_1$  та  $\beta_0$  від значень критеріїв зовнішньої валідації в різних методах ( $N=100$ ,  $sd(y) = 20$ ,  $sd(x) = 10$ ,  $k(NN-AD) = 1$ )

### 3.4. Середні за розміром вибірки ( $N=40$ , $sd(y) = 10$ , $sd(x)=0$ )

Можна побачити (рис. 3.6), що отримані залежності є якісно однакові для вибірок з та без урахування AD.

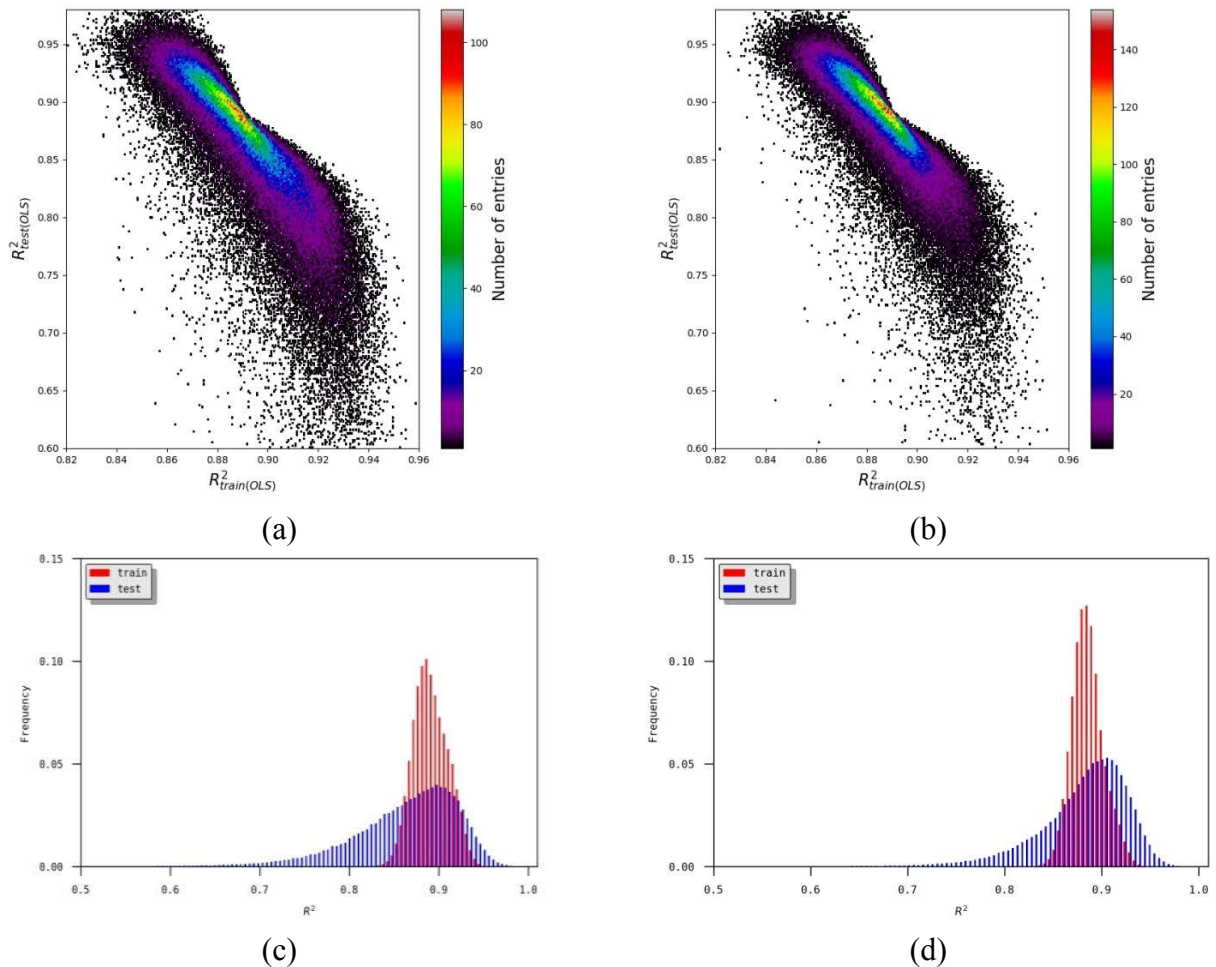
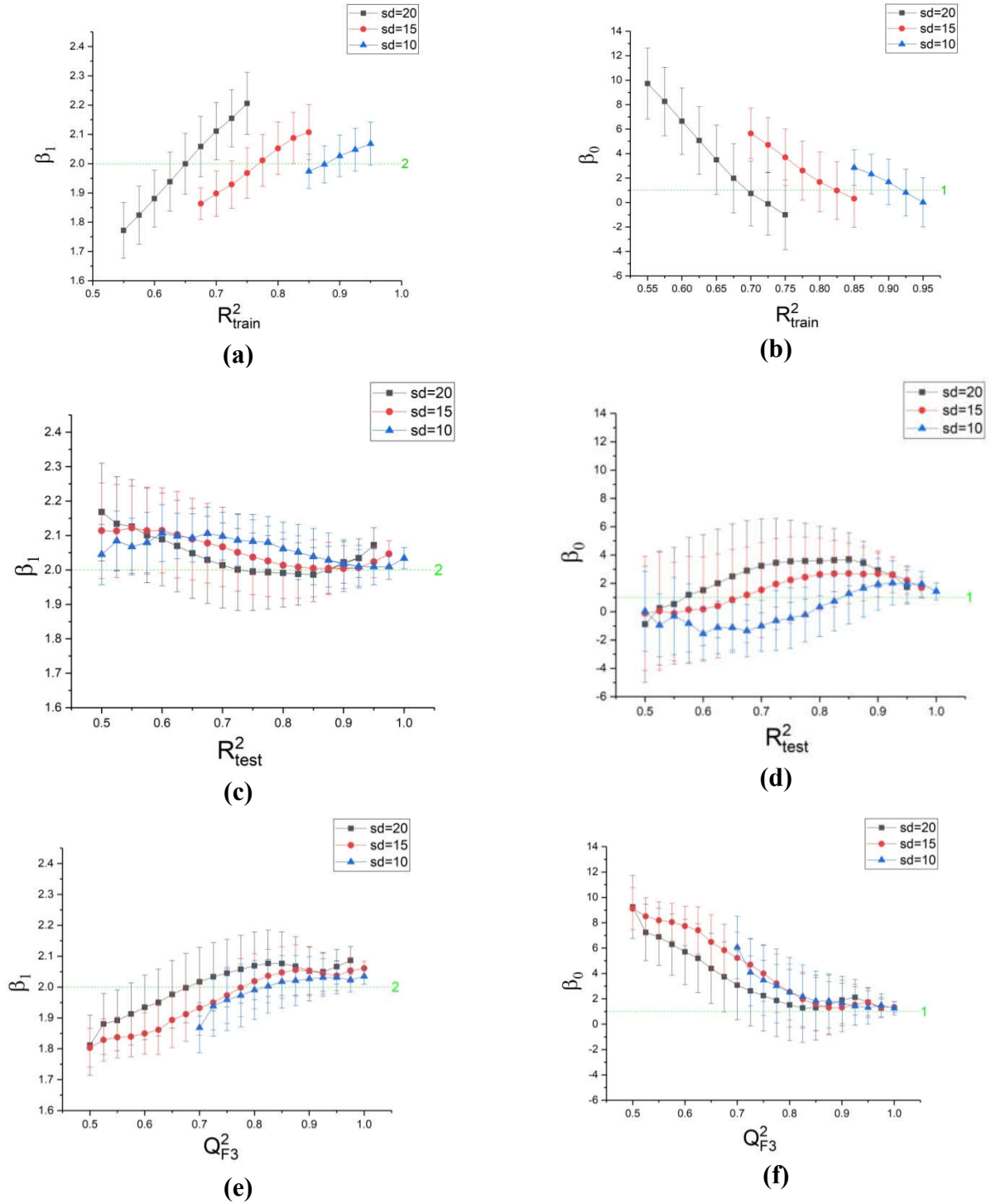


Рис. 3.6 Розподіл  $R^2_{test}$  відносно  $R^2_{train}$  ( $N=40$ ,  $sd(y)=20$ ,  $sd(x)=0$ ). Для **(a)** та **(c)** AD не контролювався, для **(b)** та **(d)**  $k(NN-AD) = 1$ . Теплові карти (a) та (b) ілюструють ділянки з найбільшою популяцією точок. Розбиття на тестову й тренувальну вибірку проводилося  $10^5$  разів

Зауважимо, що на сьогоднішній день існує уявлення (див. наприклад (55)), що треба обирати такі моделі, у яких одночасно і  $R^2_{test}$  і  $R^2_{train}$  мають високі значення. З аналізу рис. 3.6 (a, b та c, d) можна побачити, що одночасно ці дві умови, взагалі кажучи, не виконуються. Однак, знову ж таки, в інтервалі найвищої густини маємо досить близькі значення параметрів зовнішньої й внутрішньої валідацій:  $R^2_{test} \sim 0.85-0.92$ ,  $R^2_{train} \sim 0.87-0.89$ . Згідно наших розрахунків така відповідність гарантується для 30% вибірок без урахування AD і 50% вибірок з  $k(NN-AD) = 1$ .

Для більш детального розуміння цих обставин, ми розглянули зв'язок коефіцієнті рівняння (3.2) з індексами внутрішньої й зовнішньої валідації при різних значеннях  $sd(y)$  (рис. 3.7).



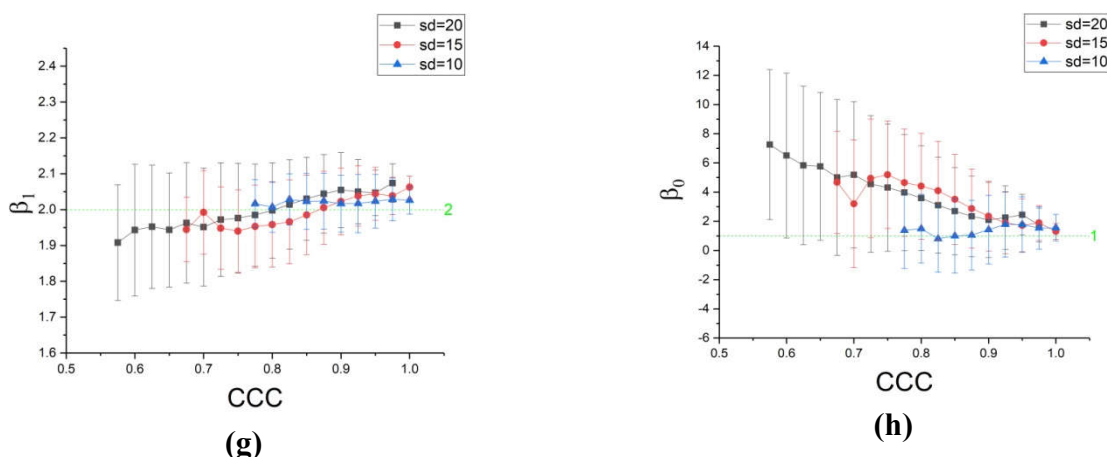


Рис. 3.7 Залежність середніх значень коефіцієнтів рівняння (3.2)  $\beta_1$  та  $\beta_0$  отриманого в OLS від значень критеріїв внутрішньої й зовнішньої валідації при різних значеннях  $sd(y)$

Можна побачити (рис. 3.7 (a, b)), що зростання параметру внутрішньої валідації ( $R^2_{train}$ ) не пов'язано з покращенням коефіцієнтів регресії. Однак для невеликої величини  $sd(y) = 10$  розраховані коефіцієнти регресії все ж таки розташовані ближче до "ідеального" значення. У цілому, виходячи з представлених графічних даних, можна стверджувати, що критерій внутрішньої валідації погано характеризує якість моделі.

У протилежність цьому, збільшення зовнішніх критеріїв валідації  $R^2_{test}$  та  $Q^2_{F3}$  у ситуаціях, коли похибка в  $x$  відсутня, тенденційно пов'язано з покращенням значень коефіцієнтів регресії OLS.

Загалом, спираючись на рис. 3.2(a), 3.3(a, b), 3.6(a, b), зазначимо, що кількість точок (тут точки це розбиття на тестову і навчаючу вибірки) з великими значеннями зовнішніх критеріїв валідації досить мала.

Порівнюючи залежності, які отримано в цих експериментах, для регресій, побудованих різними методами, бачимо (рис. 3.8), що значення коефіцієнтів OLS та LAD ближче до точних, ніж у методах ODR та LADOD. І це не дивно, оскільки похибка не вносила у незалежну змінну  $sd(x) = 0$ .

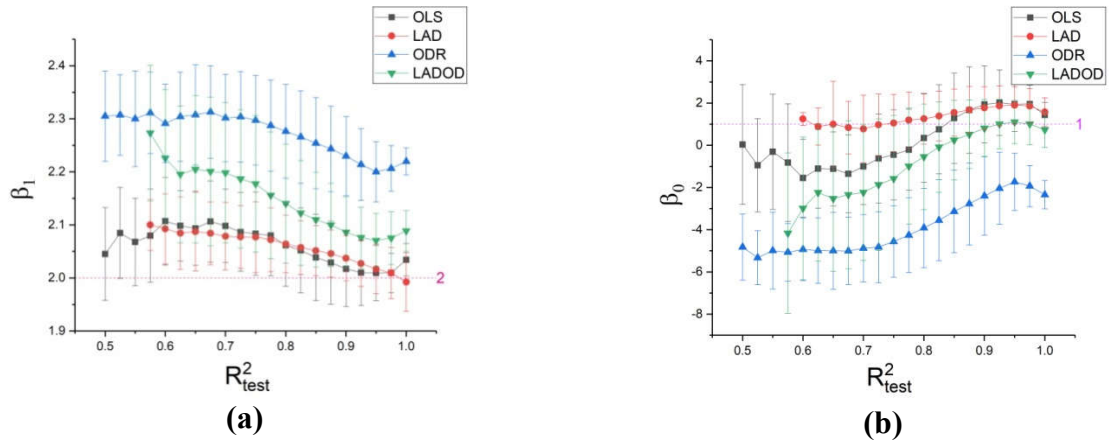
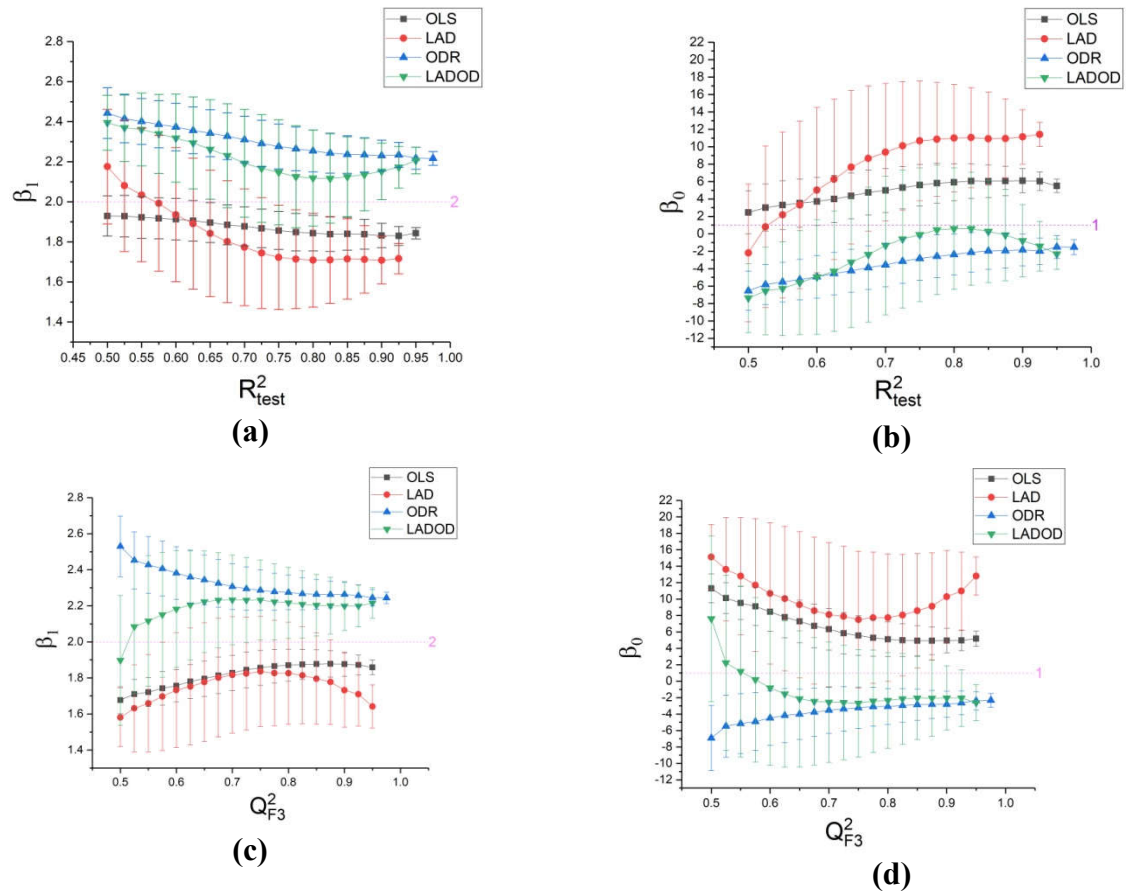


Рис. 3.8 Залежність середніх значень коефіцієнтів  $\beta_1$  та  $\beta_0$  від значень критеріїв зовнішньої валідації в різних методах ( $N = 40$ ,  $sd(y) = 10$ ,  $k(\text{NN-AD}) = 1$ )

Однак введення навіть невеликої похибки в залежну змінну  $sd(x) = 5$  може значно ускладнити ситуацію. У такому випадку вже не можна стверджувати, що методи OLS або LAD заздалегідь призводять до кращих результатів, ніж методи ODR та LADOD (рис. 3.9).





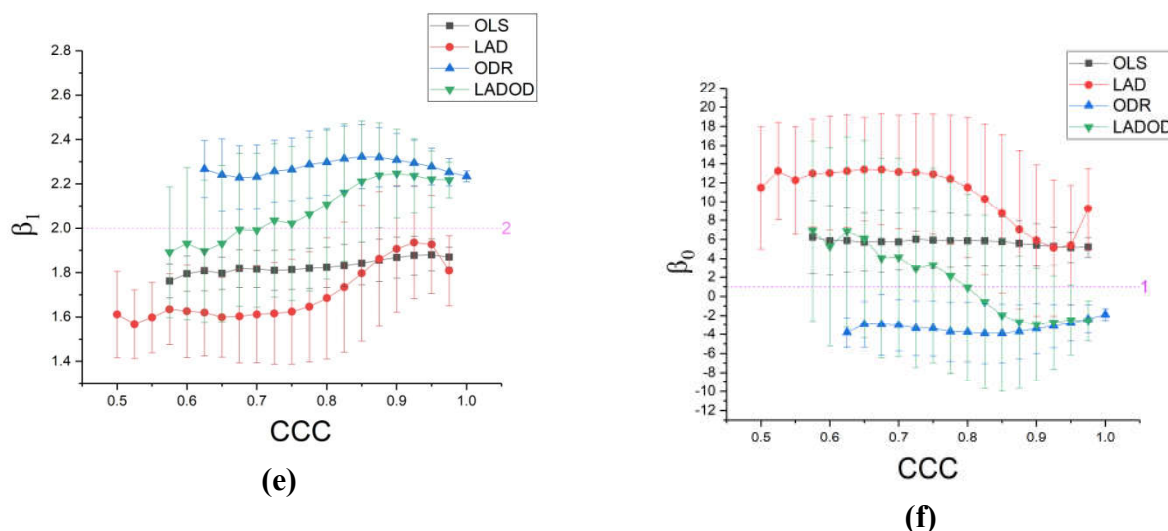


Рис. 3.9 Залежність середніх значень коефіцієнтів  $\beta_1$  та  $\beta_0$  від значень критеріїв зовнішньої валідації в різних методах ( $N=40$ ,  $sd(y) = 10$ ,  $sd(x) = 5$ ,  $k(\text{NN-AD}) = 1$ )

### 3.5. Малі за розміром вибірки ( $N=20$ , $sd(y)=5$ , $sd(x)=2.5$ )

Результати для малих вибірок, що ми вивчали, не відрізнялись суттєво від тих, що були отримані для великих та середніх вибірок. Залежність  $R_{train}^2$  від  $Q_{LOO}^2$  навіть для такої кількості точок все ще була близька до лінійної. Однак слід зауважити, що в малих вибірках  $R_{test}^2$  та  $Q_{LOO}^2$  за абсолютною величиною значно різняться. Залежність  $R_{test}^2$  від  $Q_{LOO}^2$  представлено на рис. 3.10.

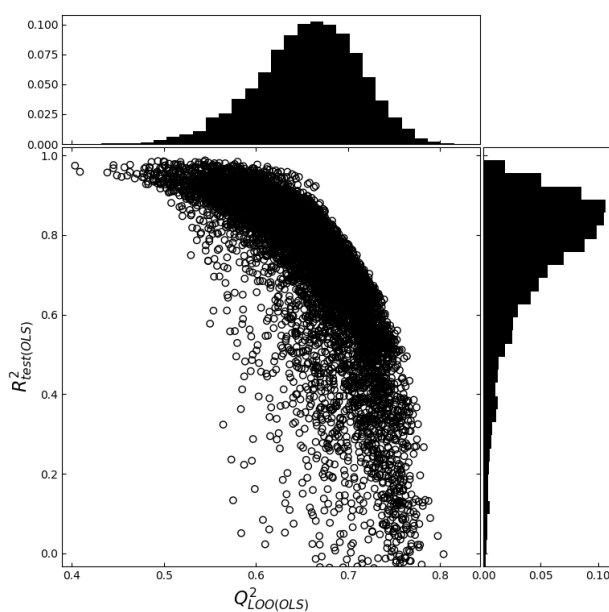


Рис. 3.10 Залежність  $R_{test}^2$  від  $Q_{LOO}^2$  для методу OLS. ( $N=20$ ,  $sd(y) = 5$ ,  $sd(x) = 2.5$ ,  $k(\text{NN-AD}) = 1$ )

На цьому рисунку представлено також гістограми розподілу величин  $R_{test}^2$  та  $Q_{LOO}^2$ . Знову відзначимо нелінійну й зворотну за тенденцією залежність  $R_{test}^2$  від  $Q_{LOO}^2$ . Однак, на відміну від великих та середніх за розміром вибірок, в області максимальної густини  $R_{test}^2 \sim 0.8$ , тоді як  $Q_{LOO}^2 \sim 0.65$ . Отже, співвідношення цих величин не можуть бути опорними для оцінок регресійної моделі побудованої на малих вибірках із відносно великим розкидом.

Цікаво, що залежність коефіцієнтів регресійних рівнянь від  $R_{test}^2$  (рис. 3.11) для OLS та LAD виявляються прийнятними. Самі коефіцієнти в середньому "достатньо" близькі до "ідеальних" значень. Разом із тим виявилось, що моделі ODR та LADOD, на відміну від даних для великих та малих вибірок, демонструють помітно гірші результати, що не узгоджується з концепцією EIV! Табл. 3.1. дає ілюстрацію цієї обставини. Тут можна побачити, що OLS дає величини коефіцієнтів ближче до "ідеальних" значень. Також помітно великою є величина  $R_{train}^2$  для OLS. Втім, величина  $Q_{LOO}^2$  для методу LADOD є найкращою з усіх представлених.

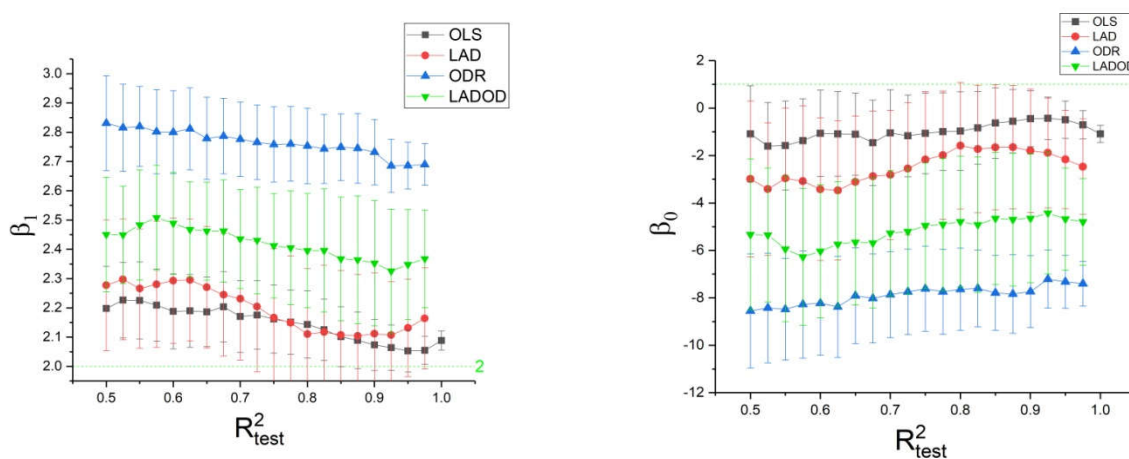


Рис. 3.11 Залежність середніх значень коефіцієнтів  $\beta_1$  та  $\beta_0$  від значень  $R_{test}^2$  у різних методах ( $N=20$ ,  $sd(y)=5$ ,  $sd(x)=2.5$ ,  $k(NN-AD) = 1$ )



Таблиця 3.1

**Результати регресійного аналізу методами OLS, LAD, ODR та LADOD для повної вибірки (N=20, sd(y)=5, sd(x)=2.5) без розділення на тестову й тренувальну**

метод	$\beta_0$	$\beta_1$	$R^2_{train}$	$Q^2_{LOO}$
OLS	-1.2541	2.1111	0.7682	0.7191
LAD	-1.1625	2.0543	0.7659	0.7119
ODR	-7.2217	2.6515	0.7179	0.6467
LADOD	-5.5569	2.4445	0.7467	0.7376

### Висновки до розділу 3

1. Раціональне розбиття початкової вибірки на тестову й тренувальну (навчаючу) є необхідним моментом в оцінках якості отриманих регресійних моделей. Випадковим чином обране одиничне розбиття (це, на жаль, є типовим прийомом сучасних QSAR досліджень) може вести або до недооцінювання якості регресійної моделі, або, що гірше, до переоцінки. У розділі 3 показано, що для якісної характеристики моделі необхідно вивчати розподіл точок у великій кількості розбивань *train-test*. Маємо зазначити, що ця процедура на сучасному рівні комп'ютерної техніки не представляється надто "важкою".

2. Покращення (збільшення) коефіцієнтів внутрішньої валідації ( $R^2_{train}$ ,  $Q^2_{LOO}$ ), взагалі кажучи, не є свідомством покращення прогностичної властивості моделі. Для вхідних даних із вираженим розкидом, типовою картиною є зворотна (суттєво нелінійна) залежність  $R^2_{train} - R^2_{test}$ . Для адекватної оцінки регресійної моделі, необхідне дослідження поведінки параметрів внутрішньої й зовнішньої валідації в області значної густини точок розбиття.

3. Урахування *Applicability Domain* (AD) є необхідним моментом дослідження лінійної моделі. У наших дослідженнях AD зміщує розподіли зовнішніх критеріїв валідації відносно внутрішніх у сторону збільшення.

4. Нещодавно запропоновані зовнішні критерії CCC та ІС у наших дослідженнях, на жаль, виявилися неінформативними. Можливості цих параметрів потребують подальшого дослідження.

5. У розділі представлено дані щодо валідації лінійних залежностей, отриманих методами OLS, LAD, ODR і LADOD. Встановлено, що OLS виявився найточнішим у більшості випадків, а саме: у ситуації, коли велика похибка  $sd(y)$  та  $sd(x)=0$ . Для великих вибірок, коли  $sd(x) \neq 0$ , метод ODR дав результати найближчі до "ідеальних" величин. Метод LADOD, який також відповідає концепції EIV, дав відносно близькі до "ідеальних" величини регресійних коефіцієнтів. Загалом, поведінка OLS і LAD виявилась подібною, хоча OLS помітно "перемагає" у точності опису лінійних залежностей. З представлених графічних даних можна вивести, що параметри LAD якісно співвідносяться до OLS, як LADOD до ODR. Цікаво однак, що в малих вибірках навіть у присутності похибок у незалежній змінній ( $sd(x) \neq 0$ ), метод OLS дав найкращі результати. Ця обставина потребує подальших ретельних досліджень.

6. Спираючись на представлені дані, можна вивести, що найбільшою проблемою є валідація рівнянь, які отримані виходячи з малих вибірок. Вочевидь, у такому разі величини  $Q_{LoO}^2$  можуть бути корисними.

7. У результаті проведення тестових досліджень на скриптовій мові **Python3**, створено програмний пакет **valid\_regression**, який увійшов як складова частина в комплекс програм **QUASAR**, що розробляється на кафедрі хімічного матеріалознавства Харківського Національного Університету імені В. Н. Каразіна.

Основні положення цього розділу викладено в публікаціях автора (117-118)

## РОЗДІЛ 4

### ***L*<sub>1</sub>-РЕГУЛЯРИЗАЦІЯ В ПОБУДОВІ КЛАСИФІКАЦІЙНИХ МОДЕЛЕЙ**

#### **4.1. Загальна проблема побудови класифікаційних функцій**

Побудова класифікаційних моделей, зокрема моделей, що дозволяють провести бінарний відбір ("так/ні", "активний/неактивний" тощо), є важливим підходом у методології QSAR / QSPR. Затребуваність у таких моделях пов'язана з кількома обставинами. **По-перше**, існує досить широка область даних, які погано "оцифровані". Тобто активність систем представлено лише на якісному рівні "активний/неактивний". Багато подібної інформації пов'язано з дослідженнями токсичності, канцерогенної активності тощо. Складність у встановленні кількісної оцінки таких характеристик призводить до представлення експериментальних даних лише на якісному рівні (наприклад: канцероген/не канцероген<sup>V</sup>). **По-друге**, доволі часто зустрічається ситуація, коли дані щодо активності/властивості представлено для систем за різних термодинамічних умов (температура, тиск), або для надто різних (у структурному відношенні) систем. У таких випадках побудова регресійних моделей виявляється неможливою. У якості прикладу, згадаємо дані щодо активності естрадіолів<sup>119,120</sup>. У цих роботах дані (*Relative Binding Affinity*, RBA) представлено для дуже різноманітних структур. До того ж умови проведення експерименту (температура, час експозиції, рецептори) для різних груп систем були різні. У такому разі отримання єдиної регресійної моделі стає неможливим, але можливо побудувати єдину класифікаційну (дискримінаційну) функцію, яка здатна виділити системи з активністю вище певної величини ( $\log RBA > \log RBA_{th}$ , див. (121)), яку завжди можливо обрати, виходячи з біохімічних міркувань.

---

<sup>V</sup> Слід зауважити, що хоча канцерогенна активність і може бути описана індексом Айболла (Iball) (122) все ж, найчастіше, використовується розділення систем на два, або три чи чотири класи відносно "сили" канцерогену.

**По-третє,** проблема направленого відбору сполук із заданими властивостями більшою мірою потребує інформації щодо систем із високою активністю (або заданою властивістю), ніж власне континуальної моделі, яка описує зв'язок будови молекули з активністю/властивістю. Отже, необхідна математична класифікаційна модель, за допомогою якої можливий відсів малоактивних чи, з іншого боку, відбір високоактивних систем. Такий відсів/відбір (скринінг) є ключовим моментом сучасних досліджень "структура-властивість".

Важливою обставиною є те, що з розвитком високопродуктивного скринінгу, почали з'являтися спеціалізовані бази даних біологічних (та інших) активностей / властивостей сполук. Наразі існує багато як безкоштовних, так і комерційно доступних баз даних, що містять різноманітні властивості хімічних сполук. Задачею QSAR/QSPR є розробка таких методів і моделей, що змогли б якісно використати ці дані. При цьому вважається, що в якості дескрипторів для таких моделей слід використовувати величини, які можна було б достатньо легко розрахувати для довільної молекули. Зазвичай це різноманітні структурні дескриптори. Певна кількість сучасних комп'ютерних програм дозволяє розрахувати ці дескриптори<sup>100-102</sup>.

При цьому слід мати на увазі, що в базах даних може міститись інформація про значну кількість молекул: тисячі, десятки тисяч або навіть значно більше. Для них може бути розрахована величезна кількість структурних дескрипторів (біля 7000 у сучасних комп'ютерних програмах). Вказані обставини характеризують задачу обробки інформації в базах даних, виділення ключової інформації та формулювання класифікаційних правил як досить складну в алгоритмічному сенсі. Отже, необхідна розробка таких методів, які б ефективно працювали з великими наборами даних. Крім того, бажано, щоб класифікаційне правило (класифікаційна функція) були достатньо прості для структурно-хімічної інтерпретації.

Для розв'язання задач класифікації на сьогоднішній день існує досить широкий набір методів. Перш за все, слід згадати метод дискримінаційної

(дискримінантної) функції, який було запропоновано Фішером ще у 1936 році та тестовано на задачі класифікації ірисів<sup>123</sup>. Значний прогрес у побудові класифікаційних функцій було досягнуто при використанні методу логістичної регресії (*Logistic Regression*, LR)<sup>124,125</sup>. В останні роки певну популярність набрали: метод опорних векторів (*Support Vector Machine*, SVM)<sup>126,127</sup>, метод випадкових лісів (*Random Forests*, RF)<sup>5,6</sup> і, звісно, метод штучних нейронних мереж (зазвичай використовують конволюційні – згорткові мережі)<sup>128</sup>.

У той час, як передбачувальна здатність вищезгаданих методів для багатьох задач може бути велика, ці методи слід інтерпретувати як методи "чорної скриньки"<sup>128,129</sup>. Це означає, що 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. І хоч методи ANN, для яких можлива інтерпретація, уже були запропоновані<sup>128</sup>, усе ще існує необхідність у розвитку методів, що можуть відповісти на питання: "Чому саме ця молекула є активною або неактивною?", "Які структурні параметри є найважливішими?".

У цьому розділі ми використовували метод LR на основі відбору предикторів методом LARS-LASSO (LR-LARS-LASSO)<sup>130</sup>. Цей метод обіцяв бути розрахунково-ефективним, а також давати компактні регресійні рівняння, що містять відносно невелику (бажано мінімальну) кількість дескрипторів. Аналіз відповідних коефіцієнтів LR дає розуміння як саме: позитивно чи негативно, той чи інший дескриптор впливає на величину (інтенсивність) активності.

У цьому розділі методом  $L_1$ -регуляризації логістичної функції (у варіанті LR-LARS-LASSO) ми розраховували компактні класифікаційні рівняння, які далі були зіставлені із моделями, що отримані з використанням ефективного методу RF (доступного в *sci-kit learn*<sup>131</sup>), а також з методом дерева класифікацій<sup>132</sup> і методом k-найближчих сусідів. Останній підхід (метод k-NN) доволі часто

згадується в літературі<sup>23,133</sup> як достатньо простий метод класифікації. Відзначимо, однак, що метод дерев класифікації (попередник методу RF) також є методом, який спроможний давати рівняння, що можуть бути інтерпретовані.

#### 4.2. Алгоритм розрахунку LR-LARS-LASSO

Процедура знаходження параметрів логістичної регресії, як і метод OLS, також може бути  $L_1$ -регуляризована. Для цього, як і раніше, задачу (1.34) розв'язують з обмеженням на коефіцієнти регресії  $\|\beta\|_1 < \xi$ . Для отримання регуляризованих вирішень звернемо увагу на те, що розв'язок рівняння (1.40) відповідає вирішенню зваженого методу OLS. Так, якщо зробити заміну:

$$\hat{X} = W^{\frac{1}{2}}X; \quad \hat{Y} = W^{\frac{1}{2}}Z, \quad (4.1)$$

то рівняння (1.40) приймає наступний вигляд:

$$\beta^{\text{new}} = (\hat{X}^+ \hat{X})^{-1} \hat{X} \hat{Y}, \quad (4.2)$$

що в точності відповідає вирішенню рівнянню OLS (1.5)! Таким чином, для того, щоб отримати рішення  $L_1$ -регуляризованої задачі LR, необхідно розв'язувати задачу (4.2) з урахуванням регуляризації методом, яким до цього розв'язувалась задача  $L_1$ -регуляризації OLS (LASSO, див. розділ 1.3). При цьому єдиною принциповою відмінністю є той факт, що рівняння (1.40) необхідно розв'язувати кілька разів до досягнення самоузгодженості рішень (тобто  $\beta$  перестає значно змінюватися на сусідніх ітераціях), оскільки як параметр  $z$ , так і параметр  $W$  залежать від  $\beta$ .

У представленій роботі ми знаходили  $L_1$ -регуляризований розв'язок LR розв'язуючи проблему (1.40) із застосуванням LARS-LASSO алгоритму на кожній ітерації<sup>33</sup>. Алгоритм LARS-LASSO представлено в підрозділі 2.2. У нашій роботі замість обмеження  $\|\beta\|_1 < \xi$  ми обирали певну кількість дескрипторів на кожній ітерації. Задача (1.40) розв'язувалась із використанням LARS-LASSO алгоритму з оновленими на кожній ітерації матрицями  $W$  та  $Z$ . У якості критерію зупинки виступала різниця  $\|y - p\|_2$ . Також різниця значень

функції (1.35), що мінімізується, на кожній ітерації має бути меншою ніж  $\varepsilon = 10^{-8}$ .

Таким чином, алгоритм, який ми будемо називати LR-LARS-LASSO, має наступний вигляд:

- 1) Для матриці  $\mathbf{X}$  та вектору властивостей ( $\mathbf{y}$ ) у якості стартового наближення покласти вектор  $\boldsymbol{\beta} = \mathbf{0}$ , а також  $\gamma^{\text{old}} = \mathbf{0}$ ;  $\ell^{\text{old}} = \mathbf{0}$ .
- 2) Виходячи з рівнянь (1.34) й (1.39), розрахувати  $p_i$  і діагональну матрицю  $\mathbf{W}$  відповідно. Розрахувати  $\gamma^{\text{new}} = \|\mathbf{y} - \mathbf{p}\|_2$ , а також  $\ell^{\text{new}}$  з рівняння (1.35).
- 3) Якщо  $|\gamma^{\text{new}} - \gamma^{\text{old}}| < \varepsilon$  та  $|\ell^{\text{new}} - \ell^{\text{old}}| < \varepsilon$  **STOP**; інакше покласти  $\ell^{\text{old}} = \ell^{\text{new}}$ ,  $\gamma^{\text{old}} = \gamma^{\text{new}}$ .
- 4) З рівняння (1.41) розрахувати вектор  $\mathbf{z}$ .
- 5) Виходячи з (4.1), розрахувати  $\hat{\mathbf{X}}$  та  $\hat{\mathbf{Y}}$ .
- 6) Знайти  $\boldsymbol{\beta}$ , розв'язав (4.2) з використанням алгоритму LARS-LASSO (див. підрозділ 2.2)
- 7) Перейти до кроку 2

Більш детально опис алгоритму LR-LARS-LASSO, який було використано в даний роботі, можна знайти в (130). Алгоритм LR-LARS-LASSO було реалізовано нами на мові програмування FORTRAN.

### 4.3. Тестові набори даних

Для проведення тестових розрахунків було використано два набори даних. У першому наборі, ми класифікували молекули у відповідності з їх експериментальними значеннями основності в газовій фазі до  $\text{Li}^+$  з високим та низьким значенням.

Літій, а також його солі, мають багато технологічних використань. Їх використовують у літій-іонних акумуляторах<sup>134-137</sup>, у каталізі органічних реакцій<sup>138-140</sup>, у дегідруванні “сполук-сховищ водню”<sup>141-144</sup>, у мас-спектрометрії іонного приєднання (*Ion Attachment Mass Spectrometry*, IAMS)<sup>145-147</sup> і т. д. У

зв'язку з цим ми вважаємо, що розробка класифікаційних моделей для сполук основних до  $\text{Li}^+$  могла би допомогти розумінню тих структурних особливостей органічних сполук, що дадуть змогу оптимізувати застосування сполук літію в технологічних задачах.

У якості характеристики спорідненості молекул основ (що позначалися як В) до  $\text{Li}^+$  іонів у газовій фазі, ми використовували літій катіон основність (*Lithium Cation Basicity*, LiCB), що визначається з константи рівноваги реакції (або з енергії Гіббсу реакції):



наступним чином:

$$\begin{aligned} \Delta_r G &= -RT \ln K; \\ \text{LiCB} &= -\Delta_r G \end{aligned} \quad (4.4)$$

Набір органічних сполук було взято з (83). Структури хімічних сполук наведено в додатку Д, найбільш типові сполуки цієї вибірки наведено на рис. 4.1. Сполуки із значенням основності до літій-катіону, що вища ніж середня за вибіркою (145.58 кДж/моль), вважалися активними, а з нижчою – неактивними. Таким чином, із загальної вибірки 113 сполук вважалися активними, а 115 – неактивними.

Друга вибірка складалась із синтетичних (стероїдних та нестероїдних) аналогів естрогену (жіночий половий гормон). В організмі жінок естрогени виконують найрізноманітніші біологічні функції. Порушення функціонування цих гормонів призводить до великої кількості захворювань, у тому числі: порушення репродуктивних функцій<sup>148,149</sup>, раку молочної залози<sup>150,151</sup>, розвитку остеопорозу<sup>152,153</sup> та інших<sup>154</sup>. Розробка моделей для таких сполук може пришвидшити дизайн лікарських сполук, що використовуються в терапії методом заміщення естрогенів синтетичними аналогами<sup>155</sup>, або ж передбачити можливі активні речовини, що можуть забруднювати середу й спричиняти шкоду як людству, так і тваринному світу<sup>156</sup>.



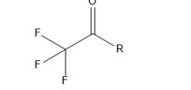
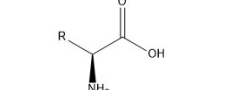
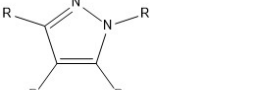
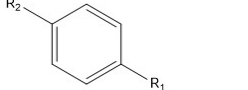
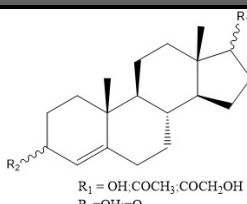
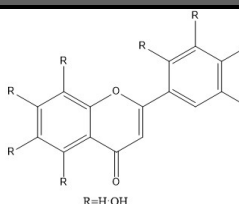
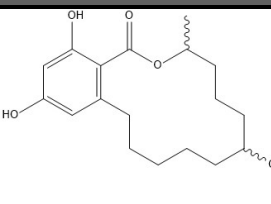
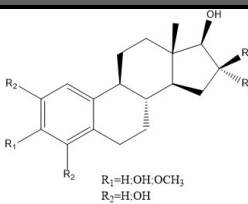
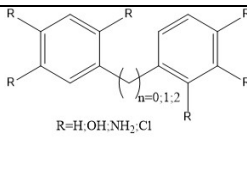
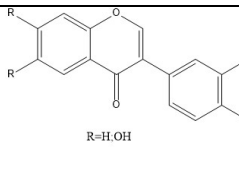
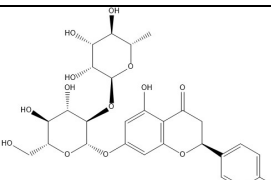
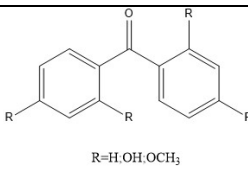
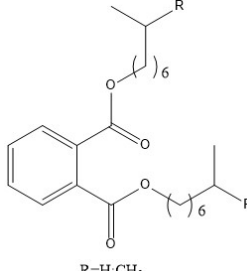
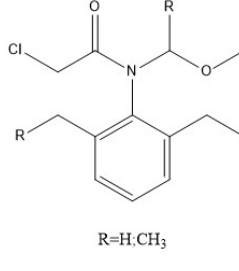
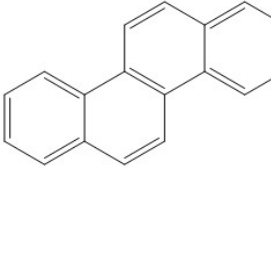
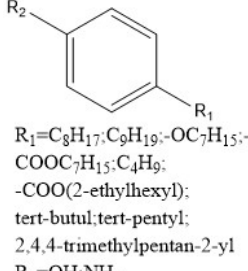
Основ- ність до $\text{Li}^+$	 $\text{R}=\text{N}(\text{CH}_3)\text{CH}_3, \text{OCH}_3, \text{OCH}_2\text{CH}_3; \dots$	 $\text{R}=\text{H}, \text{CH}(\text{CH}_3)\text{CH}_3, \text{CH}_2\text{Ph}, \text{CH}_2\text{CH}_2\text{SCH}_3$	 $\text{R}=\text{H}, \text{CH}_3$	 $\text{R}_1=\text{CH}_2\text{CH}_3, \text{S}(=\text{O})\text{CH}_3; \text{OCH}_3; \text{C}=\text{O}; \dots$ $\text{R}_2=\text{H}, \text{CH}_3$
DSSTox набір	 $\text{R}_1 = \text{OH}, \text{COCH}_3, \text{COCH}_2\text{OH}$ $\text{R}_2 = \text{OH}, =\text{O}$	 $\text{R}=\text{H}, \text{OH}$		 $\text{R}_1=\text{H}, \text{OH}, \text{OCH}_3$ $\text{R}_2=\text{H}, \text{OH}$ $\text{R}_3=\text{H}, \text{CH}_3$
	 $\text{R}=\text{H}, \text{OH}, \text{NH}_2, \text{Cl}$ $n=0,1,2$	 $\text{R}=\text{H}, \text{OH}$		 $\text{R}=\text{H}, \text{OH}, \text{OCH}_3$
	 $\text{R}=\text{H}, \text{CH}_3$	 $\text{R}=\text{H}, \text{CH}_3$		 $\text{R}_1=\text{C}_8\text{H}_{17}, \text{C}_9\text{H}_{19}, -\text{OC}_7\text{H}_{15}, -\text{COOC}_7\text{H}_{15}, \text{C}_4\text{H}_9; -\text{COO}(2\text{-ethylhexyl}); \text{tert-butyl}, \text{tert-pentyl}; 2,4,4\text{-trimethylpentan-2-yl}$ $\text{R}_2=\text{OH}, \text{NH}_2$

Рис. 4.1 Деякі типові сполуки з наборів молекул. Замісники позначалися як  $\text{R}_x$ , де  $\text{R}_1$  – тип замісника, що міг незалежно зустрітись в зазначених позиціях

У нашій роботі вивчалася відносна спорідненість зв'язування молекул до рецепторів естрогену ( $\log\text{RBA}$ ). Методика визначення і дослідження цієї величини для використаної вибірки наведені в (157, 158). Вибірка даних DSSTox (NCTRER, *National Center for Toxicological Research Estrogen Receptor Binding Database*) було знайдено у відкритому доступі на інтернет-ресурсі Pubchem<sup>159</sup>. У цьому наборі з 216 молекул, 128 було класифіковано як активні, а залишок – 88 як неактивні, відповідно до (159). Молекули вибірки складались з багатьох класів хімічних сполук, а саме: стероїдних сполук, фітоестрогенів, похідних дифенілметану, похідних дифенілу, похідних фенолу тощо. Найбільш типові представники цієї вибірки наведено на рис. 4.1.

Геометрію молекул у цьому розділі було трансформовано в 3D та оптимізовано з використанням методу MMFF94 силового поля, доступного в

бібліотеці **rdkit** (160) для мови програмування Python. Після чого було розраховано набір дескрипторів з використанням програми PaDEL-descriptor<sup>102</sup>. Для вивчення зв'язування до рецепторів естрогенів використовувалися як 2D, так і 3D дескриптори. Для вивчення основності органічних сполук до катіону Li використовувалися лише 2D дескриптори, оскільки більшість молекул цієї вибірки були планарними. Для того, щоб гарантувати раціональне розбиття вибірки на тестову й навчальну, ми використали кластеризацію методом k-середніх<sup>74</sup>, доступного в пакеті **Scikit learn** для формування кластерів із схожими структурними властивостями (формування кластерів на прикладі вибірки літій катіон основних сполук наведено у додатку Д). Довільним параметром у методі k-середніх є кількість кластерів (k-середніх). Оскільки ми хотіли отримати тестову вибірку, що складала 30% від загальної кількості молекул у вибірці, ми обирали k рівним 30% від кількості молекул у вибірці. На жаль, молекули, що дуже сильно відрізнялись від інших молекул вибірки, за таких обставин, як правило, цілком займали кластер і були єдиною молекулою на кластер. Нашим завданням було отримати таку тестову вибірку, яка б добре описувалась тренувальним набором. Інакше кажучи, вибірки формувались таким чином, що кожна молекула тестового набору містила хоча б одну структурно-схожу молекулу тренувального набору. Тому з кластерів, що містили більш ніж одну молекулу, вибиралась одна молекула до тестової вибірки. Оскільки при цьому кластери з однією молекулою ігнорувались, кількість відібраних молекул становила менший відсоток ніж 30%. Щоб компенсувати недостатню кількість молекул в тестовій вибірці з кластерів, що містили більш ніж три молекули, ми додатково відбирали другу молекулу до тестової вибірки до тих пір, доки 30% молекул не було віднесено до тестової вибірки.

Слід зазначити, що, виходячи з рис. 4.1, можна зробити висновок, що набори молекул, використані в цьому дослідженні, були досить структурно-різноманітними.

#### 4.4. Результати LR-LARS-LASSO розрахунків

Для аналізу й порівняння результатів, отриманих нами в методах LR-LARS-LASSO, а також RF, ми використовували ROC-криві (*Receiver Operating Characteristic*), оскільки обидва методи в якості результату отримують не просто бінарні класифікаційні величини (активна молекула чи ні), а вірогідність того, що молекула є активною. Для таких методів, у залежності від значення порогу, що дискримінує молекули на активні та неактивні, одні й ті ж молекули можуть бути класифіковані по-різному. При цьому для побудови ROC-кривої досліджують відношення між частиною правильно визначених активних молекул від загальної кількості активних молекул (*True positive rate*, TPR), а також частиною молекул помилково віднесених до активних від загальної кількості неактивних молекул (FPR, *False positive rate*). Для побудування ROC-кривої значення порогу зменшують від одиниці до 0, фіксуючи точки на графіку, коли змінюються TPR або FPR. Більше інформації стосовно ROC-кривих може бути знайдено в (48).

На рис. 4.2 та 4.3 наведено ROC-криві, отримані для тестової вибірки, для набору даних з Li<sup>+</sup>-основності сполук та набору DSSTox відповідно. На графіках також наведено чисельне значення площі під ROC-кривими (*Area Under Curve*, AUC). На рис. 4.2 (а) та 4.3 (а) порівнюються зміни якості рівнянь LARS-LASSO у залежності від кількості дескрипторів, а на рис. 4.2 (б) та 4.3 (б) порівнюється якість моделей LR-LARS-LASSO з найкращою моделлю отриманою в методі RF.

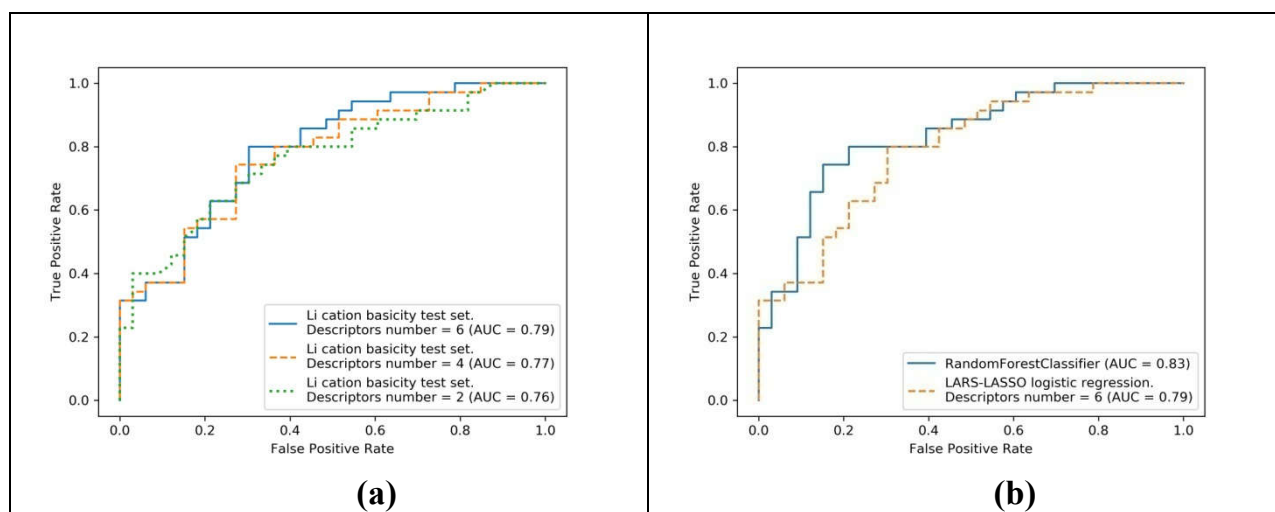


Рис. 4.2 ROC-криві, отримані для класифікації тестової вибірки Li-катион основності. (a) Порівняння результатів, отриманих з використанням методу LR-LARS-LASSO, за різної кількості дескрипторів у моделі. (b) Порівняння ROC-кривих, отриманих у методі LR-LARS-LASSO з шістьма дескрипторами та методі RF з використанням найкращої моделі

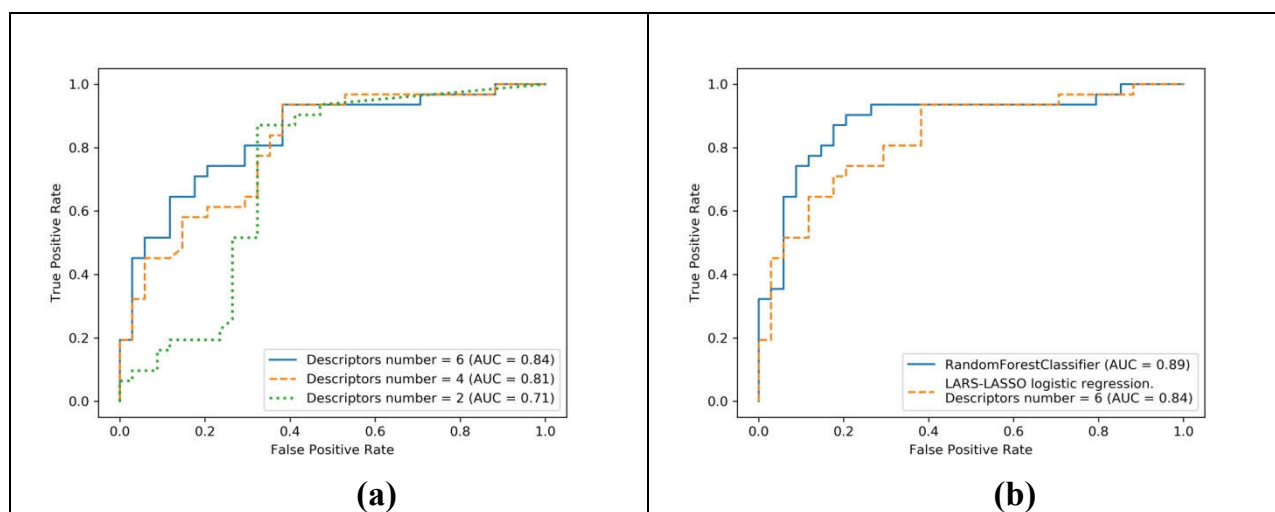


Рис. 4.3 ROC-криві, отримані для класифікації тестової вибірки DSSTOX. (a) Порівняння результатів, отриманих з використанням методу LR-LARS-LASSO за різної кількості дескрипторів у моделі. (b) Порівняння ROC-кривих, отриманих у методі LR-LARS-LASSO з шістьма дескрипторами та методі RF з використанням найкращої моделі

З аналізу ROC-кривих можна бачити, що для вибірки  $\text{Li}^+$  основності (рис. 4.2 (a)) включення більше двох дескрипторів у модель незначним чином

покращує якість моделей, у той час як для вибірки DSSTox більша кількість дескрипторів була необхідна для покращення моделі. Для обох тестових вибірок включення більшої кількості дескрипторів до моделі призводило до покращення передбачувальної здатності моделей відповідно до критерію AUC для тестової вибірки. Таким чином, послідовно додаючи дескриптори до регресійної моделі, стає можливим послідовно покращувати якість відповідних моделей (див. табл. 4.1). Слід зазначити, що використання більш ніж трьох дескрипторів для обох тестових завдань не збільшувало кількість правильних класифікацій для тестової вибірки. Для тренувальної вибірки збільшення кількості дескрипторів у моделі завжди покращувало кількість коректно класифікованих молекул для тренувальної вибірки DSSTox, у той час як для тренувальної вибірки  $\text{Li}^+$  основності сполук покращення видалися незначними. Рівняння логістичної регресії з відповідними характеристиками AUC наведено у таблиці 4.1.

Таблиця 4.1

**Дескриптори та відповідні коефіцієнти регресії LR-LARS-LASSO у порядку зниження важливості для класифікації: а) вибірка основності до  $\text{Li}^+$ ; б) DSSTox вибірка**

а) основність до $\text{Li}^+$									
Кількість дескрипторів	Назва дескрипторів							AUC train	AUC test
	Intercept	TIC1	GATS4s	MLFER_L	MLFER_BH	nHBAcc	ETA_Shape_Y		
2	-0.49	0.010	0.13					0.91	0.76
4	-0.98	0.014	0.24	0.060	0.079			0.93	0.77
6	-1.55	0.016	0.29	0.14	0.19	0.039	0.39	0.94	0.79
б) DSSTox									
	Intercept	minsOH	maxHsOH	GATS1i	Elu	Elp	TDB2i		
2	-0.598	0.097	0.491					0.76	0.71
4	-0.253	0.100	0.557	-0.145	-0.475			0.84	0.81
6	6.76	0.092	1.02	-0.69	-4.21	-0.310	-0.0135	0.88	0.84

З табл. 4.1 можна побачити, що згідно з моделями, отриманими в LR-LARS-LASSO, основність органічних сполук залежить від загального індексу вмісту інформації (*Total information content index*, TIC1)<sup>89</sup>, а також GATS4S (*Geary autocorrelation weighted by intrinsic state*). Індекс GATS4S відповідає

локальній просторовій кореляції в молекулі. Більше інформації за цим індексом може бути знайдено в (161). Спорідненість молекул до рецепторів естрогену залежить від індексів: minsOH (*Minimum atom-type Electrotopological State: OH*) та maxHsOH (*maximum atom-type H Electrotopological State: OH*), що описують вплив електронів сусідніх груп на відповідні групи. Більше інформації за цими дескрипторами може бути знайдено в (162, 163).

З рис. 4.2 (b), а також 4.3 (b) можна побачити, що рівняння класифікації, отримані методом LR-LARS-LASSO, практично не програють у якості класифікації моделям, отриманим у RF. Порівнюючи результати моделей LR-LARS-LASSO з точки зору кількості правильних класифікацій з методом KNN, LR з двома й більше дескрипторами класифікувала тестову вибірку завжди краще. Найкраща модель KNN для тестової вибірки DSSTox класифікувала правильно 69.2% молекул за  $k=2$ , у той час як модель логістичної регресії з двома дескрипторами – 72.3%. Для тестової вибірки основності до  $\text{Li}^+$  різниця була більш помітною з лише 58.8 % коректних класифікацій за kNN з  $k=5$  та 73.5 % вірних класифікація з використанням моделей LR з 4 дескрипторами. При порівнянні результатів з **деревами рішень**, найкращі результати для DSSTox – 78.5% правильних класифікацій, а для  $\text{Li}$ -катіон основності – 75%. Слід зазначити, що обидва методи (як метод KNN, так і метод дерев класифікацій) мають доволі емпіричні параметри, варіюванням яких може бути досягнуто найкращий розв'язок. При цьому обидва методи дуже швидко перенавчаються. Результати для дерев класифікації з різною кількістю дерев для набору DSSTox наведено в табл. 4.2. Приклад **дерева рішень** наведено на рис. 4.4. При цьому метод LR-LARS-LASSO не перенавчається поки кількість дескрипторів, що приймають участь у рівнянні, є адекватною.

Таблиця 4.2

**Процент правильних класифікацій, отриманих з використанням дерева рішень з різною глибиною, для тестової вибірки DSSTox**

Максимальна глибина	% правильних класифікацій для тестового набору
2	76.9
3	81.5
4	75.4
5	78.5
7	78.5
10	73.8

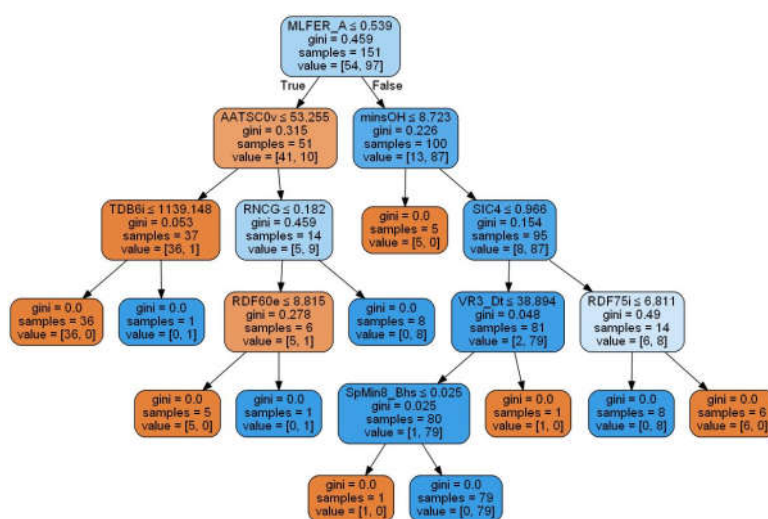


Рис. 4.4 Дерево класифікації з максимальною глибиною = 5, отримане для DSSTox вибірки. Тут після кожного вузла розв'язку наводиться кількість активних та неактивних сполук (у квадратних дужках)

Аналізуючи дерево рішень на рис. 4.4, можна побачити, що, хоча всі молекули тренувального набору класифіковано ідеально правильно, деякі вузли вирішень ґрунтуються на лише одній молекулі. Очевидно, що таке вирішення є не обґрунтованим і, у загальному випадку, скоріш за все, є некоректним, що може призвести до погіршення результатів передбачення тестового набору.

## Висновки до розділу 4

1. Зазвичай вважається, що логістична регресія не може бути застосована для обробки великих наборів даних, розмір яких, на теперішній час, продовжує

зростати. Згідно наших розрахунків, модифікація LR-LARS-LASSO виявилась спроможною подолати цей недолік. Розроблений алгоритм, реалізований у комп'ютерній програмі, виявився розрахунково-ефективним і спроможним конкурувати з існуючими алгоритмами машинного навчання

2. Явними перевагами методу LR-LARS-LASSO є однозначність розв'язку, чого нема в більшості сучасних методів (наприклад, у методі дерев рішень, RF, ANN тощо), а також інтерпретуємість отриманих рівнянь. При цьому передбачувальна здатність отриманих моделей практично не відрізняється від результатів інших методів.

3. У розділі представлено прості класифікаційні функції для відбору органічних систем різної природи: за основністю по відношенню до катіону літію (класифікувалися як сильні та слабкі основи) та спорідненості до рецепторів естрогену (активні або неактивні). Отримані рівняння є досить простими, із якістю прогнозу, що може бути порівняна з відомими, громіздкими підходами.

Основні положення цього розділу викладено в публікації автора (164).



## РОЗДІЛ 5

### МОЖЛИВОСТІ ВИКОРИСТАННЯ $L_1$ -РЕГУЛЯРИЗАЦІЇ В КВАНТОВІ ХІМІЇ

Одним з ключових питань сучасної квантової хімії є проблема скорочення кількості параметрів хвильової функції і її "довжини". Особливо це стосується неемпіричних (*ab initio*) розрахункових методів, які явно включають ефекти електронної кореляції. Серед таких методів теорія збурень Меллера-Плесета (*Møller–Plesset perturbation theory*, MP), метод конфігураційної взаємодії, теорія зв'язаних кластерів (*Coupled Cluster*, CC) та інші часом комбіновані підходи.

Надмірно великі сучасні базиси атомних орбіталей ведуть до того, що кількість електронно-збуджених конфігурацій різної кратності значно зростає. Це веде до необхідності побудови ефективних алгоритмів мінімізації будови хвильових функцій, які здатні гарантувати якісні оцінки молекулярних характеристик за мінімальними затратами. Таке скорочення орбітального простору або простору електронно-збуджених конфігурацій може бути досягнуто доволі різними методами. Деякі з них можна знайти в роботах (165-172).

Пошуки нових підходів до цієї проблеми продовжуються і, вочевидь, будуть продовжуватись найближчими роками. Зауважимо, що наразі, незважаючи на певні успіхи теорії функціоналу густини, стає зрозумілим, що найближчим часом відбудеться значний зсув у сторону класичної квантової хімії. Об'єм розрахунків такими методами як MP та CC буде зростати. Ми вважаємо, що метод  $L_1$ -регуляризації також може бути застосований для розв'язку проблеми скорочення "довжини" хвильової функції методів MP та CC. Спираючись на цей регуляризаційний підхід, можливо як скорочення кількості вакантних орбіталей, так і скорочення кількості електронно-збуджених конфігурацій.

У представленій роботі зроблено перші спроби  $L_1$ -регуляризованих розв'язків рівнянь багаточастинкової теорії MP і теорії CC. Ці високоточні

методи урахування електронної кореляції можуть бути надто затратними в розрахунковому сенсі і тому пошук послідовної системи спрощень є актуальною проблемою, розв'язання якої обіцяє створення нових ефективних підходів.

У даному розділі буде розглянуто ключові питання теорії СС як-от: відбір ефективних алгоритмів, що реалізують розв'язок рівняння Шредингера з хвильовою функцією СС, скорочення набору електронно-збуджених конфігурацій та тестування розроблених методів.

### 5.1. Градієнтні алгоритми розв'язку нелінійних рівнянь теорії СС

У зв'язку з тим, що процедура розв'язку рівнянь теорії СС – це ітераційна процедура, розрахункова складність якої є лімітуючою для широкого застосування методу, нами, перш за все, було досліджено використання різноманітних градієнтних методів знаходження хвильової функції СС. Для цього були вивчені кілька багатокрокових методів першого порядку – методів, які використовують інформацію про наближені розв'язки, що були отримані на попередніх ітераціях. При цьому інформація щодо других похідних, розрахунки яких є надто затратними, не використовується. Було розглянуто наступні методи багатокрокової оптимізації першого порядку:

- метод прямого обернення в ітераційному підпросторі (*Direct Inverse in Iterative Subspace*, DIIS)
- метод «важкої кульки» (*Heavy Ball*, HB)
- алгоритми засновані на підходах Нестерова.

У даному дослідженні використовувався напівемпіричний  $\pi$ -електронний варіант методу СС з урахуванням лише двократних збуджень (CCD) відносно референсного Гартрі-Фоківського стану. Але вони елементарно узагальнюються й на інші рівні наближень.

Слід також зазначити, що для вивчення методів оптимізації вибір між напівемпіричним та неемпіричним (*ab initio*) методом не є суттєвим. Але ми вважаємо, що саме напівемпіричний метод теорії СС є перспективним у

проблемі вивчення електронної структури великих систем, що включають сотні або навіть тисячі важких атомів. Сам метод CCD цікавий тому, що є найпростішим кластерним підходом, що ефективно враховує чотирьохкратні збудження. Наразі в літературі інтенсивно вивчається можливості підходів оснований на CCD у використанні до опису типових мультиреференсних проблем (квазівироджених станів, статичних кореляціях і *т.д.*, дивись наприклад (173-175)).

Загалом хвильові функції методів СС (з урахуванням лише двократних збуджень – CCD й однократних та двократних збуджень – CCSD) можуть бути представлені наступним чином:

$$\begin{aligned} |\Psi_{\text{CCD}}\rangle &= \exp(T_2)|0\rangle = \left(1 + T_2 + \frac{1}{2}T_2^2 + \dots\right)|0\rangle, \\ |\Psi_{\text{CCSD}}\rangle &= \exp(T_1 + T_2)|0\rangle = \left(1 + T_1 + T_2 + \frac{1}{2}T_1^2 + \frac{1}{2}T_2^2 + T_1T_2 + \dots\right)|0\rangle. \end{aligned} \quad (5.1)$$

Тут  $|0\rangle$  – референсний, зазвичай Гартрі-Фоківський, детермінант, а кластерні оператори  $T_1$  і  $T_2$  генерують суперпозиції відповідно однократно- та двократно-збуджених конфігурацій відносно  $|0\rangle$ . Ці оператори можуть бути виражені через відповідні матриці амплітуд  $t$  наступним чином:

$$T = \sum t_J |J\rangle. \quad (5.2)$$

Символом  $|J\rangle$  позначено електронно-збуджену конфігурацію відповідної кратності, а  $t_J$  - відповідна амплітуда.

Розв'язок рівнянь теорії CCD зводиться до розрахунків термів, які включають лінійну та нелінійну компоненти. З формальної точки зору, ці рівняння для  $t_J$  можуть бути знайдені, виходячи з рівняння:

$$\Delta_J(t) = \langle J | H - E | \Psi_{\text{CCD}} \rangle = \langle J | H | \Psi_{\text{CCD}} \rangle_C = \left( A_J + \sum_I B_{JI} t_I + \frac{1}{2} \sum_{I,K} C_{JKL} t_K t_L \right) / \Delta \epsilon_J = 0, \quad (5.3)$$

де  $A$ ,  $B$ , і  $C$  – одно- та двохелектронні інтеграли,  $\Delta \epsilon_J$  – зміна одноелектронних енергій молекулярних орбіталей (*Molecular Orbitals*, MO), які відповідають переходу до електронно-збудженої конфігурації, а індекс "C" позначає, що у формулі (5.3) враховуються лише зв'язані (*connected*) терми. При цьому

компоненти розкладу, що включають енергію системи ( $E_{\text{CCD}}$ ), анулюються згідно до теореми про зв'язані кластери<sup>176</sup>. Детальні рівняння для (5.3) можуть бути знайдені в (177).

У процесі ітераційного розв'язку рівнянь (5.3) ми вимушені багаторазово розраховувати амплітудну матрицю  $t$  і матрицю  $\Delta$ . На  $k$ -тій ітерації такі матриці будемо позначати як  $t^{(k)}$  та  $\Delta^{(k)} = \Delta(t^{(k)})$ . Отже, стандартний градієнтний метод (*Standard Gradient Approach*, SGA)<sup>178</sup> знаходження амплітудної матриці  $t$  описується як ітераційний процес:

$$t^{(k+1)} = t^{(k)} - \alpha^{(k)} \Delta^{(k)}, \quad (5.4)$$

зупинка котрого відбувається за умови  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$ . Тобто в наших розрахунках величина  $\|\Delta^{(k)}\|_2$  відповідає  $L_2$ -нормі. У найпростішому варіанті SGA, який використовувався в даній роботі, параметр  $\alpha^{(k)} = \alpha = \text{const}$ . Зазначимо, що оптимізація величини  $\alpha^{(k)}$  на кожній ітерації, що відповідає методу найшвидшого спуску (*Steepest Descent*), призводить до суттєвого зростання розрахункових витрат, оскільки потребує розрахунку вектору  $\Delta(\Delta^{(k)})$ .

У схемі інтерполяції DIIS, застосованої до задачі CCD, будується коригована амплітудна матриця  $t$ . На  $\ell$ -тій ітерації вона виражається як суперпозиція амплітудних матриць, отриманих на поточній ітерації  $\ell$ , а також на  $m$  попередніх ітераціях:

$$t^{(\ell)} \leftarrow c_0 t^{(\ell)} + c_1 t^{(\ell-1)} + c_2 t^{(\ell-2)} + \dots + c_m t^{(\ell-m)}. \quad (5.5)$$

При цьому коефіцієнти обираються таким чином, щоб мінімізувати  $\|\Delta^{(\ell)}\|_2$ , при умові:

$$\sum_{i=0}^m c_i = 1. \quad (5.6)$$

Більш детальний опис методу DIIS у теорії Гартрі-Фока може бути знайдено в (179,180). Застосування DIIS до теорії CC дивись (181, 182).

Метод НВ (178) розроблено виходячи з механічної аналогії. Процес оптимізації системи можна інтерпретувати як рух шарику по гіперповерхні до точки мінімуму. При цьому типовою є ситуація, коли кулька, рухаючись по вузькому

жолобу, періодично б'ється о стінки, що відповідає збільшенню числа ітерацій. Для того, щоб уникнути таких ситуацій у методі НВ, до рівнянь SGA (5.4) додається моментна компонента:

$$t^{(k+1)} = t^{(k)} - \alpha \Delta^{(k)} + \beta^{(k)}(t^{(k)} - t^{(k-1)}), \quad (5.7)$$

Тут фактор  $t^{(k)} - t^{(k-1)}$  якби підштовхує шарик у напрямку попередньої ітерації. Ідеологія Нестерова<sup>183</sup> (NEST) породжує низку алгоритмів, що наразі досить інтенсивно обговорюються в літературі<sup>183,184</sup>. У нашій роботі ми використовували два найбільш розповсюджені алгоритми<sup>185</sup>.

Алгоритм NEST1:  $y^{(0)} = t^{(0)}, \theta_0 \in (0,1)$ ,

$$\begin{aligned} &\text{for } k = 0, 1, \dots, \text{ do} \\ &\quad t^{(k+1)} = y^{(k)} - \alpha \Delta(y^{(k)}); \\ &\quad \theta_{k+1} \in (0,1) \text{ from } \theta_{k+1}^2 = (1 - \theta_{k+1})\theta_k^2 + q\theta_{k+1}; \beta_{k+1} = \theta_k(1 - \theta_k) / (\theta_k^2 + \theta_{k+1}); \\ &\quad y^{(k+1)} = t^{(k+1)} + \beta_{k+1}(t^{(k+1)} - t^{(k)}); \\ &\text{end} \end{aligned} \quad (5.8)$$

Алгоритм NEST2 відповідає незмінному в ітераційному процесі параметру  $\beta_k = \beta^*$ :

$$\begin{aligned} &y^{(0)} = t^{(0)}; \\ &\text{for } k = 0, 1, \dots, \text{ do} \\ &\quad t^{(k+1)} = y^{(k)} - \alpha \Delta(y^{(k)}); \\ &\quad y^{(k+1)} = t^{(k+1)} + \beta (t^{(k+1)} - t^{(k)}); \\ &\text{end} \end{aligned} \quad (5.9)$$

### 5.1.1 Тестові оцінки ефективності різних алгоритмів

Для визначення збіжності методів багатокрокової мінімізації першого порядку ми використовували напівемпіричний  $\pi$ -електронний варіант теорії CCD. У розрахунках використовувався модельний гамільтоніан Попла-Парізера-Парра, ППП, (*Pariser-Parr-Pople*, PPP)<sup>186,187</sup>). Розраховувались дві  $\pi$ -системи. Лінійний полієн дека-1,3,5,7,9-пентаєн,  $C_{10}H_{12}$ , та модельна планарна циклічна молекула [14]аннулен,  $C_{14}H_{14}$ .

У так званих "м'якій" параметризації гамільтоніана ППП, стандартний резонансний інтеграл пари зв'язаних вуглецевих атомів дорівнює  $b_0 = -2.274$  eВ,

одноцентровий кулонівський інтеграл  $\Gamma_0=11.13$  еВ. Двоцентрові кулонівські інтеграли оцінювались виходячи з відомої формули Оно<sup>188</sup>. У розрахунках використовувалася ідеалізована геометрія довжин усіх –C–C– зв’язків рівних 1.4 Å, кути в лінійному *trans*-полієні – 120°. Циклополієн C<sub>14</sub>H<sub>14</sub> розглядався як правильний багатокутник. При розрахунку C<sub>10</sub>H<sub>12</sub> вводилося альтернування резонансних інтегралів для подвійних (зі знаком “+”) та одинарних зв’язків (зі знаком “-”):  $b_{\pm} = b_0(1 \pm 0.1)$ . У розрахунках циклічної системи C<sub>14</sub>H<sub>14</sub> резонансний інтеграл зменшували від стандартного  $b_0=-2.274$  еВ до величин - 1.3 еВ. У цьому випадку збіжність ітераційного процесу звичайних методів погіршується в зв’язку з квазівиродженням молекулярних орбіталей.

У методі DIIS швидкість знаходження розв’язку визначається вибором параметру  $\alpha$  та кількістю кроків інтерполяції  $n_{\text{DIIS}}=m+1$  див. рівняння (5.5). У табл. 5.1 та 5.2 показано як змінюється швидкість збіжності (кількості ітерацій) до фіксованої точності  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при зміні параметру  $\alpha$  для методу DIIS з різною розмірністю вектора (5.5)  $n_{\text{DIIS}}$ .

Таблиця 5.1

**Кількість ітераційних кроків методу DIIS для досягнення збіжності згідно з критерієм  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при різних значеннях параметру  $\alpha$ . Система C<sub>10</sub>H<sub>12</sub>**

$\alpha$	$n_{\text{DIIS}} = 2$	$n_{\text{DIIS}} = 3$	$n_{\text{DIIS}} = 4$	$n_{\text{DIIS}} = 8$
0.7	69	48	47	32
0.8	61	44	40	31
0.9	52	52	45	32
1	62	54	42	32
1.1	67	51	46	33
1.2	67	49	46	33
1.3	66	50	46	34

Наведені таблиці демонструють відносну незалежність кількості ітерацій від вибору параметру  $\alpha$ . Така незалежність стає більш помітною при збільшенні  $n_{\text{DIIS}}$ . Так, при  $n_{\text{DIIS}}=8$ , вибір  $\alpha$  практично не впливає на кількість ітерацій. Ця обставина є цінною властивістю методу DIIS.

При переході до другої системи  $C_{14}H_{14}$  зі зруйнованими зв'язками ( $b = -1.4$  eV), збіжність при малих  $n_{DIIS}$  значно погіршується, що характерно для квазівироджених ситуацій. Але при  $n_{DIIS}=8$  характеристики збіжності для обох систем виявляються ідентичними!

Таблиця 5.2

**Кількість ітераційних кроків методу DIIS для досягнення збіжності згідно з критерієм  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при різних значеннях параметру  $\alpha$ . Система  $C_{14}H_{14}$**

**( $b = -1.4$  eV)**

$\alpha$	$n_{DIIS} = 2$	$n_{DIIS} = 3$	$n_{DIIS} = 4$	$n_{DIIS} = 8$
0.7	118	100	63	36
0.8	104	108	54	32
0.9	92	72	75	32
1	136	76	73	32
1.1	187	76	59	34

На жаль, при роботі з великими системами, збереження матриць в оперативній пам'яті ЕОМ стає неможливим, а витрати на зчитування та перезапис на жорсткий диск будуть суттєво впливати на час розрахунку. Тому практичний інтерес також мають методи, що утримують якомога менше даних в оперативній пам'яті. Серед таких методів метод НВ, збіжність якого визначається вибором двох параметрів. Параметр  $\alpha$  визначає величину кроку в напрямку градієнту, а  $\beta$  – релаксаційний параметр (5.7).

Результати розрахунку двох  $\pi$ -систем методом НВ наведено в табл. 5.3 та 5.4.

Таблиця 5.3

**Кількість ітераційних кроків методу НВ для досягнення збіжності згідно з критерієм  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при різних значеннях параметрів  $\alpha$  та  $\beta$ . Система**

**$C_{10}H_{12}$**

$\alpha$	$\beta$				
	0	0.1	0.2	0.3	0.4
1.6	нз <sup>vi)</sup>	нз	нз	57	54

<sup>vi</sup> нз – нема збіжності ітераційної процедури.

Продовження таблиці 5.3

1.5	нз	нз	107	42	54
1.4	нз	206	37	42	55
1.3	627	62	33	42	55
1.2	92	56	39	43	55
1	84	71	55	43	56
0.9	94	81	65	43	56
0.8	108	93	76	53	56

Таблиця 5.4

Кількість ітераційних кроків методу НВ для досягнення збіжності згідно з критерієм  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при різних значеннях параметрів  $\alpha$  та  $\beta$ . Система

$C_{14}H_{14}$  ( $b = -1.4$  eV)

$\alpha$	$\beta$			
	0.3	0.4	0.5	0.6
1	195	156	105	98
1.1	175	138	83	99
1.3	144	110	74	99
1.4	132	98	74	100
1.5	121	88	74	99
1.6	111	78	75	100
1.7	103	67	75	102
1.8	261	58	76	101
1.9	нз	86	76	104

Із наведених даних можна бачити, що при введенні невеликої величини для релаксаційного параметру  $\beta$ , швидкість збіжності може значно покращитись навіть при значній величині  $\alpha$  ( $>1.3$  для лінійного полієну). Очевидно, що ситуація, коли  $\beta=0$ , відповідає SGA. Введення релаксаційного параметру стабілізує збіжність ітераційної процедури, дозволяючи отримувати рішення за обмежену кількість ітерацій у широкому діапазоні зміни  $\alpha$ . Це велика перевага методу НВ у порівнянні з методами Нестерова, які, як буде показано нижче, розбігаються при невдалому виборі параметрів ітераційного процесу. Слід зазначити, однак, що кількість ітерацій відрізняється для  $C_{10}H_{12}$  та  $C_{14}H_{14}$  (при  $b=-1.4$  eV) у методі НВ, на відміну від DIIS з  $n_{DIIS} = 8$ . Проте для  $C_{14}H_{14}$  (при  $b=-1.4$  eV), метод НВ гарантує знаходження CCD рішень за 74-75 ітерацій.

Розрахунки в рамках методів Нестерова (алгоритми NEST1 та NEST2) проявили значну залежність від значень параметрів (аж до повної відсутності



збіжності при невеликому варіюванні параметрів). У табл. 5.5 та 5.6 продемонстрована зміна швидкості збіжності в методах Нестерова при змінах параметрів ітераційної процедури. Наведені дані вказують на значний розкид кількості ітерацій при різних значеннях "моментних" параметрів ( $q$  для NEST1 та  $\beta$  для NEST2).

Таблиця 5.5

**Кількість ітераційних кроків методу NEST1 для досягнення збіжності згідно з критерієм  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при різних значеннях параметрів  $\alpha$  та  $q$ .**

Система $C_{14}H_{14}$ ( $b=-1.4$ eV)				
$\alpha$	$q$			
	0.04	0.05	0.06	0.1
0.6	110	134	161	222
0.7	102	95	127	186
0.8	102	97	90	158
0.9	89	87	87	136
1	701	419	308	165

Таблиця 5.6

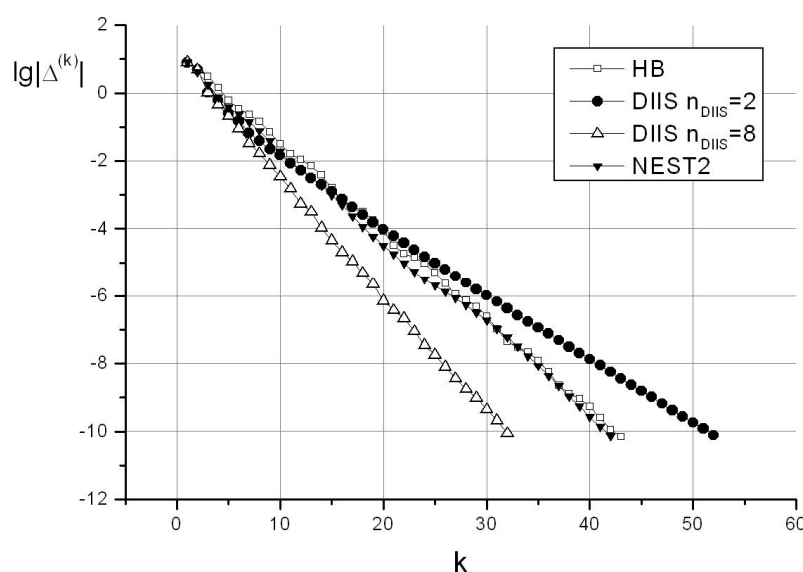
**Кількість ітераційних кроків методу NEST1 для досягнення збіжності згідно з критерієм  $\|\Delta^{(k)}\|_2 \leq 10^{-10}$  при різних значеннях параметрів  $\alpha$  та  $q$ .**

Система $C_{14}H_{14}$ ( $b = -1.4$ eV)								
$\alpha$	$\beta$							
	0.4	0.5	0.55	0.57	0.58	0.59	0.6	0.62
0.8	215	167	140	127	119	111	99	92
0.9	188	145	118	104	93	84	86	89
1	203	152	203	234	253	274	300	367

Задача знаходження рішення CCD надзвичайно чутлива до ступеню квазівиродження молекул. Для циклічної молекули  $C_{14}H_{14}$  при достатньо малих значеннях резонансних інтегралів ( $|b| < 1.3$  eV) розв'язки отримати не вдалося. При значеннях резонансного інтегралу ( $|b| \sim 1.3$  eV) усі методи приблизно з однаковою швидкістю збігались до точності  $\|\Delta^{(k)}\|_2 \sim 10^{-8}$ . При подальшому пошуку розв'язків методи НВ та NEST розбігались, але метод DIIS, ( $n_{DIIS} = 2$ ) з помітним сповільненням швидкості збіжності та з появою осциляцій, збігся до

вказаної точності. При цьому в методі DIIS 88 ітерацій було витрачено на досягнення точності  $\|\Delta^{(k)}\|_2 \sim 10^{-8}$ , а також ще 184 на досягнення точності  $\|\Delta^{(k)}\|_2 \sim 10^{-10}$ . При розв'язку цієї задачі методом DIIS із більшою розмірністю (5.5) ця проблема зникла. Так, при розмірності  $n_{\text{DIIS}}=8$  метод ефектно збігався до точності  $\|\Delta^{(k)}\|_2 \sim 10^{-10}$  усього за 38 ітерацій!

У цілому порівняння ітераційних методів на прикладі розрахунку полієна  $C_{10}H_{12}$  наведено на рис. 5.1.



**Рис. 5.1** Залежність  $\lg\|\Delta^{(k)}\|_2$  від номеру ітерації

Із наведеної залежності можна зробити висновок, що збіжність ітераційних процедур NEST2 і HB, з оптимально визначеними параметрами, у випадку розрахунку "звичайних" (не вироджених) систем, виявилась навіть краще, ніж для DIIS (з  $n_{\text{DIIS}}=2$ ). Проте DIIS виявився більш надійним методом розрахунку, що гарантує рішення задачі навіть у "важких" обставинах, зв'язаних із квазивиродженням. Але слід зазначити також, що для метода DIIS потрібен значний обсяг операцій вводу–виводу, велика кількість розрахунків скалярних добутків векторів великої розмірності, а також обернення матриці. Остання операція може містити додаткові проблеми, пов'язані із виродженністю матриці

в ситуаціях, коли задача достатньо близька до розв'язку. У цьому випадку має бути використано псевдообернення за Муром-Пенроузом<sup>18,19</sup>.

## 5.2. Теорія зв'язаних кластерів і регуляризація

Однією з найважливіших проблем практичної квантової хімії є вибір форми хвильової функції, яка могла б адекватно описати досліджуваний стан системи. У багатьох випадках під таким вибором мається на увазі відбір мінімального орбітального набору, а також вибір репрезентативної вибірки електронно-збуджених конфігурацій. Так звана “хімічна інтуїція” може допомогти у виборі відповідного орбітального базисного набору, а також вибірки конфігурацій, але також існують й інші методи, що можуть допомогти у виборі. Подібні підходи часто використовуються в розрахунках методом багатоконфігураційного самоузгодженого поля (*Multi-Configurational Self-consistent Field Theory*, MCSCF). Серед таких підходів слід зазначити: попереднє оцінювання конфігураційного складу хвильової функції в рамках теорії збурень<sup>189</sup>, а також метод індексів електронної кореляції [див. наприклад (190-192)].

Існує також багато підходів для скорочення кількості параметрів у багаточастинкових моделях. Серед них локальні методи<sup>165-169</sup>, а також методи оптимізації віртуального орбітального простору<sup>170,171</sup>, що були успішно використані на практиці<sup>172</sup>.

В останні десятиріччя було визнано, що ідея локального трактування кореляційних ефектів може мати суттєву практичну перевагу при розрахунку енергетичних характеристик молекул<sup>193,194</sup>. Застосування локального підходу в теорії CC<sup>195,196</sup> зробило можливим *ab initio* розрахунки систем із значною кількістю важких атомів (дивись наприклад (197)).

У теорії CC було реалізовано кілька регуляризаційних підходів. Так, метод сингулярного розкладу (*Singular Values Decomposition*, SVD), що методологічно пов'язаний з підходами регуляризації, було використано в (198,199) для стиснення  $T_3$ -амплітудної матриці. Безпосереднє застосування

регуляризації може бути знайдено в (200), де  $L_2$ -регуляризація була використана для видалення проблеми сингулярності в лінеаризованій теорії CCSD. Інший регуляризований підхід до різних рівнів теорії CC було реалізовано в (201,202). Дивись також обговорення регуляризованих розв'язків теорії CC у (203,204). Слід відзначити, що так звані ренормалізовані теорії CC також можуть бути інтерпретовані як різновид регуляризаційних підходів<sup>205-209</sup>.

У нашій роботі в якості альтернативи існуючим підходам для відбору конфігураційного набору ми розглядали  $L_1$ -регуляризацію. Ми вважаємо, що  $L_1$ -регуляризація буде корисною для квантової хімії оскільки:

1)  $L_1$ -рішення дозволяють трансформувати хвильову функцію до компактного вигляду. Це дає можливість інтерпретувати й звузити часом громіздкий набір електронно-збуджених конфігурацій, що включає сотні тисяч (і більше) членів.

2)  $L_1$ -рішення, отримані в рамках певного наближеного методу, дозволяють сортувати компоненти хвильової функції за їх “значимістю” і сформувати шуканий набір функцій. Наприклад, у випадку багатоконфігураційного наближення з'являється можливість формування списку електронно-збуджених конфігурацій необхідного певного розміру. Надалі цей набір конфігурацій можна використовувати в точних методах (CASSCF, мультиреференсній теорії). Наразі в цих методах практикується попередній ручний відбір активних конфігурацій.

3)  $L_1$ -регуляризовані розв'язки для багаточастинкових методів дозволяють створити ієрархію апроксимацій методу. Цю ієрархію можна отримати варіюванням параметру регуляризації  $\lambda$ . Апроксимація може бути як дуже грубою, але з меншими розрахунковими ресурсами, так і більш точною, але й більш затратною в розрахунковому сенсі.

У представлений дисертаційній роботі нами було вперше запропоновано використання  $L_1$ -регуляризації для квантовохімічної задачі (див. однак (210)). Цей підхід було реалізовано як у відносно-простій теорії збурень MP2 (відповідний регуляризований метод будемо називати  $L_1$ -MP2), так і в більш

складній теорії СС з урахуванням тільки двократних збуджень ( $L_I$ -CCD), а також у методі з урахуванням двократних і однократних збуджень ( $L_I$ -CCSD). Розрахунки в теорії СС проводилися як з напівемпіричним гамільтоніаном ППП, так і з неемпіричним (*ab initio*).

### 5.3. Теоретичні засади $L_I$ -СС розрахунків

$L_I$ -регуляризований вираз для загального випадку квантово-механічної моделі може бути отримано виходячи із відношення Релея:

$$W^{(\lambda)} = \langle \Psi | H | \Psi \rangle / \langle \Psi | \Psi \rangle + \lambda |\Psi|_1. \quad (5.10)$$

Формально, модуль хвильової функції можна представити у вигляді скалярного добутку:

$$|\Psi|_1 = \langle \Psi | \text{sign}(\Psi) \rangle. \quad (5.11)$$

Але детальний вираз для  $|\Psi|_1$  суттєво залежить від обраного представлення хвильової функції. Виходячи з (5.10), для субградієнту може бути записаний наступним чином:

$$\partial W = (H - W_\lambda(\Psi))|\Psi\rangle + \lambda |\text{sign}(\Psi)\rangle. \quad (5.12)$$

Тут символ  $|\text{sign}(\Psi)\rangle$ , позначає вектор, що містить знаки коефіцієнтів базисних функцій.

Для багатоконфігураційних методів CCD та CCSD рівняння (5.11) може бути перетворене наступним чином:

$$|\Psi|_1 = \sum_J t_J \text{sign}(t_J) + \sum_{J>I} t_J t_I \text{sign}(t_J t_I) + \sum_{J>I>K} t_J t_I t_K \text{sign}(t_J t_I t_K) + \sum_{J>I>K>L} t_J t_I t_K t_L \text{sign}(t_J t_I t_K t_L), \quad (5.13)$$

Виходячи з (5.3) та (5.12), для методу CCD можна записати вираз для  $J$ -проекції субградієнту:

$$\partial W_J^{(\lambda)} = \Delta_J + d_J, \quad (5.14)$$

Цей вираз включає в собі регуляризуючий член у представленні “*soft-threshold*” [дивись наприклад (11, 103)]:

$$d_J = \lambda \langle J | \text{sign}(\Psi_{\text{CCD}}) \rangle, \quad (5.15)$$

$$d_J = \begin{cases} \lambda \operatorname{sign}(t_J), & \text{if } |t_J| > 0, \\ +\lambda, & \text{if } t_J = 0, \& \Delta_J < -\lambda, \\ -\lambda, & \text{if } t_J = 0, \& \Delta_J > +\lambda, \\ 0, \Delta_J = 0, & \text{if } t_J = 0, \& -\lambda \leq \Delta_J \leq +\lambda. \end{cases} \quad (5.16)$$

І, виходячи з рівнянь (5.14) та (5.16), ітераційна процедура для  $L_I$ -CCD може бути наведена наступним чином:

$$t_J^{(k+1)} = P_0[t_J^{(k)} - \xi \partial W_J^{(k)} / \Delta \epsilon_J], \quad (5.17)$$

де  $t_J^{(k)}$  амплітуда на  $k$ -тій ітерації,  $\xi$  – крок ітераційної процедури, а  $\Delta \epsilon_J$  – зміна енергії молекулярних орбіталей відповідно до енергії електронного збудження  $|0\rangle \rightarrow |J\rangle$ . На відміну від стандартного градієнтного методу, ми використовуємо оператор  $P_0$  у рівнянні (5.17), що дозволяє уникнути осциляцій у точках, де значення амплітуд наближується до 0,  $t_J \approx 0$ . Коли такі осциляції відбуваються, ми використовуємо:

$$P_0[X_J^{(k)}] = \begin{cases} t_J^{(k+1)}, & \text{if } t_J^{(k)} \cdot X_J^{(k)} > 0, \\ 0 & \text{if } t_J^{(k)} \cdot X_J^{(k)} < 0. \end{cases} \quad (5.19)$$

Ітераційна процедура (5.17) повторюється доки норма суб-градієнту не буде більше, ніж обране граничне значення  $|\partial W| > \eta \sim 10^{-10}$ . Такий критерій гарантує занулення всіх незначних амплітуд. Метод  $L_I$ -CCSD реалізовано схожим чином.

Отже, розв'язок рівнянь регуляризованого методу СС має два етапи. На першому етапі проводиться стандартна процедура проектування СС. Після чого на другому етапі проводиться мінімізація  $L_I$ -норми СС коефіцієнтів. У граничному випадку другий етап призводить до скорочення амплітудного набору, у тому числі й до пустої множини амплітуд при достатньо великому значенні регуляризуючого параметру  $\lambda$ . Таким чином, у залежності від величини  $\lambda$  може бути отримано амплітудний набір зі скороченою кількістю ненульових амплітуд.

$L_I$ -регуляризовані розрахунки можуть проводитися двома підходами в залежності від напрямку зміни регуляризуючого параметру. У першому підході ми скануємо регуляризуючий параметр від 0 (цей граничний випадок

відповідає не регуляризованому розрахунку відповідної теорії) до якогось позитивного значення. Це дозволяє аналізувати структуру хвильової функції. Або ж у другому підході  $\lambda$  змінюється від великих позитивних значень параметру до нуля, що генерує набір апроксимацій до точного розрахунку СС, що відповідає  $\lambda = 0$ .

Тут слід розглянути специфічний для методу СС момент, пов'язаний з регуляризацією. А саме – розмірну узгодженість методу. Відомо, що для методу СС розмірна узгодженість виконується, якщо при розкладі експоненти (5.1) за ступеневим рядом ніякі терми не виключаються з розкладу. Проте при включенні регуляризуючої добавки до рівняння Шредингера, а також при обнулінні деяких амплітуд за рахунок  $L_I$ -регуляризації, ця властивість методу може не виконуватись. Особливо це стосується розрахунків із великими значеннями коефіцієнту  $\lambda$ . Втім, за відносно малих значень  $\lambda$  результуючі значення енергій близькі до тих, що отримані в точному методі CCD або CCSD. Отже, розмірна екстенсивність принаймні частково виконується. Досить повний опис проблеми розмірної узгодженості в наближених теоріях СС може бути знайдено в (205, 206).

#### 5.4. Напівемпіричні розрахунки $L_I$ -СС

У тестових напівемпіричних розрахунках цього розділу ми використовували  $\pi$ -електронне наближення Попла-Паризера-Парра (ППП) (186,187). Була використана ідеалізована геометрія. Вважалося, що довжина зв'язку  $-C-C-$  дорівнювала 1.4 Å. Для циклічних фрагментів була використана геометрія правильного многокутника. Стандартний резонансний інтеграл пари зв'язаних атомів ( $\mu, \nu$ ):

$$\beta_0 = \langle \mu | H | \nu \rangle = -2.274 \text{ eV}. \quad (5.20)$$

Двоцентрові кулонівські інтеграли розраховувались за допомогою емпіричної формули Оно<sup>188</sup>:

$$\Gamma_{\mu\nu} = \langle \mu\nu | \mu\nu \rangle = \frac{\zeta\Gamma}{\sqrt{1 + (\Gamma \cdot R_{\mu\nu})^2}}, \quad \Gamma = \langle \mu\mu | \mu\mu \rangle, \quad (5.21)$$

де  $R_{\mu\nu}$  – між'ядерна відстань, а одноцентровий кулонівський інтеграл для вуглецю  $\Gamma = 11.13$  еВ. Параметр  $\zeta$  гарантує відповідність розмірності величин, що входять до формули (5.21).

Тестові напівемпіричні розрахунки проводилися для наступних молекул (рис. 5.2):

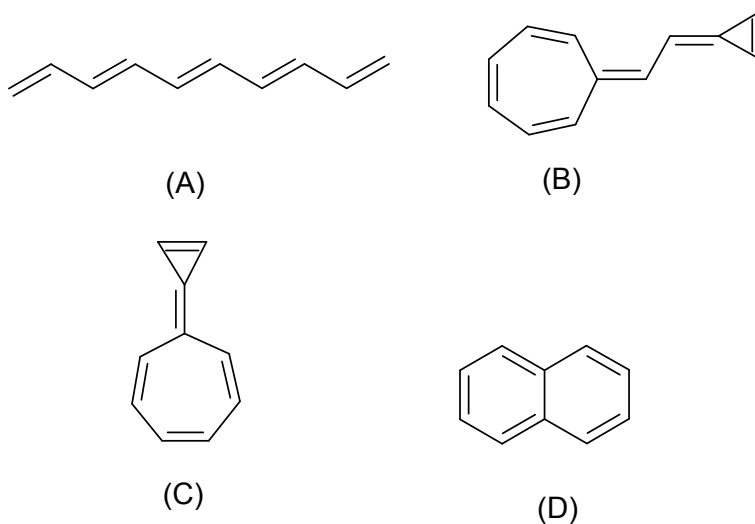


Рис. 5.2  $\pi$ -спряжені тестові системи

Перша система (A) дека-1,3,5,7,9-пентаєн – типова квазіодновимірна молекула, молекули (B та C) є аналогами молекули каліцену, дослідження яких має практичну цінність у зв'язку із значними нелінійно-оптичними характеристиками<sup>211</sup>, молекула нафталіну (D) обрана як типовий представник  $\pi$ -спряжених систем.

Орбіталі молекул A і B позначалися таким чином, щоб перша орбіталь (1) відповідала найвищій зайнятій МО (*Highest Occupied Molecular Orbital*, НОМО), а 1' – найнижчій вакантній МО (*Lowest Unoccupied Molecular Orbital*, LUMO):

$$....5^2 4^2 3^2 2^2 1^2 | 1' 2' 3' 4' 5'.... \quad (5.22)$$

Для ідентифікації МО молекул C та D використовувалися симетрійні позначення (HF) рішень. У відповідності із цим, послідовність HF орбіталей за енергією може бути представлено наступним чином:

$$C \quad 1b_1^2 2b_1^2 1a_2^2 3b_1^2 4b_1^2 | 2a_2 3a_2 5b_1 4a_2 6b_1$$



$$\mathbf{D} \quad 1b_{3u}^2 1b_{1g}^2 1b_{2g}^2 2b_{3u}^2 1a_u^2 | 2b_{1g} 2b_{2g} 3b_{3u} 2a_u 3b_{2g}$$

Двократно-збуджені конфігурації, що виникають у теорії CCD, відповідають переходу двох електронів із зайнятих на вакантні МО. Наприклад, перехід пари електронів з НОМО орбіталі на LUMO, у молекулі **C**  $(4b_1)_\alpha(4b_1)_\beta \rightarrow (2a_2)_\alpha(2a_2)_\beta$  породжує двократно-збуджену конфігурацію, що позначається як  $\left| \begin{smallmatrix} 2a_2 2a_2 \\ 4b_1 4b_1 \end{smallmatrix} \right\rangle$  (спін орбіталі з рискою відповідають  $\beta$ -спіну).

При введенні регуляризуючого параметру в розрахунок CCD величини амплітуд монотонно знижуються (рис. 5.3). При чому для різних амплітуд з різними вагами швидкість зменшення модулів амплітуд виявляється різною. За деякого значення параметру  $\lambda$ , який є різним для різних амплітуд, амплітуди скорочуються.

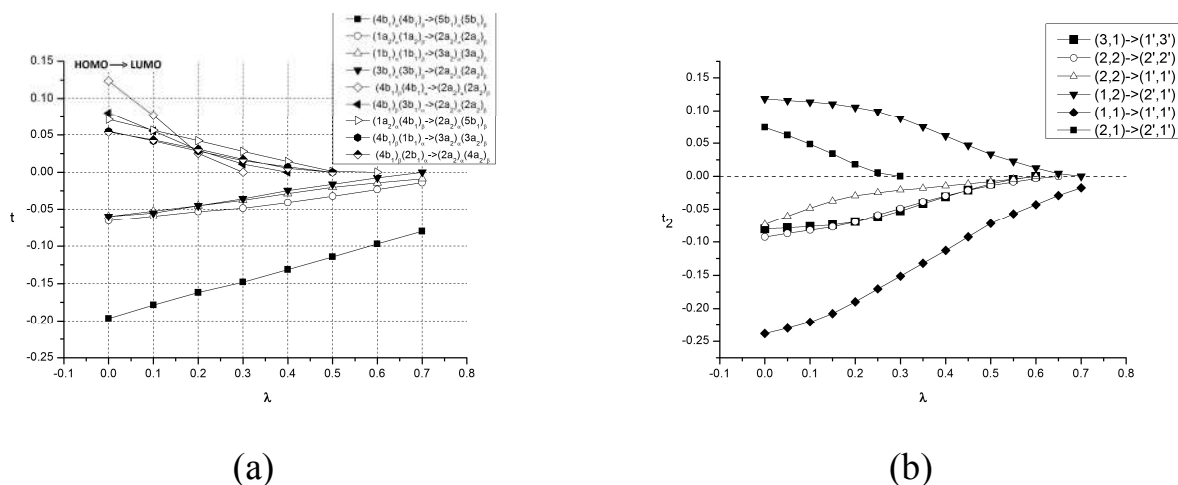


Рис. 5.3 Залежність амплітуд від величини параметру  $\lambda$  для методу  $L_I$ -CCD.

Графік а - молекула каліцену **C**, b - молекула полієну **A**

Так, для молекули **C** із зростанням параметра  $\lambda$ , величина, що відповідає другій по значимості амплітуді у не регуляризованому розрахунку (збудження НОМО  $\rightarrow$  LUMO)  $(4b_1)_\beta(4b_1)_\alpha \rightarrow (2a_2)_\alpha(2a_2)_\beta$ , швидко зменшується (рис. 5.3 (a)). У той же час, спочатку невеликі за величиною амплітуди, які відповідають електронним переходам  $(3b_1)_\alpha(3b_1)_\beta \rightarrow (2a_2)_\alpha(2a_2)_\beta$  та  $(1a_2)_\alpha(1a_2)_\beta \rightarrow (2a_2)_\alpha(2a_2)_\beta$ , зменшуються значно повільніше. У кінці-кінців, при значенні  $\lambda \sim 0.7$ ,

залишаються лише три амплітуди в лінійній частині рівняння для CCD (5.1) і відповідна хвильова функція отримує наступний вигляд:

$$|\Psi_{L_I\text{-CCD}}(\lambda \sim 0.7)\rangle \approx |0\rangle + c_1 \left| \begin{smallmatrix} 5b_1 5\bar{b}_1 \\ 4b_1 4\bar{b}_1 \end{smallmatrix} \right\rangle + c_2 \left| \begin{smallmatrix} 3a_2 3\bar{a}_2 \\ 1b_1 1\bar{b}_1 \end{smallmatrix} \right\rangle + c_3 \left| \begin{smallmatrix} 2a_2 2\bar{a}_2 \\ 1a_2 1\bar{a}_2 \end{smallmatrix} \right\rangle + \text{n. c.} \quad (5.23)$$

Тут **n.c.** позначає сукупність нелінійних компонент, породжених вказаними амплітудами відповідно до (5.1). Виходячи з рис. 5.3(a), можна побачити, що величини амплітуд співвідносяться наступним чином:  $|c_1| \gg |c_2| \approx |c_3|$ . А функцію (5.23) можна інтерпретувати як найпростіше представлення (наближення) хвильової функції CCD для даної системи.

Подібна поведінка характерна й для інших молекул. Так, наприклад, для молекули **A** при значенні  $\lambda = 0.55$  існують лише п'ять конфігурацій з ненульовим внеском у хвильову функцію. Залежності амплітуд для цієї молекули наведено на рис. 5.3 (b), а відповідну наближену хвильову функцію можна представити наступним чином:

$$\begin{aligned} |\Psi_{L_I\text{-CCD}}(\lambda \sim 0.55)\rangle \approx & |0\rangle + t_{31}^{1'3'} \left( \left| \begin{smallmatrix} 1'3' \\ 31 \end{smallmatrix} \right\rangle + \left| \begin{smallmatrix} 3'1 \\ 13 \end{smallmatrix} \right\rangle \right) + t_{22}^{2'2'} \left| \begin{smallmatrix} 2'2' \\ 22 \end{smallmatrix} \right\rangle + t_{22}^{1'1'} \left| \begin{smallmatrix} 1'1' \\ 22 \end{smallmatrix} \right\rangle \\ & + t_{12}^{2'1'} \left( \left| \begin{smallmatrix} 2'1' \\ 12 \end{smallmatrix} \right\rangle + \left| \begin{smallmatrix} 1'2 \\ 21 \end{smallmatrix} \right\rangle \right) + t_{11}^{1'1'} \left| \begin{smallmatrix} 1'1' \\ 11 \end{smallmatrix} \right\rangle + \text{n.c.} \end{aligned} \quad (5.24)$$

П'ять конфігурацій у рівнянні (5.24) включають у себе збудження НОМО-LUMO  $\left| \begin{smallmatrix} 1'1' \\ 11 \end{smallmatrix} \right\rangle$ . Просте представлення (5.24) може бути корисним при розрахунках великих квазі-одновимірних полімерних систем, таких як нанотрубки.

Звісно, що в різних методах величини амплітуд можуть по різному залежати від  $\lambda$ . Для порівняння розрахунків, що виконано за допомогою різних методів, проведено розрахунки молекул **B** та **D** (див. рис. 5.2). Тут ми використовували методи  $L_I$ -MP2 та  $L_I$ -CCD (рис. 5.4).

З рис. 5.4 (a) можна побачити, що в розрахунку  $L_I$ -MP2 за малих значень регуляризуючого параметру або за його відсутності, амплітуда конфігурації  $(1,1) \rightarrow (4',4')$  за абсолютним значенням більша, за амплітуду конфігурації  $(6,6) \rightarrow (3',3')$ , але із збільшенням  $\lambda$  відносно положення амплітуд цих

конфігурацій змінюється й становиться таким самим, як і в  $L_I$ -CCD розрахунку (рис. 5.4 (b)).

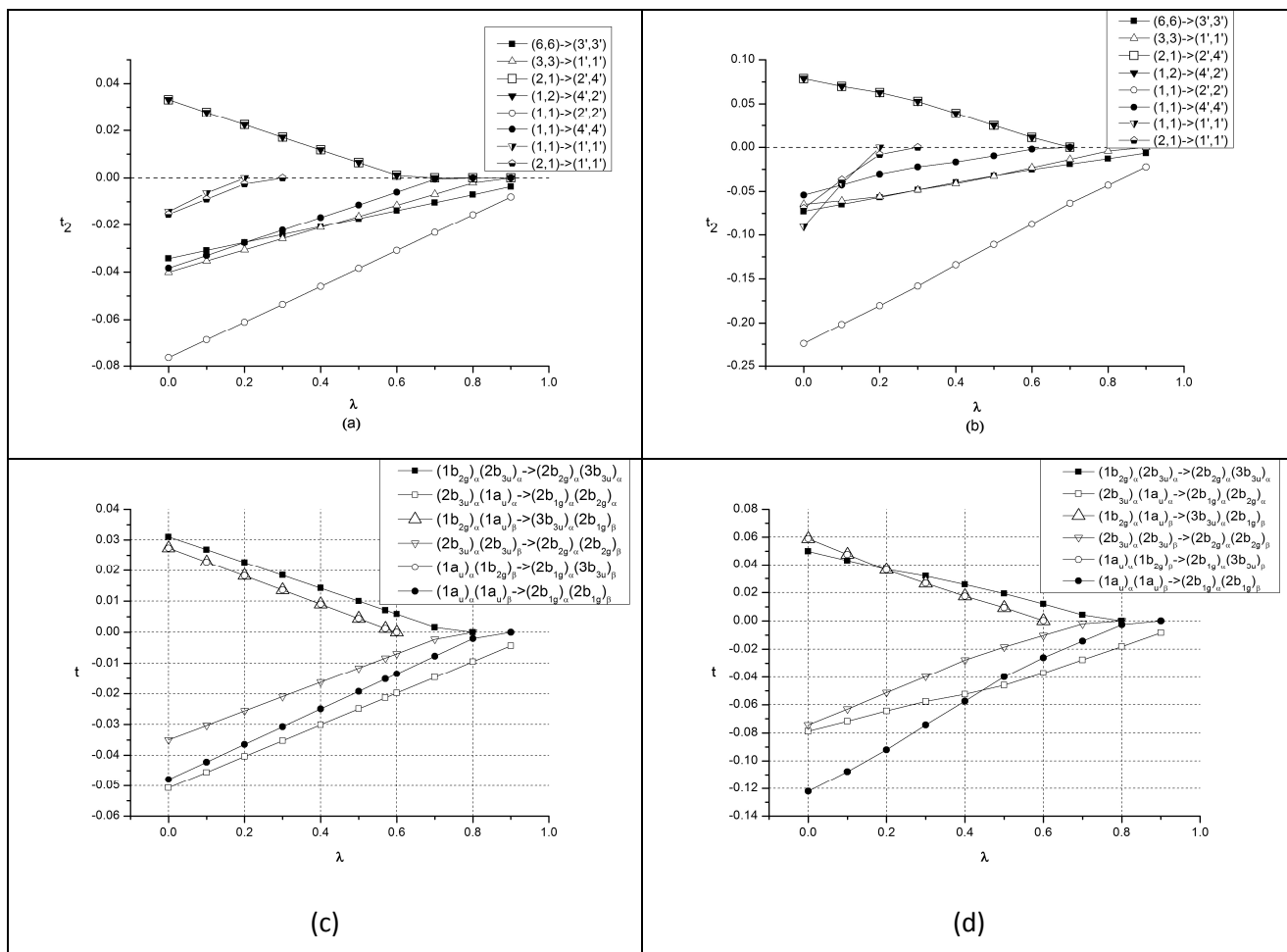


Рис. 5.4 Залежність величин амплітуд від величини параметру  $\lambda$  у методах  $L_I$ -MP2 (a, c) та  $L_I$ -CCD (b, d). Розрахунки молекули каліцену **B** (a, b), молекули нафталіну **D** (c, d)

Для молекули нафталіну (**D**) можна побачити, що амплітуда при конфігурації  $(1a_u)_\alpha(1a_u)_\beta \rightarrow (2b_{1g})_\alpha(2b_{1g})_\beta$  у  $L_I$ -CCD розрахунку (рис. 5.4 (d)) зменшується швидше, ніж друга за величиною амплітуда конфігурації  $(2b_{3u})_\alpha(1a_u)_\alpha \rightarrow (2b_{1g})_\alpha(2b_{2g})_\alpha$ . У  $L_I$ -MP2 розрахунку (рис. 5.4 (c)) ці амплітуди дещо різняться за відсутності регуляризуючого параметру, але, після введення  $\lambda$ , поведінка, а також швидкість зменшення цих амплітуд, є такою ж, як і в  $L_I$ -CCD.

Слід також зазначити, що із наведених залежностей (рис. 5.3 (c,d)) можна побачити, що однакові з причин спінової та просторової симетрії амплітуди  $(1b_{2g})_\alpha(1a_u)_\beta \rightarrow (3b_{3u})_\alpha(2b_{1g})_\beta$  та  $(1a_u)_\alpha(1b_{2g})_\beta \rightarrow (2b_{1g})_\alpha(3b_{3u})_\beta$  виявляються незмінно однаковими при скануванні  $\lambda$ . Таким чином, симетрія хвильової функції при внесенні регуляризуючої добавки не порушується.

У цілому для молекули **B** відповідно до рис. 5.4 (a,b) існує три конфігурації (як у методі  $L_I$ -CCD, так і в методі  $L_I$ -MP2), що дають найбільший внесок у формування хвильової функції. Таким чином, відповідне наближення для хвильової функції може бути представлено наступним чином:

$$|\Psi\rangle \approx |0\rangle + t_{11}^{2'2'} \left| \begin{smallmatrix} 2'2' \\ 11 \end{smallmatrix} \right\rangle + t_{66}^{3'3'} \left| \begin{smallmatrix} 3'3' \\ 66 \end{smallmatrix} \right\rangle + t_{33}^{1'1'} \left| \begin{smallmatrix} 1'1' \\ 33 \end{smallmatrix} \right\rangle + \text{n.c.} \quad (5.25)$$

Слід зазначити, що в даному рівнянні відсутній внесок НОМО-LUMO конфігурації!

Для молекули нафталіну (**D**), відповідно до рис. 5.4 (d), одним із найпростіших представлень CCD є:

$$\begin{aligned} |\Psi_{L_I\text{-CCD}}(\lambda \sim 0.7)\rangle \approx & |0\rangle + c_1 \left( \left| \begin{smallmatrix} 2b_{1g} 2b_{2g} \\ 2b_{3u} 1a_u \end{smallmatrix} \right\rangle + \left| \begin{smallmatrix} 2\bar{b}_{1g} 2\bar{b}_{2g} \\ 2\bar{b}_{3u} 1\bar{a}_u \end{smallmatrix} \right\rangle \right) + c_2 \left| \begin{smallmatrix} 2b_{1g} 2\bar{b}_{1g} \\ 1a_u 1\bar{a}_u \end{smallmatrix} \right\rangle + c_3 \left| \begin{smallmatrix} 2b_{2g} 2\bar{b}_{2g} \\ 2b_{3u} 2\bar{b}_{3u} \end{smallmatrix} \right\rangle \\ & + c_4 \left( \left| \begin{smallmatrix} 2b_{2g} 3b_{3u} \\ 1b_{2g} 2b_{3u} \end{smallmatrix} \right\rangle + \left| \begin{smallmatrix} 2\bar{b}_{2g} 3\bar{b}_{3u} \\ 1\bar{b}_{2g} 2\bar{b}_{3u} \end{smallmatrix} \right\rangle \right) + \text{n.c.} \end{aligned} \quad (5.26)$$

тут  $|c_1| > |c_2| > |c_3| > |c_4|$ .

З того ж рисунку також можна побачити, що найпростіша хвильова функція при  $\lambda \approx 0.9$ , окрім референтного стану, також включає в себе лише одну (просторову) двократно-збуджену конфігурацію:

$$|\Psi_{L_I\text{-CCD}}(\lambda \sim 0.9)\rangle \approx |0\rangle + c_1 \left( \left| \begin{smallmatrix} 2b_{1g} 2b_{2g} \\ 2b_{3u} 1a_u \end{smallmatrix} \right\rangle + \left| \begin{smallmatrix} 2\bar{b}_{1g} 2\bar{b}_{2g} \\ 2\bar{b}_{3u} 1\bar{a}_u \end{smallmatrix} \right\rangle \right). \quad (5.27)$$

Це практично відповідає методу обмеженої конфігураційної взаємодії за умови незалежного варіювання коефіцієнту  $c_1$ .

Перерізи, отримані за різних значень параметру  $\lambda$ , дозволяють отримувати компактні набори конфігурацій, що відповідають різним рівням наближень до точної хвильової функції.

Слід зазначити, що для обох молекул, за великих значень параметру  $\lambda$ , певні набори амплітуд мали ненульові значення як в методі  $L_I$ -MP2, так і в

методі  $L_I$ -CCD. Це цінна властивість, оскільки за таких обставин можна отримати наближений амплітудний набір у теоріях низького рівня. Після цього використовувати його в більш складних теоріях.

### 5.5. Неємпіричні розрахунки $L_I$ -СС

У неємпіричних (*ab initio*) розрахунках ми досліджували кілька простих модельних задач. Перша – дослідження молекули LiH, для якої параметр  $\lambda$  сканувався від нуля до достатньо великих позитивних значень (до зникнення з рівнянь усіх конфігурацій окрім однієї) з використанням методу  $L_I$ -CCD. Це дозволило зробити аналіз структури хвильової функції й сформулювати прості моделі для дослідження ефектів електронної кореляції. Друга модельна задача – дослідження молекули BH з використанням методу  $L_I$ -CCSD із зворотнім скануванням: від великих позитивних значень  $\lambda$  до нуля. При цьому була отримана ієрархія наближень до точного розв'язку рівнянь CCSD. Усі розрахунки проводилися в наближенні замороженого остова (*frozen core*).

Орбіталі Гартрі-Фока, а також одно- та двохелектронні інтеграли були розраховані для обраного базисного набору з використанням програмного пакету GAMESS<sup>99</sup>.

#### 5.5.1. $L_I$ -CCD та $L_I$ -CCSD розрахунки дисоціації молекули LiH

Розрахунок гетероядерної системи LiH є класичним тестом для дослідження якості результатів квантовохімічного методу щодо опису дисоціації одинарного зв'язку. У цих розрахунках нами використовувався стандартний базис 6-31G.

Перший розрахунок було виконано за рівноважної геометрії з довжиною зв'язку  $R_e = 1.64 \text{ \AA}$ . При цьому енергетична послідовність орбіталей Гартрі-Фоку виявилась наступною:

$$1\sigma^2 2\sigma^2 | 3\sigma 1p_x 1p_y 4\sigma 5\sigma 2p_x 2p_y 6\sigma 7\sigma. \quad (5.28)$$

Тут перші дві  $\sigma$ -орбіталі у хвильовій функції Гартрі-Фоку є зайнятими, усі інші – вакантні.

Для цієї системи ми спочатку проводимо розрахунки з використанням більш легкого рівня теорії з урахуванням тільки двократних збуджень  $L_I$ -CCD. Багатоконфігураційна хвильова функція для  $L_I$ -CCD методу була отримана за рахунок промотування двох електронів з найвищої зайнятої МО на незайняті МО:  $(2\sigma^2) \rightarrow (a,b)$ . Збудження з найнижчої остовної, зайнятої двома електронами МО  $1\sigma^2$ , не дуже впливають на результат розрахунку. Найбільша CCD амплітуда відповідає  $t_2(2\sigma^2 \rightarrow 6\sigma^2) = -0.0625$  конфігурації. Як і раніше, були отримані залежності кластерних амплітуд оператора збудження  $T_2$  від параметру  $\lambda$  (рис. 5.5). На графіку приводяться лише незайняті спін-орбіталі (a,b), на котрі відбувається двократний електронний перехід. При збільшенні регуляризуючого параметру вище значень  $\lambda=0.04$ , початкова відносна складна композиція хвильової функції спрощується, оскільки лише дві конфігурації  $(2\sigma^2) \rightarrow (6\sigma_\alpha, 6\sigma_\beta)$  та  $(2\sigma^2) \rightarrow (7\sigma_\alpha, 7\sigma_\beta)$  мають ненульовий внесок у хвильову функцію. При подальшому збільшенні регуляризуючого параметру, перша амплітуда з вказаних також зникає, таким чином, за значень регуляризуючого параметру  $\lambda > 0.055$  хвильова функція формується лише з однієї конфігурації  $(2\sigma^2) \rightarrow (7\sigma_\alpha, 7\sigma_\beta)$ , а також референсного стану.

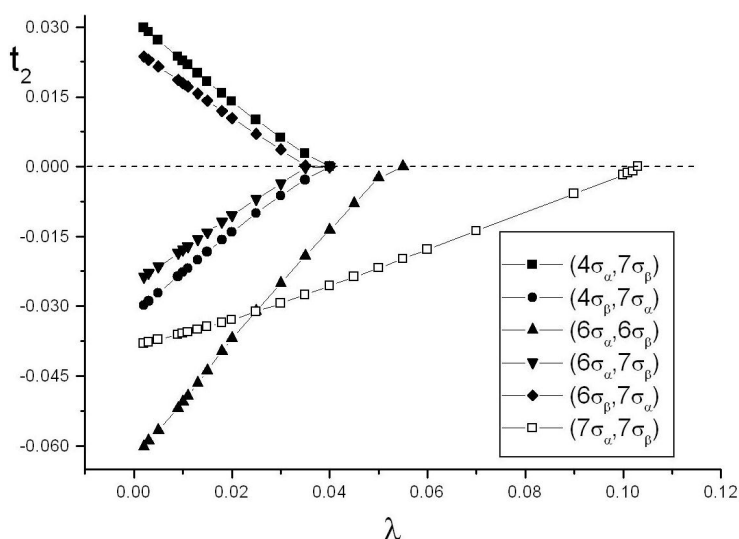


Рис. 5.5 Зміни найважливіших кластерних амплітуд  $t_2$ , що відповідають двократним збудженням з валентної зайнятої орбіталі  $2\sigma^2$  для молекули LiH у рівноважному стані ( $R_e = 1.64 \text{ \AA}$ ). Розрахунок  $L_I$ -CCD

Коли між'ядерна відстань зростає до  $R = 2R_e$ , енергетична послідовність орбіталей залишається такою ж, як і при  $R = R_e$  (5.28). Проте структура хвильової функції СС ускладнюється. Розрахунок  $L_I$ -CCD з орбіталями Гартрі-Фоку демонструє ускладнення в структурі збуджень (див. рис. 5.6). Найбільшою амплітудою ССД розрахунку є  $t_2(2\sigma^2 \rightarrow 4\sigma^2) = -0.195$ . І тільки за відносно великих значень параметру  $\lambda \geq 0.05$  у структурі хвильової функції залишаються лише дві домінантні конфігурації з переходом на вакантні орбіталі  $4\sigma$  та  $7\sigma$ . Відповідні значення кластерних амплітуд при цьому виявляються близькими за значеннями, але доволі малими.

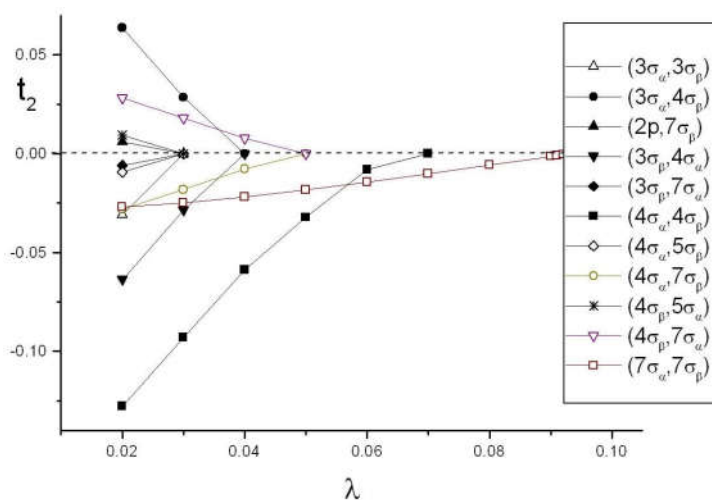


Рис. 5.6 Залежність  $t_2$  амплітуд від  $\lambda$  для молекули LiH за подвійної між'ядерної відстані  $R = 2R_e$ . Розрахунок  $L_I$ -CCD

Варіюванням регуляризуючого параметру, виходячи з графіків залежності амплітуд від  $\lambda$ , можна отримати різні наближенні розв'язки методу ССД. При цьому, при малих значеннях  $\lambda \leq 0.02$ , хвильова функція включає усі двократно-збуджені конфігурації, але вже при  $\lambda = 0.04$  хвильова функція стає більш компактною:

$$|\Psi'_{\text{CCD}}\rangle \approx |0\rangle - t_{2\sigma 2\sigma}^{4\sigma 4\sigma} \left| \begin{smallmatrix} 4\sigma_a & 4\sigma_p \\ 2\sigma_a & 2\sigma_p \end{smallmatrix} \right\rangle - t_{2\sigma 2\sigma}^{7\sigma 7\sigma} \left| \begin{smallmatrix} 7\sigma_a & 7\sigma_p \\ 2\sigma_a & 2\sigma_p \end{smallmatrix} \right\rangle + t_{2\sigma 2\sigma}^{4\sigma 7\sigma} \left( \left| \begin{smallmatrix} 4\sigma_a & 7\sigma_p \\ 2\sigma_a & 2\sigma_p \end{smallmatrix} \right\rangle + \left| \begin{smallmatrix} 7\sigma_a & 4\sigma_p \\ 2\sigma_a & 2\sigma_p \end{smallmatrix} \right\rangle \right) + \text{n. c.} \quad (5.29)$$

Для молекули LiH нелінійні компоненти (н.с.) – це чотирьохелектронні збудження, що включають у себе збудження з внутрішніх  $1\sigma$  спин-орбіталей. При  $0.05 \leq \lambda < 0.07$  хвильова функція спрощується до:

$$|\Psi''_{\text{CCD}}\rangle \approx |0\rangle - t_{2\sigma_2\sigma}^{4\sigma_4\sigma} \left| \begin{smallmatrix} 4\sigma_\alpha & 4\sigma_\beta \\ 2\sigma_\alpha & 2\sigma_\beta \end{smallmatrix} \right\rangle - t_{2\sigma_2\sigma}^{7\sigma_7\sigma} \left| \begin{smallmatrix} 7\sigma_\alpha & 7\sigma_\beta \\ 2\sigma_\alpha & 2\sigma_\beta \end{smallmatrix} \right\rangle + \text{n.c.} \quad (5.30)$$

І нарешті при  $\lambda > 0.07$  існує тільки одна двократно-збуджена конфігурація у хвильовій функції, яка має наступний вигляд:

$$|\Psi'''_{\text{CCD}}\rangle \approx |0\rangle - t_{2\sigma_2\sigma}^{7\sigma_7\sigma} \left| \begin{smallmatrix} 7\sigma_\alpha & 7\sigma_\beta \\ 2\sigma_\alpha & 2\sigma_\beta \end{smallmatrix} \right\rangle + \text{n.c.} \quad (5.31)$$

Оскільки наближення CCD не включає в себе одноелектронні збудження, CCD результати, отримані для між'ядерної відстані більшої ніж рівноважна, відрізняються від результатів, отриманих у наближенні CCSD, оскільки CCD не враховує ефектів орбітальної релаксації (які описуються амплітудами  $t_1$ ), що мають у цьому разі значний вплив. Результати  $L_I$ -CCSD розрахунку для молекули LiH при  $R=2R_e$  наведено на рис. 5.7.

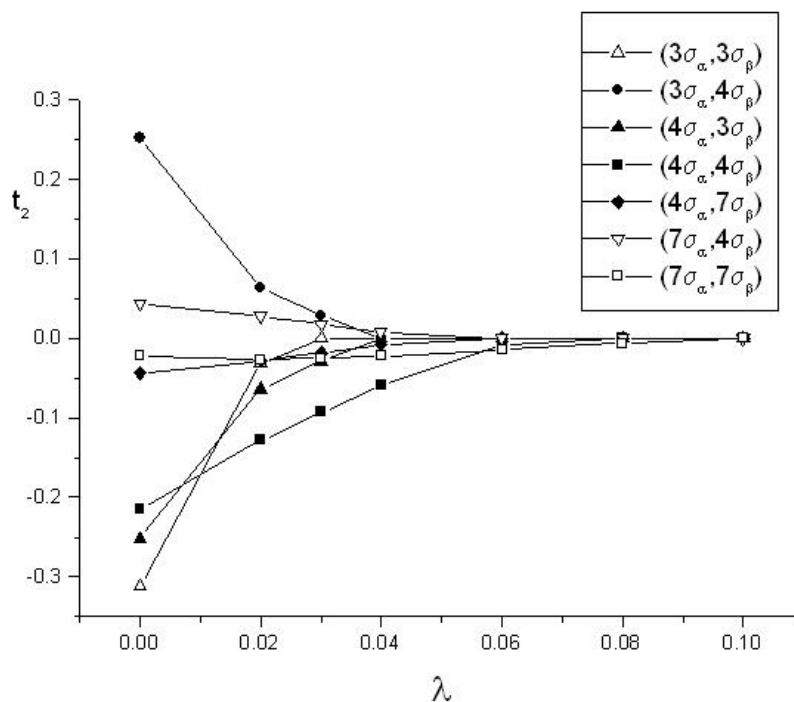


Рис. 5.7 Залежність CCSD  $t_2$  амплітуд від  $\lambda$  для молекули LiH за подвійної між'ядерної відстані  $R = 2R_e$ . Розрахунок  $L_I$ -CCSD

Тут можна бачити групу збуджень з відносно великими значеннями амплітуд ( $\sim 0.2$ - $0.3$ ). У цих збудженнях усі вакантні орбіталі, на які відбувається збудження з зайнятих орбіталей, мають розпушуючий характер. Результати розрахунку  $L_I$ -CCSD демонструють, що при  $R=2R_e$  існує кілька конфігурацій із



значним внеском у хвильову функцію навіть із збільшенням регуляризуючого параметру  $\lambda$ . Проте при  $\lambda > 0.02$ , домінуючий внесок до  $L_I$ -CCSD хвильової функції має збудження  $2\sigma_\alpha 2\sigma_\beta \rightarrow 4\sigma_\alpha 4\sigma_\beta$ .

У цьому розрахунку існує багато вакантних  $\sigma$ -орбіталей з розпушуючим або незв'язуючим характером. Із збільшенням параметру  $\lambda$  збудження НОМО-LUMO зникає з розрахунку, але в розрахунку все ще існують деякі вакантні орбіталі, що гарантують належний баланс "зв'язування-розпушування" у розрахунку.

Значно інша картина виникає при використанні орбіталей багатоконфігураційного методу самоузгодженого поля (*Many Configurational Selfconsistent field*, MCSCF) у розрахунку CCD. Просторові орбіталі MCSCF забезпечують компактне представлення хвильової функції в методі MCSCF. Отже, дві активні МО ( $2\sigma$  та  $3\sigma$ ) створюють найбільш компактний орбітальний набір і для розрахунків CCD. При використанні просторових орбіталей MCSCF навіть за малих значень  $\lambda$   $L_I$ -CCD хвильова функція має форму (рис. 5.8). Тут двократне збудження НОМО-LUMO, як й очікувалось, має домінантний внесок у формування хвильової функції. Внесок другої важливої конфігурації ( $2\sigma^2$ )  $\rightarrow (7\sigma_\alpha, 7\sigma_\beta)$  доволі малий.

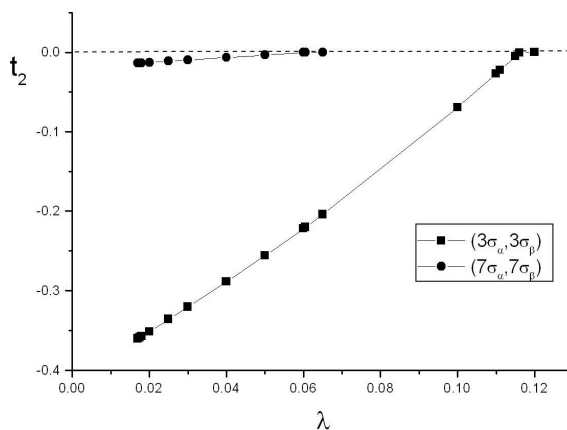


Рис. 5.8 Залежність  $L_I$ -CCD  $t_2$  амплітуд від  $\lambda$  для молекули LiH за подвійної між'ядерної відстані  $R = 2R_e$ . Розрахунок на основі орбіталей MCSCF

Для аналізу можливостей використання натуральних орбіталей MCSCF у мультиреференсній теорії СС заданого стану з повним урахуванням активного простору (CASCC) дивись (212).

### 5.5.2. $L_I$ -CCSD розрахунки молекул ВН

Двохатомна молекула ВН була розрахована з використанням  $L_I$ -CCSD підходу, а також стандартного базисного набору 6–31(d,p). Розрахунки були виконані як і для LiH: за рівноважної геометрії ( $R_e=1.23 \text{ \AA}$ ) та подовженої  $R = 2R_e$ . У цьому прикладі ми вивчали зміни енергії молекули за різного значення регуляризуючого параметру  $\lambda$ . Розрахунки спочатку проводилися за великих значень  $\lambda$ . Такий підхід дозволяв отримувати найбільш наближені, але не тривіальні (у розрахунках була присутня хоча б одна ненульова амплітуда  $t_j$ ) розв'язки методу CCSD. Після чого параметр  $\lambda$  знижувався до нуля. Таким чином, генерувалася ієрархія наближених розв'язків CCSD починаючи з дуже наближених та закінчуючи абсолютно точними. У розрахунках значення  $t_1$  та  $t_2$  амплітуд, отриманих для більш великих значень  $\lambda$ , використовувались у розрахунках з меншими  $\lambda$  у якості стартового наближення.

Результати розрахунків наведено в табл. 5.7. Розрахунки було проведено починаючи з  $\lambda = 0.01$ . У таблиці наведено процентну кількість ненульових  $t_1$  та  $t_2$  амплітуд (відповідно  $N(t_1)$  та  $N(t_2)$ ) у порівнянні з нерегуляризованим розрахунком. Також представлено процент урахування енергії кореляції ( $\epsilon_{\text{corr}}$ ) відносно стандартного CCSD розрахунку. Наведено різницю повних енергій для молекули із розтягнутим зв'язком ( $2R_e$ ) та молекули із рівноважною геометрією:

$$\Delta E = E(2R_e) - E(R_e). \quad (5.32)$$

Аналізуючи табл. 5.7, можна бачити, що лише 26%  $t_2$  амплітуд залишаються ненульовими за найбільших значень  $\lambda$ , у той час як  $t_1$  амплітуди майже зникають зовсім. За таких умов розрахунком  $L_I$ -CCSD було отримано 65 % енергії кореляції за нормальної між'ядерної відстані. За розтягнутої між'ядерної відстані в методі  $L_I$ -CCSD було відтворено 79% енергії кореляції.

Слід зазначити, що навіть за такої наближеної моделі до точного CCSD розрахунку було отримано  $\Delta E = 0.099$  ат. од., у той час як точний розрахунок дає  $\Delta E = 0.104$  а.о. Для найменшого наведеного значення  $\lambda$  кількість  $L_I$ -CCSD  $N(t_2)$  амплітуд дорівнювала 74 % за рівноважної геометрії і 52% для відстані  $2R_e$ , при цьому енергія кореляції була розрахована доволі точно: 96% та 98% відповідно.

Таблиця 5.7

**Проценти ненульових амплітуд ( $N(t_2)$  та  $N(t_1)$ ) і частки енергії електронної кореляції ( $\epsilon_{\text{corr}}$ ) відносно величин методу CCSD для двох між'ядерних відстаней, а також їх різниця для молекули ВН (метод  $L_I$ -CCSD)**

$\lambda$	$R_e = 1.23 \text{ \AA}$			$2R_e$			$\Delta E,$ (ат. од.)
	$N(t_2)$	$N(t_1)$	$\epsilon_{\text{corr}}$	$N(t_2)$	$N(t_1)$	$\epsilon_{\text{corr}}$	
0.01	26	0	65	13	15	79	0.099
0.006	39	6	78	19	23	87	0.101
0.005	43	12	81	25	27	89	0.101
0.002	65	34	92	42	42	95	0.103
0.001	74	40	96	52	54	98	0.104
0	100	100	100	100	100	100	0.104

Таким чином, ми робимо висновок, що метод  $L_I$ -CCSD практично для всіх значень параметру  $\lambda$  достатньо точно відтворив значення енергії кореляції, що отримані в CCSD розрахунку. При цьому розрахункові витрати відповідно до кількості амплітуд, що приймають участь у розрахунку, значно знижуються.

Фрагмент кривої потенційної енергії для молекули ВН за різних між'ядерних відстаней ( $R_e - 4R_e$ ) розрахований з використанням методу  $L_I$ -CCSD представлено на рис. 5.9.

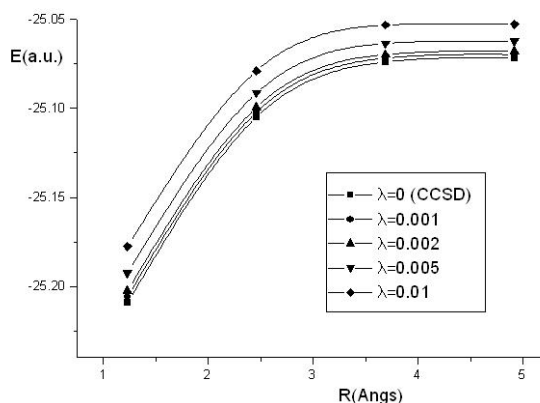


Рис. 5.9 Крива потенційної енергії ВН розрахована в методі  $L_1$ -CCSD за різних значень параметру  $\lambda$

Можна побачити, що криві потенціальної енергії, що отримані за різних значень параметру  $\lambda$  якісно схожі. Розрахунки демонструють доволі низьку похибку непаралельності (*Nonparallelity error*, NPE), що визначається як різниця між найбільшим та найменшим відхиленням від точної кривої, розрахованої в стандартному CCSD розрахунку ( $\lambda=0$ ):

$\lambda$	0.001	0.002	0.005	0.006	0.01
NPE (ат. од.)	0.0015	0.0031	0.0073	0.0086	0.0127

Включення збуджень вищої кратності (трьохкратних збуджень  $T_3$ , метод  $L_1$ -CCSDT) у розрахунок також демонструє непогану стабільність результатів (см. табл. 5.8).

Таблиця 5.8

**Середній процент ненульових амплітуд  $N$ , а також енергетична різниця (у ат. од.) для двох між'ядерних відстаней ( $R_e$ ,  $2R_e$ ) для молекули ВН (метод  $L_1$ -CCSDT)**

$\lambda$	0	0.002	0.003	0.004	0.005	0.01
$\Delta E$ , ат. од.	0.101	0.100	0.100	0.099	0.099	0.098
$N$ , $R_e$	100	65	57	48	43	26
$N$ , $2R_e$	100	44	35	30	27	14

Навіть за доволі наближеного розрахунку ( $\lambda = 0.01$ ), величина (5.32) дорівнює  $\Delta E = 0.098$  ат. од., що досить близько до точного значення  $\Delta E = 0.101$  ат. од.

### 5.5.3. Симетрична дисоціація молекули води в теорії $L_I$ -CCSD

Симетрична дисоціація води є також яскравим прикладом ефективності  $L_I$ -регуляризації.  $L_I$ -CCSD розрахунки молекули води були виконані з використанням стандартного базисного набору cc-pVDZ. За рівноважній геометрії довжина зв'язку О-Н дорівнює  $R_e = 0.95 \text{ \AA}$ , а кут між зв'язками –  $104.5^\circ$ . Залежності отримані за симетричного збільшення довжин зв'язків обох О-Н груп до довжин  $R = 1.5R_e, 2R_e, 2.5R_e$  та  $3R_e$  від рівноважної  $R_e$  наведено на рис. 5.10.

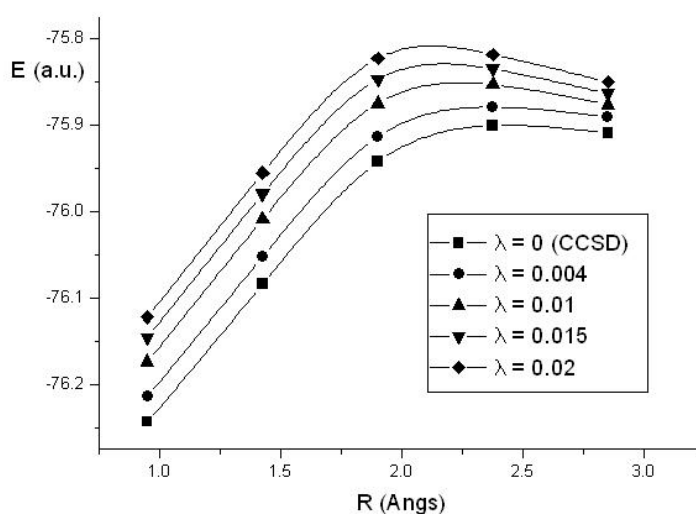


Рис. 5.10 Симетрична дисоціація молекули води в розрахунку  $L_I$ -CCSD

Тут можна побачити більші відхилення між повним CCSD розрахунком, а також наближеними моделями ( $\lambda > 0$ ).  $L_I$ -CCSD ( $\lambda = 0.004$ ) – крива потенціальної енергії з  $NPE = 0.014$  ат. од., більш паралельна до повного CCSD розрахунку, ніж за значення  $\lambda = 0.02$  ( $NPE = 0.069$  ат. од.). Тут слід зазначити, що лише 16%  $t_2$ -амплітуд приймали участь в останньому розрахунку.

Таким чином, із зростанням  $\lambda$  величина NPE для  $L_I$ -CCSD розрахунків може також зростати, що є зрозумілим наслідком зменшення конфігураційного складу. Слід зазначити, що такі властивості як розмірна узгодженість або NPE відіграють більш важливу роль при розтягнутих зв'язках. І хоч розрахунки, виконані для молекул з розтягнутими зв'язками, у порівнянні з рівноважною

геометрією, не є типовими для методу CC, наш підхід продемонстрував задовільну поведінку методу навіть у важких структурно-хімічних ситуаціях як-от: квазівиродження. Ми вважаємо, що значну перспективу метод регуляризації може отримати про використання в мультиреференсних теоріях як-от: метод CASCSD<sup>173</sup>. Таке узагальнення потребує подальшого дослідження.

## Висновки до розділу 5

1. Детально досліджено різні багатокрокові методи першого порядку для розв'язків рівнянь теорії зв'язаних кластерів. Встановлено ефективність простого методу "важкої кульки". Разом із тим, алгоритм DIIS з вірно підбраною довжиною інтерполяційного вектору дозволяє розв'язати навіть задачі зі значним орбітальним квазівиродженням.

2. За допомогою  $L_I$ -регуляризації, може бути створено ранжований набір електронно-збуджених конфігурацій. Його використання дозволяє створити прогресивну систему наближень квантовохімічного методу (зокрема теорії CC). Такі наближення, від найпростішого та "дешевого" у розрахунковому сенсі, до більш точного, але й потребуючого більших комп'ютерних витрат, створюють підґрунтя для систематичного використання прецизійних підходів до середніх за розміром систем. Загалом  $L_I$ -регуляризовані підходи надають універсальний спосіб скорочення комп'ютерних витрат без значних модифікацій розрахункових алгоритмів.

3. У рамках багаточастинкових підходів MP2, CCD, CCSD, CCSDT, розроблені в дисертації  $L_I$ -регуляризовані алгоритми продемонстрували свою ефективність як в напівемпіричних дослідженнях  $\pi$ -електронних систем (полієни, ароматичні системи), так і в *ab initio* розрахунках дисоціації малих молекул. Показано, що навіть короткі розклади хвильової функції  $L_I$ -CC здатні достатньо точно описати енергетичні характеристики молекул. Такі підходи можуть бути корисними, зокрема в теоретичних дослідженнях систем для молекулярної електроніки, де урахування електронної кореляції є критичним

при описі ефектів передачі електронних збуджень уздовж ланцюгу  $\pi$ -спряження<sup>213</sup> та нелінійно оптичних властивостей<sup>211</sup>.

4. Результати, що представляють систематику електронно-збуджених конфігурацій, яка базується на ідеї регуляризації, добре узгоджуються з хімічною інтуїцією.  $L_I$ -розвязки дозволяють створити достатньо короткі розклади хвильової функції, що надає змогу адекватно описати систему на якісному рівні. Разом з тим,  $L_I$ -MP2 метод дає систему наближень для високоточних методів і дозволяє створити адекватний активний простір для мультиреференсних підходів.

Основні положення цього розділу викладено в публікаціях автора (214-218).

## ЗАГАЛЬНІ ВИСНОВКИ

Встановлено ефективність  $L_1$ -регуляризації як універсального підходу до скорочення набору незалежних параметрів-змінних основних рівнянь задач, які стосуються прогнозу фізико-хімічних характеристик молекул.

1. Використання  $L_1$ -регуляризації в задачах побудови прогностичних моделей фізико-хімічних властивостей молекул дозволяє створити однозначний впорядкований ряд дескрипторів, систематичне включення яких генерує послідовність альтернативних, достатньо простих, регресійних OLS чи LAD рівнянь зі зростаючою кількістю параметрів. Показано, що отримані малопараметричні рівняння в досліджених випадках мають значно кращі прогностичні характеристики, ніж результати стандартних PCR та PLS методів, у яких не виконується відбір дескрипторів.

2. Показано, що розвинений підхід дозволяє сформулювати адекватні регресійні рівняння для ряду важливих фізико-хімічних параметрів, серед яких: рКа, температури кипіння, в'язкість. Отримані рівняння були побудовані для ситуацій, коли навчаючі вибірки включали структурно-різноманітні системи.

3. Обрані за допомогою процедури  $L_1$ -регуляризації мінімальні набори дескрипторів можуть з успіхом бути використані для створення штучної нейронної мережі. Це значно полегшує процедуру налаштування мережі й запобігає перенавчанню. Дані щодо прогнозу рКа різних органічних сполук характеризують представлений підхід до формування нейронної мережі як ефективний.

4. Розглянуті в дисертаційній роботі альтернативні способи побудови лінійної регресії (OLS, LAD, ODR та LADOD) дають можливість знаходження "ідеальної", з точки зору прогностичної здатності, моделі, яка може бути створена на заданому наборі дескрипторів. "Прагматичні" оцінки, які ґрунтуються на зіставленні результатів, що отримані різними методами, дозволяють не тільки обрати найкращу модель опису фізико-хімічних та біохімічних властивостей, але й оцінити якість дескрипторного набору.



5. Виходячи з модельних досліджень розбиттів на навчаючу і тестову вибірки, показано, що адекватна оцінка прогностичної здатності розробленої хемометричної моделі може бути досягнута лише за умови потрапляння розбиття близько до центру густини точок на залежності  $R_{test}^2 - R_{train}^2$ . При цьому бажано урахування границь застосування параметрів моделі (AD).

6. Досліджений  $L_1$ -регуляризований варіант методу логістичної регресії дозволяє створити ефективну класифікаційну функцію, що залежить від малого числа параметрів. Отримані для тестових експериментів результати (зокрема функції ROC) продемонстрували точність класифікації на рівні сучасних, складних для інтерпретації, багатопараметричних підходів.

7. Отримано класифікаційні функції для відбору органічних систем різної природи за основністю по відношенню до катіону літію та спорідненості до рецепторів естрогену. Представлені функції є досить простими й гарантують якість прогнозу.

8. Запропонований на основі  $L_1$ -регуляризації новий підхід до скорочення "довжини" розкладу хвильової функції дав можливість створити послідовність наближень до точного квантовохімічного методу. Підхід реалізовано в рамках багаточастинкової теорії збурень ( $L_1$ -MP2) та теорії зв'язаних кластерів ( $L_1$ -CCD,  $L_1$ -CCSD,  $L_1$ -CCSDT). Встановлено, зокрема, що в рамках теорії зв'язаних кластерів, на прикладі малих систем (молекули BH, LiH, H<sub>2</sub>O), наближені розв'язки рівняння Шредингера за малої кількості параметрів доволі точно відтворюють значення енергетичних параметрів системи.

9. Напівемпіричні розрахунки  $\pi$ -спряжених систем, що були проведені з використанням запропонованих у дисертації методів  $L_1$ -MP2,  $L_1$ -CCD,  $L_1$ -CCSD, продемонстрували перспективність нових методів у проблемі опису електронної будови систем для молекулярної електроніки.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] Sivanandam, S. N.; Deepa, S. N. Genetic Algorithms BT - Introduction to Genetic Algorithms; Sivanandam, S. N., Deepa, S. N., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 15–37. [https://doi.org/10.1007/978-3-540-73190-0\\_2](https://doi.org/10.1007/978-3-540-73190-0_2).
- [2] Sastry, K.; Goldberg, D.; Kendall, G. Genetic Algorithms BT - Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques; Burke, E. K., Kendall, G., Eds.; Springer US: Boston, MA, 2005; pp 97–125. [https://doi.org/10.1007/0-387-28356-0\\_4](https://doi.org/10.1007/0-387-28356-0_4).
- [3] Dorigo, M.; Birattari, M.; Stutzle, T. Ant Colony Optimization. *IEEE Comput. Intell. Mag.* **2006**, 1 (4), 28–39. <https://doi.org/10.1109/MCI.2006.329691>.
- [4] Dorigo, M.; Stützle, T. Ant Colony Optimization: Overview and Recent Advances. *Int. Ser. Oper. Res. Manag. Sci.* **2019**, 272, 311–351. [https://doi.org/10.1007/978-3-319-91086-4\\_10](https://doi.org/10.1007/978-3-319-91086-4_10).
- [5] Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. <https://doi.org/10.1198/tast.2009.08199> **2012**, 63 (4), 308–319. <https://doi.org/10.1198/TAST.2009.08199>.
- [6] Vladimir Svetnik, \*,†; Andy Liaw, †; Christopher Tong, †; J. Christopher Culberson, ‡; Robert P. Sheridan, § and; Feuston‡, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (6), 1947–1958. <https://doi.org/10.1021/CI034160G>.
- [7] Hawkins\*, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2003**, 44 (1), 1–12. <https://doi.org/10.1021/CI0342472>.
- [8] Демиденко, Е. З. *Линейная и Нелинейная Регрессии*; Финансы и статистика, 1981.
- [9] Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. Москва: Наука, 1986. С. 232.

- [10] Björck, Å. Least Squares Methods. *Handb. Numer. Anal.* **1990**, *1*, 465–652. [https://doi.org/10.1016/S1570-8659\(05\)80036-5](https://doi.org/10.1016/S1570-8659(05)80036-5).
- [11] Hastie, T.; Tibshirani, R.; Wainwright, M. Statistical Learning with Sparsity : The Lasso and Generalizations. *Stat. Learn. with Sparsity Lasso Gen.* **2015**, 1–337. <https://doi.org/10.1201/B18401>.
- [12] Cao, W.; Sun, J.; Xu, Z. Fast Image Deconvolution Using Closed-Form Thresholding Formulas of  $L_q(Q=1/2,2/3)$  Regularization. *J. Vis. Commun. Image Represent.* **2013**, *24* (1), 31–41. <https://doi.org/10.1016/J.JVCIR.2012.10.006>.
- [13] Xu, Z.; Chang, X.; Xu, F.; Zhang, H.  $L_{1/2}$  Regularization: A Thresholding Representation Theory and a Fast Solver. *IEEE Trans. Neural Networks Learn. Syst.* **2012**, *23* (7), 1013–1027. <https://doi.org/10.1109/TNNLS.2012.2197412>.
- [14] Willoughby, R. A. Solutions of Ill-Posed Problems (A. N. Tikhonov and V. Y. Arsenin). *SIAM Rev.* **1979**, *21* (2), 266–267. <https://doi.org/10.1137/1021044>.
- [15] Morozov, V. A.; Stessin, M. I. Regularization Methods for Ill-Posed Problems. **1993**, 257.
- [16] Marquardt, D. W.; Snee, R. D. Ridge Regression in Practice. *Am. Stat.* **1975**, *29* (1), 3–20. <https://doi.org/10.1080/00031305.1975.10479105>.
- [17] McDonald, G. C. Ridge Regression. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1* (1), 93–100. <https://doi.org/10.1002/WICS.14>.
- [18] Penrose, R. A Generalized Inverse for Matrices. *Math. Proc. Cambridge Philos. Soc.* **1955**, *51* (3), 406–413. <https://doi.org/10.1017/S0305004100030401>.
- [19] Rao, C.; Mitra, S. Further Contributions to the Theory of Generalized Inverse of Matrices and Its Applications. *Sankhya* **1971**, *33* (3), 289–300.
- [20] Chattefuee, S.; Hadi, A. S. *Regression Analysis by Example: Fourth Edition*; Wiley Blackwell, 2006. <https://doi.org/10.1002/0470055464>.
- [21] Schreiber-Gregory, D. N. Ridge Regression and Multicollinearity: An in-Depth Review. *Model Assist. Stat. Appl.* **2018**, *13* (4), 359–365. <https://doi.org/10.3233/MAS-180446>.

- [22] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, 58 (1), 267–288. <https://doi.org/10.1111/J.2517-6161.1996.TB02080.X>.
- [23] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
- [24] Hastie, T.; Tibshirani, R.; Tibshirani, R. J. Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. **2017**.
- [25] Cui, A.; Peng, J.; Li, H.; Wen, M.; Jia, J. Iterative Thresholding Algorithm Based on Non-Convex Method for Modified Lp-Norm Regularization Minimization. *J. Comput. Appl. Math.* **2019**, 347, 173–180. <https://doi.org/10.1016/J.CAM.2018.08.021>.
- [26] Xu, H.-K. Properties and Iterative Methods for the Lasso and Its Variants. *Chinese Ann. Math. Ser. B* 2014 353 **2014**, 35 (3), 501–518. <https://doi.org/10.1007/S11401-014-0829-9>.
- [27] Gauraha, N. Introduction to the LASSO. *Reson.* 2018 234 **2018**, 23 (4), 439–464. <https://doi.org/10.1007/S12045-018-0635-X>.
- [28] Singh, D.; Singh, B. Investigating the Impact of Data Normalization on Classification Performance. *Appl. Soft Comput.* **2020**, 97, 105524. <https://doi.org/10.1016/J.ASOC.2019.105524>.
- [29] Schmidt, M. *Least Squares Optimization with L1-Norm Regularization*; 2005; Vol. 98. <https://people.duke.edu/~hpgavin/SystemID/References/Schmidt-LASSO-2005.pdf> (accessed 1 April 2019)
- [30] Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, 96 (456), 1348–1360. <https://doi.org/10.1198/016214501753382273>.
- [31] Beck, A.; Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.* **2009**, 2 (1), 183–202. <https://doi.org/10.1137/080716542>.

- [32] Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R.; Ishwaran, H.; Knight, K.; Loubes, J. M.; Massart, P.; Madigan, D.; Ridgeway, G.; Rosset, S.; Zhu, J. I.; Stine, R. A.; Turlach, B. A.; Weisberg, S.; Johnstone, I.; Tibshirani, R. Least Angle Regression. *Ann. Stat.* **2004**, *32* (2), 407–499. <https://doi.org/10.1214/009053604000000067>.
- [33] Tibshirani, R. J. The Lasso Problem and Uniqueness. *Electron. J. Stat.* **2013**, *7* (1), 1456–1490. <https://doi.org/10.1214/13-EJS815>.
- [34] Wesolowsky, G. O. A New Descent Algorithm for The Least Absolute Value Regression Problem. *Commun. Stat. - Simul. Comput.* **1981**, *10* (5), 479–491. <https://doi.org/10.1080/03610918108812224>.
- [35] Bloomfield, P.; Steiger, W. L. *Least Absolute Deviations*; Birkhäuser Boston, 1984. <https://doi.org/10.1007/978-1-4684-8574-5>.
- [36] Ahn, S. J. *Least Squares Orthogonal Distance Fitting of Curves and Surfaces in Space*; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; Vol. 3151. <https://doi.org/10.1007/B104017>.
- [37] Lenth, R. V.; Huffel, S. Van; Vandewalle, J.; Lawson, C. L.; Hanson, R. J.; Gauss, C. F.; Stewart, G. W. *The Total Least Squares Problem: Computational Aspects and Analysis*; 1999; Vol. 94. <https://doi.org/10.2307/2670017>.
- [38] Onizhuk, M. O.; Ivanov, V. V; Panteleimonov, A. V; Kholin, Y. V. Alternative Methods for Constructing of Linear Regressions. *Methods Objects Chem. Anal.* **2017**, *12* (3), 105–111. <https://doi.org/10.17721/moca.2017.105-111>.
- [39] Berdnyk, M. I.; Onizhuk, M. O.; Ivanov, V. V; Karazin, V. N. Methods for Building Linear Regression Equations in the “Structure-Property” Problems. *Kharkov Univ. Bull. Chem. Ser.* **2018**, No. 30, 6–17. <https://doi.org/10.26565/2220-637x-2018-30-01>.
- [40] *Statisticians of the Centuries*; Springer New York, 2001. <https://doi.org/10.1007/978-1-4613-0179-0>.
- [41] Мудров, В. И.; Кушко, В. Л. Метод Наименьших Модулей. *М. Знание* **1971**, *64*.
- [42] Розенфельд, Б. А. *Многомерные Пространства*. Рипол Классик **2013**.

- [43] Abdi, H.; Williams, L. J. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. John Wiley & Sons, Ltd July 1, 2010, pp 433–459. <https://doi.org/10.1002/wics.101>.
- [44] Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, 2 (1–3), 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9).
- [45] Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, 185 (C), 1–17. [https://doi.org/10.1016/0003-2670\(86\)80028-9](https://doi.org/10.1016/0003-2670(86)80028-9).
- [46] Andersson, M. A Comparison of Nine PLS1 Algorithms. *J. Chemom.* **2009**, 23 (10), 518–529. <https://doi.org/10.1002/cem.1248>.
- [47] Wold, S.; Ruhe, A.; Wold, H.; W. J. Dunn, I. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. <http://dx.doi.org/10.1137/0905052> **2006**, 5 (3), 735–743. <https://doi.org/10.1137/0905052>.
- [48] Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, 27 (8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
- [49] Frédéric Bonnans, J.; Charles Gilbert, J.; Lemaréchal, C.; Sagastizábal, C. A. *Numerical Optimization: Theoretical and Practical Aspects*; Springer Berlin Heidelberg, 2006. <https://doi.org/10.1007/978-3-540-35447-5>.
- [50] Veerasamy, R.; Rajak, H.; Jain, A.; Sivadasan, S.; Varghese, C. P.; Agrawal, R. K. Ravichandran Veerasamy, et Al : Validation of QSAR Models-Strategies and Importance Validation of QSAR Models -Strategies and Importance. *Int. J. Drug Des. Discov.* **2011**, 2 (3), 511–519.
- [51] Toropova, A. P.; Toropov, A. A. Does the Index of Ideality of Correlation Detect the Better Model Correctly? *Mol. Inform.* **2019**, 38 (8–9), 1800157. <https://doi.org/10.1002/minf.201800157>.
- [52] Toropov, A. A.; Raška, I.; Toropova, A. P.; Raškova, M.; Veselinović, A. M.; Veselinović, J. B. The Study of the Index of Ideality of Correlation as a New Criterion of Predictive Potential of QSPR/QSAR-Models. *Sci. Total Environ.* **2019**, 659, 1387–1394. <https://doi.org/10.1016/j.scitotenv.2018.12.439>.

- [53] Roy, K.; Mitra, I.; Kar, S.; Ojha, P. K.; Das, R. N.; Kabir, H. Comparative Studies on Some Metrics for External Validation of QSPR Models. *J. Chem. Inf. Model.* **2012**, 52 (2), 396–408. <https://doi.org/10.1021/CI200520G>.
- [54] Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. *J. Chem. Inf. Model.* **2012**, 52 (8), 2044–2058. <https://doi.org/10.1021/ci300084j>.
- [55] Andrada, M. F.; Vega-Hissi, E. G.; Estrada, M. R.; Garro Martinez, J. C. Impact Assessment of the Rational Selection of Training and Test Sets on the Predictive Ability of QSAR Models. *SAR QSAR Environ. Res.* **2017**, 28 (12), 1011–1023. <https://doi.org/10.1080/1062936X.2017.1397056>.
- [56] Kar, S.; Roy, K.; Leszczynski, J. Applicability Domain: A Step toward Confident Predictions and Decidability for QSAR Modeling. In *Methods in Molecular Biology*; Humana Press, New York, NY, 2018; Vol. 1800, pp 141–169. [https://doi.org/10.1007/978-1-4939-7899-1\\_6](https://doi.org/10.1007/978-1-4939-7899-1_6).
- [57] Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, 52 (10), 2570–2578. <https://doi.org/10.1021/ci300338w>.
- [58] Golbraikh, A.; Tropsha, A. Beware of Q<sup>2</sup>! In *Journal of Molecular Graphics and Modelling*; Elsevier, 2002; Vol. 20, pp 269–276. [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1).
- [59] Kubinyi, H. CINF 84-Herman Skolnik Award Lecture: Why Models Fail. In *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*; 2006; Vol. 232.
- [60] Hawkins, D. M. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*. American Chemical Society January 2004, pp 1–12. <https://doi.org/10.1021/ci0342472>.
- [61] Wiedermann, W.; Hagmann, M. Asymmetric Properties of the Pearson Correlation Coefficient: Correlation as the Negative Association between Linear



- Regression Residuals. *Commun. Stat. - Theory Methods* **2016**, 45 (21), 6263–6283. <https://doi.org/10.1080/03610926.2014.960582>.
- [62] Kaneko, H. Estimation of Predictive Performance for Test Data in Applicability Domains Using Y-Randomization. *J. Chemom.* **2019**, 33 (9), e3171. <https://doi.org/10.1002/cem.3171>.
- [63] Rücker, C.; Rücker, G.; Meringer, M. Y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, 47 (6), 2345–2357. <https://doi.org/10.1021/ci700157b>.
- [64] Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R<sup>2</sup>: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, 55 (7), 1316–1322. <https://doi.org/10.1021/acs.jcim.5b00206>.
- [65] Schüürmann, G.; Ebert, R. U.; Chen, J.; Wang, B.; Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* **2008**, 48 (11), 2140–2145. <https://doi.org/10.1021/ci800253u>.
- [66] Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q<sub>2</sub> Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, 49 (7), 1669–1678. <https://doi.org/10.1021/ci900115y>.
- [67] Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, 41 (1), 186–195. <https://doi.org/10.1021/ci000066d>.
- [68] Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of Model Predictive Ability by External Validation Techniques. *J. Chemom.* **2010**, 24 (3–4), 194–201. <https://doi.org/10.1002/cem.1290>.
- [69] Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models: How to Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J. Chem. Inf. Model.* **2011**, 51 (9), 2320–2335. <https://doi.org/10.1021/ci200211n>.



- [70] Gramatica, P.; Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. *J. Chem. Inf. Model.* **2016**, *56* (6), 1127–1131. <https://doi.org/10.1021/ACS.JCIM.6B00088>.
- [71] Roy, K.; Kar, S. Importance of Applicability Domain of QSAR Models. In *Pharmaceutical Sciences*; IGI Global, 2016; pp 1012–1043. <https://doi.org/10.4018/978-1-5225-1762-7.ch039>.
- [72] Leonard, J. T.; Roy, K. On Selection of Training and Test Sets for the Development of Predictive QSAR Models. *QSAR Comb. Sci.* **2006**, *25* (3), 235–251. <https://doi.org/10.1002/qsar.200510161>.
- [73] Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel K-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *J. Cheminform.* **2013**, *5* (5), 1–9. <https://doi.org/10.1186/1758-2946-5-27>.
- [74] Lloyd, S. P. Least Square Quantization in PCM. Bell Telephone Laboratories Paper. Published in Journal Much Later: Lloyd, SP: Least Squares Quantization in PCM. *IEEE Trans. Inform. Theor.*(1957/1982) **1957**, *18*.
- [75] Xu, J.; Hagler, A. Chemoinformatics and Drug Discovery. *Molecules*. Molecular Diversity Preservation International August 30, 2002, pp 566–600. <https://doi.org/10.3390/70800566>.
- [76] Cynthia D. Selassie, \*; Rajni Garg; Sanjay Kapur; Alka Kurup; Rajeshwar P. Verma; Suresh Babu Mekapati, and; Hansch, C. Comparative QSAR and the Radical Toxicity of Various Functional Groups. *Chem. Rev.* **2002**, *102* (7), 2585–2605. <https://doi.org/10.1021/CR940024M>.
- [77] Hansch, C.; Gao, H. Comparative QSAR: Radical Reactions of Benzene Derivatives in Chemistry and Biology. *Chem. Rev.* **1997**, *97* (8), 2995–3060. <https://doi.org/10.1021/cr9601021>.
- [78] M. Reinhard and A. Drefahl. *Handbook for Estimating Physicochemical Properties of Organic Compounds*; Wiley, 1999.

- [79] Mackay, D.; Shiu, W.-Y.; Lee, S. C. *Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals*; CRC press, 2006.
- [80] Quinones-Torrelo, C.; Martin-Biosca, Y.; Martinez-pla, J.; Sagrado, S.; Villanueva-Camanas, R.; Medina-Hernandez, M. QRAR Models for Central Nervous System Drugs Using Biopartitioning Micellar Chromatography. *Mini-Reviews Med. Chem.* **2002**, 2 (2), 145–161. <https://doi.org/10.2174/1389557024605519>.
- [81] Ghomishah, Z.; Gorji, A. E.; Sobati, M. A. Prediction of Critical Properties of Sulfur-Containing Compounds: New QSPR Models. *J. Mol. Graph. Model.* **2020**, 101, 107700. <https://doi.org/10.1016/j.jmgm.2020.107700>.
- [82] Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C. Predicting Explosibility Properties of Chemicals from Quantitative Structure-Property Relationships. *Process Saf. Prog.* **2010**, 29 (4), 359–371. <https://doi.org/10.1002/prs.10379>.
- [83] Jesús Jover; Ramón Bosque, A.; Sales\*, J. Determination of Lithium Cation Basicity from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, 44 (5), 1727–1736. <https://doi.org/10.1021/CI0498362>.
- [84] Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; Methods and Principles in Medicinal Chemistry; Wiley, 1993. <https://doi.org/10.1002/9783527616824>.
- [85] Marini, F. *Chemometrics in Food Chemistry*; Elsevier Science, 2013.
- [86] Roy, K.; Kar, S.; Das, R. N. *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*; Springer, 2015.
- [87] Roy, K. *Advances in QSAR Modeling*; Roy, K., Ed.; Challenges and Advances in Computational Chemistry and Physics; Springer International Publishing: Cham, 2017; Vol. 24. <https://doi.org/10.1007/978-3-319-56850-8>.
- [88] Kubinyi, H. From Narcosis to Hyperspace: The History of QSAR. *Quant. Struct. Relationships* **2002**, 21 (4), 348–356.
- [89] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry; Wiley, 2000; Vol. 11. <https://doi.org/10.1002/9783527613106>.

- [90] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Methods and Principles in Medicinal Chemistry; Wiley, 2009; Vol. 41. <https://doi.org/10.1002/9783527628766>.
- [91] Filzmoser, P.; Gschwandtner, M.; Todorov, V. Review of Sparse Methods in Regression and Classification with Application to Chemometrics. *Journal of Chemometrics*. John Wiley & Sons, Ltd March 1, 2012, pp 42–51. <https://doi.org/10.1002/cem.1418>.
- [92] Ing, C. K.; Lai, T. L. A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models. *Stat. Sin.* **2011**, 21 (4), 1473–1513. <https://doi.org/10.5705/ss.2010.081>.
- [93] Hosseinpour, M.; Sharifi, H.; Sharifi, Y. Stepwise Regression Modeling for Compressive Strength Assessment of Mortar Containing Metakaolin. *Int. J. Model. Simul.* **2018**, 38 (4), 207–215. <https://doi.org/10.1080/02286203.2017.1422096>.
- [94] Roberts, J.; Bursten, J. R. S.; Risko, C. Genetic Algorithms and Machine Learning for Predicting Surface Composition, Structure, and Chemistry: A Historical Perspective and Assessment. *Chem. Mater.* **2021**, 33 (17), 6589–6615. <https://doi.org/10.1021/acs.chemmater.1c00538>.
- [95] Niazi, A.; Leardi, R. Genetic Algorithms in Chemometrics. *Journal of Chemometrics*. John Wiley & Sons, Ltd June 1, 2012, pp 345–351. <https://doi.org/10.1002/cem.2426>.
- [96] Fuhrmann, J.; Rurainski, A.; Lenhof, H. P.; Neumann, D. A New Lamarckian Genetic Algorithm for Flexible Ligand-Receptor Docking. *J. Comput. Chem.* **2010**, 31 (9), 1911–1918. <https://doi.org/10.1002/jcc.21478>.
- [97] Tibshirani, R. The LASSO method for variable selection in the COX model. *Stat. Med.* **1997**, 16 (4), 385–395. [https://doi.org/10.1002/\(SICI\)1097-0258\(19970228\)16:4<385::AID-SIM380>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3).
- [98] Vidaurre, D.; Bielza, C.; Larrañaga, P. A Survey of L1 Regression. *Int. Stat. Rev.* **2013**, 81 (3), 361–387. <https://doi.org/10.1111/insr.12023>.
- [99] Barca, G. M. J.; Bertoni, C.; Carrington, L.; Datta, D.; Silva, N. De; Deustua, J. E.; Fedorov, D. G.; Gour, J. R.; Gunina, A. O.; Guidez, E.; Harville, T.; Irle, S.;

- Ivanic, J.; Kowalski, K.; Leang, S. S.; Li, H.; Li, W.; Lutz, J. J.; Magoulas, I.; Mato, J.; Mironov, V.; Nakata, H.; Pham, B. Q.; Piecuch, P.; Poole, D.; Pruitt, S. R.; Rendell, A. P.; Roskop, L. B.; Ruedenberg, K.; Sattasathuchana, T.; Schmidt, M. W.; Shen, J.; Slipchenko, L.; Sosonkina, M.; Sundriyal, V.; Tiwari, A.; Vallejo, J. L. G.; Westheimer, B.; Włoch, M.; Xu, P.; Zahariev, F.; Gordon, M. S. Recent Developments in the General Atomic and Molecular Electronic Structure System. *J. Chem. Phys.* **2020**, *152* (15), 154102. <https://doi.org/10.1063/5.0005188>.
- [100] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual Computational Chemistry Laboratory – Design and Description. *J. Comput. Mol. Des.* **2005**, *19* (6), 453–463. <https://doi.org/10.1007/S10822-005-8694-Y>.
- [101] Virtual Computational Chemistry Laboratory <http://www.vcclab.org/> (accessed Sep 27, 2021).
- [102] Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474. <https://doi.org/10.1002/JCC.21707/ABSTRACT>.
- [103] Bahmani, S. *Algorithms for Sparsity-Constrained Optimization*; Springer Science & Business Media, 2013; Vol. 261.
- [104] Albert, A. *The Determination of Ionization Constants: A Laboratory Manual*; Springer Netherlands, 2012.
- [105] Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. Absolute PKa Determinations for Substituted Phenols. *J. Am. Chem. Soc.* **2002**, *124* (22), 6421–6427. <https://doi.org/10.1021/ja012474j>.
- [106] Jensen, J. H.; Swain, C. J.; Olsen, L. Prediction of PKa Values for Druglike Molecules Using Semiempirical Quantum Chemical Methods. *J. Phys. Chem. A* **2017**, *121* (3), 699–707. <https://doi.org/10.1021/ACS.JPCA.6B10990>.
- [107] NeuPy — NeuPy <http://neupy.com/pages/home.html> (accessed Sep 27, 2021).
- [108] Zefirov, N. S.; Palyulin, V. A. QSAR for Boiling Points of “Small” Sulfides. Are the “High-Quality Structure-Property-Activity Regressions” the Real High

- Quality QSAR Models? *J. Chem. Inf. Comput. Sci.* **2001**, *41* (4), 1022–1027. <https://doi.org/10.1021/ci0001637>.
- [109] Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 835–845. <https://doi.org/10.1021/ci980339t>.
- [110] Rücker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points. *J. Chem. Inf. Model.* **2005**, *45* (1), 74–80. <https://doi.org/10.1021/ci0497298>.
- [111] List of molecular descriptors calculated by Dragon [http://www.taletе.mi.it/products/dragon\\_molecular\\_descriptor\\_list.pdf](http://www.taletе.mi.it/products/dragon_molecular_descriptor_list.pdf) (accessed Sep 27, 2021).
- [112] Suzuki, T.; Ohtaguchi, K.; Koide, K. Computer-Assisted Approach to Develop a New Prediction Method of Liquid Viscosity of Organic Compounds. *Comput. Chem. Eng.* **1996**, *20* (2), 161–173. [https://doi.org/10.1016/0098-1354\(94\)00012-D](https://doi.org/10.1016/0098-1354(94)00012-D).
- [113] Berdnyk, M. I.; Zakharov, A. B.; Ivanov, V. V. Application Of  $L_1$ -Regularization Approach In QSAR Problem. Linear Regression And Artificial Neural Networks. *Methods Objects Chem. Anal.* **2019**, *14* (2), 79–90. <https://doi.org/10.17721/moca.2019.79-90>.
- [114] Бердник, М.И.; Дяченко, А.В.; Иванов, В. В; Регрессионные модели QSAR, *Збірник тез доповідей, Хімічні Проблеми Сьогодення (ХІПС-2018)*; Вінниця, 2018; p 177.
- [115] Berdnyk, M.; Ivanov, V.; Zakharov, A;  $L_1$ -Regularization In Different Applications Of Chemical Modeling, *Molecular Engineering And Computational Modelling For Nano- And Biotechnology: From Nanoelectronics To Biopolymers* : Book of Abstracts International Scientific Conference (Cherkasy, September 25–26 , 2018 ); Cherkasy, 2018; P. 30-33.
- [116] Бердник, М.І.;  $L_1$ -регуляційний підхід у розрахунках фізикохімічних властивостей молекул, *Сучасні Проблеми Хімії* : тези доповідей XX Міжнародної конференції студентів та аспірантів (м. Київ, 15–17 травня, 2019); Київ, 2019; С. 140.

- [117] Berdnyk, M. I.; Denysenko, K. A.; Zakharov, A. B.; Ivanov, V. V.; Validation Of Regression Equations In QSAR Problem, *Сучасні Тенденції 2020* : Тези доповідей Київської Конференції з аналітичної хімії (м. Київ, 21-23 жовтня, 2020); Київ, 2020; Р.79-80.
- [118] Денисенко, К. А.; Бердник, М. И.; Захаров, А. Б.; Метод валидации уравнений линейной регрессии, *Хімічні Каразінські читання - 2021* : тези доп. XIII всеукр. наук. конф. студентів та аспірантів (м. Харків, 20–21 квітня, 2021 р.); Харків, 2021; С. 122-123.
- [119] Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of Quantitative Structure-Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor †. *Journal of Chemical Information and Computer Sciences*. American Chemical Society 1998, pp 669–677. <https://doi.org/10.1021/ci980008g>.
- [120] Sippl, W. Receptor-Based 3D QSAR Analysis of Estrogen Receptor Ligands – Merging the Accuracy of Receptor-Based Alignments with the Computational Efficiency of Ligand-Based Methods. *J. Comput. Mol. Des.* 2000 146 **2000**, 14 (6), 559–572. <https://doi.org/10.1023/A:1008115913787>.
- [121] Иванов, В. В.; Слета, Л. А.; Толстая, Дискриминантный Анализ Биологической Активности Производных Эстрадиола. *Ж. Органічної та фармацевтичної хімії*. **2004**, т. 2, 4 (8), С. 66-68.
- [122] Herndon, W. C. THEORY OF CARCINOGENIC ACTIVITY OF AROMATIC HYDROCARBONS\*. *Trans. N. Y. Acad. Sci.* **1974**, 36 (2 Series II), 200–217. <https://doi.org/10.1111/J.2164-0947.1974.TB01566.X>.
- [123] FISHER, R. A. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Ann. Eugen.* **1936**, 7 (2), 179–188. <https://doi.org/10.1111/J.1469-1809.1936.TB02137.X>.
- [124] Cox, D.; Hinkley, D.; Rubin, D.; Silverman, B. *Analysis of Binary Data, Second Edition*; Routledge, 2018; Vol. 7. <https://doi.org/10.1201/9781315137391>.
- [125] Hosmer Jr, D. W.; Lemeshow, S.; Sturdivant, R. X. *Applied Logistic Regression*; John Wiley & Sons, 2013; Vol. 398.



- [126] Suthaharan, S. Support Vector Machine; Springer, Boston, MA, 2016; pp 207–235. [https://doi.org/10.1007/978-1-4899-7641-3\\_9](https://doi.org/10.1007/978-1-4899-7641-3_9).
- [127] Noble, W. S. What Is a Support Vector Machine? *Nature Biotechnology*. Nature Publishing Group December 2006, pp 1565–1567. <https://doi.org/10.1038/nbt1206-1565>.
- [128] Zeng, X.; Zhu, S.; Lu, W.; Liu, Z.; Huang, J.; Zhou, Y.; Fang, J.; Huang, Y.; Guo, H.; Li, L.; Trapp, B. D.; Nussinov, R.; Eng, C.; Loscalzo, J.; Cheng, F. Target Identification among Known Drugs by Deep Learning from Heterogeneous Networks. *Chem. Sci.* **2020**, *11* (7), 1775–1797. <https://doi.org/10.1039/c9sc04336e>.
- [129] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 463–477. <https://doi.org/10.1038/s41573-019-0024-5>.
- [130] Lee, S.-I.; Lee, H.; Abbeel, P.; Ng, A. Y. Efficient  $L_1$  Regularized Logistic Regression. In *Aaai*; 2006; Vol. 6, pp 401–408.
- [131] scikit-learn: machine learning in Python — scikit-learn 1.0 documentation <https://scikit-learn.org/stable/> (accessed Sep 28, 2021).
- [132] Gupta, D. L.; Malviya, A. K.; Singh, S. Performance Analysis of Classification Tree Learning Algorithms. *Int. J. Comput. Appl.* **2012**, *55* (6).
- [133] Jiang, L.; Cai, Z.; Wang, D.; Jiang, S. Survey of Improving K-Nearest-Neighbor for Classification. *Proc. - Fourth Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2007* **2007**, *1*, 679–683. <https://doi.org/10.1109/FSKD.2007.552>.
- [134] Whittingham, M. S. Lithium Batteries and Cathode Materials. *Chem. Rev.* **2004**, *104* (10), 4271–4302. <https://doi.org/10.1021/cr020731c>.
- [135] Armand, M.; Tarascon, J.-M. Building Better Batteries. *Nature* **2008**, *451* (7179), 652–657. <https://doi.org/10.1038/451652a>.
- [136] Ellis, B. L.; Lee, K. T.; Nazar, L. F. Positive Electrode Materials for Li-Ion and Li-Batteries. *Chem. Mater.* **2010**, *22* (3), 691–714. <https://doi.org/10.1021/cm902696j>.

- [137] Goodenough, J. B.; Kim, Y. Challenges for Rechargeable Li Batteries. *Chem. Mater.* **2010**, *22* (3), 587–603. <https://doi.org/10.1021/cm901452z>.
- [138] Fujiki, K.; Ikeda, S.; Kobayashi, H.; Mori, A.; Nagira, A.; Nie, J.; Sonoda, T.; Yagupolskii, Y. Evaluation of Lewis Acidity of “Naked” Lithium Ion through Diels-Alder Reaction Catalyzed by Lithium TFPB in Nonpolar Organic Solvents. *Chem. Lett.* **2000**, *29* (1), 62–63. <https://doi.org/10.1246/cl.2000.62>.
- [139] Moss, S.; King, B. T.; de Meijere, A.; Kozhushkov, S. I.; Eaton, P. E.; Michl, J. LiCB<sub>11</sub>Me<sub>12</sub>: A Catalyst for Pericyclic Rearrangements. *Org. Lett.* **2001**, *3* (15), 2375–2377. <https://doi.org/10.1021/ol0161864>.
- [140] Volkis, V.; Mei, H.; Shoemaker, R. K.; Michl, J. LiCB<sub>11</sub>(CH<sub>3</sub>)<sub>12</sub> -Catalyzed Radical Polymerization of Isobutylene: Highly Branched Polyisobutylene and an Isobutylene–Ethyl Acrylate Copolymer. *J. Am. Chem. Soc.* **2009**, *131* (9), 3132–3133. <https://doi.org/10.1021/ja807297g>.
- [141] Li, L.; Yao, X.; Sun, C.; Du, A.; Cheng, L.; Zhu, Z.; Yu, C.; Zou, J.; Smith, S. C.; Wang, P.; Cheng, H.-M.; Frost, R. L.; (Max) Lu, G. Q. Lithium-Catalyzed Dehydrogenation of Ammonia Borane within Mesoporous Carbon Framework for Chemical Hydrogen Storage. *Adv. Funct. Mater.* **2009**, *19* (2), 265–271. <https://doi.org/10.1002/adfm.200801111>.
- [142] Lim, K. L.; Kazemian, H.; Yaakob, Z.; Daud, W. R. W. Solid-State Materials and Methods for Hydrogen Storage: A Critical Review. *Chem. Eng. Technol.* **2010**, *33* (2), 213–226. <https://doi.org/10.1002/ceat.200900376>.
- [143] Li, A.; Lu, R.-F.; Wang, Y.; Wang, X.; Han, K.-L.; Deng, W.-Q. Lithium-Doped Conjugated Microporous Polymers for Reversible Hydrogen Storage. *Angew. Chemie Int. Ed.* **2010**, *49* (19), 3330–3333. <https://doi.org/10.1002/anie.200906936>.
- [144] Stergiannakos, T.; Tylanakis, E.; Klontzas, E.; Trikalitis, P. N.; Froudakis, G. E. Hydrogen Storage in Novel Li-Doped Corrole Metal-Organic Frameworks. *J. Phys. Chem. C* **2012**, *116* (15), 8359–8363. <https://doi.org/10.1021/jp210975x>.
- [145] Fujii, T. Alkali-Metal Ion/Molecule Association Reactions and Their Applications to Mass Spectrometry. *Mass Spectrom. Rev.* **2000**, *19* (3), 111–138. [https://doi.org/10.1002/1098-2787\(200005/06\)19:3<111::AID-MAS1>3.0.CO;2-K](https://doi.org/10.1002/1098-2787(200005/06)19:3<111::AID-MAS1>3.0.CO;2-K).



- [146] Sablier, M.; Fujii, T. Mass Spectrometry of Free Radicals. *Chem. Rev.* **2002**, *102* (9), 2855–2924. <https://doi.org/10.1021/cr010295e>.
- [147] Takahashi, S.; Nakamura, M.; Fujii, T. Design and Performance of a Compact Li + Ion Attachment Mass Spectrometry System with an Atmospheric Sampling Device. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (3), 547–552. <https://doi.org/10.1007/s13361-011-0302-x>.
- [148] Rogan, W. J.; Ragan, N. B. Some Evidence of Effects of Environmental Chemicals on the Endocrine System in Children. *Int. J. Hyg. Environ. Health* **2007**, *210* (5), 659–667. <https://doi.org/10.1016/j.ijheh.2007.07.005>.
- [149] Ma, L. Endocrine Disruptors in Female Reproductive Tract Development and Carcinogenesis. *Trends Endocrinol. Metab.* **2009**, *20* (7), 357–363. <https://doi.org/10.1016/j.tem.2009.03.009>.
- [150] Brody, J. G.; Moysich, K. B.; Humblet, O.; Attfield, K. R.; Beehler, G. P.; Rudel, R. A. Environmental Pollutants and Breast Cancer. *Cancer* **2007**, *109* (S12), 2667–2711. <https://doi.org/10.1002/cncr.22655>.
- [151] Rudel, R. A.; Attfield, K. R.; Schifano, J. N.; Brody, J. G. Chemicals Causing Mammary Gland Tumors in Animals Signal New Directions for Epidemiology, Chemicals Testing, and Risk Assessment for Breast Cancer Prevention. *Cancer* **2007**, *109* (S12), 2635–2666. <https://doi.org/10.1002/cncr.22653>.
- [152] Liu, J.; Zhang, X.; Zhao, M.; Peng, S. Synthesis, Evaluation and 3D QSAR Analysis of Novel Estradiol–RGD Octapeptide Conjugates with Oral Anti-Osteoporosis Activity. *Eur. J. Med. Chem.* **2009**, *44* (4), 1689–1704. <https://doi.org/10.1016/j.ejmech.2008.09.036>.
- [153] Weitzmann, M. N.; Pacifici, R. Estrogen Deficiency and Bone Loss: An Inflammatory Tale. *J. Clin. Invest.* **2006**, *116* (5), 1186–1194. <https://doi.org/10.1172/JCI28550>.
- [154] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19* (14), 1639–

1662. [https://doi.org/10.1002/\(SICI\)1096-987X\(19981115\)19:14<1639::AID-JCC10>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B).
- [155] Sexton, M. J.; Gherman, R. B. Selective Estrogen Receptor Modulators: The Ideal Estrogen Replacement? *Prim. Care Update Ob. Gyns.* **2001**, 8 (1), 25–30. [https://doi.org/10.1016/S1068-607X\(00\)00066-4](https://doi.org/10.1016/S1068-607X(00)00066-4).
- [156] Wang, P.; McInnes, C.; Zhu, B. T. Structural Characterization of the Binding Interactions of Various Endogenous Estrogen Metabolites with Human Estrogen Receptor  $\alpha$  and  $\beta$  Subtypes: A Molecular Modeling Study. *PLoS One* **2013**, 8 (9), e74615. <https://doi.org/10.1371/journal.pone.0074615>.
- [157] Blair, R. M.; Fang, H.; Branham, W. S.; Hass, B. S.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. The Estrogen Receptor Relative Binding Affinities of 188 Natural and Xenochemicals: Structural Diversity of Ligands. *Toxicol. Sci.* **2000**, 54 (1), 138–153. <https://doi.org/10.1093/toxsci/54.1.138>.
- [158] Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B. S.; Xie, Q.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure - Activity Relationships for a Large Diverse Set of Natural, Synthetic, and Environmental Estrogens. *Chem. Res. Toxicol.* **2001**, 14 (3), 280–294. <https://doi.org/10.1021/tx000208y>.
- [159] AID 1204 - DSSTox (NCTRER) National Center for Toxicological Research Estrogen Receptor Binding Database - PubChem <https://pubchem.ncbi.nlm.nih.gov/bioassay/1204> (accessed Sep 28, 2021).
- [160] RDKit <https://www.rdkit.org/> (accessed Sep 28, 2021).
- [161] Fagerberg, J. H.; Bergström, C. A. Intestinal Solubility and Absorption of Poorly Water Soluble Compounds: Predictions, Challenges and Solutions. *Therapeutic Delivery*. Future Science Ltd London, UK August 28, 2015, pp 935–939. <https://doi.org/10.4155/tde.15.45>.
- [162] Hall\*, L. H.; Kier†, L. B. The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *J. Chem. Inf. Comput. Sci.* **2000**, 40 (3), 784–791. <https://doi.org/10.1021/CI990140W>.

- [163] Hu, Q. N.; Liang, Y. Z.; Yin, H.; Peng, X. L.; Fang, K. T. Structural Interpretation of the Topological Index. 2. The Molecular Connectivity Index, the Kappa Index, and the Atom-Type E-State Index. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1193–1201. <https://doi.org/10.1021/ci049973z>.
- [164] Berdnyk, M.; Ivanov, V.; Application Of Lasso Logistic Regression To Classification Problems In Chemistry, *Modern Chemistry Problems : Book of abstracts XXII International Conference for Students, PhD Students and Young Scientists* (м. Київ, 19–21 травня, 2021); Київ, 2021; С. 9.
- [165] Pulay, P. Localizability of Dynamic Electron Correlation. *Chem. Phys. Lett.* **1983**, *100* (2), 151–154. [https://doi.org/10.1016/0009-2614\(83\)80703-9](https://doi.org/10.1016/0009-2614(83)80703-9).
- [166] Saebø, S.; Pulay, P. Local Treatment of Electron Correlation. *Annu. Rev. Phys. Chem.* **1993**, *44* (1), 213–236. <https://doi.org/10.1146/annurev.pc.44.100193.001241>.
- [167] Knab, R.; Förner, W.; Čížek, J.; Ladik, J. Numerical Application of the Coupled Cluster Theory with Localized Orbitals to Polymers II. Optimal Localization of Wannier Functions and the Correlation Energy in Different Approximations. *J. Mol. Struct. THEOCHEM* **1996**, *366* (1–2), 11–33. [https://doi.org/10.1016/0166-1280\(96\)04518-6](https://doi.org/10.1016/0166-1280(96)04518-6).
- [168] Ivanov, V. V.; Zakharov, A. B.; Adamowicz, L. Molecular Dipole Static Polarisabilities and Hyperpolarisabilities of Conjugated Oligomer Chains Calculated with the Local  $\pi$ -Electron Coupled Cluster Theory. <http://dx.doi.org/10.1080/00268976.2013.788742> **2013**, *111* (24), 3779–3792. <https://doi.org/10.1080/00268976.2013.788742>.
- [169] Zakharov, A. B.; Ivanov, V. V.; Adamowicz, L.  $\pi$ -Electron Calculations Using the Local Linear-Response Coupled-Cluster Singles and Doubles Theory. *J. Phys. Chem. C* **2015**, *119* (52), 28737–28748. <https://doi.org/10.1021/ACS.JPCC.5B09496>.
- [170] Adamowicz, L.; Bartlett, R. J. Optimized Virtual Orbital Space for High-Level Correlated Calculations. *J. Chem. Phys.* **1987**, *86* (11), 6314–6324. <https://doi.org/10.1063/1.452468>.

- [171] Adamowicz, L. Optimized Virtual Orbital Space (OVOS) in Coupled-Cluster Calculations. In *Molecular Physics*; Taylor & Francis Group, 2010; Vol. 108, pp 3105–3112. <https://doi.org/10.1080/00268976.2010.520752>.
- [172] Neogrády, P.; Pitoňák, M.; Granatier, J.; Urban, M. Coupled Cluster Calculations: Ovos as an Alternative Avenue Towards Treating Still Larger Molecules. In *Challenges and Advances in Computational Chemistry and Physics*; Springer, Dordrecht, 2010; Vol. 11, pp 429–454. [https://doi.org/10.1007/978-90-481-2885-3\\_16](https://doi.org/10.1007/978-90-481-2885-3_16).
- [173] Ivanov, V. V.; Adamowicz, L. CASCCD: Coupled-Cluster Method with Double Excitations and the CAS Reference. *J. Chem. Phys.* **2000**, *112* (21), 9258. <https://doi.org/10.1063/1.481547>.
- [174] Henderson, T. M.; Bulik, I. W.; Stein, T.; Scuseria, G. E. Seniority-Based Coupled Cluster Theory. *J. Chem. Phys.* **2014**, *141* (24), 244104. <https://doi.org/10.1063/1.4904384>.
- [175] Bulik, I. W.; Henderson, T. M.; Scuseria, G. E. Can Single-Reference Coupled Cluster Theory Describe Static Correlation? *J. Chem. Theory Comput.* **2015**, *11* (7), 3171–3179. <https://doi.org/10.1021/acs.jctc.5b00422>.
- [176] Brueckner, K. A. Many-Body Problem for Strongly Interacting Particles. II. Linked Cluster Expansion. *Phys. Rev.* **1955**, *100* (1), 36–45. <https://doi.org/10.1103/PhysRev.100.36>.
- [177] Bartlett, R. J. Many-Body Perturbation Theory and Coupled Cluster Theory for Electron Correlation in Molecules. *Annu. Rev. Phys. Chem.* **1981**, *32* (1), 359–401. <https://doi.org/10.1146/annurev.pc.32.100181.002043>.
- [178] Поляк, Б. Т. *Введение в Оптимизацию*; Наука. Гл. ред. физ.-мат. лит., 1983.
- [179] Pulay, P. J. *Comput. Chem.* **1982**, *3*, 556–560
- [180] Cancés, E.; Le Bris, C. Can We Outperform the DIIS Approach for Electronic Structure Calculations? *Int. J. Quantum Chem.* **2000**, *79* (2), 82–90. [https://doi.org/10.1002/1097-461X\(2000\)79:2<82::AID-QUA3>3.0.CO;2-I](https://doi.org/10.1002/1097-461X(2000)79:2<82::AID-QUA3>3.0.CO;2-I).

- [181] Scuseria, G. U.; Lee, T. J.; Schaefer III, H.; III. *Chem. Phys. Lett* **1986**, *130*(3), 236-239.
- [182] Иванов, В.В.; Лях, Д.И. Автоматическая генерация диаграмм в методе связанных кластеров. Кластерное разложение, включающее трехчастичные возбуждения. *Вісник Харківського національного університету* **2002**, 8 (31), №549, С. 21-23.
- [183] Nesterov, Y. E. One Class of Methods of Unconditional Minimization of a Convex Function, Having a High Rate of Convergence. *USSR Comput. Math. Math. Phys.* **1984**, *24* (4), 80–82. [https://doi.org/10.1016/0041-5553\(84\)90234-9](https://doi.org/10.1016/0041-5553(84)90234-9).
- [184] Nesterov, Y.; Nemirovskii, A. *Interior-Point Polynomial Algorithms in Convex Programming*; SIAM, 1994.
- [185] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*; Springer Science & Business Media, 2003; Vol. 87.
- [186] Pariser, R.; Parr, R. G. A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. I. *J. Chem. Phys.* **1953**, *21* (3), 466. <https://doi.org/10.1063/1.1698929>.
- [187] Pople, J. A. Electron Interaction in Unsaturated Molecules. *Trans Faraday Soc* **1953**, *49*, 1375–1385.
- [188] Ohno, K. Some Remarks on the Pariser-Parr-Pople Method. *Theor. Chim. acta* **1964**, *23* (3), 219–227. <https://doi.org/10.1007/BF00528281>.
- [189] Harrison, R. J. Approximating Full Configuration Interaction with Selected Configuration Interaction and Perturbation Theory. *J. Chem. Phys.* **1991**, *94* (7), 5021–5031. <https://doi.org/10.1063/1.460537>.
- [190] Luzanov, A. V.; Prezhdo, O. V. Analysis of Multiconfigurational Wave Functions in Terms of Hole-Particle Distributions. *J. Chem. Phys.* **2006**, *124* (22), 224109. <https://doi.org/10.1063/1.2204608>.
- [191] Ivanov, V. V.; Adamowicz, L.; Lyakh, D. I. Multireference State-Specific Coupled-Cluster Theory and Multiconfigurationality Index. BH Dissociation. *Collect. Czechoslov. Chem. Commun.* **2005**, *70* (7), 1017–1033. <https://doi.org/10.1135/cccc20051017>.

- [192] Ivanov, V. V.; Lyakh, D. I.; Adamowicz, L. New Indices for Describing the Multi-Configurational Nature of the Coupled Cluster Wave Function. *Mol. Phys.* **2005**, *103* (15–16), 2131–2139. <https://doi.org/10.1080/00268970500083283>.
- [193] Korona, T.; Werner, H.-J. Local Treatment of Electron Excitations in the EOM-CCSD Method. *J. Chem. Phys.* **2003**, *118* (7), 3006–3019. <https://doi.org/10.1063/1.1537718>.
- [194] Mata, R. A.; Werner, H. J.; Schütz, M. Correlation Regions within a Localized Molecular Orbital Approach. *J. Chem. Phys.* **2008**, *128* (14), 144106. <https://doi.org/10.1063/1.2884725>.
- [195] Werner, H. J.; Schütz, M. An Efficient Local Coupled Cluster Method for Accurate Thermochemistry of Large Systems. *J. Chem. Phys.* **2011**, *135* (14), 144116. <https://doi.org/10.1063/1.3641642>.
- [196] Knowles, P.; Schütz, M.; Werner, H. J. Modern Methods and Algorithms of Quantum Chemistry. In *Proceedings*,; 2000; Vol. 3, p 97.
- [197] Adler, T. B.; Werner, H.-J. An Explicitly Correlated Local Coupled Cluster Method for Calculations of Large Molecules Close to the Basis Set Limit. *J. Chem. Phys.* **2011**, *135* (14), 144117. <https://doi.org/10.1063/1.3647565>.
- [198] Kinoshita, T.; Hino, O.; Bartlett, R. J. Singular Value Decomposition Approach for the Approximate Coupled-Cluster Method. *J. Chem. Phys.* **2003**, *119* (15), 7756–7762. <https://doi.org/10.1063/1.1609442>.
- [199] Hino, O.; Kinoshita, T.; Bartlett, R. J. Singular Value Decomposition Applied to the Compression of T3 Amplitude for the Coupled Cluster Method. *J. Chem. Phys.* **2004**, *121* (3), 1206. <https://doi.org/10.1063/1.1763575>.
- [200] Taube, A. G.; Bartlett, R. J. Rethinking Linearized Coupled-Cluster Theory. *J. Chem. Phys.* **2009**, *130* (14), 144112. <https://doi.org/10.1063/1.3115467>.
- [201] Kowalski, K.; Fan, P.-D. Generating Functionals Based Formulation of the Method of Moments of Coupled Cluster Equations. *J. Chem. Phys.* **2009**, *130* (8), 084112. <https://doi.org/10.1063/1.3076138>.
- [202] Kowalski, K.; Valiev, M. Extensive Regularization of the Coupled Cluster Methods Based on the Generating Functional Formalism: Application to Gas-Phase



- Benchmarks and to the SN2 Reaction of CHCl<sub>3</sub> and OH<sup>−</sup> in Water. *J. Chem. Phys.* **2009**, *131* (23), 234107. <https://doi.org/10.1063/1.3270957>.
- [203] Arponen, J. S.; Bishop, R. F. A Holomorphic Representation Approach to the Regularization of Model Field Theories in Coupled Cluster Form. *Theor. Chim. acta* **1991**, *804* **1991**, *80* (4), 289–305. <https://doi.org/10.1007/BF01117414>.
- [204] Lawler, K. V.; Parkhill, J. A.; Head-Gordon, M. The Numerical Condition of Electron Correlation Theories When Only Active Pairs of Electrons Are Spin-Unrestricted. *J. Chem. Phys.* **2009**, *130* (18), 184113. <https://doi.org/10.1063/1.3134223>.
- [205] Kowalski, K.; Piecuch, P. Renormalized CCSD(T) and CCSD(TQ) Approaches: Dissociation of the N<sub>2</sub> Triple Bond. *J. Chem. Phys.* **2000**, *113* (14), 5644–5652. <https://doi.org/10.1063/1.1290609>.
- [206] Kowalski, K.; Piecuch, P. Extensive Generalization of Renormalized Coupled-Cluster Methods. *J. Chem. Phys.* **2005**, *122* (7), 074107. <https://doi.org/10.1063/1.1848093>.
- [207] Piecuch, P.; Włoch, M. Renormalized Coupled-Cluster Methods Exploiting Left Eigenstates of the Similarity-Transformed Hamiltonian. *J. Chem. Phys.* **2005**, *123* (22), 224105. <https://doi.org/10.1063/1.2137318>.
- [208] Pimienta, I. S. O.; Kowalski, K.; Piecuch, P. Method of Moments of Coupled-Cluster Equations: The Quasivariational and Quadratic Approximations. *J. Chem. Phys.* **2003**, *119* (6), 2951–2962. <https://doi.org/10.1063/1.1589001>.
- [209] Nooijen, M.; Le Roy, R. J. Orbital Invariant Single-Reference Coupled Electron Pair Approximation with Extensive Renormalized Triples Correction. *J. Mol. Struct. THEOCHEM* **2006**, *768* (1–3), 25–43. <https://doi.org/10.1016/j.theochem.2006.05.017>.
- [210] Ozoliņš, V.; Lai, R.; Caflisch, R.; Osher, S. Compressed Plane Waves Yield a Compactly Supported Multiresolution Basis for the Laplace Operator. *Proc. Natl. Acad. Sci.* **2014**, *111* (5), 1691–1696. <https://doi.org/10.1073/pnas.1323260111>.
- [211] Zakharov, A. B.; Ivanov, V. V.; Adamowicz, L. Molecular Nonlinear Optical Parameters of  $\pi$ -Conjugated Nonalternant Hydrocarbons Obtained in Semiempirical

- Local Coupled-Cluster Theory. *J. Phys. Chem. C* **2014**, *118* (15), 8111–8121. <https://doi.org/10.1021/jp5002176>.
- [212] Zaporozhets, I. A.; Ivanov, V. V.; Lyakh, D. I.; Adamowicz, L. Discontinuities-Free Complete-Active-Space State-Specific Multi-Reference Coupled Cluster Theory for Describing Bond Stretching and Dissociation. *J. Chem. Phys.* **2015**, *143* (2), 024109. <https://doi.org/10.1063/1.4926392>.
- [213] Zakharov, A. B.; Ivanov, V. V.; Adamowicz, L. Electronic Perturbation Effects in the Presence of Electric Field for  $\pi$ -Conjugated Systems: An Electron-Correlation Study. *Int. J. Quantum Chem.* **2020**, *120* (16), e26260. <https://doi.org/10.1002/qua.26260>.
- [214] Бердник, М. И.; Иванов, В. В. Многошаговые Методы Первого Порядка в Решении Уравнений Теории Связанных Кластеров. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія Хімія* **2015**, No. 25, 39–45.
- [215] Бердник, М. И.; Иванов, В. В. L1-Регуляризация в Квантовой Химии.  $\pi$ -Электронная Теория Связанных Кластеров с Учетом Двукратных Возбуждений. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія Хімія* **2016**, No. 26, 58–64.
- [216] Ivanov, V. V.; Berdnyk, M. I.; Adamowicz, L. L<sub>1</sub>-Regularisation of the Coupled-Cluster Solutions. *Mol. Phys.* **2017**, *115* (21–22), 2892–2902. <https://doi.org/10.1080/00268976.2017.1359345>.
- [217] Бердник, М. И.; Иванов, В. В. L<sub>1</sub>-регуляризация. от статистики до квантовой химии, *Хімічні Каразінські читання - 2016* : тези доп. VIII всеукр. наук. конф. студентів та аспірантів (м. Харків, 18–20 квітня 2016 р.); Харків, **2016**; С. 132-133.
- [218] Бердник, М. И.; Иванов, В. В. Применение l<sub>1</sub>-регуляризации в неэмпирических и полуэмпирических расчетах квантовой химии, *XII Всеукраїнська конференція молодих вчених та студентів з актуальних питань хімії* : збірка праць всеукр. наук. конф. (м. Харків, 11-13 травня 2016 р.); Харків, **2016**; С. 32.



## ДОДАТОК А

## СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

**Наукові праці у наукових фахових виданнях України, що входять до міжнародної наукометричної бази Scopus:**

- [1] Berdnyk, M. I.; Zakharov, A. B.; Ivanov, V. V. Application Of  $L_1$ -Regularization Approach In QSAR Problem. Linear Regression And Artificial Neural Networks. *Methods Objects Chem. Anal.* **2019**, *14* (2), 79–90. <https://doi.org/10.17721/moca.2019.79-90>.

(Особистий внесок здобувача: програмна реалізація застосованих методів регресії а також штучних нейронних мереж, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, участь у обговоренні результатів, написання публікації).

**Наукові праці у наукових фахових виданнях України**

- [2] Бердник, М. И.; Иванов, В. В. Многошаговые Методы Первого Порядка в Решении Уравнений Теории Связанных Кластеров. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія Хімія* **2015**, № 25, 39-45.

(Особистий внесок здобувача: програмна реалізація методу зв'язаних кластерів, а також методів оптимізації до впроваджених ітеративних схем, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання публікації).

- [3] Бердник, М. И.; Иванов, В. В.  $L_1$ -Регуляризация в Квантовой Химии.  $\pi$ -Электронная Теория Связанных Кластеров с Учетом Двукратных Возбуждений. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія Хімія* **2016**, № 26, 58–64.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованого методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованого методу, участь у обговоренні результатів, написання публікації).

[4] Berdnyk, M. I.; Onizhuk, M. O.; Ivanov, V. V Methods for Building Linear Regression Equations in the “Structure-Property” Problems. *Kharkov Univ. Bull. Chem. Ser.* **2018**, № 30, 6–17. <https://doi.org/10.26565/2220-637x-2018-30-01>.

(Особистий внесок здобувача: програмна реалізація застосованих методів регресії, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання публікації).

**Наукові праці, в яких опубліковані основні наукові результати дисертації у періодичних наукових виданнях закордонних держав, що входять до ОЕСР, і реферуються у міжнародній наукометричній базі Scopus**

[5] Ivanov, V. V.; Berdnyk, M. I.; Adamowicz, L.  $L_1$ -Regularisation of the Coupled-Cluster Solutions. *Mol. Phys.* **2017**, *115* (21–22), 2892–2902. <https://doi.org/10.1080/00268976.2017.1359345>.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованого методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованого методу, участь у обговоренні результатів, написання публікації).

**Наукові праці, які засвідчують апробацію матеріалів дисертації**

[6] Бердник, М. И.; Иванов, В. В.  $L_1$ -регуляризация. от статистики до квантовой химии, *Хімічні Каразінські читання - 2016* : тези доп. VIII всеукр. наук. конф. студентів та аспірантів, Харків, Україна, квітень 18–20, 2016; ХНУ імені В. Н. Каразіна: Харків, 2016; С. 132-133.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованих статистичних методів і методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[7] Бердник, М. И.; Иванов, В. В. Применение  $l_1$ -регуляризации в неэмпирических и полуэмпирических расчетах квантовой химии, *XII Всеукраїнська конференція молодих вчених та студентів з актуальних питань хімії* : збірка праць всеукр. наук. конф., Харків, Україна, травень 11-13, 2016; ДНУ НТК ІМК НАНУ: Харків, 2016; С. 32.

(Особистий внесок здобувача: програмна реалізація  $L_1$ -регуляризованого методу зв'язаних кластерів, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[8] Бердник, М.И.; Дяченко, А.В.; Иванов, В. В; Регрессионные модели QSAR, *Збірник тез доповідей, Хімічні Проблеми Сьогодення (ХПС-2018)*, Вінниця, Україна, березень 27-29, 2018; Донецький національний університет імені Василя Стуса: Вінниця, 2018; С. 177.

(Особистий внесок здобувача: програмна реалізація методу LARS-LASSO а також методів регресії, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[9] Berdnyk, M.; Ivanov, V.; Zakharov, A;  $L_1$ -Regularization In Different Applications Of Chemical Modeling, *Molecular Engineering And Computational Modelling For Nano- And Biotechnology: From Nanoelectronics To Biopolymers* : Book of Abstracts International Scientific Conference, Cherkasy, Ukraine, September 25–26, 2018; Bohdan Khmelnytsky Cherkasy National University: Cherkasy, 2018; P. 30-33.

(Особистий внесок здобувача: програмна реалізація методів регресії, регуляризованих квантовохімічних методів, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[10] Бердник, М.І.;  $L_1$ -регуляційний підхід у розрахунках фізикохімічних властивостей молекул, *Сучасні Проблеми Хімії* : тези доповідей XX Міжнародної конференції студентів та аспірантів, Київ, Україна, травень 15–17, 2019; Київський національний університет імені Тараса Шевченка: Київ, 2019; С. 140.

(Особистий внесок здобувача: програмна реалізація використаних методів, розрахунки фізико-хімічних властивостей молекул, участь у обговоренні результатів, написання тез, доповідь на конференції).

[11] Berdnyk, M. I.; Denysenko, K. A.; Zakharov, A. B.; Ivanov, V. V.; Validation Of Regression Equations In QSAR Problem, *Сучасні Тенденції 2020* : Тези доповідей Київської Конференції з аналітичної хімії, Київ, Україна, жовтень 21-23, 2020; Київський національний університет імені Тараса Шевченка: Київ, 2020; С.79-80.

(Особистий внесок здобувача: програмна реалізація методів валідації, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

[12] Денисенко, К. А.; Бердник, М. И.; Захаров, А. Б.; Метод валидации уравнений линейной регрессии, *Хімічні Каразінські читання - 2021* : тези доп. XIII всеукр. наук. конф. студентів та аспірантів, Харків, Україна, квітень 20–21, 2021; ХНУ імені В. Н. Каразіна: Харків, 2021; С. 122-123.

(Особистий внесок здобувача: програмна реалізація методів валідації, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, участь у написанні тез).

[13] Berdnyk, M.; Ivanov, V.; Application Of Lasso Logistic Regression To Classification Problems In Chemistry, *Modern Chemistry Problems* : Book of abstracts XXII International Conference for Students, PhD Students and Young Scientists, Київ, Україна, травень 19–21, 2021; Київський національний університет імені Тараса Шевченка: Київ, 2021; С. 9.

(Особистий внесок здобувача: програмна реалізація методів класифікації, розрахунки з використанням програмно-реалізованих методів, участь у обговоренні результатів, написання тез, доповідь на конференції).

**ДОДАТОК Б**  
**ПРОСТИЙ ПРИКЛАД РЕГУЛЯРИЗАЦІЇ: РОЗРАХУНОК  $L_1$ -**  
**РЕГУЛЯРИЗОВАНОГО СЕРЕДНЬОГО ЗНАЧЕННЯ**  
 (програма на скриптовій мові Python)

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
#
'''
-----
Toy example for article
Authors: V.V.Ivanov, M.I.Berdnik
-----
'''
import statistics

lamb=1    # regularization Parameter
aksi=0.1  # step of iteration
eps1=1.0e-6 # accuracy
#----->

x=[0.62, 0.6, 0.75, 0.65, 0.62, 1.1] # sample
#----->
zero=0.00000000000000000000e0
def subg(nn,lamb,x,a):
    delta1=zero
    for i in range(nn):
        delta1=delta1+(a-x[i])

    delta1=delta1*2.0e0
    if abs(a) > eps1: w=delta1+lamb*a/abs(a)
    if abs(a) <= eps1:
        if delta1 < -lamb: w=delta1+lamb
        if delta1 > lamb: w=delta1-lamb
        if delta1 >= -lamb and delta1 <= lamb: w=zero

    return delta1,w

def func1(nn,lamb,x,a):
    fu =zero
    for i in range(nn):
        fu= fu + (a-x[i])**2
```

```

    fu=fu+lamb*abs(a)
    return fu

format1='%s %3.0f %s %4.4f %s %3.2e %s %2.2e %s %2.2e'

nn=len(x)
print ('Size of sample=',nn)

av=a=sum(x)/nn
#----->
w=1
it=0
while abs(w) > eps1:
    it=it+1

    fu= func1(nn,lamb,x,a)
    delta1,w=subg(nn,lamb,x,a)

    a=a-aksi*w # step

    text1=('iter=',it,'a=',a,'func=',fu,'Delta=',delta1,'|SubGrd|=',abs(w))
    print(format1 %text1)

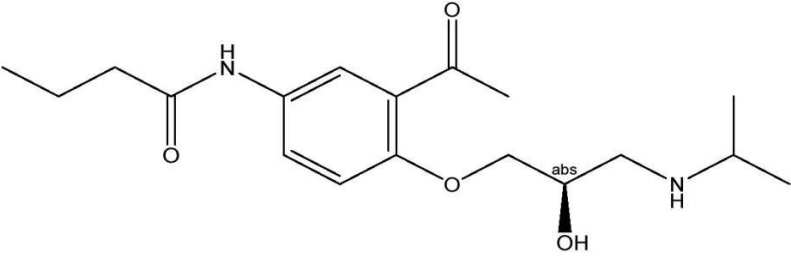
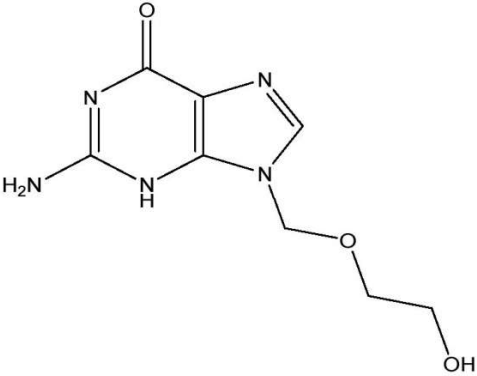
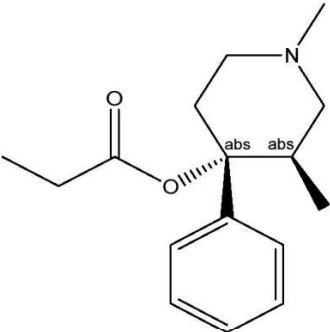
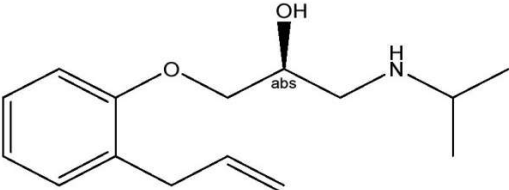
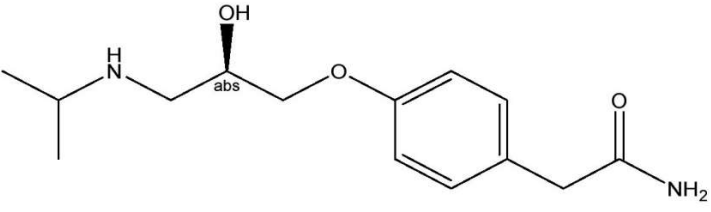
print (70*'-','\n Sample:',x)
print ('Lambda=',lamb )
print ('Avarage =',av )
print (' Median =',statistics.median(x) )
print ('L1-Regularized Avarage =',a )

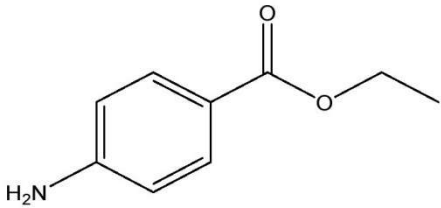
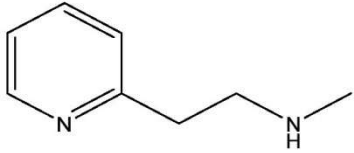
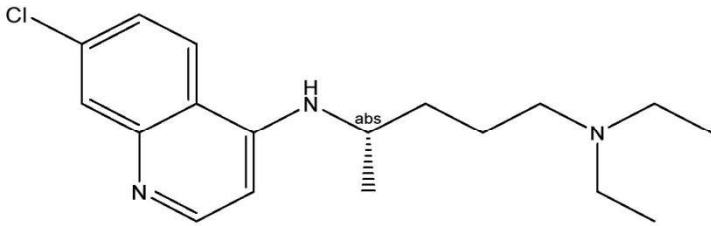
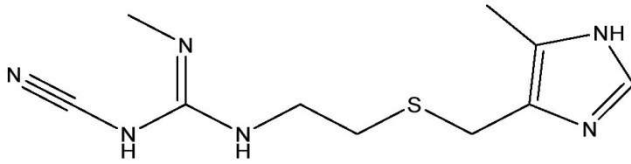
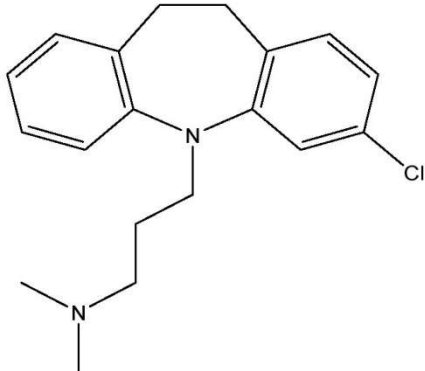
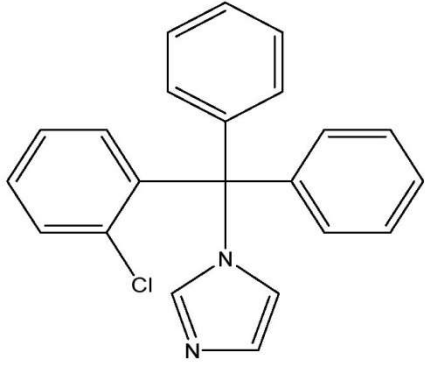
print (__doc__)

```

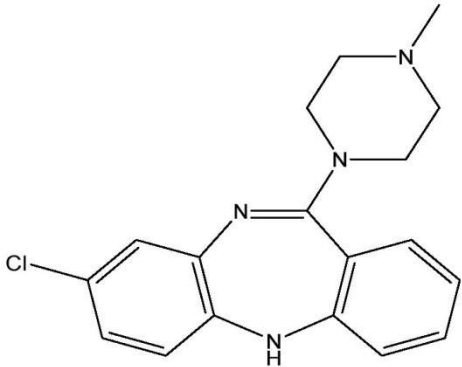
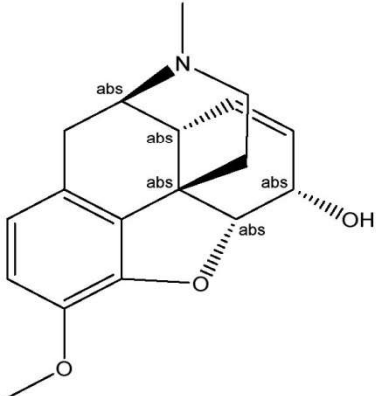
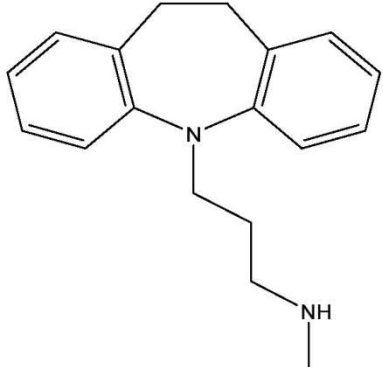
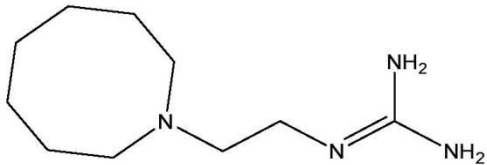
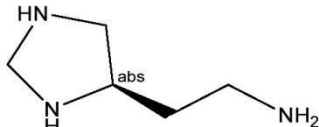
# ДОДАТОК В

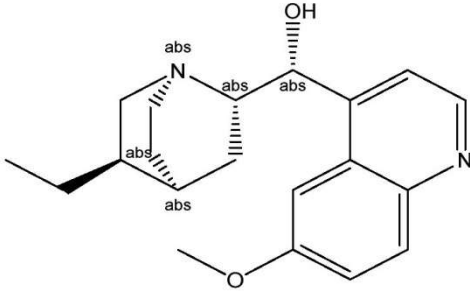
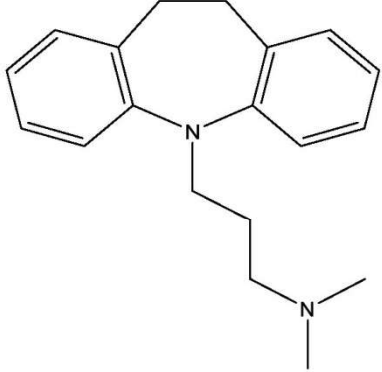
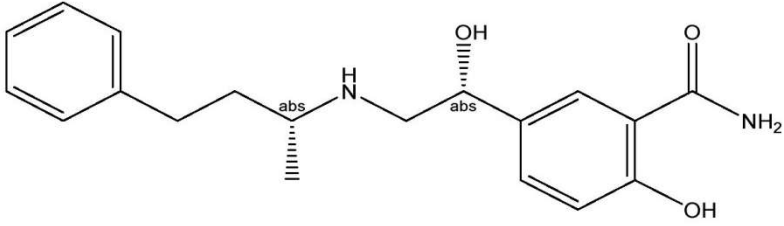
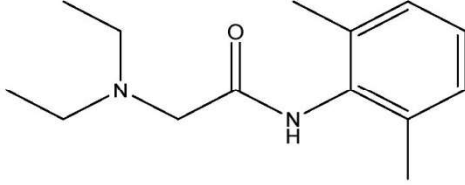
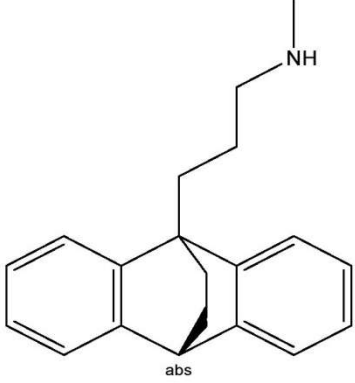
## КОНСТАНТИ ІОНІЗАЦІЇ ОРГАНІЧНИХ СПОЛУК РІЗНОЇ ПРИРОДИ

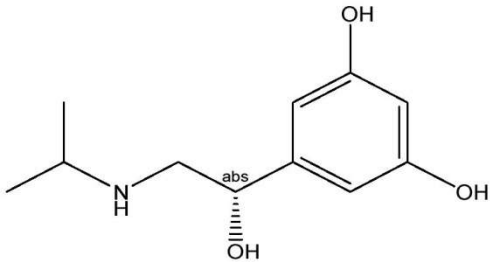
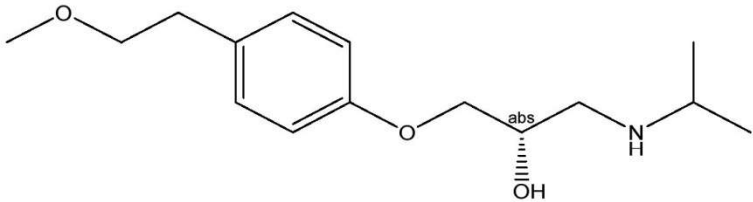
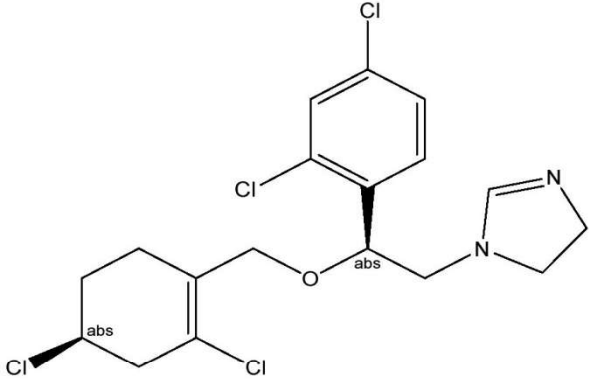
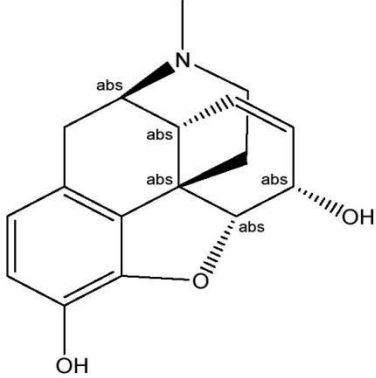
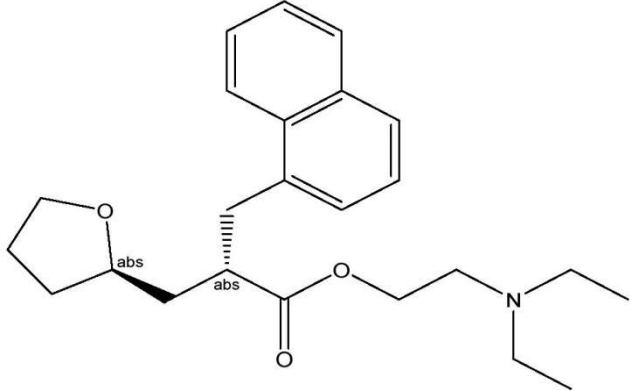
Structure	№	Name	pKa(exp)
	1	Acebutolol	9.5
	2	Acyclovir	2.2
	3	Alphaprodine	8.7
	4	Alprenolol	9.6
	5	Atenolol	9.6

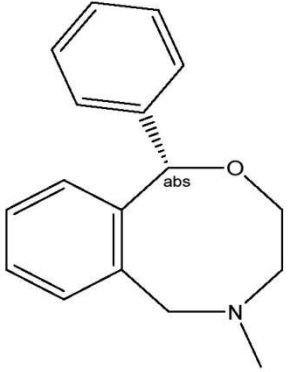
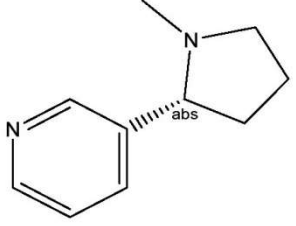
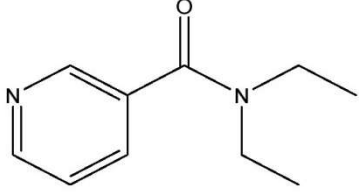
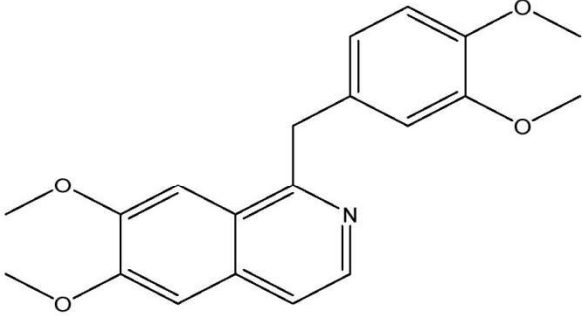
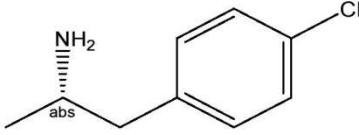
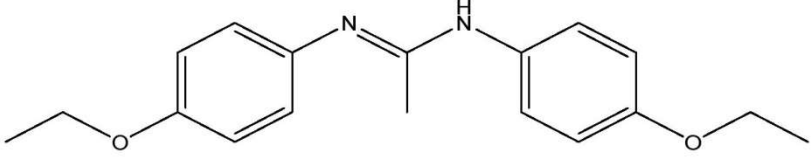
		6 Benzocaine	2.5
		7 Betahistine	10
		10 Chloroquine	10.6
		11 Cimetidine0	6.8
		12 Clomipramine	9.4
		13 Clotrimazole	5.8

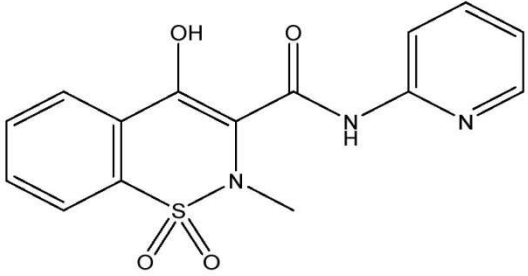
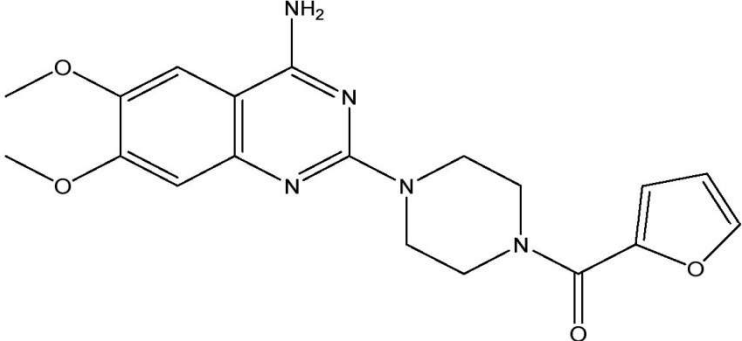
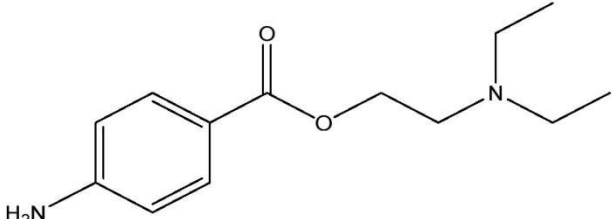
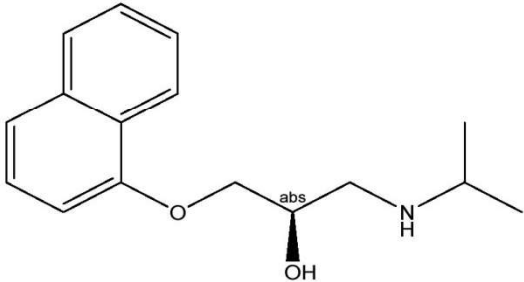
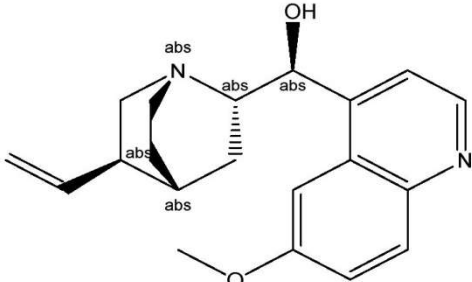


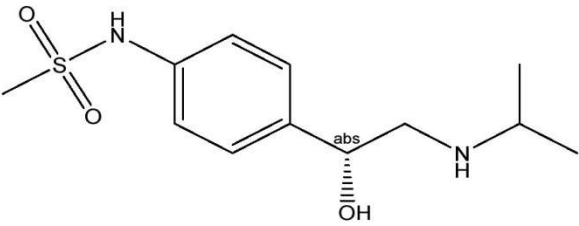
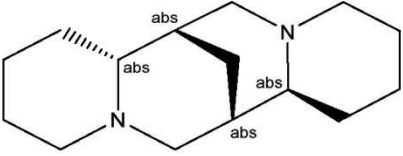
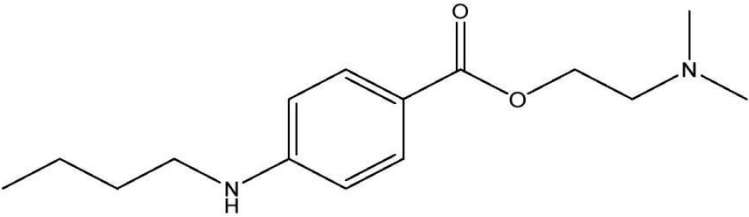
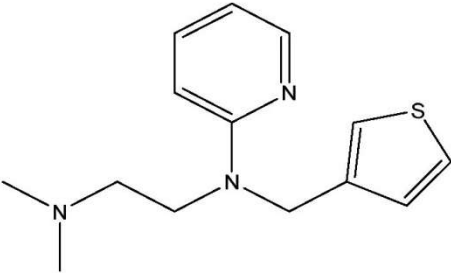
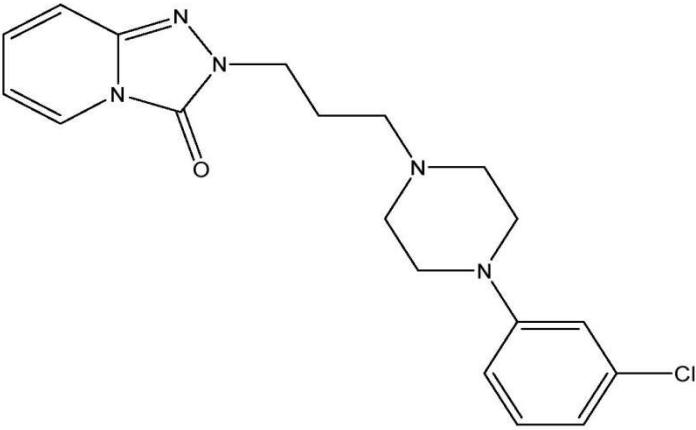
	14	Clozapine	7.5
	16	Codeine	8.1
	17	Desipramine	10.3
	18	Guanethidine	11.4
	19	Histamine	9.7

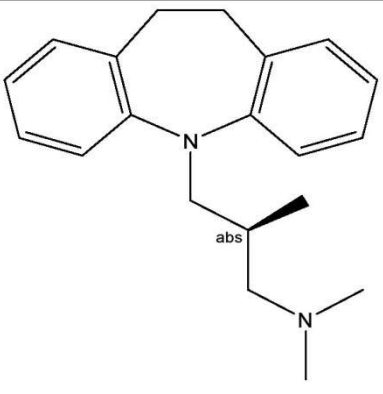
	20	Hydroquinine	9.1
	22	Imipramine	9.6
	23	Labetalol	7.3
	24	Lidocaine	7.9
	25	Maprotiline	10.3

	27	Metaproterenol	9.9
	28	Metoprolol	9.6
	29	Miconazole	6.4
	30	Morphine	8.2
	31	Nafronyl	9.1

	32	Nefopam	8.5
	34	Nicotine	8.1
	36	Nikethamide	3.5
	37	Papaverine	6.4
	38	p-Cl-amphetamine	9.9
	39	Phenacaine	9.3

	41 Piroxicam	5.3
	42 Prazosin	7
	43 Procaine	9.1
	45 Propranolol	9.6
	46 Quinine	8.5

	47	Sotalol	9.3
	48	Sparteine	12
	49	Tetracaine	8.5
	50	Thenyldiamine	8.9
	51	Trazodone	6.8

 <chem>CN(C)CC[C@H](C)CN1Cc2ccccc2C3CCCCC13</chem>	52	Trimipramine	9.4
--	----	--------------	-----

# ДОДАТОК Г

## ТЕМПЕРАТУРИ КИПІННЯ ФЛґУОРОАЛКАНІВ

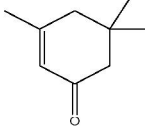
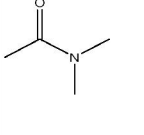
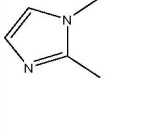
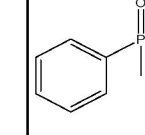
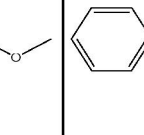
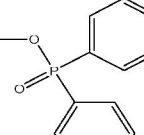
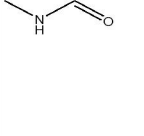
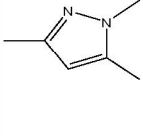
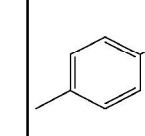
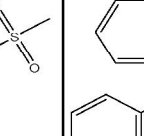
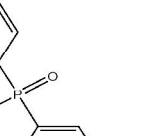
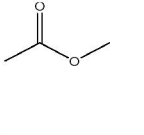
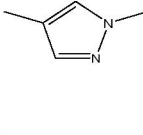
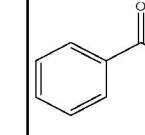
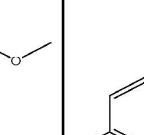
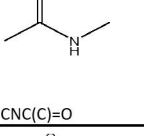
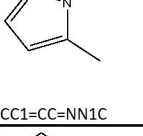
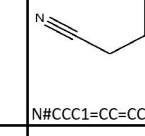
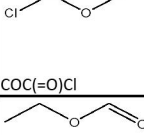
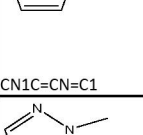
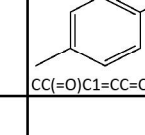
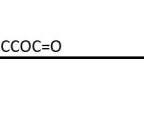
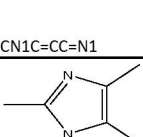
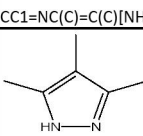
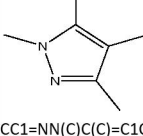

№	BP, °C	Structure	№	BP, °C	Structure	№	BP, °C	Structure
1	-15	F <sub>3</sub> C-CF <sub>2</sub> -CH <sub>3</sub>	29	18	F <sub>3</sub> C-CH <sub>2</sub> -CF <sub>2</sub> -CF <sub>3</sub>	56	22	F <sub>2</sub> HC-CHF-CH <sub>3</sub>
2	12	F <sub>3</sub> C-CH(CH <sub>3</sub> ) <sub>2</sub>	30	44	FH <sub>2</sub> C-CF <sub>2</sub> -CF <sub>2</sub> -CHF <sub>2</sub>	57	55	F <sub>2</sub> HC-CHF-CH <sub>2</sub> F
3	24.5	F <sub>3</sub> C-CH <sub>2</sub> -CH <sub>2</sub> -CF <sub>3</sub>	31	15	F <sub>2</sub> HC-CF <sub>2</sub> -CF <sub>2</sub> -CF <sub>3</sub>	58	40.5	F <sub>2</sub> HC-CHF-CHF <sub>2</sub>
4	32.5	F <sub>2</sub> HC-CHF-CF <sub>2</sub> -CF <sub>3</sub>	32	-1.7	F <sub>3</sub> C-CF <sub>2</sub> -CF <sub>2</sub> -CF <sub>3</sub>	59	-0.8	F <sub>2</sub> HC-CF <sub>2</sub> -CH <sub>3</sub>
5	35	F <sub>2</sub> HC-CF <sub>2</sub> -CHF-CF <sub>3</sub>	33	-21	F <sub>2</sub> HC-CHF <sub>2</sub>	60	15	F <sub>2</sub> HC-CH <sub>2</sub> -CF <sub>3</sub>
6	26.5	FH <sub>2</sub> C-CF <sub>2</sub> -CF <sub>2</sub> -CF <sub>3</sub>	34	-25	F <sub>2</sub> HC-CH <sub>3</sub>	61	-0.5	F <sub>3</sub> C-CHF-CH <sub>3</sub>
7	-51.6	CH <sub>2</sub> F <sub>2</sub>	35	-26.2	F <sub>3</sub> C-CH <sub>2</sub> F	62	-48.5	F <sub>3</sub> C-CHF <sub>2</sub>
8	41	FH <sub>2</sub> C-CH <sub>2</sub> -CH <sub>2</sub> F	36	-37.5	H <sub>3</sub> C-CH <sub>2</sub> F	63	17	F <sub>3</sub> C-CH <sub>2</sub> -CH <sub>2</sub> -CH <sub>3</sub>
9	21	F <sub>3</sub> C-CHF-CH <sub>2</sub> F	37	-47	H <sub>3</sub> C-CF <sub>3</sub>	64	0	F <sub>3</sub> C-CF <sub>2</sub> -CH <sub>2</sub> F
10	-78.5	CH <sub>3</sub> F	38	22	FH <sub>2</sub> C-CH(CH <sub>3</sub> ) <sub>2</sub>	65	45	F <sub>2</sub> HC-CH <sub>2</sub> -CH <sub>2</sub> F
11	11	F <sub>2</sub> HC-CF <sub>2</sub> -CHF <sub>2</sub>	39	12	FC(CH <sub>3</sub> ) <sub>3</sub>	66	14.5	F <sub>3</sub> C-CF <sub>2</sub> -CF <sub>2</sub> -CH <sub>3</sub>
12	7.5	F <sub>2</sub> HC-CH <sub>2</sub> -CH <sub>3</sub>	40	21.5	F <sub>3</sub> C-CH(CH <sub>3</sub> )-CF <sub>3</sub>	67	110	FH <sub>2</sub> C-CHF-CHF-CH <sub>2</sub> F
13	6	F <sub>3</sub> C-CHF-CHF <sub>2</sub>	41	40	F <sub>3</sub> C-CH(CF <sub>3</sub> )-CH <sub>2</sub> F	68	78	F <sub>2</sub> HC-CH(CHF <sub>2</sub> )-CH <sub>2</sub> F
14	-9.7	H <sub>3</sub> C-CHF-CH <sub>3</sub>	42	12	F <sub>3</sub> C-CH(CF <sub>3</sub> ) <sub>2</sub>	69	56.5	F <sub>2</sub> HC-CH <sub>2</sub> -CHF-CH <sub>3</sub>
15	-13	F <sub>3</sub> C-CH <sub>2</sub> -CH <sub>3</sub>	43	-0.3	F <sub>3</sub> C-CF(CF <sub>3</sub> ) <sub>2</sub>	70	57	F <sub>2</sub> HC-CH <sub>2</sub> -CF <sub>2</sub> -CH <sub>3</sub>
16	-19	F <sub>3</sub> C-CHF-CF <sub>3</sub>	44	77	FH <sub>2</sub> C-CH <sub>2</sub> -CH <sub>2</sub> -CH <sub>2</sub> F	71	46.5	F <sub>2</sub> HC-CHF-CH <sub>2</sub> -CH <sub>3</sub>
17	-78	F <sub>3</sub> C-CF <sub>3</sub>	45	31	H <sub>3</sub> C-CF <sub>2</sub> -CH <sub>2</sub> -CH <sub>3</sub>	72	90	F <sub>2</sub> HC-CHF-CHF-CH <sub>2</sub> F
18	-82	CHF <sub>3</sub>	46	10.45	FH <sub>2</sub> C-CH <sub>2</sub> F	73	64	F <sub>2</sub> HC-CF <sub>2</sub> -CHF-CH <sub>2</sub> F
19	-128	CF <sub>4</sub>	47	25	F <sub>2</sub> HC-CF <sub>2</sub> -CH <sub>2</sub> F	74	57.5	F <sub>2</sub> HC-CF <sub>2</sub> -CHF-CHF <sub>2</sub>
20	-2.5	FH <sub>2</sub> C-CH <sub>2</sub> -CH <sub>3</sub>	48	40	F <sub>3</sub> C-CH <sub>2</sub> -CF <sub>2</sub> -CH <sub>3</sub>	75	53.5	F <sub>2</sub> HC-CF <sub>2</sub> -CF <sub>2</sub> -CH <sub>2</sub> F
21	-38	F <sub>3</sub> C-CF <sub>2</sub> -CF <sub>3</sub>	49	17	H <sub>3</sub> C-CF <sub>2</sub> -CF <sub>2</sub> -CH <sub>3</sub>	76	59	F <sub>2</sub> HC-CH(CF <sub>3</sub> )-CH <sub>2</sub> F
22	-0.5	H <sub>3</sub> C-CF <sub>2</sub> -CH <sub>3</sub>	50	34.3	F <sub>3</sub> C-CHF-CHF-CF <sub>3</sub>	77	53.5	F <sub>3</sub> C-CH <sub>2</sub> -CH <sub>2</sub> -CH <sub>2</sub> F
23	29.4	FH <sub>2</sub> C-CH <sub>2</sub> -CF <sub>3</sub>	51	32	FH <sub>2</sub> C-CHF-CH <sub>3</sub>	78	45.5	F <sub>3</sub> C-CH <sub>2</sub> -CHF-CH <sub>3</sub>
24	-0.8	F <sub>3</sub> C-CH <sub>2</sub> -CF <sub>3</sub>	52	66	FH <sub>2</sub> C-CHF-CH <sub>2</sub> F	79	23	F <sub>3</sub> C-CF(CH <sub>2</sub> F)-CF <sub>3</sub>
25	-17	F <sub>3</sub> C-CF <sub>2</sub> -CHF <sub>2</sub>	53	18.7	FH <sub>2</sub> C-CF <sub>2</sub> -CH <sub>3</sub>	80	42	F <sub>3</sub> C-CHF-CF <sub>2</sub> -CH <sub>2</sub> F
26	4	FH <sub>2</sub> C-CHF <sub>2</sub>	54	27	FH <sub>2</sub> C-CF <sub>2</sub> -CH <sub>2</sub> F	81	43.5	F <sub>3</sub> C-CF <sub>2</sub> -CH <sub>2</sub> -CH <sub>2</sub> F
27	32.5	FH <sub>2</sub> C-CH <sub>2</sub> -CH <sub>2</sub> -CH <sub>3</sub>	55	39	F <sub>2</sub> HC-CH <sub>2</sub> -CHF <sub>2</sub>	82	41	H <sub>3</sub> C-CHF-CHF-CH <sub>3</sub>
28	25	H <sub>3</sub> C-CH <sub>2</sub> -CHF-CH <sub>3</sub>						



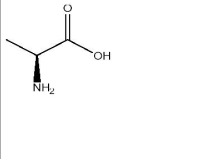
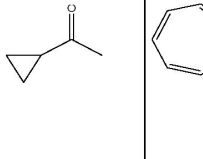
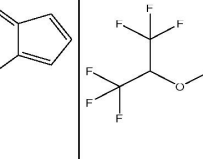
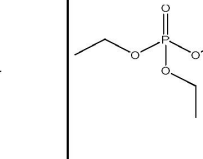
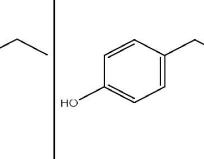
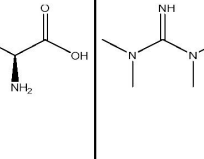

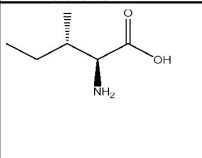
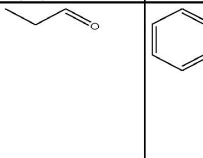
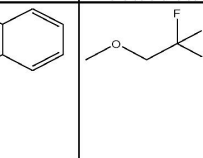
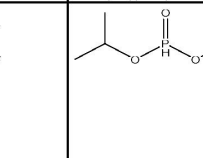
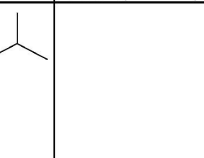
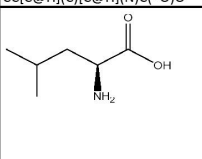
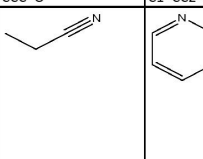
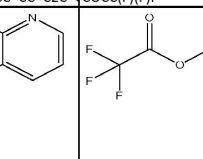
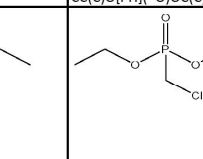
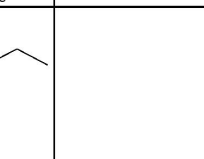
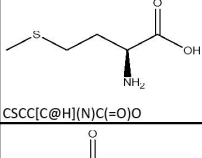
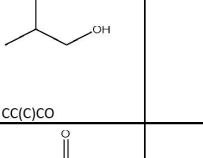
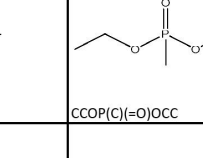
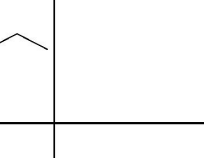
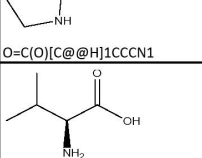
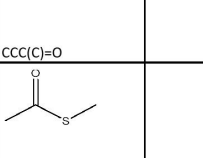
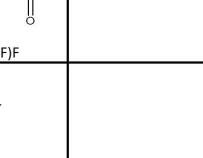
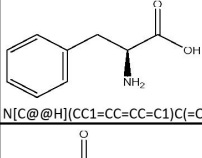
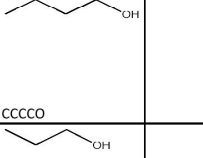
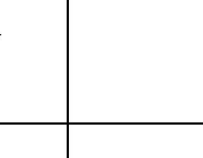
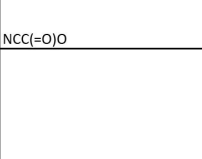
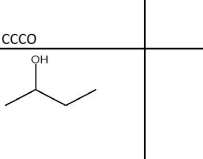


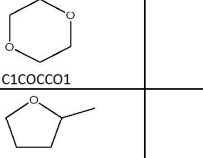
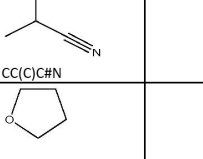


# ДОДАТОК Д

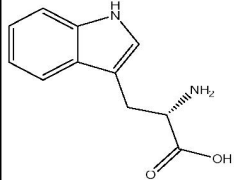
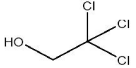
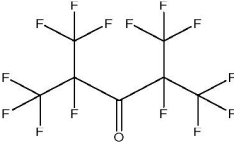
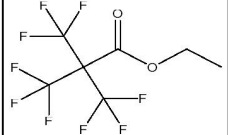
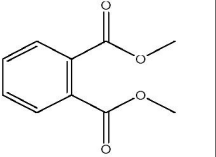
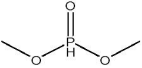
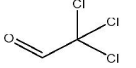
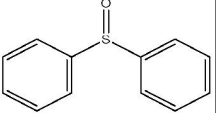
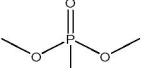
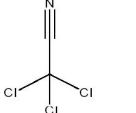
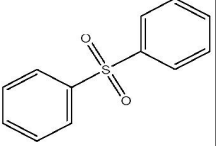
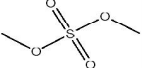
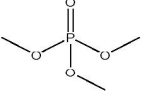
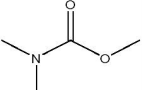
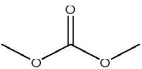
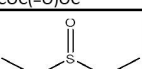
## РОЗБИТТЯ НА КЛАСТЕРИ ОРГАНІЧНИХ СПОЛУК ОСНОВНИХ ДО КАТІОНУ Li

Кожен стовпчик відображає кластер відповідно до структурних характеристик. Розбиття з використанням методу k-середніх

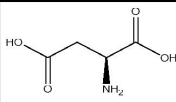
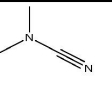
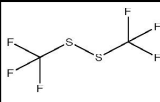
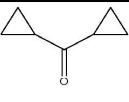

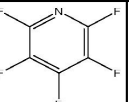
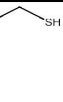
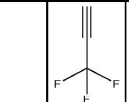
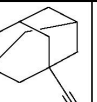
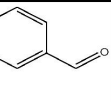
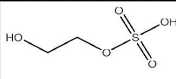


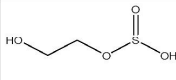
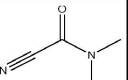







0	1	2	3	4	5
					
<chem>CC1=CC(=O)CC(C)(C)C1</chem>	<chem>CC(=O)N(C)C</chem>	<chem>CC1=NC=CN1C</chem>	<chem>COP(C)(=O)C1=CC=CC=C1</chem>	<chem>O=P(OC1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=CC=C1</chem>	<chem>NC(C(F)(F)F)(C(F)(F)F)C(F)(F)F</chem>
					
	<chem>CNC=O</chem>	<chem>CC1=NN(C)C(C)=C1</chem>	<chem>CC1=CC=C(S(C)(=O)=O)C=C1</chem>	<chem>O=P(C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=CC=C1</chem>	<chem>OC(C(F)(F)F)(C(F)(F)F)C(F)(F)F</chem>
					
	<chem>COC(C)=O</chem>	<chem>CC1=CN(C)N=C1</chem>	<chem>COC(=O)C1=CC=CC=C1</chem>	<chem>O=P(OC1=CC=C(F)C=C1)(C1=CC=CC=C1)C1=CC=CC=C1</chem>	
					
	<chem>CNC(C)=O</chem>	<chem>CC1=CC=NN1C</chem>	<chem>N#CCCC1=CC=CC=C1</chem>		
					
	<chem>COC(=O)Cl</chem>	<chem>CN1C=CN=C1</chem>	<chem>CC(=O)C1=CC=C(C)C=C1</chem>		
					
	<chem>CCOC=O</chem>	<chem>CN1C=CC=N1</chem>			
					
		<chem>CC1=NC(C)=C(C)[NH]1</chem>			
					
		<chem>CC1=N[NH]C(C)=C1C</chem>			
					
		<chem>CC1=NN(C)C(C)=C1C</chem>			

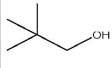
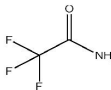

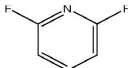
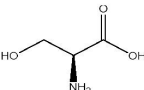
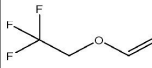
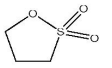
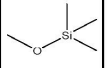
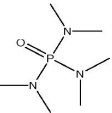
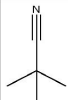
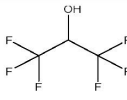


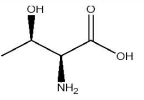


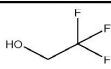

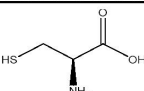
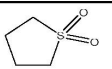
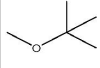
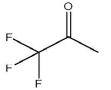
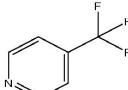
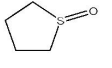
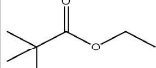
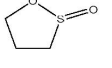



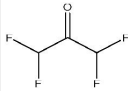


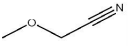

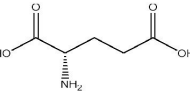
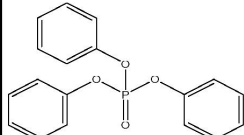
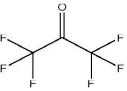
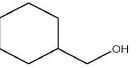
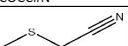
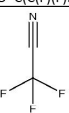
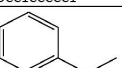
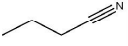
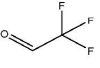
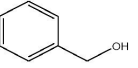
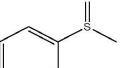
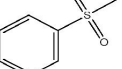
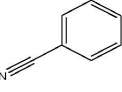
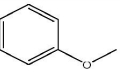
14	15	16	17	18	19	20
						
<chem>C[C@H](N)C(=O)O</chem>	<chem>CC(=O)C1CC1</chem>	<chem>C1=CC=C2C=CC=C2C=C1</chem>	<chem>COC(C(F)(F)F)C(F)(F)F</chem>	<chem>CCOP(=O)(OCC)OCC</chem>	<chem>N[C@@H](CC1=CC=C(C(O)C=C1)C(=O)O</chem>	<chem>CN(C)C(=N)N(C)C</chem>
						
<chem>CC[C@H](C)[C@H](N)C(=O)O</chem>	<chem>CCC=O</chem>	<chem>C1=CC2=CC=CC=C2C=C1</chem>	<chem>COCC(F)(F)F</chem>	<chem>CC(C)O[PH](=O)OC(C)C</chem>		
						
<chem>CC(C)[C@H](N)C(=O)O</chem>	<chem>CCC#N</chem>	<chem>C1=CC2=CC=CN=C2N=C1</chem>	<chem>CCOC(=O)C(F)(F)F</chem>	<chem>CCOP(=O)(CCl)OCC</chem>		
						
<chem>CSCC[C@H](N)C(=O)O</chem>	<chem>CC(C)CO</chem>		<chem>COC(=O)C(F)(F)F</chem>	<chem>CCOP(C)(=O)OCC</chem>		
						
<chem>O=C(O)[C@@H]1CCCCN1</chem>	<chem>CCC(C)=O</chem>		<chem>CC(=O)COC(=O)C(F)(F)F</chem>			
						
<chem>CC(C)[C@H](N)C(=O)O</chem>	<chem>CSC(C)=O</chem>		<chem>CN(C)C(=O)C(F)(F)F</chem>			
						
<chem>N[C@@H](CC1=CC=CC=C1)C(=O)O</chem>	<chem>CCCCO</chem>		<chem>CSC(C)=O)C(F)(F)F</chem>			
						
<chem>NCC(=O)O</chem>	<chem>CCC(O)O</chem>					
						
	<chem>CCC(C)O</chem>					
						
	<chem>C1COCCO1</chem>					
						
	<chem>CC1CCCCO1</chem>					
	<chem>CC(C)C#N</chem>					
	<chem>C1CCOC1</chem>					

21	22	23	24	25	26	27
 <chem>N[C@@H](CC1=C[NH]C2=CC=CC=C2)C(=O)O</chem>	H <sub>2</sub> O	 <chem>OCC(Cl)(Cl)Cl</chem>	 <chem>O=C(C(F)(C(F)(F)F)C(F)(F)F)C(F)(F)F</chem>	 <chem>CCOC(=O)C(C(F)(F)F)(C(F)(F)F)C(F)(F)F</chem>	 <chem>COC(=O)C1=CC(=O)OC=C1</chem>	 <chem>CO[PH](=O)OC</chem>
	NH <sub>3</sub>	 <chem>O=CC(Cl)(Cl)Cl</chem>			 <chem>O=S(C1=CC=CC=C1)C1=CC=CC=C1</chem>	 <chem>CO[PH](=O)OC</chem>
	N	 <chem>N#CC(Cl)(Cl)Cl</chem>			 <chem>O=S(=O)(C1=CC=CC=C1)C1=CC=CC=C1</chem>	 <chem>COS(=O)(=O)OC</chem>
						 <chem>COP(=O)(OC)OC</chem>
						 <chem>COC(=O)N(C)C</chem>
						 <chem>COC(=O)OC</chem>
						 <chem>COS(=O)OC</chem>

28	29	30	31	32	33	34
<chem>O=P(OC1=CC=C(C(F)(F)F)C=C1)(C1=CC=CC=C1)C1=CC=CC=C1</chem>	<chem>COC(=O)C1=CC(C(=O)OC)=CC=C1</chem>	<chem>C1=N[NH]C=N1</chem>	<chem>CS(=O)(=O)C1=CC=C([N+](=O)[O-])C=C1</chem>	<chem>CCO</chem>	<chem>C1=CC2=CC3=CC=CC=C3C=C2C=C1</chem>	<chem>N#CCBr</chem>
		<chem>CC1=N[NH]C=C1</chem>		<chem>NC=O</chem>	<chem>C1=CC2=CC=CC3=CC=CC3=C2C=C1</chem>	
		<chem>C1=C[NH]C=N1</chem>		<chem>CC(C)O</chem>		
		<chem>C1=C[NH]N=C1</chem>		<chem>CNC</chem>		
		<chem>C1=NN=N[NH]1</chem>		<chem>CC=O</chem>		
				<chem>N#CC</chem>		
		<chem>C1=C[NH]N=N1</chem>		<chem>CC#N</chem>		
		<chem>CC1=C[NH]N=C1</chem>		<chem>CC(N)=O</chem>		
				<chem>CO</chem>		
				<chem>CC(C)=O</chem>		
				<chem>CC(=O)O</chem>		
				<chem>CC(N)=O</chem>		
				<chem>CN</chem>		

35	36	37	38	39	40	41	42	43	44
									
<chem>N[C@@H](CC(=O)O)C(=O)O</chem>	<chem>CN(C)C#N</chem>	<chem>FC(F)(F)SSC(F)(F)F</chem>	<chem>O=C(C1CC1)C1CC1</chem>	<chem>C#N</chem>	<chem>FC1=C(F)C(F)=C(F)C(F)=C1</chem>	<chem>CCS</chem>	<chem>C#CC(F)(F)F</chem>	<chem>N#CC12CC3CC(CO)C1C2</chem>	<chem>O=CC1=CC=CC=C1</chem>
									
<chem>O=S(=O)(O)OCCO</chem>	<chem>COC</chem>					<chem>CC(C)CS</chem>			
									
<chem>O=S(=O)(O)OCCO</chem>	<chem>CN(C)C(=O)C#N</chem>					<chem>CC(C)S</chem>			
									
	<chem>CN(C)C=O</chem>					<chem>CS</chem>			
									
	<chem>COC=O</chem>					<chem>CCCCS</chem>			
									
	<chem>CN(C)C</chem>					<chem>CCCS</chem>			

 <chem>CC(C)(C)CO</chem>	 <chem>NC(=O)C(F)(F)F</chem>	 <chem>O=C(F)F</chem>	 <chem>FC1=NC(F)=CC=C1</chem>	 <chem>N[C@@H](CO)C(=O)O</chem>	 <chem>C=COCC(F)(F)F</chem>	 <chem>O=S1(=O)CCCC1</chem>	 <chem>CO[Si](C)(C)C</chem>	 <chem>CN(C)P(=O)(N(C)C)N(C)C</chem>
 <chem>CC(C)(C)C#N</chem>	 <chem>OC(C(F)(F)F)C(F)(F)F</chem>	 <chem>O=S=O</chem>	 <chem>FC1=CC=CC=N1</chem>	 <chem>C[C@@H](O)[C@H](N)C(=O)O</chem>		 <chem>C1CCSC1</chem>		
 <chem>CC(C)(C)O</chem>	 <chem>OCC(F)(F)F</chem>		 <chem>ClC1=CC=CC=N1</chem>	 <chem>N[C@@H](CS)C(=O)O</chem>		 <chem>O=S1(=O)CCCC1</chem>		
 <chem>COC(C)(C)C</chem>	 <chem>CC(=O)C(F)(F)F</chem>		 <chem>FC(F)(F)C1=CC=CC=N1</chem>			 <chem>O=S1CCCC1</chem>		
 <chem>CCOC(=O)C(F)(F)F</chem>						 <chem>O=S1CCCC1</chem>		
 <chem>CC(C)(C)S</chem>								

54	55	56	57	58	59	60	61	62
								
<chem>O=C(C(F)F)C(F)F</chem>	<chem>C1=CON=C1</chem>	<chem>O=P(Cl)(Cl)Cl</chem>	<chem>COCC#N</chem>	<chem>C1=CSC=N1</chem>	<chem>N[C@@H](CCC(=O)O)C(=O)O</chem>	<chem>O=P(OC1=CC=CC=C1)(OC1=CC=CC=C1)O</chem>	<chem>O=C(C(F)F)C(F)F</chem>	<chem>OCC1CCCCC1</chem>
								
			<chem>CSCC#N</chem>				<chem>N#CC(F)F</chem>	<chem>CCC1=CC=CC=C1</chem>
								
			<chem>CCCC#N</chem>				<chem>O=CC(F)F</chem>	<chem>OCC1=CC=CC=C1</chem>
								
								<chem>CS(=O)C1=CC=CC=C1</chem>
								
								<chem>CS(=O)(=O)C1=CC=CC=C1</chem>
								
								<chem>N#CC1=CC=CC=C1</chem>
								
								<chem>COC1=CC=CC=C1</chem>



63	64	65	66	67
N#CCC#N	COC(=O)C1=CC=C(C(=O)OC)C=C1	C=O	CSS(C)(=O)=O	C1CCSCC1
			CS(C)(=O)=O	CC1CCC(C)O1
			CS(C)=O	CCC(=O)CC
			CP(C)(C)=O	CCOCC
				CCSCC
				CCP(=O)(CC)CC
				CSCC(C)C
				CC(C)C(=O)C(C)C
				CC(=O)CC(=O)C
				CCCOCCC
				CCCSCCC
				CCOC(C)(C)C
				CCC(=O)OC
				CC(C)OC(C)C
				CCOC(C)C=O