МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
імені В. Н. Каразіна

Кафедра хімічного матеріалознавства

УДК 541.35

До захисту допускаю
_____ Завідувач кафедри
«___» _____ 2023р. д.х.н., професор
О.І. Коробов

НОВИЙ ПІДХІД ДО УРАХУВАННЯ ГЕТЕРОАТОМІВ В МЕТОДІ QSAR ДЛЯ ОПИСУ ФІЗИКО-ХІМІЧНІХ ВЛАСТИВОСТЕЙ

Кваліфікаційна робота магістра
II курсу групи ХМ-63
хімічного факультету
ДЕНИСЕНКА КИРИЛА АНДРІЙОВИЧА

Науковий керівник

к.х.н., доцент кафедри хімічного матеріалознавства А. Б. Захаров

ХАРКІВ 2023

РЕФЕРАТ

Дана кваліфікаційна робота містить 40 сторінок, 2 розділи, 23 підрозділи, 12 рисунків, 12 таблиць, 24 посилання та 37 формул.

Об’єктом дослідження в даній роботі є нові підходи розрахунку топологічних індексів для урахування гетероатомів, їх перевірка, визначення їх якості та порівняння запропонованих підходів.

Мета роботи: запровадження нових підходів до розрахунку топологічних індексів з урахуванням гетероатомів. Перший метод базується на заміні топологічної матриці на матрицю порядків зв’язків. Другий метод базується на використанні реберних графів та оптимізації відповідних значень для обраних ребер. Зробити перевірку можливостей нових підходів, провести порівняння та визначити покращення якості регресійних моделей при використанні дескрипторів розрахованих за допомогою запропонованих методів.

В ході даної роботи були розглянуті теорія графів, реберні графи, види дескрипторів та топологічних індексів з урахуванням гетероатомів. Запропоновані нові підходи в розрахунках топологічних індексів з урахуванням гетероатомів, написано програмне забезпечення з реалізацією розрахунку за даними підходами. Проведено зіставлення з класичним методом розрахунку та порівняння двох нових методів.

Ключові слова: QSPR, МОЛЕКУЛЯРНИЙ ГРАФ, ТОПОЛОГІЧНІ ІНДЕКСИ, МЕТОДИ УРАХУВАННЯ ГЕТЕРОАТОМІВ, ПОРЯДОК ЗВ'ЯЗКУ.

ABSTRACT

This qualification work contains 40 pages, 2 section, 23 subsections, 12 figures, 12 tables, 24 references and 37 formulas.

The object of research in this paper are new approaches to topological indices calculation to properly account heteroatomic systems, their verification, quality of determination, and comparison of former.

Purpose of the work: to introduce new approaches to topological indices calculation those take heteroatoms into account. The first method is based on replacing the topological matrix with the bond order. The second method is based on the use of edge graphs and the optimization of the corresponding values for the selected edges. Verify the capabilities of new approaches, compare, and determine the improvement in the quality of regression models when using descriptors calculated using the proposed methods.

In the framework of this work, graph theory, edge graphs, types of descriptors and topological indices with regard to heteroatoms were considered. New approaches to the calculation of topological indices with heteroatoms were proposed, and software was written to implement the calculation of these approaches. A comparison with the classical method of calculation and a comparison of the two new methods are made.

Keywords: QSPR, MOLECULAR GRAPH, TOPOLOGICAL INDICES, METHODS OF ACCOUNTING FOR HETEROATOMS PRESENCE, BOND ORDER.

ЗМІСТ

ВСТУП 5
1. ЛІТЕРАТУРНІЙ ОГЛЯД 7
1.1. QSPR 7
1.2. Теорія графів 8
1.2.1. Молекулярний граф 9
1.2.2. Реберні графи 12
1.3. Молекулярні дескриптори 14
1.4. Топологічні індекси 16
1.4.1 Індекси Гарольда-Вінера 16
1.4.2 Індекс Платта та індекс Гордона-Скентлбері 17
1.4.3 Індекси Хосоя 18
1.4.4 Індекси Загребської групи 18
1.4.5 Індекс молекулярної зв'язності Рандіча 19
1.5. Методи урахування гетероатомів 20
1.5.1. Молекулярно-топологічний індекс Шульца 20
1.5.2. Індекси Загребської групи 21
1.5.3. Характерний кореневий індекс 21
1.5.4. Індекси розширеної матриці суміжності 22
2. РОЗРАХУНКОВА ЧАСТИНА 23
2.1 Методика проведення 23
2.1.1. Перший метод 23
2.1.2. Другий метод 26
2.2 Результати та обговорення 29
2.2.1. Перший метод 29
2.2.2. Другий метод 34
2.3. Порівняння покращення якості моделей 35
ВИСНОВКИ 38
СПИСОК ЛІТЕРАТУРИ 39

ВСТУП

Зі стрімким розвитком промисловості XIX — XX ст. та становленням хімії як окремої великої дисципліни, важливою складовою ставали матеріали певних експлуатаційних параметрів. Виникала гостра потреба мати змогу отримувати матеріали з заданими характеристиками та сполуки з особливими властивостями. Для розв’язку такого важливого практичного питання, було необхідно напрацювання фундаментальної теоретичної бази для розуміння взаємозв'язку між властивостями молекул та їх структурою.

Розвиток таких досліджень в певний момент перейшов від побудови залежностей емпіричними шляхами до все більш теоретично-зважених моделей. Багаторічний розвиток математичних методів опису молекул привів до появи нового напряму — QSAR («Quantitative Structure-Activity Relationship») або QSPR («Quantitative Structure-Property Relationship»).

Оглядаючись на історію розвитку, можна побачити серед різноманіття методів, як був створений підхід, котрий базується на представленні молекули як графу. Використання такого підходу та топологічних індексів — потужніший інструмент при побудові моделей типу структура-властивість.

Найперші такі методи якісно описували молекули з певних сторін, але все одно мали певні недоліки, як наприклад, не враховували вплив гетероатомів. Впродовж розвитку збільшувалась кількість методів та їх якість. Але питання врахування гетероатомів при розрахунку топологічних індексів не закрите повністю. Та можливо знаходити нові, кращі, розв'язання цього питання.

У зв'язку з цим в даній роботі пропонується два нових підходи для розрахунку топологічних індексів з урахуванням гетероатомі, для покращення якості регресійних рівнянь. А саме, заміна значень топологічної матриці суміжності на квантово-хімічно розраховані порядки зв'язку як перший підхід. Та використання реберних графів для обчислення молекулярних дескрипторів з подальшою оптимізацією значень ребер графу у якості другого методу урахування гетероатомів. Програмно реалізовані алгоритми та проаналізована можливість використання цих підходів. Проведені кількісні порівняння з класичними підходами.

Актуальність даної роботи та напрямку досліджень підтверджується сотнями публікацій у різних журналах за останнє десятиріччя підтверджують актуальність розглянутих тем. Можна виділити найбільш цікаві роботи: [1-15].

1. ЛІТЕРАТУРНІЙ ОГЛЯД

1.1. QSPR

Кількісне співвідношення структура-властивість (*Quantitative Structure-Property Relationships,* QSPR) — математична процедура побудови моделей для прогнозування властивостей хімічних сполук за їх молекулярними предикторами (структурні, фізико-хімічні, біологічні тощо). Також такий підхід дозволяє розв'язувати обернену задачу до таких математичних моделей, що дозволяє знаходити структури хімічних сполук з потрібними властивостями.

Основними шляхами побудови моделей кількісного співвідношення структура-властивість є підходи математичної статистики та використання машинного навчання.

Частіш за все, математично моделі QSPR мають вигляд рівнянь, тобто співвідношення між різноманітними макроскопічними властивостями молекул та параметрами опису структури (дескрипторами), саме вони в певних моделях характеризують властивість, яка розглядається. Дескриптори можуть бути різної природи, як даними, отриманими експериментальним шляхом, так і різноманітними розрахунковими параметрами [16].

Отримане статистично-значиме рівняння ілюструє зв’язок відгуку (**R**) та шуканої властивості від структурних чи фізико-хімічних параметрів молекули (**p**). У найбільш загальному вигляді математичний вираз моделі має наступний вигляд:

$$R = f(p) \tag{1.1}$$

Отже, математично формалізувавши, отримаємо функцію властивості від набору дескрипторів:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_i \cdot x_i, \tag{1.2}$$

де y — деяка шукана властивість, залежна змінна,

$x_1, x_2, \ldots, x_i$ — незалежні змінні, дескриптори або предиктори, на базі яких проводиться побудова моделі,

$\beta_0, \beta_1, \beta_2, \ldots, \beta_i$ — коефіцієнти регресії, внески конкретних дескрипторів [17].

До дескрипторів висувають певні вимоги, без виконання яких неможлива побудова QSPR моделі. Вони повинні мати низьку кореляцію з іншими дескрипторами при високій кореляції з досліджуваною властивістю, за можливості мати фізичну інтерпретацію та кількісно відрізнятися для структур із малими структурними відмінностями.

Зазвичай виділяються такі стадії розв'язання проблеми QSPR:

- Підготовка вхідних даних: набору молекул та відповідних їм значень шуканої властивості. Розділення цього набору на навчальну підвибірку (на основі якої будується модель) та тестову підвибірку (вона використовується для перевірки якості отриманої моделі).
- Проведення оптимізації геометрії молекул.
- Обчислення молекулярних дескрипторів, таких як топологічні, електронні, геометричні або використання властивостей молекул, наприклад, фізико-хімічних.
- Побудова регресійних рівнянь. Проводиться розрахунок математичного рівняння, тобто регресійний аналіз.
- Проведення перевірки якості прогностичних рівнянь за допомогою вже визначеної тестової вибірки та валідаційних метрик. У випадку гарних показників тестування така QSPR модель може буди використана для прогнозування модельованої властивості для інших, нових, молекул.

Такий підхід дозволяє працювати навіть у випадках, якщо у наявності мала за розміром початкова вибірка.

1.2. Теорія графів

Граф — це сукупність об'єктів та бінарних відношень між цими об'єктами.

Хімічний граф — це модель хімічної системи, що використовується для характеристики взаємодій між хімічними об'єктами. Велика кількість різноманітних хімічних об'єктів може бути представлена в спрощеному вигляді графів: групи атомів, молекули, ансамблі молекул, полімери, кластери, реакції, механізми реакцій. Графічне представлення хімічного об'єкта повинно зберігати всі важливі особливості досліджуваної властивості та давати якісні або кількісні висновки, що узгоджуються з

тими, які пропонуються складнішими методами. Для отримання такого представлення виходять з того, що хімічну систему утворюють елементи, між якими існує взаємодія. Елементами, представленими у вигляді вершин графа, можуть бути молекулярні орбіталі, електрони, атоми, групи атомів, молекули або ізомери. Взаємодією між елементами, представленими ребрами графа, можуть бути хімічні зв'язки, незв'язані взаємодії, стадії реакцій, формальні зв'язки між групами атомів або формальні перетворення функціональних груп [18].

Також можна сказати що топологічні індекси — це числа, пов'язані з конституційними формулами шляхом математичних операцій над графом, що представляють ці формули. Необхідність використання таких інструментів, як топологічні індекси, пов'язана з тим, що фізико-хімічні властивості виражаються у вигляді чисел і, таким чином, мають метрику, що дозволяє проводити порівняння та кореляції. На відміну від цього, хімічні структури, навіть виражені в математичній формі графів, є дискретними об'єктами. Для того, щоб кількісно оцінити ступінь подібності або відмінності хімічних структур або знайти кореляції між структурами та властивостями (QSAR або QSPR), необхідно перевести структури в числа. Для електронних факторів, квантової хімії або лінійних співвідношень вільної енергії використовують числові дані. Для стеричних факторів або для гідрофобності/гідрофільності існують добре встановлені числові значення. Однак, топологічні індекси надають простіше рішення, яке, на відміну від методів, таких як методи Мезея, Перлмана, Бердена або Крамера, не потребує тривалого комп'ютерного часу [19].

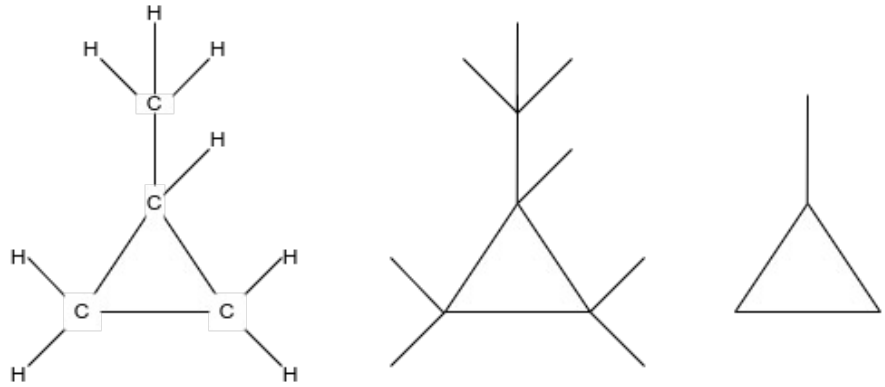
1.2.1. Молекулярний граф

Хімічні структури можна розглядати як графи, тобто як дві непорожні множини V (**V**ertex) та E (**E**dges) елементів, тобто граф (формула 1.3):

$$G = (E, V) \tag{1.3}$$

Елементи V називаються вершинами та символізують атоми. Елементи E називаються ребрами: вони є двосторонніми відношеннями між елементами V, тобто невпорядкованими парами, і символізують ковалентні зв'язки між атомами. Якщо не вказано інше, атоми водню ігноруються. Наприклад, метилциклопропан (Т), рисунок

1.1.а, безводневий граф, який включає тільки неводневі атоми, рисунок 1.1.в. Також він може включати атоми водню, рисунок 1.1.б. [20]



а б в

Рисунок 1.1 . а — Метилциклопропан,
б — водневий граф,
в — безводневий граф.

Дві вершини, з'єднані ребром, називаються суміжними, і граф може бути однозначно описаний матрицею суміжності A, формула 1.4. Ця матриця має елементи a_{ij}, що дорівнюють 1 для суміжних вершин i та j, та нулю в іншому випадку.

$$A(T) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{1.4}$$

Шлях - це послідовність ребер, що не повторюються, така, що між ребрами немає розриву. Шляхи можуть мати повторювані ребра. Хімічні графи є зв'язними графами, оскільки існує принаймні один шлях від будь-якої вершини до будь-якої іншої вершини графа.

Топологічна відстань d_{ij} між двома вершинами i та j - це кількість ребер на найкоротшому шляху між цими вершинами. Граф також однозначно визначається матрицею відстаней D, елементами якої є відстані d_{ij}, формула 1.5.

$$D(T) = \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix} \tag{1.5}$$

Легко бачити, що матриці *A* і *D* симетричні відносно головної діагоналі, та те, що вони мають спільні елементи 1 на відповідних суміжних позиціях та нулі на головній діагоналі, але всі інші нулі *A* замінені в *D* цілими числами, більшими за 1.

Сума рядку *i* або стовпцю *i* у *A* відповідає ступеню вершини - DEG_i або VD_i (vertex degree), формула 1.6. Та називаються ступенями вершин v_{ij}, вони представляють кількість вершин, які є суміжними з вершиною *i*. Котрі називаються за атомами вуглецю первинними, вторинними, третинними або четвертинними, коли вони відповідають вершинам ступеня 1, 2, 3 або 4 відповідно.

$$VD(T) = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 2 \end{bmatrix} \tag{1.6}$$

Подібно до ступенів вершин, можна визначити ступінь ребра e_i як кількість всіх ребер, суміжних з даним ребром *i*. Звичайно, потрібно рахувати обидві кінцеві точки ребра, та це призводить до взаємозв'язку між ступенями ребер та ступенями вершин v_i та v_j цих кінцевих точок, формула 1.7:

$$e_i = v_i + v_j - 2 \tag{1.7}$$

Суми записів по рядках або стовпцях *D* називаються сумами відстаней s_i. Можна показати, що для дерев — центроїд точка з мінімальним значенням суми відстаней, що може бути, або вершина, або пара сусідніх вершин. Чим менша сума відстаней, тим ближче вершина *i* до центроїда графа. Контурами називаються шляхи, початкова і кінцева вершини яких збігаються. Граф без циклів називається деревом, та в хімії такий граф відповідає ациклічній сполуці. Найменша кількість контурів (тобто найменша кількість зв'язків, які необхідно видалити, щоб отримати ациклічну структуру) називається цикломатичним числом μ графа, формула 1.8. Позначивши кількість ребер — *e* та позначивши кількість вершин — *n*, можна отримати наступний вираз:

$$\mu = e - v + 1 \tag{1.8}$$

Вершини графа можна позначати цілими числами від 1 до n - кількістю вершин (це число називають також порядком графа). Два графи називаються ізоморфними, якщо існує таке позначення одного з них, котре зберігає всі суміжності іншого.

1.2.2. Реберний граф

У теорії графів реберним графом L(G) неорієнтованого графа G називається граф L(G), що представляє сусідство ребер графа G.

З найбільш ранніх і найбільш важливих теорем про реберні графи потрібно згадати теорему Вітні, який довів, що за одним винятком структура графа G повністю визначається реберним графом. Іншими словами, за одним винятком, увесь граф може бути відновлений із реберного графа.

Формальне визначення реберного графа:

Нехай задано граф G, тоді його реберний граф L(G) - це такий граф, що

- будь-яка вершина графа L(G) представляє ребро графа G
- дві вершини графа L(G) суміжні тоді та тільки тоді, коли їхні відповідні ребра мають спільну вершину ("суміжні") в G.

Впорядкована послідовність лінійних графів L(G), отриманих за допомогою ітераційної процедури, починаючи з молекулярного графа G:

$$\langle L_0, L_1, L_2, \ldots, L_m \rangle \tag{1.9}$$

де L_0 - лінійний граф нульового порядку, що збігається з вихідним молекулярним графом G;

L_1 - лінійний граф G;

L_2 - лінійний граф першого лінійного графа L_1;

L_m - m-й лінійний граф G, тобто $L_m(G) = L(L_{m-1}(G))$.

Числа вершин A_m та ребер B_m графа L_m задаються наступними співвідношеннями:

$$A_m = B_{m-1} \tag{1.10}$$

$$B_m = \sum_{i=1}^{A_{m-1}} \frac{\delta_i}{2} = \frac{1}{2} \cdot \sum_{i=1}^{A_{m-1}} \delta_i^2 - B_{m-1} \tag{1.11}$$

де A_{m-1} та B_{m-1} - кількість вершин та ребер у лінійному графі (m-1)-го порядку відповідно;

δ_i - ступінь вершини;

Можна зауважити, що кількість ребер у m-му лінійному графі L_m збігається з номером зв'язку (m-1)-го лінійного графа L_{m-1}.

Кожна вершина i_m поточного лінійного графа L_m позначає пару вершин графа нижчого порядку L_{m-1}:

$$i_m = (j_{m-1}, k_{m-1}) \tag{1.12}$$

де дві вершини j та k в L_{m-1} обов'язково з'єднані ребром графа і самі є парами вершин графа L_{m-2}.

Зв'язок між і-м ребром та відповідними вершинами j та k можна записати у вигляді:

$$j_{m-1} \in i_m \ ; \ k_{m-1} \in i_m \tag{1.13}$$

Зв'язаність вершин в процесі лінійної похідної може бути представлена дельтою Кронекера у вигляді:

$$\delta(i_m, i_{m+1}) = \begin{cases} 0, & \text{if } (i_m \in i_{m+1}) \\ 1, & \text{otherwise} \end{cases} \tag{1.14}$$

Це співвідношення можна поширити на будь-який довільний ранг похідної m та *n*, (*m* > *n*), стверджуючи, що $\delta(i_m, i_n) = 1$ тільки в тому випадку, якщо вершина i_m, з'являється хоча б в одній з підмножин, що визначають вершину i_n.

Руїдж і Вільф розглянули послідовність графів, формула 1.9. Вони показали, що для кінцевого зв'язного графа G можливі тільки чотири види поведінки цієї послідовності:

- Якщо G - циклічний граф, то L(G) і всі — наступні ізоморфні самому графу G. Це єдине сімейство зв'язних графів, для яких L(G) ізоморфне G.
- Якщо G - клешня, то L(G) і всі наступні графи є трикутниками.
- Якщо G - шлях, то кожен наступний реберний граф - укорочений шлях, поки він не перетвориться на порожній граф.
- У всіх інших випадках розмір графів збільшується необмежено.

Якщо G незв'язний, то ця класифікація застосовна до кожної окремої компоненти зв'язності графа G.

Приклад отримання реберного графа. Якщо G початковий граф, то L_1 та L_2 реберні графи першого та другого порядку відповідно, Рисунок 1.2:

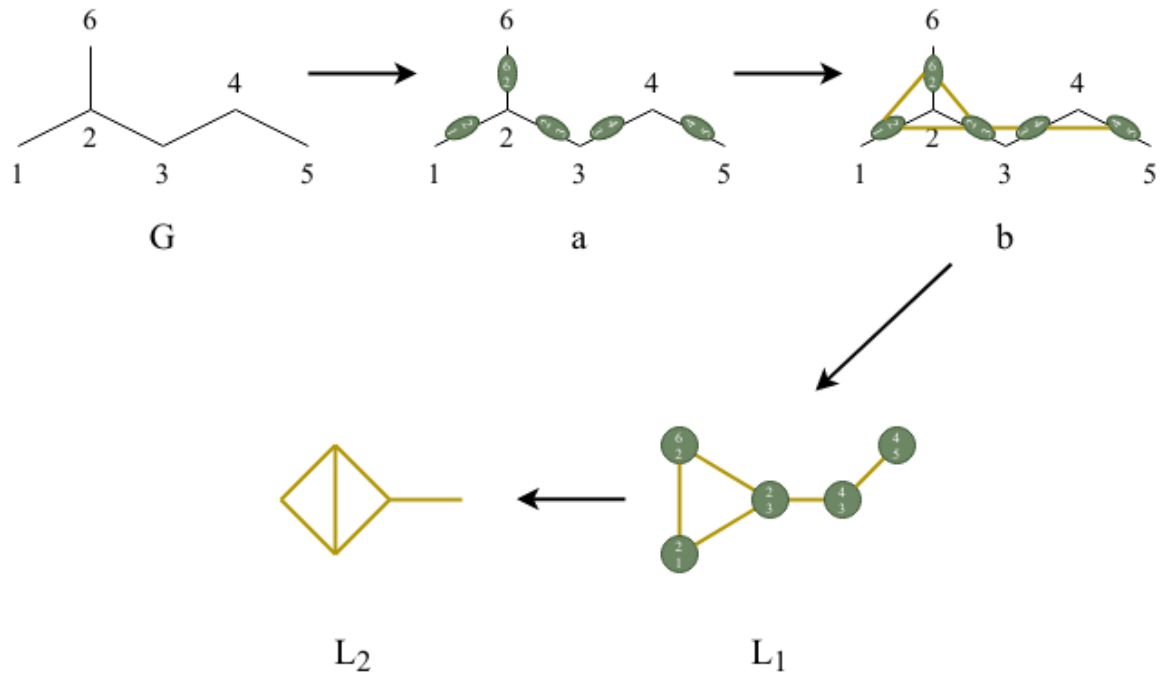


Рисунок 1.2 Схема утворення реберного графа 1-го порядку (L_1) та другого порядку (L_2)

a — граф з відміченими вершинами реберного графа

b — граф з відміченими вершинами та ребрами реберного графа L_1

1.3. Молекулярні дескриптори

Молекулярний дескриптор — це кінцевий результат логічної та математичної процедури, яка перетворює хімічну інформацію, закодовану в символічному представленні молекули, в корисне число або результат деякого стандартизованого експерименту [18].

Класифікують дескриптори:

- за походженням
  - “Експериментальні” (logP, розчинність у воді) - здебільшого розраховуються за комп'ютерними моделями
  - Розраховані (отримуються з 1D, 2D та інших представлень)
- по описуваному об'єкту
  - Глобальний (що описують молекулу в цілому (молекулярний об'єм, молекулярна поверхня, дипольний момент, топологічні індекси, ...)

- Локальний (описують окремі атоми або фрагменти молекул (атомні заряди, поляризованості зв'язків)
- За полем (описують молекулярні поля в області, що оточує молекулу: електростатичний потенціал, ...)

- за розмірністю представлення молекулярної структури
  - 0D-дескриптори. Об'єднує всі молекулярні дескриптори, які не надають ніякої інформації про молекулярну структуру або зв'язок атомів. Наприклад, кількість атомів, кількість зв'язаних атомів або молекулярна маса є 0D-дескрипторами. Перевагою цих дескрипторів є те, що їх легко отримати. Однак, взяті один за одним, ці дескриптори не містять багато інформації про структурні особливості молекул і тому їх часто комбінують з іншими дескрипторами.
  - 1D-дескриптори. Цей клас об'єднує молекулярні дескриптори, які можуть бути розраховані на основі набору підструктур, таких як функціональні групи. Найбільш поширеними 1D-дескрипторами є відбитки пальців. Наприклад, відбиток може бути бінарним вектором, де 1 вказує на наявність структурної ознаки, а 0 - на її відсутність.
  - 2D-дескриптори. Цей клас об'єднує всі дескриптори, які надають інформацію про молекулярну топологію на основі графового представлення молекул. Типовими 2D-дескрипторами є матриця суміжності, кулонівська матриця або матриця відстаней. У випадку матриці суміжності дескриптор вказує, які атоми пов'язані в молекулі. Оскільки 2D-дескриптори чутливі до структурних особливостей молекули (розмір, форма і симетрія), вони є загальноприйнятим вибором в якості молекулярних дескрипторів.
  - 3D-дескриптори. Цей клас об'єднує всі геометричні дескриптори, які надають інформацію про просторові координати атомів молекули. Найбільш відомими 3D-дескрипторами є молекулярна матриця та 3D-MoRSE дескриптори. У випадку молекулярної матриці дескриптор представляє декартові координати (x, y, z) кожного атома.
  - 4D-дескриптори. 4D-дескриптори також називають "сітковими дескрипторами". Ці дескриптори, на додаток до молекулярної

геометрії, вводять четвертий вимір. Цей новий вимір зазвичай характеризує взаємодію між молекулою (молекулами) та активним центром (центрами) рецептора або множинні конформаційні стани молекули (молекул).

1.4. Топологічні індекси

Топологічні індекси — це 2D-дескриптори, отримані на основі топологічного представлення молекулярних структур (тобто молекулярного графа); при розрахунку цих дескрипторів не враховується інформація про просторовий розподіл атомів. Існують різні категорії топологічних індексів.

Можна виділити наступні різновиди топологічних індексів:

- Топологічні індекси першого покоління (Індекси Гарольда Вінеру, Індекс Платта, індекс Гордона-Скентлбері)
- Топологічні індекси другого покоління (Індекс молекулярної зв'язності Рандіча, індекс Хола, інформаційно-теоретичні індекси)
- Топологічні індекси третього покоління (Інформаційні індекси на основі розподілу сум відстаней, хіміко-тополочогічні відстані)

1.4.1 Індекси Гарольда-Вінера.

Приблизно в той же час, коли Норберт Вінер розробив свою революційну роботу з кібернетики, Гарольд Вінер запропонував у 1947-1948 роках один з перших молекулярних дескрипторів топологічної природи для ациклічних насичених вуглеводнів (алканів). Суму числа зв'язків, що з'єднують усі пари атомів, він назвав числом кількості шляхів. Насправді ця первісна назва була неправильною, оскільки в графах шляхи можуть мати різну довжину, і в наш час цей молекулярний дескриптор, перейменований в індекс Вінера, позначається *W*. Чим компактніший граф, тим менше значення *W*. Пізніше Хосой показав, що *W* є півсумою всіх елементів матриці відстаней, або сумою суми дистанції, формула 1.15:

$$W = 0.5 \cdot \sum_{i}^{n} \sum_{j}^{n} d_{ij} = 0.5 \cdot \sum_{i}^{n} s_i \qquad (1.15)$$

Рувре перевідкрив ці ідеї та запропонував як топологічний індекс подвійне значення *W*. Було запропоновано кілька аналітичних формул для обчислення *W* для різних класів графів, таких як лінійні алкани, бензеноїди тощо. Крім *W*, Вінер запропонував також так зване число полярності *p* графа як число пар вершин, розділених трьома ребрами, або, іншими словами, число відстаней довжини три в *D*. Як *W*, так і *p* були застосовані Вінером і Платтом для кореляції структури алканів з термодинамічними властивостями, такими як тиск, теплоти утворення і випаровування, молекулярний об'єм та молярна рефракція. Індекс *W* важливий не тільки тому, що він був винайдений першим, але й тому, що він простий в обчисленні та досить успішно використовується для описання багатьох властивостей. Його основним недоліком є висока виродженість, тобто той факт, що багато неізоморфних графів мають одне і те ж значення *W*.

1.4.2 Індекс Платта та індекс Гордона-Скентлбері .

Практично одночасно з Вінером Платт ввів індекс *F* як суму кількостей зв'язків, що прилягають до кожного зі зв'язків у молекулі, або в графо-теоретичній термінології *F* - це сума всіх степенів ребер. Ймовірно, через ширше використання індексу Вінера, ніж індексу Платта, пріоритет Платта ігнорується.

Професор Манфред Гордон, який розробив застосування теорії графів в хімії полімерів, запропонував разом з доктором Скентлбері в 1964 році індекс, який вони назвали *N*, що визначається як кількість різних способів, якими шлях *P* довжиною два (тобто пропановий підграф) може бути накладений на даний молекулярний граф. Можна продемонструвати, що *F=2N*, що ще раз вказує на те, що в цей період кілька незалежних хіміків прийшли до подібних ідей. Існують також інші співвідношення з дескрипторами графів, про які йшлося вище:

$$2N_2 = \sum_{i}^{n} v_i(v_i - 1) \tag{1.16}$$

$$2N_2 = \sum_{i,j}^{n} (v_i + v_j) - 2 \tag{1.17}$$

1.4.3 Індекси Хосоя.

На початку сімдесятих років Харуо Хосоя запропонував ще один молекулярний дескриптор під назвою *Z*. Його визначення має вигляд, формула 1.18:

$$Z = \sum_{k}^{[n/2]} p(G,k) \tag{1.18}$$

де *p(G, k)* - кількість способів, якими можна вибрати k ребер графа так, щоб жодні два з них не були суміжними, а квадратні дужки Гаусса вказують на найменше ціле число, що не перевищує числа, узятого в ці дужки.

Кількість *e* ребер дорівнює *p(G, 1)* і за визначенням *p(G,0)* дорівнює 1. Для дерев існує простіший зв'язок між *Z* і абсолютними значеннями коефіцієнтів характеристичного полінома графа (отриманого шляхом введення змінної x на головній діагоналі матриці суміжності *A* і прирівнювання отриманого таким чином визначника до нуля; корені цього характеристичного полінома є власними значеннями). Сума цих абсолютних значень дорівнює *Z*. Хосоя ввів термін топологічний індекс, який зараз став загальноприйнятим у науковому товаристві. Було показано, що індексу Z притаманна відмінна кореляційна здатність з багатьма фізичними властивостями.

1.4.4 Індекси Загребської групи.

Приблизно в 1975 році хіміки, які цікавилися хімічними застосуваннями теорії графів у столиці Хорватії, Загребі, запропонували два топологічні індекси, отримані з матриці суміжності:

$$M_1 = \sum_{i=1}^{A} \delta_i^2 \tag{1.19}$$

де δ_i - ступінь вершини і-го атома

$$M_2 = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\delta_i \cdot \delta_j) \tag{1.20}$$

де δ_i та δ_j - пари суміжних вершин молекулярного графа.

Ця група складалася з квантових хіміків і математиків: Ненад Тринайстич, який був завідувачем лабораторії в Інституті Руджера Бошковича в Загребі, Хорватія, з кількома співробітниками, та Іван Гутман, який був професором в Університеті Крагуєваца, Сербія. Та ще один хорватський хімік, який взаємодіяв з цією групою, - Мілан Рандіч. Показали, що індекси Платта, Гордона-Скентлбері та перший з індексів Загребської групи пов'язані між собою:

$$F = 2N_2 = M_1 - e \tag{1.21}$$

За допомогою матриці суміжності індекси Загребської групи також можуть бути виражені як:

$$M_1 = \sum_{i}^{n} [A^2]_{ii} [A^2]_{ii} \tag{1.22}$$

$$M_2 = \sum_{i,j}^{n} [A^2]_{ii} [A^2]_{jj} \tag{1.23}$$

де $[A^2]_{ii} = \delta_i$

1.4.5 Індекс молекулярної зв'язності Рандіча

Близьким до другого індексу Загребської групи, але набагато більш вдалим та на сьогодні широко використовуваним і прийнятим є так званий індекс Рандіча, введений ним у 1975 році:

$$\chi = \sum_{i,j} (\nu_i \cdot \nu_j)^{-1/2} \tag{1.24}$$

З цього індексу починається група топологічних індексів другого покоління, хоча топологічні індекси першого покоління будуть продовжувати з'являтися протягом наступних кількох років. Мало того, що для суми оберненого квадратного кореня з добутків ступенів кінцевих точок для всіх ребер виродження є низьким, так ще й кореляційна здатність індексу χ є досить хорошою для багатьох фізичних та біохімічних властивостей. Цілком ймовірно, що цей індекс досі є найбільш широко використовуваним серед усіх топологічних індексів і дав початок кільком іншим, спорідненим індексам.

1.5. Методи урахування гетероатомів

Існує велика кількість різноманітних методів ураховування гетероатомів у структурі при розрахунку топологічних молекулярних дескрипторів. Як індивідуальних, так і модифікацій методів без урахування гетероатомів.

1.5.1 Молекулярно-топологічний індекс Шульца

Топологічний індекс, спочатку названий Щульцем як “Молекулярний топологічний індекс”. Індекс Шульца визначається за формулою 1.25: [21]

$$MTI = \sum_{i=1}^{A} \left[(A + D) \cdot v \right]_i = \sum_{i=1}^{A} t_i \tag{1.25}$$

де A - матриця суміжності;

D - матриці відстаней;

v - A-вимірний вектор-стовпець, утворений ступенями вершини атомів A в безводневому молекулярному графі;

t_i - елементи, які називаються числами складності, A-вимірного вектора-стовпця t отримані наступним чином:

$$t = (A + D)v \tag{1.26}$$

Тобто матриці D і A сумуються, а потім сума множиться на вектор v.

Запропоновано узагальнення молекулярно-топологічного індексу Шульца для врахування гетероатомів та кратних зв'язків на основі матриці відстаней Бариша.

Матриця відстаней Бариша $^{Z}\mathbf{D}$ - симетрична зважена матриця відстаней, що враховує одночасно наявність гетероатомів та кратних зв'язків у молекулі, визначається за формулою 1.27 та 1.28:

$$\left[{}^{Z}D \right]_{ij} = \begin{cases} d_{ij}(Z, \pi^*), & \text{if } i \neq j \\ 1 - \frac{Z_C}{Z_i}, & \text{if } i = j \end{cases} \tag{1.27}$$

$$d_{ij}(Z, \pi^*) = \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \cdot \frac{Z_C^2}{Z_{b(1)} \cdot Z_{b(2)}} \right) \tag{1.28}$$

де Z_C - атомний номер атома вуглецю;

Z_i - атомний номер i-го атома;

π^* - порядок умовного зв'язку;

$d_{ij}(Z,\pi^*)$ - зважена топологічна відстань, обчислена шляхом підсумовування ваг ребер над усіма зв'язками d_{ij}, що беруть участь у найкоротшому шляху між вершинами v_i та v_j;

d_{ij} - топологічна відстань.

1.5.2 Індекси Загребської групи

Також і для індексів Загребської групи, що визначаються за формулами 1.19 та 1.20, існують підходи для урахування гетероатомів. Таким є використання хімічної матриці суміжності.

Матриця хімічної суміжності є прикладом несиметричних зважених матриць суміжності, які визначаються як:

$$[A]_{ij} = \begin{cases} \frac{m_j}{m_C}, & \text{if } (i,j) \in E(G) \\ 0, & \text{otherwise} \end{cases} \tag{1.29}$$

де m_j - атомна маса пов'язана з j-ю вершиною, зв'язаною з вершиною i,

m_C - атомна маса атома вуглецю.

Хімічні матриці суміжності, засновані на відносних атомних масах, використовуються для розрахунку індексу атомно-молекулярного зв'язку, Загребських топохімічних індексів та топохімічного індексу суперадсорбції.

1.5.3 Характерний кореневий індекс (CRI)

Це сума позитивних власних значень λ матриці шляху X_p, заснована на зв'язності шляху, розрахованої за ступенем валентності вершини δ^v атомів у шляху.

$$CRI = \sum_{i=1}^{n^+} \lambda_i^+ \tag{1.30}$$

Де матриця шляху X_p, визначається за формулою 1.29:

$$[X_p]_{ij} = \begin{cases} d_{ij}(\delta), & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{1.31}$$

де $d_{ij}(d)$ - зв'язність найкоротшого шляху між вершинами v_i та v_j.

Діагональні записи дорівнюють нулю, за припущення, що немає зв'язності атома з самим собою. Ця матриця є розширенням матриці **X**, яка враховує тільки пари суміжних вершин.

Дескриптор CRI зберігає інформацію про всі зв'язки в безводневому молекулярному графі і є чутливим до наявності гетероатомів в молекулі.

1.5.4 Індекси розширеної матриці суміжності.

Індекси розширених матриць суміжності (**EA**-індекси) являють собою два молекулярних дескриптори, що розраховуються з розширених матриць суміжності **EA**, запропонований Янгом у 1994 р. Перший - це сума абсолютних власних значень матриці **EA**, яка називається індексом EAS:

$$EA = \sum_{i=1} \lambda_i^{EA} \tag{1.32}$$

Другий молекулярний дескриптор - максимальне абсолютне власне значення матриці EA, що називається індексом EAmax:

$$EAmax = max_i \left| \lambda_i^{EA} \right| \tag{1.33}$$

Ці дескриптори враховують гетероатоми та кратні зв'язки, мають високу дискримінаційну здатність і добре корелюють з рядом фізико-хімічних властивостей і біологічною активністю органічних сполук.

Розширена матриця суміжності (**EA**) - це зважена матриця суміжності A*A, елементи якої визначаються як функції локальних інваріантів вершин матриці суміжності A та деяких атомних властивостей. Визначені функції спрямовані на усунення виродження елементів матриці суміжності, яка є бінарною матрицею.

Розширеною матрицею суміжності вершин (або просто розширеною матрицею суміжності) називається матриця суміжності **EA**, елементи якої визначаються за формулою 1.34.

$$[EA]_{ij} = \begin{cases} a_{ij} \cdot \frac{\delta_i/\delta_j + \delta_j/\delta_i}{2}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{1.34}$$

де a_{ij} - елементи матриці суміжності, а d - ступінь вершини.

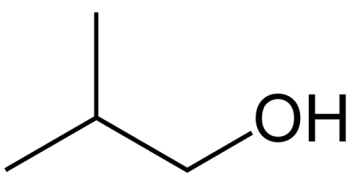
2. РОЗРАХУНКОВА ЧАСТИНА

2.1 Методика проведення розрахунку

Нами запропоновано 2 різні підходи для урахування гетероатомів.

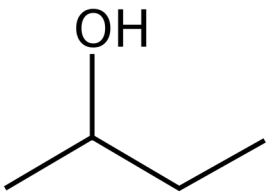
2.1.1 Перший метод

Через те що у класичних молекулярних графах усі неводневі атоми розглядаються як однакові, то маємо ситуацію, при котрій ніяким чином не ураховується природа атомів. Ця проблема стає помітною коли у сполуці присутні гетероатоми, в такому випадку можливе виникнення колізій — випадків коли для різних сполук маємо однакові матриці суміжності, Рисунок 2.1, або маємо однакові ступені вершин, що призводить до отримання однакових значень індексів для різних молекул.



	At	1	2	3	4	5
C	1	0	1	0	0	0
C	2	1	0	1	1	0
C	3	0	1	0	0	0
C	4	0	1	0	0	1
O	5	0	0	0	1	0

а

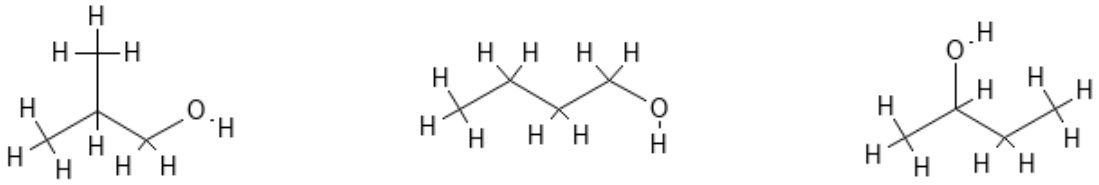


	At	1	2	3	4	5
C	1	0	1	0	0	0
C	2	1	0	1	1	0
C	3	0	1	0	0	0
C	4	0	1	0	0	1
O	5	0	0	0	1	0

б

Рисунок 2.1 Порівняння матриць суміжності для а) 2-метилпропан-1-олу та б) бутан-2-олу

Якщо ураховувати водневі атоми, розглянувши ізомери бутан-1-олу рисунок 2.2, то для класичного підходу спостерігається збільшення кількості колізій.



а б в

Рисунок 2.2 Структурна формула з водневими атомами а) 2-метилпропан-1-ол, б) бутан-1-ол та в) бутан-2-ол

Таким чином отримаємо для 2-метилпропан-1-олу, бутан-2-олу та бутан-1-олу ідентичні матриці суміжності, які наведені на рисунках 2.3-2.5, а також топографічні індекси. При використанні таких дескрипторів у регресійному аналізі отримані моделі будуть мати низькі як описові, так і прогностичні здібності.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	VD
1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	4
2	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	4
3	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	4
4	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	4
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
11	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1

Рисунок 2.3 Матриця суміжності та ступенів вершин 2-метилпропан-1-олу

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	VD
1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	4
2	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	4
3	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	4
4	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	4
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1

Рисунок 2.4 Матриця суміжності та ступенів вершин бутан-1-ол

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	VD
1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	4
2	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	4
3	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	2
4	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	4
5	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	4
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1

Рисунок 2.5 Матриця суміжності та ступенів вершин бутан-2-ол

Альтернативним варіантом було розглянуто використання порядку зв'язку у якості матриці суміжності.

Їх розрахунок проводився у декілька етапів. Проводилася передоптимізація геометрії молекули з використанням силового поля, що імплементовано в бібліотеці RDKit [22].

Наступним кроком після передоптимізації — розрахунок електронної густини та інтегралів перекривання. Був використаний програмний пакет GAMESS [23], та оптимізація з розрахунком виконувалися з його використанням.

У якості метода розрахунку було використано напівемпіричний метод AM1.

Після отримання електронної густини та інтегралів перекривання необхідно поелементно перемножити ці дві матриці. Результуюча матриця є розгорнутою матрицею порядків зв'язків, для її використання проводиться складання по блоках, у межах одного атому.

На Рисунку 2.6 наведена загальна схема проведення розрахунку.

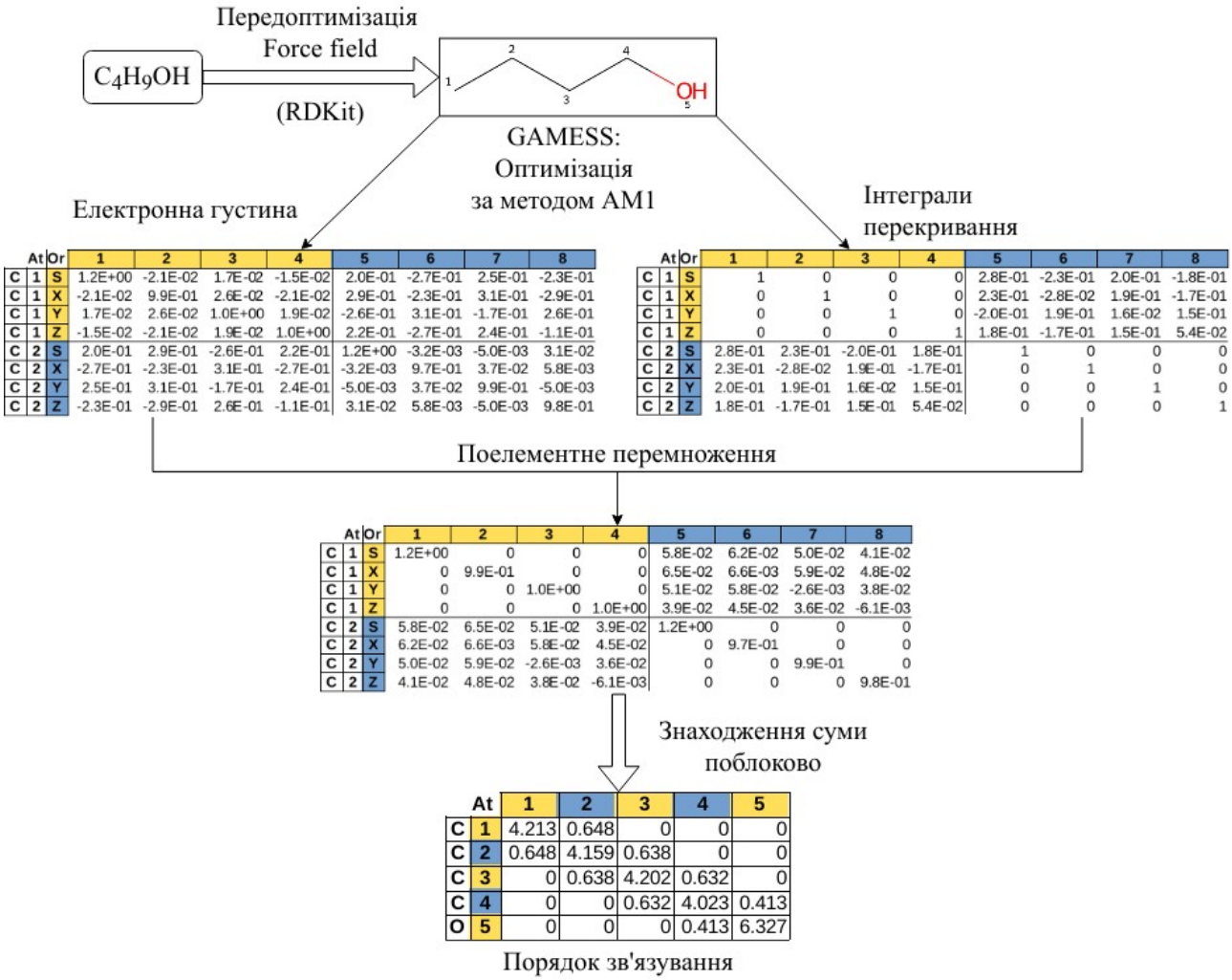


Рисунок 2.6 Схема проведеного розрахунку

За отриманими значеннями порядку зв'язку були розраховані наступні індекси: ZM1, ZM2, ZM3 та χ.

2.1.2 Другий метод

Існуючі підходи розрахунків молекулярних дескрипторів пропонують різноманітні шляхи описання властивостей молекул. Та їх об’єднує одне — розрахунок дескриптора за певним правилом без можливості подальшої оптимізації отриманих значень.

Запропонований підхід базується на двох ідеях

- Реберні графи різних порядків містять інформацію про молекулу
- Можливість оптимізувати параметри матриці суміжності дозволить отримати якісніші результати.

Схема реалізації підходу представлена на Рисунку 2.7. Та проводиться наступним чином:

1. Вибираємо початкові параметри:
    1. Який максимальний ступінь реберного графа буде розраховуватися.
    2. Ребра графу, які будуть оптимізуватися.
    3. Точність оптимізації.
2. Встановлюються початкові значення матриці суміжності для кожного графа.
3. Розраховуються дескриптори з початкових значень.
4. За методом найменших квадратів розраховувались значення коефіцієнтів рівняння:

$$Y = A_0 + A_1 \cdot X_1 + A_2 \cdot X_2 + \cdots + A_n \cdot X_n \tag{2.1}$$

де X_i — дескриптори розраховані з реберного графа (i-1) ступеню

5. Використано у якості мінімізованої функції RSS, формула 2.2:

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{2.2}$$

6. Проведення оптимізації для обраних типів ребер, методом Пауелла, та перевірка чи була отримана бажана точність. Якщо ні — перерахунок з новими значеннями матриці суміжності з пункту 3.

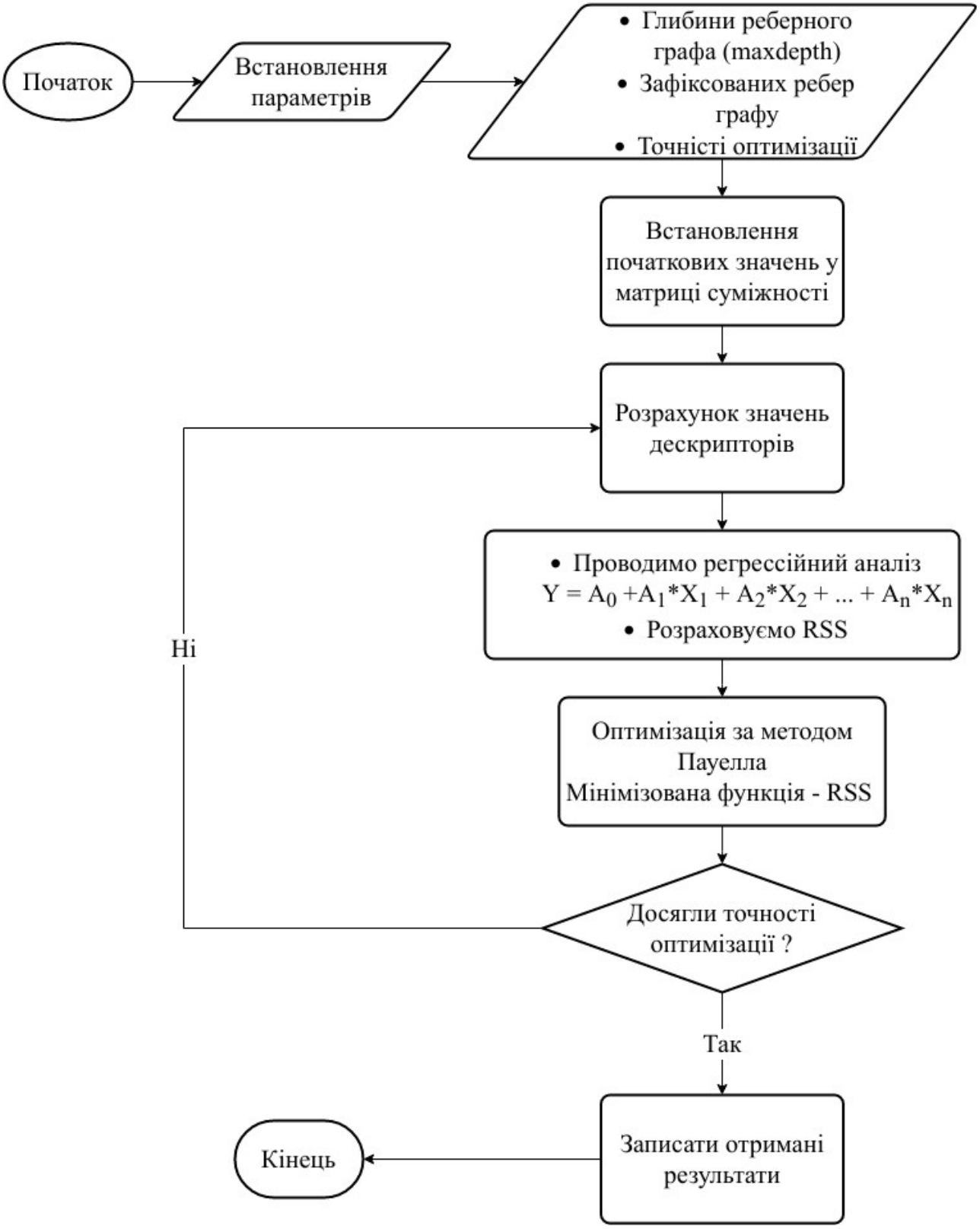


Рисунок 2.7 Схема розрахунку

2.2 Результати та обговорення

Досліджувана вибірка являє собою набір з 269 молекул з гетероатомами та температури кипіння для цих сполук. Молекули представлені наступними класами: спирти, етери, тіоетери, первинні та вторинні аміни.

2.2.1 Перший метод

Отримані індекси використовувалися для побудови регресійних моделей для описання температури кипіння. Були побудовані як однопараметричні, так і багатопараметричні регресійні моделі, у якості дескрипторів виступають розраховані попередньо індекси у різних комбінаціях. З максимально можливою чотирьох параметричною моделлю, дескриптори — ZM1, ZM2, ZM3, χ.

Приведені значення на графіках відповідають коефіцієнту детермінації (R^2), розрахованого за формулою 2.3, для рівнянь, з використанням дескрипторних наборів.

$$R^2 = 1 - \frac{\sum_n^i (y_i - \hat{y}_i)^2}{\sum_n^i (y_i - \bar{y}_i)^2} \tag{2.3}$$

де $\sum_n^i (y_i - \hat{y}_i)^2$ – сума квадратів залишків регресії;

$\sum_n^i (y_i - \bar{y}_i)^2$ – загальна сума квадратів відхилень від середнього значення;

y_i – експериментальні значення;

$\hat{y}_i$ – значення, знайдені за рівнянням регресії;

$\bar{y}_i$ – середнє значення експериментальних значень [24].

У стовпцях наступних таблиць міститься такі дані:

Стовпчик №2 (Topol) відповідає значенням R^2 у випадку розрахунку дескрипторів класичним методом, за топологічною матрицею.

Стовпчик №3 (QS, quantum supervised) — значення R^2, де дескриптори розраховувалися через заміну топологічної матриці на порядок зв’язку.

Стовпчик №4 (Topol_wH) — відповідає значенням R^2 у випадку розрахунку дескрипторів класичним чином, за топологічною матрицею з урахуванням атомів водню.

Стовпчик №5 (QS_wH) — значення R^2, де дескриптори розраховувалися через заміну топологічної матриці на порядок зв'язку з урахуванням атомів водню.

Стовпчик (QS - Topol) різниця між стовпчиком №2 та №3.

При розбитті вибірки за класами отримали наступні значення покращення якості моделі. У Таблиці 2.1 наведені значення для спиртів. А у Таблиці 2.2 для етерів:

Таблиця 2.1 Покращення якості моделі на прикладі спиртів

Alcohols (48)			
	Topol	**QS**	**QS – Topol**
ZM1	0.470	0.821	0.351
ZM2	0.385	0.708	0.324
ZM3	0.071	0.806	0.735
RANDICH	0.927	0.863	-0.064
ZM1+ZM2	0.598	0.955	0.357
ZM1+ZM3	0.936	0.906	-0.031
ZM1+RANDICH	0.958	0.943	-0.015
ZM2+ZM3	0.831	0.964	0.133
ZM2+RANDICH	0.960	0.950	-0.009
ZM3+RANDICH	0.957	0.935	-0.022
ZM1+ZM2+ZM3	0.937	0.972	0.034
ZM1+ZM2+RANDICH	0.960	0.959	-0.001
ZM1+ZM3+RANDICH	0.959	0.985	0.026
ZM2+ZM3+RANDICH	0.960	0.968	0.008
ZM1+ZM2+ZM3+RANDICH	0.961	0.985	0.024

Таблиця 2.2 Покращення якості моделі на прикладі етерів

Ethers (69)			
	Topol	**QS**	**QS – Topol**
ZM1	0.761	0.947	0.187
ZM2	0.733	0.899	0.166
ZM3	0.373	0.943	0.569
RANDICH	0.978	0.961	-0.017
ZM1+ZM2	0.777	0.977	0.199
ZM1+ZM3	0.968	0.962	-0.007
ZM1+RANDICH	0.978	0.970	-0.007
ZM2+ZM3	0.927	0.979	0.052
ZM2+RANDICH	0.978	0.974	-0.004
ZM3+RANDICH	0.978	0.969	-0.009
ZM1+ZM2+ZM3	0.969	0.979	0.010
ZM1+ZM2+RANDICH	0.979	0.979	0.000
ZM1+ZM3+RANDICH	0.979	0.979	0.000
ZM2+ZM3+RANDICH	0.980	0.979	0.000
ZM1+ZM2+ZM3+RANDICH	0.980	0.980	-0.001

У випадку не розділення вибірки за класами, а використання її усієї, також спостерігається покращення якості моделей, Таблиця 2.3:

Таблиця 2.3 Покращення якості моделей на прикладі усієї вибірки

Different compound classes (269)

	Topol	QS	QS – Topol
ZM1	0.337	0.518	0.181
ZM2	0.316	0.447	0.131
ZM3	0.159	0.535	0.376
RANDICH	0.446	0.409	-0.037
ZM1+ZM2	0.357	0.631	0.275
ZM1+ZM3	0.445	0.539	0.094
ZM1+RANDICH	0.447	0.584	0.137
ZM2+ZM3	0.416	0.609	0.193
ZM2+RANDICH	0.448	0.447	0.000
ZM3+RANDICH	0.447	0.554	0.107
ZM1+ZM2+ZM3	0.446	0.645	0.199
ZM1+ZM2+RANDICH	0.449	0.700	0.251
ZM1+ZM3+RANDICH	0.447	0.695	0.249
ZM2+ZM3+RANDICH	0.449	0.612	0.163
ZM1+ZM2+ZM3+RANDICH	0.449	0.912	0.463

Найбільше покращення спостерігається для однопараметричних моделей з дескрипторами ZM1, ZM2, ZM3. Та для тіоетерів, первинних та вторинних амінів на Рисунках 2.7-10 наведені графіки покращення однопараметричних моделей:

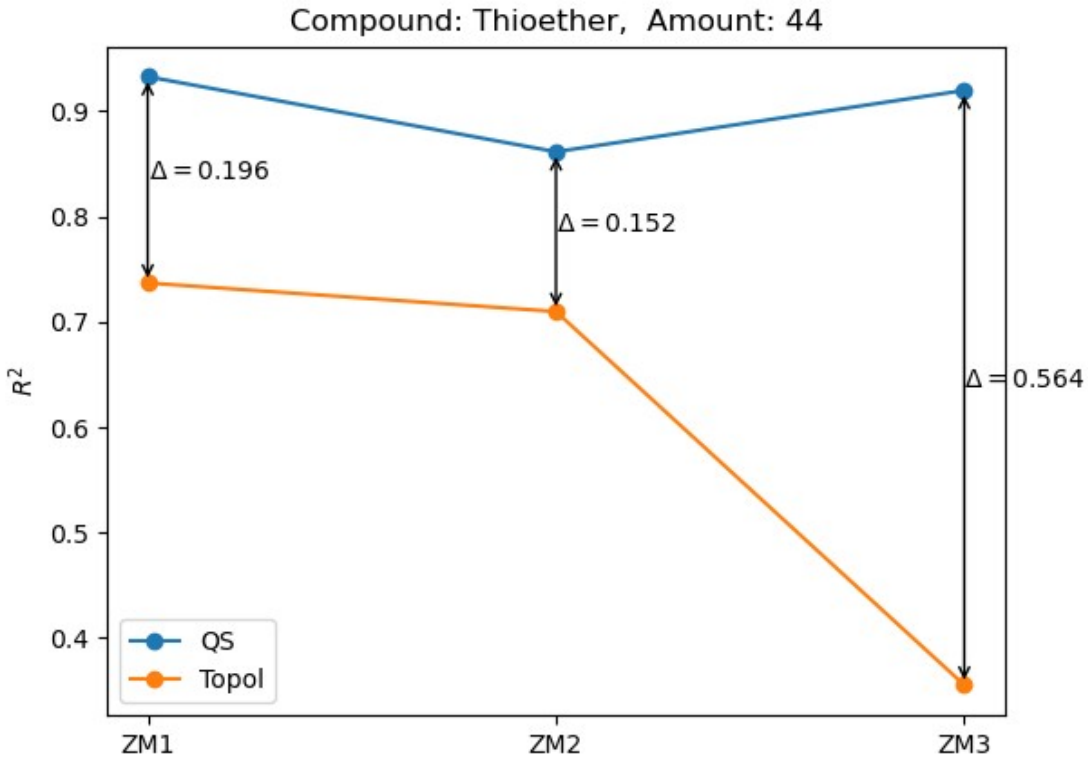


Рисунок 2.7 Графік покращення якості моделей на прикладі тіоетерів

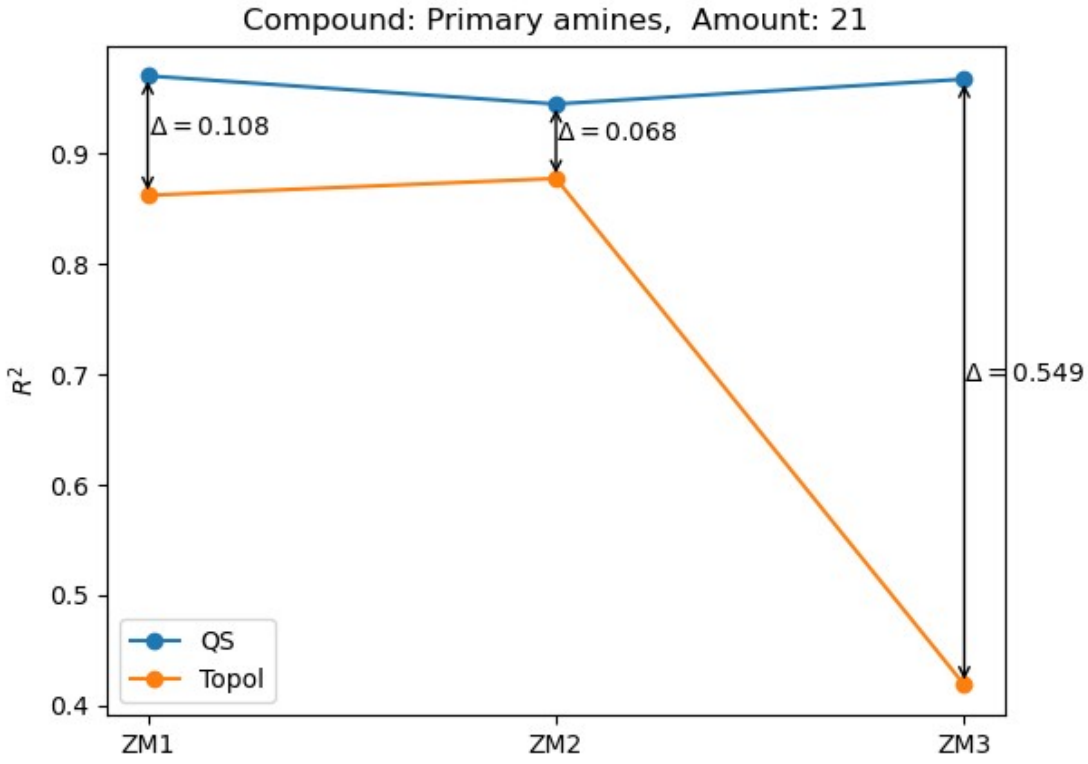


Рисунок 2.8 Графік покращення якості моделей на прикладі первинних амінів

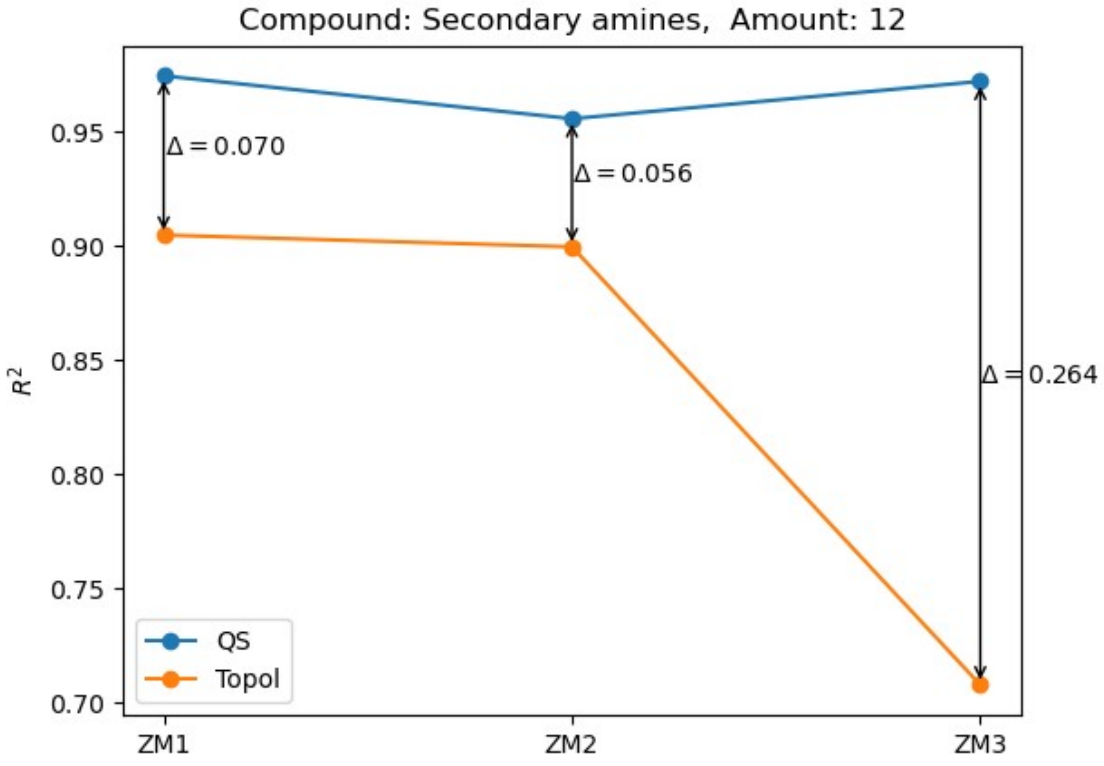


Рисунок 2.9 Графік покращення якості моделей на прикладі вторинних амінів

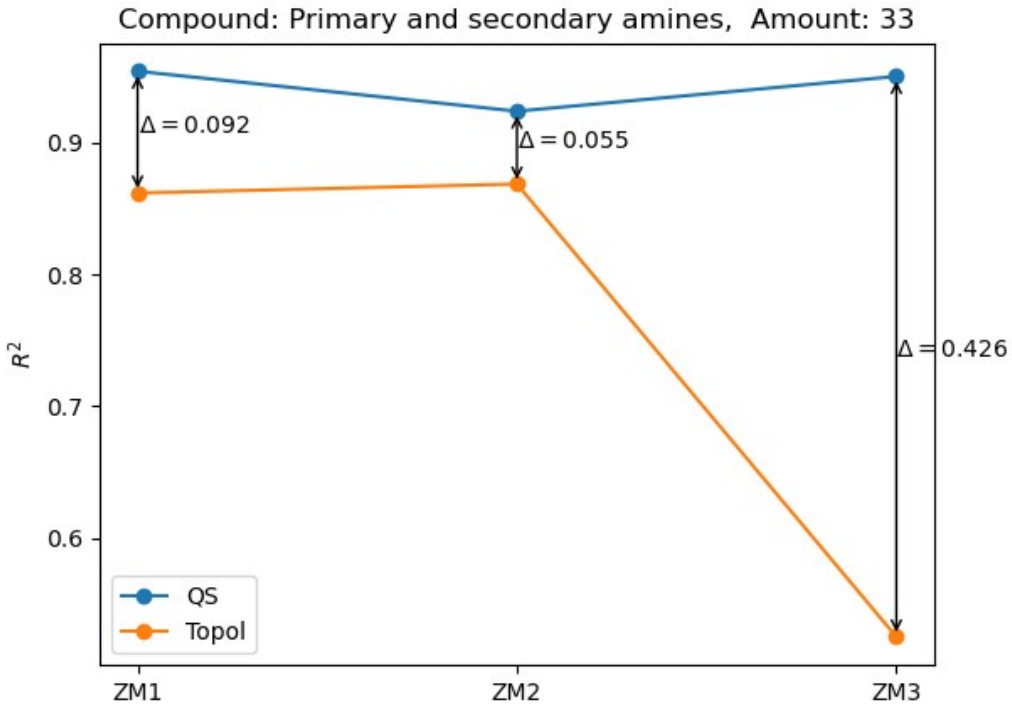


Рисунок 2.10 Графік покращення якості моделей на прикладі первинних та вторинних амінів

Спостерігаємо суттєве покращення якості описання температури кипіння при використанні однопараметричних рівнянь для дескрипторів ZM1, ZM2, ZM3. Для χ спостерігається невелике погіршення.

При використанні підходу з урахуванням атомів водню, спостерігаємо ще більше покращення при використанні порядку зв'язку, замість топологічної матриці, у випадку однопараметричних рівнянь для дескрипторів ZM1, ZM2, ZM3. Але для їх комбінацій (багатопараметричні моделі) спостерігаємо незначний приріст, а у певних випадках суттєве погіршення якості описання.

Урахування атомів водню для малих вибірок та близьких сполук, веде до утворення критичної кількості колізій. Спостерігаємо однакове значення коефіцієнту детермінації для однопараметричних моделей.

Отримати якісне рівняння регресії можливо не завжди, а іноді взагалі не можливо ні яке, через неможливість обернути матрицю. Такий випадок наведений у Таблиці 2.4:

Таблиця 2.4 Значення коефіцієнтів детермінації для тіоетерів

	Topol	QS	Topol_wH	QS_wH
ZM1	0.737	0.932	0.935	0.937
ZM2	0.710	0.861	0.935	0.939
ZM3	0.355	0.919	0.935	0.939
RANDICH	0.982	0.957	0.935	0.935
ZM1+ZM2	0.766	0.986	-	0.966
ZM1+ZM3	0.974	0.985	-212.184	0.980
ZM1+RANDICH	0.986	0.985	-1.419	0.980
ZM2+ZM3	0.952	0.985	0.812	0.943
ZM2+RANDICH	0.986	0.986	-3.844	0.979
ZM3+RANDICH	0.986	0.985	-25.073	0.980
ZM1+ZM2+ZM3	0.974	0.986	-	0.981
ZM1+ZM2+RANDICH	0.987	0.986	0.238	0.980
ZM1+ZM3+RANDICH	0.987	0.985	-0.068	0.980
ZM2+ZM3+RANDICH	0.987	0.986	-	0.981
ZM1+ZM2+ZM3+RANDICH	0.987	0.987	-2.303	0.981

2.2.2 Другий метод

Були проведені розрахунки та отримані значення до та після оптимізації значень ребер, для окремих класів сполук та для усієї вибірки в цілому.

У Таблицях 2.5 - 2.9 наведені значення для коефіцієнту детермінації та RMSE для проведених розрахунків. Перевірка проводилася на прикладі трьох індексів — ZM1, ZM2, ZM3.

Таблиця 2.5 Значення для спиртів, сполук - 48

R2	ZM1	ZM2	ZM3
Usual	0.9212	0.8848	0.5844
Opt	0.9597	0.9668	0.7528

RMSE	ZM1	ZM2	ZM3
Usual	8.5290	10.3133	19.5888
Opt	6.1022	5.5337	15.1091

Таблиця 2.6 Значення для тіоетерів, сполук - 44

R2	ZM1	ZM2	ZM3
Usual	0.9643	0.9489	0.8068
Opt	0.9741	0.9733	0.8396

RMSE	ZM1	ZM2	ZM3
Usual	7.9316	9.4888	18.4438
Opt	6.7495	6.8524	16.8055

Таблиця 2.7 Значення для амінів, сполук - 33

R2	ZM1	ZM2	ZM3
Usual	0.9586	0.9530	0.8664
Opt	0.9896	0.9865	0.9536

RMSE	ZM1	ZM2	ZM3
Usual	10.4366	11.1217	18.7574
Opt	5.2326	5.9552	11.0588

Таблиця 2.8 Значення для етерів, сполук - 69

R2	ZM1	ZM2	ZM3
Usual	0.9634	0.9460	0.8089
Opt	0.9720	0.9654	0.8352

RMSE	ZM1	ZM2	ZM3
Usual	8.6557	10.5152	19.7805
Opt	7.5731	8.4113	18.3679

Таблиця 2.9 Значення для усієї вибірки, сполук - 269

R2	ZM1	ZM2	ZM3
Usual	0.4441	0.4335	0.3519
Opt	0.6942	0.6713	0.5935

RMSE	ZM1	ZM2	ZM3
Usual	36.5784	36.9281	39.4963
Opt	27.1316	28.1275	31.2795

Як видно з наведених таблиць, для будь-якого з розглянутих випадків спостерігається покращення якості моделі.

2.3. Порівняння покращення якості моделей

Задля перевірки стабільності та правильного порівняння двох запропонованих методів проведена наступна перевірка. Вона полягала у розбитті вибірки на навчальну та тестову, з розміром тестової вибірки 25%. Розбиття на навчальну та тестову вибірки проводилося 100 разів, бралося медіанне значення коефіцієнту детермінації. Наведено порівняння та перевірка з використанням індексу ZM3.

Та у Таблиці 2.10 та 2.11 наведені значення для розрахунку на індексі ZM3 для навчальної та тестової вибірки для.

Таблиці 2.10 Медіанні значення коефіцієнтів детермінації для навчальної та тестової вибірки при 100 розбиттів для першого методу.

	Alcohols	Thioethers	Amines	Ethers	Different classes
R^2	0.806	0.919	0.950	0.942	0.535
Train	0.801	0.919	0.953	0.944	0.532
Test	0.801	0.903	0.927	0.930	0.533

Таблиці 2.11 Медіанні значення коефіцієнтів детермінації для навчальної та тестової вибірки при 100 розбиттів для другого методу.

	Alcohols	Thioethers	Amines	Ethers	Different classes
R^2(opt)	0.753	0.840	0.954	0.835	0.594
Train	0.758	0.850	0.954	0.844	0.595
Test	0.660	0.709	0.854	0.750	0.537

Як бачимо з таблиць — отримані моделі достатньо добре описують відповідні вибірки.

Порівнюючи квантово-хімічний, перший метод, та реберний, другий, за допомогою даних наведений у Таблиці 2.10 та 2.11 можна помітити, що перший метод має невелику перевагу для кожної з вибірок, що представленні окремими класами сполук. Натомість для вибірки у котрій об’єднанні усі наведені класи сполук вже бачимо перевагу другого методу.

Та через не суттєву перевагу, другий запропонований метод, може знайти використання через легкість його застосування, порівняно з першим методом.

Для порівняння першого методу з його модифікацією, розрахунком з урахуванням атомів водню, наведенні медіанні значення, розраховані аналогічно за немодифікованим методом, у Таблиці 2.12.

Таблиця 2.12 Медіанні значення коефіцієнтів детермінації для навчальної та тестової вибірки при 100 розбиттів для першого методу з модифікацією.

	Alcohols	Thioethers	Amines	Ethers	Different classes
R^2	0.827	0.939	0.956	0.949	0.491
Train	0.823	0.939	0.958	0.950	0.485
Test	0.820	0.927	0.933	0.936	0.497

Можна зробити висновок, що для окремих класів сполук такий підхід може застосовуватися та давати кращі результати. Однак, у випадку з останньою вибіркою,

маємо погіршення значень. Вони пояснюються появою певної кількості ізомерних сполук одного складу, що призводить до появи колізій при розрахунках дескрипторів ZM3 у цьому випадку, а це своєю чергою зменшує якість початкових даних. Бо складається випадок, коли для декількох сполук отримуємо занадто близькі, але не однакові, значення дескрипторів.

Особливо це помітно для звичайного підходу з ураховуванням атомів водню, Таблиця 2.4 стовпчик Topol_wH, там спостерігаємо саме рівність дескрипторів для ізомерних сполук.

ВИСНОВКИ

1. Розроблені методи дають можливість отримувати більш якісні моделі опису властивостей, ніж класичні підходи. Сутність першого полягає у використанні порядку зв’язку як елементи топологічної матриці суміжності при розрахунках топологічних індексів, а другого — у мінімізації доцільної функції за рахунок варіювання параметрів матриці суміжності, що відповідають вершинам реберних графів n порядків.
2. На прикладі реальних систем отримали суттєве покращення моделі, для деяких моделей з коефіцієнтом детермінації менше 0.5 отримали моделі з коефіцієнтами вище 0.95.
3. Отримані рішення є стабільними, що доводить зроблена перевірка. Та підтверджує прогностичну можливість моделей.
4. На відміну від класичного QSAR, продемонстрована можливість використання атомів гідрогену для розрахунку топологічних дескрипторів, результати виявили ефективність, відповідні коефіцієнти детермінації підтверджують непогану прогностичну здатність.
5. Створили програмне забезпечення для розрахунку запропонованих методів урахування гетероатомів.

СПИСОК ЛІТЕРАТУРИ

1 Raza Z., Imran M. The reverse Zagreb indices of F-sum of graphs, Journal of Discrete Mathematical Sciences and Cryptography **2022**, *25*(6), 1885-1897.

2 Furtula, B., Gutman, I. A forgotten topological index. J Math Chem. **2015**, *53*, 1184–1190.

3 Zakharov, A. B., Ivanov, V. V. A new approach in topological descriptors usage. Iterated line graphs in the theoretical prediction of physico-chemical properties of saturated hydrocarbons. Kharkiv University Bulletin. Chemical Series **2019**, (32), 38-45.

4 Ali A., Trinajstić N. A Novel/Old Modification of the First Zagreb Index. Molecular Informatics **2018**, *37*(6) 1-8.

5 Noureen S., Ahmad A., Akbar B. Extremum Modified First Zagreb Connection Index of n-Vertex Trees with Fixed Number of Pendent Vertices. Discrete Dynamics in Nature and Society **2020**, 6 .

6 Raza Z., Ali A. Bounds on the Zagreb indices for molecular (n,m)-graphs. Int J Quantum Chem. 2020, 120.

7 Hussain M, Rehman Au, Shekhovtsov A, Asif M, Sałabun W. Study of Transformed ηζ Networks via Zagreb Connection Indices. Information **2022**, *13*(4), 179.

8 Yousaf A., Nadeem M. An Efficient Technique to Construct Certain Counting Polynomials and Related Topological Indices for 2D-Planar Graphs, Polycyclic Aromatic Compounds **2022**, *42*(7), 4328-4342

9 Dehmer M., Warmuza K., Bonchev D. Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Wiley-VCH, Weinheim **2012**, 436.

10 Raza Z. Zagreb Connection Indices for Some Benzenoid Systems, Polycyclic Aromatic Compounds **2022** , *42*(4), 1814-1827,

11 Muhammad J., Ahmed A., Aqsa S., Topological Aspects of Dendrimers via Connection-Based Descriptors, Computer Modeling in Engineering & Sciences **2023**, *135*(2), 1649-1667.

12 Wang Z., Mao Y., Li Y. On relations between Sombor and other degree-based indices. J. Appl. Math. Comput. **2022**, *68*, 1–17.

13 Mehdi E., Ghalavand A. Trees with the minimal second Zagreb index. Kragujevac Journal of Mathematics **2018**, *42*, 325-333.

14 Hosamani S., Perigidad D., Jamagoud S., Maled Y., Gavade S. QSPR Analysis of Certain Degree Based Topological Indices. Journal of Statistics Applications & Probability **2017**, *6(*2), 361-371.

15 Neelu S., Shaik B., Neeraj A., Anita K., Agrawal V.K., Satya G.,. QSAR and Molecular Modeling Studies on a Series of Indole-based Pyridone Analogues as HCV NS5B Polymerase Inhibitors. Letters in Drug Design & Discovery **2016**, *13*, 757-770.

16 Cynthia S., Rajeshwar P. V. History of Quantitative Structure-Activity Relationships. Burger's medicinal Chemistry and Drug Discovery: 2010; pp 1-96.

17 Roy K., Kar S., Das R. N. A Primer on QSAR/QSPR Modeling Fundamental Concepts. Springer: 2015; pp 1-47.

18 O. Ivanciuc, A.T. Balaban. Graph Theory in Chemistry. In: The Encyclopedia of Computational Chemistry, Eds.: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner. John Wiley & Sons, Chichester 1998; pp 1169-1190

19 Topological Indices and Related Descriptors in QSAR and QSPR (J. Devillers and AT Balaban, Eds.) Gordon and Breach Science Publishers. The Netherlands 1999; pp 21-57.

20 Roy K., Kar S., Das R. A primer on QSAR/QSPR modeling: fundamental concepts. Springer International Publishing: 2015; pp 67-78.

21 Roberto Todeschini, Viviana Consonni Molecular Descriptors for Chemoinformatics. Wiley-VCH Verlag GmbH & Co. KgaA: 2009; pp 273-1061.

22 https://www.rdkit.org/ (дата звернення 13 лютого 2023)

23 https://www.msg.chem.iastate.edu/gamess/ (дата звернення 13 лютого 2023)

24 Faulon J.L., Bender, A. Handbook of Chemoinformatics Algorithms. Chapman and Hall CRC: 2010, pp 211-232.