УДК 547.587:547.898:661.865+544.25+535.37

А.Г. Артеменко, Л.М. Огніченко, В.Є. Кузьмін, В.І. Недоступ

ПРОГНОЗУВАННЯ ТРАНСПОРТНИХ ВЛАСТИВОСТЕЙ ОРГАНІЧНИХ СПОЛУК В ГАЗОВІЙ ФАЗІ НА ОСНОВІ QSPR АНАЛІЗУ

Фізико-хімічний інститут ім. О.В. Богатського НАН України, м. Одеса, Україна

В роботі показано, що такі транспортні властивості, як коефіцієнти в'язкості і теплопровідності різноманітних органічних речовин у газовій фазі, можна оцінити з прийнятною точністю у межах 1D-QSPR моделей, які дозволяють здійснювати прогноз досліджуваної властивості на основі тільки хімічного складу молекули. Об'єктом дослідження були транспортні властивості (коефіцієнти в'язкості та теплопровідності) органічних сполук із досить представницьких баз даних, що налічують близько 5 тисяч вуглець-, галоген-, кисень-, азот- і сірковмісних сполук. При використанні симплексного підходу моделювання молекулярної структури і таких методів машинного навчання, як метод множинної лінійної регресії (MLR) і метод випадкового лісу (RF), для сформованих баз даних побудовано адекватні 1D QSPR моделі щодо транспортних властивостей індивідуальних речовин у газовій фазі. Виконано аналіз впливу деяких структурних і фізико-хімічних чинників на досліджувані транспортні властивості органічних сполук. На основі побудованих 1D RF QSPR моделей створено комп'ютерну експертну систему для прогнозу коефіцієнтів в'язкості і теплопровідності нових речовин.

Ключові слова: 1D, QSPR, симплексне надання молекулярної структури, транспортні властивості, коефіцієнт в'язкості, коефіцієнт теплопровідності.

DOI: 10.32434/0321-4095-2024-157-6-17-24

Вступ

В'язкість і теплопровідність – це транспортні властивості речовин у газовій фазі, які є фундаментальними характеристиками для виконання розрахунків при впровадженні та проектуванні робочих середовищ як енергохімічних технологічних процесів. Знання цих властивостей є значимим для створення обладнання хіміко-технологічних процесів. Як відомо, ці властивості значною мірою визначаються середньою довжиною вільного пробігу молекул, середньою тепловою швидкістю молекул, щільністю газу. Оцінювання впливу будови речовини на її транспортні властивості у газовій фазі є вельми складним завданням, вирішення якого потребує застосування специфічних підходів, зокрема, методів хемоінформатики.

Експериментальні методи, що використовуються для визначення транспортних властивостей, мають велику трудомісткість (вимоги до виконання аналізу, чистоти реагентів тощо) та собівартість, а також незастосовні для великих молекул. Більшість з відомих методів прогнозування транспортних властивостей можуть бути застосовані до сполук одного класу або потребують додаткових експериментальних даних. Все більшої популярності набирають QSPR (Quantitative Structure–Property Relationships – кількісні співвідношення «структура–властивість») методи прогнозування термодинамічних властивостей. Так, можна виділити два типи моделей: локальні [1] (здатні прогнозувати досліджувані властивості для сполук одного класу; вони мають обмежену сферу застосування) і глобальні (використовують хімічно різнорідні вибірки органічних сполук). До глобальних можна віднести, наприклад: метод Godavarthy [2], метод NIST [3]. У всіх цих мето-

дах моделі будуються на основі топологічних, геометричних, електростатичних і квантово-хімічних дескрипторів для тривимірних структур. Варто відзначити трудомісткість таких методологій побудови моделей: в рамках 3D-моделювання проводиться конформаційний пошук з використанням напівемпіричних квантово-хімічних методів. Окрім того, в деяких випадках, дескриптори не мають прозору інтерпретацію.

Предметом даного дослідження є розробка простих QSPR моделей для прогнозування коефіцієнтів теплопровідності та динамічної в'язкості органічних сполук у газовій фазі в області помірних тисків. Методи 1D, 2D QSPR, що використовують дескриптори молекулярної структури, з нашої точки зору, дозволяють вирішити ту задачу, яка не може бути вирішена методами феноменологічної термодинаміки, визначити структурні фактори, які впливають на коефіцієнти теплопровідності і в'язкості багатоатомних молекул.

Матеріали і методи

Об'єктом дослідження були транспортні властивості (коефіцієнти в'язкості η та теплопровідності λ) з великого масиву накопичених експериментальних даних для різноманітних органічних сполук. Експериментальні дані були отримані з довідника [4], де присутні дані для 4937 сполук щодо коефіцієнтів в'язкості (η, мкПа·с) та дані для 4938 сполук щодо теплопровідності (λ, Вт/(м·К)) при температурі 25^0С і нормальному атмосферному тиску для органічних сполук в газовій фазі. У вибірках надані: вуглець-, галоген-, кисень-, азот- і сірковмісні органічні сполуки.

При використанні комплексу програм ChemAxon[1] були змодельовані і стандартизовані всі молекулярні структури вибірок. Здійснено пошук помилкових структур, неправильних імен сполук, перевірка структур на зв'язність і унікальність, відсів дублікатів. Після верифікації сформованих вибірок в них залишилось 4690 сполук з даними щодо в'язкості η (вибірка А) та 4668 сполук з даними щодо теплопровідності λ (вибірка Б).

У даній роботі для вирішення завдань використовувався симплексний підхід надання молекулярної структури (СПМС) [5,6], який базується на тому, що кожна молекула надається у вигляді системи різних симплексів – чотириатомних молекулярних фрагментів фіксованого складу і будови. Даний метод добре себе зарекомендував при вирішенні ряду різноманітних QSAR/QSPR завдань [5]. Симплексний підхід дає можливість проводити в межах однієї ви- бірки аналіз молекул, які значно відрізняються за структурою, і, крім того, дає можливість проводити прозору структурну і фізико-хімічну інтерпретації [6].

Молекулярна структура може бути описана на різних рівнях моделювання, умовно, від 1D до 4D [5]. В даній роботі використовувалися найбільш прості 1D і 2D-підходи. На 1D рівні молекула надається брутто-формулою і симплекс визначається кількістю фіксованих четвірок атомів, а на 2D рівні молекула надається структурною формулою, при цьому враховуються зв'язність атомів в симплексі, тип атомів і природа зв'язку (простий, подвійний, потрійний, ароматичний).

У рамках будь-якого рівня надання молекулярної структури для n-атомної молекули загальна кількість всіх можливих симплексів (N) дорівнює $N = \frac{n!}{(n-1)!\cdot 4!}$.

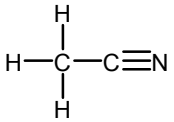
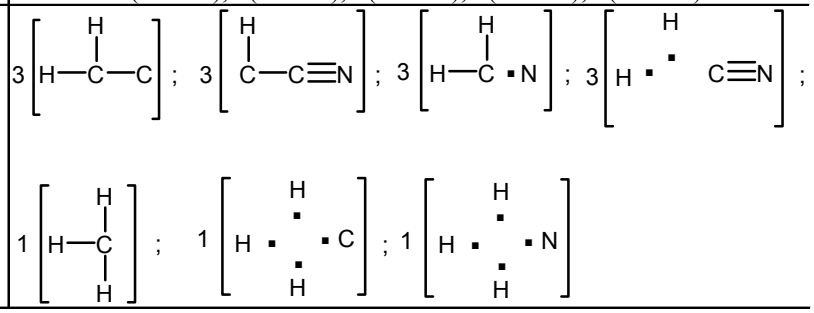
На рисунку наведено приклад генерування 1D і 2D дескрипторів для молекули ацетонітрилу, коли використовується диференціація атомів за їх природою.

Слід зазначити, що в даній роботі деталізація атомів (вершин симплексу) проводиться не лише за їх природою, а використовуються варіанти диференціації атомів на основі різних атомних характеристик, що є принциповою особливістю СПМС.

Використовуючи програмний комплекс для рішення задач структура–властивість «HIT QSAR» [7], для кожної молекули були розраховані 1D симплексні дескриптори (двійки, трійки, четвірки), атоми в яких були диференційовані не тільки за їхніми мітками (символами атомів (Elm)), а й на основі таких атомних характеристик, як електронегативність (En), електронна поляризуємість (Rf), характеристики ван-дер-ваальсових взаємодій (ВДВ) універсального силового поля [8], зокрема параметри притягнення (Attr) та відштовхування (Rep).

Для атомних характеристик, які мають реальні значення, діапазон усіх можливих значень попередньо розбивається на певну кількість дискретних груп. Мітка атома визначається з точки зору його належності до певної групи. При роз-

[1]ChemAxon JChem 6.1.3 Standardizer: http://www.chemaxon.com. – 2013.

Рівень	Структура	Загальна кількість дескрипторів	Кількість дескрипторів (певного віду)
1D	C_2H_3N	15	3(CCHH), 3(CCHN), 2(CHHH), 6(CHHN), 1(HHHN)
2D		15	

Приклад генерації симплексних дескрипторів для молекули ацетонітрилу на 1D–2D рівнях деталізації молекулярної структури

рахунку симплексних дескрипторів були використані наступні діапазони поділу значень атомних властивостей на інтервали:
електронегативність (En): A<2,19≤B<2,5≤C<3≤D,
електронна поляризуємість (Rf):
A<1,5≤B<3≤C<8≤D,
ВДВ притягнення (Attr):
A<50≤B<100≤C<250≤D<400≤E<650≤F<2000≤G,
ВДВ відштовхування (Rep) :
A<20000≤B<32000≤C<50000≤D<100000.

При розрахунку 2D симплексних дескрипторів додатково використовували атомний заряд (Chg):

A<–0,16≤B<–0,07≤C<0,02≤D<0,11≤E<0,21≤F<0,30≤G.

ліпофільність (Lip):

A<–1,51≤B<–0,96≤C<–0,42≤D<0,13≤E<0,68≤F<1,23≤G.

Крім того, усі атоми, що відповідають вершинам симплексів, також були поділені мітками на три групи: D – донори потенційного Н-зв'язку, A – акцептори потенційного Н-зв'язку та I – індиферентні. Як дескриптори також були використані інтегральні параметри цілої молекули, такі як молекулярна маса (I.AW), молекулярна рефракція (I.Rf), сумарна електронегативність (I.En), ліпофільність (I.Lip).

В даній роботі застосовані популярні методи машинного навчання, які добре себе зарекомендували для вирішення задач QSPR:

– метод випадкового лісу (Random Forest – RF) [9];

– метод множинної лінійної регресії (Multiple Linear Regression – MLR) [7,10].

Для оцінки надійності QSAR моделей велике значення має їх валідація з використанням зовнішньої тестової вибірки, що складається із сполук, які не увійшли в навчальну вибірку. QSAR моделі, що використовуються для прогнозу властивості ще недосліджених молекул, повинні мати достатньо адекватні статистичні характеристики [11]. Для оцінки прогностичної здатності використовуються різні методи формування «тестової вибірки». В ході таких процедур частина молекул виключається з процесу побудови моделі, а молекули, що залишаються, формують нову навчальну вибірку. В подальшому ця модель використовується для прогнозу властивості сполук тестової вибірки.

Модель випадкового лісу являє собою ансамбль окремих дерев рішень, при цьому для побудови кожного з дерев з усього набору сполук формується нова навчальна вибірка для даного конкретного дерева. Сполуки, що не ввійшли в навчальну вибірку потрапляють в так звану «out-of-bag» вибірку і використовуються для оцінювання прогнозуючої здатності моделі. В методі випадкового лісу обсяг контрольної вибірки (out of bag-oob) становить приблизно 33% від кількості об'єктів в навчальній вибірці.

Для оцінювання прогностичної здатності моделей, побудованих методом множинної лінійної регресії, було використано процедуру п'ятикратної зовнішньої крос-валідації (five-folds) [6], коли молекули вибірки упорядковуються згідно зі значеннями їх властивостей, і потім кожна п'ята молекула в даному ряду відбирається в тестову вибірку, тобто формуються п'ять наборів, кожний з яких містить навчальну і тестову вибірки. Серед одержаних п'яти наборів кожна молекула потрапляє в тестову вибірку тільки один раз. Тестова вибірка (20% від загальної кількості молекул) виключається з процесу побудови моделі. На даних навчальної вибірки будується модель, яка використовується для прогнозу властивостей

сполук тестової вибірки. Таким чином, для кожного з п'яти наборів будується модель, яка далі використовується в основі консенсусної моделі.

Для регресійних моделей найбільш відомим показником, що характеризує апроксимаційну здатність, є коефіцієнт детермінації (R^2), що є відношенням «пояснювальної частини» дисперсії властивості до повної дисперсії цієї властивості на вибірці:

$$R^2 = \frac{SS - RSS}{SS}, \qquad RSS = \sum_{i=1}^{N} (y_i^{розр} - y_i^{спостер})^2,$$

$$SS = \sum_{i=1}^{N} (y_i^{спостер} - \overline{y}^{спостер})^2,$$

$$RSS = \sum_{i=1}^{N} (y_i^{розр} - y^{спостер})^2, \quad \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i^{спостер},$$

де $y_i^{розр}$ і $y_i^{спостер}$ – значення властивості, що розраховане за допомогою побудованої регресійної моделі, і спостережуване значення властивості для і-того об'єкту навчальної вибірки відповідно, N – кількість об'єктів навчальної вибірки, $\overline{y}$ – середнє значення властивості для вибірки з N сполук.

Також важливим статистичним критерієм є середньоквадратична помилка (RMSE), яка розраховується за формулою

$$RMSE = \sqrt{\frac{RSS}{N}}.$$

Зрозуміло, що RMSE характеризує описову здатність регресійної моделі по відношенню до сполук, що входять в навчальну вибірку.

Для «out-of-bag» вибірок в RF моделях і тестових вибірок в MLR моделях формули для розрахунку коефіцієнта детермінації R^2_{oob} і R^2_{ts}, а також середньоквадратичної помилки $RMSE_{oob}$ і $RMSE_{ts}$, ідентичні наведеним вище, тільки замість $y_i^{розр}$ використовують прогнозоване значення $y_i^{прогн}$ для і-того об'єкту контрольної вибірки.

Результати та їх обговорення

Аналіз вибірки А щодо коефіцієнта в'язкості η показав, що серед 4690 сполук вибірки лише 766 сполук (16,2%) не мають структурних ізомерів в даній вибірці. 3604 сполук (77% від загальної кількості сполук) – це сполуки, які мають в даному наборі структурні ізомери, при цьому, серед таких сполук 459 мають однакові значення властивості, а для 3145 середня відносна відмінність від максимального значення складає 5%. У вибірці також присутні 160 пар геометричних ізомерів, що складає 6,8% від загальної кількості сполук вибірки. Для 104 пар геометричних ізомерів дані щодо властивості однакові, а для 56 пар середня відносна відмінність від максимального значення складає 2%.

Аналогічні результати спостерігаються для вибірки Б щодо коефіцієнта теплопровідності λ. В даному випадку більша частина вибірки – 3592 сполуки (77% від загальної кількості сполук) мають в досліджуваному наборі структурні ізомери, серед таких сполук 481 мають однакові значення властивості, а для 3111 середня відносна відмінність від максимального значення складає менш ніж 3%. Серед 4668 сполук даної вибірки лише 760 сполуки (16,0%) не мають структурних ізомерів в даній вибірці. Також присутні 158 пар геометричних ізомерів (6,8% від усієї вибірки). Для 109 пар геометричних ізомерів дані щодо показників теплопровідності однакові, а для 49 пар середня відносна відмінність від максимального значення властивості для пари геометричних ізомерів складає 3,8%.

На основі результатів аналізу вибірок А і Б можна зробити припущення про можливість адекватного опису транспортних властивостей вже на 1D рівні, що і буде продемонстровано далі.

У результаті аналізу вибірок А і Б також виявлено, що 4450 сполук в них збігаються, причому значення властивостей η і λ для спільних сполук даних вибірок між собою майже не корелюють, коефіцієнт кореляції (R) дорівнює 0,385. Слід зазначити, що при побудові QSPR моделей для показників теплопровідності органічних сполук з метою рівномірного розподілу значення властивості були конвертовані у $-\log \lambda$.

При використанні симплексного підходу для молекул вибірки А було розраховано 834 1D та 8278 2D структурних дескрипторів, а для вибірки Б – 840 1D та 7879 2D структурних дескрипторів, відповідно.

У результаті QSPR аналізу транспортних властивостей для сполук вибірок А і Б з використанням статистичного методу випадкового лісу (Random Forest – RF) [9] було побудовано ряд адекватних моделей (I–VI) (табл. 1). Як видно з табл. 1, вже на 1D рівні деталізації молекулярної структури вдається добитися досить високого рівня адекватності моделей та їх прогнозуючої здатності. Перехід до більш високого рівня деталізації (2D) не приводить до значного покращення якості моделей, тобто у межах 1D-QSPR моделей можливо з прийнятною точністю оцінювати коефіцієнти

Таблиця 1

Статистичні показники QSPR моделей, отриманих методом випадкового лісу

Властивість	Модель		R^2	$R^2_{(oob)}$	$RMSE_{(oob)}$
В'язкість	I	1D	0,986	0,958	3,9
	II	2D	0,995	0,962	3,7
	III	1D+2D	0,995	0,964	3,6
Теплопровідність	IV	1D	0,975	0,930	0,05
	V	2D	0,991	0,931	0,05
	VI	1D+2D	0,991	0,932	0,05

в'язкості і теплопровідності різноманітних органічних речовин у газовій фазі, при цьому достатньо знати тільки хімічний склад відповідної молекули.

Для побудованих 1D-QSPR моделей було виконано аналіз впливовості структурних дескрипторів до опису досліджуваних властивостей. У табл. 2 надано структурні параметри, які мають найбільший вплив на коефіцієнти η та λ

Як і слід було очікувати, для обох властивостей впливовими є параметри, що обумовлюють міжмолекулярні взаємодії, зокрема ван-дер-ваальсові (Attr, Rep), електростатичні (EN) та дисперсійні (Rf) (табл. 2). Для теплопровідності ключову роль грає молекулярна маса (I.AW), а ван-дер-ваальсове притягнення майже не впливає.

Для здійснення більш детальної інтерпретації було використано статистичний метод множинної лінійної регресії (MLR) [7]. У результаті QSPR аналізу на 1D рівні для навчальних вибірок А і Б було отримано цілком адекватні моделі (табл. 3). У табл. 4 наведені коефіцієнти внесків для нормованих дескрипторів для розроблених моделей VII і VIII. Для оцінювання прогностичної здатності моделей було проведено процедуру п'ятикратної зовнішньої крос-валідації (five-folds). З табл. 3 можна побачити, що моделі VII–VIII є високоякісні з цілком прийнятною прогнозуючою здатністю.

Аналіз даних табл. 4 демонструє, що якісно картина впливу структурних параметрів на транспортні властивості приблизно така ж, як і для моделей RF (див. вище), але в випадку моделей VII–VIII можна оцінити не тільки ступінь впливу, але і його напрямок. Так для моделі VII щодо в'язкості характеристики ван-дер-ваальсових взаємодій позитивно впливають на відповідний коефіцієнт, а параметри електронегативності та електронної поляризованості – негативно. Щодо теплопровідності, то вплив більшості характеристик міжмолекулярної взаємодії позитивний, але молекулярна маса має суттєвий негативний внесок. Це відповідає фізичним механізмам процесів, в яких реалізуються вказані транспортні властивості.

На основі моделей I та IV створено комп'ютерну експертну систему «TransProp Expert» [13] для прогнозування термодинамічних транспортних властивостей (коефіцієнтів в'язкості та теплопровідності) органічних сполук у газовій фазі.

Висновки

Таким чином, у межах 1D-QSPR моделей виявлено можливість оцінювання з прийнятною точністю коефіцієнтів в'язкості і теплопровідності різноманітних органічних речовин (навчальна вибірка близько 5000 сполук) у газовій фазі. Фактично, для такого оцінювання достатньо знати тільки хімічний склад відповідної молекули. Інтерпретація побудованих QSPR моделей цілком відповідає щодо фізичних уявлень для в'язкості і теплопровідності органічних речовин у газовій фазі. На основі побудованих 1D RF QSPR моделей створено комп'ютерну експертну систему для прогнозування транспортних властивостей нових речовин.

СПИСОК ЛІТЕРАТУРИ

1. *Support* vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes / Yang S., Lu W., Chen N., Hu Q. // J. Mol. Struct. Theochem. – 2005. – Vol.719. – No.1-3. – P.119-127.

2. *Godavarthy S.S., Robinson R.L., Gasem K.A.* Improved structure–property relationship models for prediction of critical properties // Fluid Phase Equilib. – 2008. – Vol.264. – No. 1-2. – P.122-136.

3. *Predictive* correlations based on large experimental datasets: critical constants for pure compounds / Kazakov A., Muzny C.D., Diky V., Chirico R.D., Frenkel M. // Fluid Phase Equilib. – 2010. – Vol.298. – P.131-142.

4. *Yaws C.L.* Yaws' handbook of thermodynamic and physical properties of chemical compounds (electronic edition): physical, thermodynamic and transport properties for 5,000 organic chemical compounds. – Knovel: Norvich, 2003. – 2078 p.

Таблиця 2

Структурні параметри з найбільшими відносними внесками до 1D QSPR моделей

В'язкість (модель I)		Теплопровідність (модель IV)	
Дескриптор*	Внесок, %	Дескриптор	Внесок**, %
Cnk(Rep)/A,B,B	3,6	I.AW	21,2
Cnk(Attr)/B,E	2,9	I.Rf	4,0
Cnk(Rf)/B,B	2,5	Cnk(Rf)/B,B,B	2,6
Cnk(Rf)/A,B,B,B	2,2	Cnk(En)/C,C,C	2,4
Cnk(Elm)/C,C,H	2,2	Cnk(Rf)/B	2,4
Cnk(Rf)/B	2,2	Cnk(Rf)/B,B,B,B	2,3
Cnk(Attr)/B,E,E	1,9	Cnk(En)/C,C	2,2
Cnk(Attr)/B,E,E,E	1,8	Cnk(Rf)/B,B	2,1
Cnk(Rep)/B,B	1,8	Cnk(Rep)/B,E	2,1
Cnk(Rep)/A,B,B,B	1,7	Cnk(Rep)/B,B	2,0
Cnk(Rf)/A,B,B	1,7	Cnk(Rep)/E	1,8
Cnk(Elm)/C,C,C,H	1,6	Cnk(En)/C	1,7
Cnk(Rf)/B,B,B	1,5	Cnk(En)/B,C,C,C	1,3
I.EN	1,5	Cnk(Rf)/D	1,3
Cnk(Rep)/A,B	1,4	Cnk(Rep)/B,B,B	1,1
Cnk(Rep)/B,B,B	1,4		
Cnk(Elm)/C,H,H	1,4		
Cnk(Rf)/A,B	1,3		
Cnk(Attr)/E,E	1,3		
Cnk(Attr)/B,B,E,E	1,3		
Cnk(Rf)/B,B,B,B	1,3		
Cnk(Elm)/C,H	1,2		
Cnk(Elm)/C,C,H,H	1,2		
Cnk(Rep)/B	1,2		
Cnk(Rep)/A,A,B,B	1,2		
Cnk(Attr)/E,E,E	1,1		

Примітки: * – Cnk означає 1D дескриптор, в дужках відмічена властивість атомів, за якою атоми було диференційовано, після «/» представлено мітки атомів згідно з певним інтервалом значень властивості; ** – Внесок оцінюється із моделі випадкового лісу за оригінальним методом [12]. Вони не мають знаку, тобто не відображають напрямок впливу дескриптора, внаслідок нелінійності моделі.

5. *Simplex* representation of molecular structure as universal QSAR/QSPR tool / Kuz'min V., Artemenko A., Ognichenko L., Hromov A., Kosinskaya A., Stelmakh S., et al. // Struct. Chem. – 2021. – Vol.32. – No. 4. – P.1365-1392.

6. *Virtual* screening and molecular design based on hierarchical QSAR technology / Kuz'min V.E., Artemenko A.G., Muratov E.N., Polischuk P.G., Ognichenko L.N., Liahovsky A.V., et al. // Recent advances in QSAR studies. – London: Springer, 2010. – P.127-176.

7. *Свідоцтво* про реєстрацію авторського права на твір № 66633. Комп'ютерна програма «Программный комплекс для решения задач структура–свойство «HIT QSAR» (Программный комплекс «HIT QSAR») / В.Є. Кузьмін, А.Г. Артеменко (Україна). – Заявл. 13.07.2016; Опубл. 28.10.2016, Бюл. № 42.

8. *UFF*, a full periodic table force field for molecular mechanics and molecular dynamics simulations / Rappe A.K., Casewit C.J., Colwell K.S., Goddard III W.A., Skiff W.M. // J. Am. Chem. Soc. – 1992. – Vol.114. – No. 25. – P.10024-10035.

9. *Breiman L.* Random forest // Mach. Learn. – 2001. – Vol.45. – P.5-32.

10. *Forster E., Ronz B.* Methoden der Korrelations- und Regressionsanalyse. – Berlin: Verlag Die Wirtschaft, 1979. – 324 p.

11. *Report* on the regulatory uses and applications in OECD member countries of (quantitative) structure-activity relationship [(Q)SAR] models in the assessment of new and existing chemicals // OECD Pap. – 2006. – Vol.6. – P.79-157.

12. *Interpretation* of QSAR models based on random forest method / Kuz'min V.E., Polishchuk P.G., Artemenko A.G., Andronati S.A. // Mol. Inf. – 2011. – Vol.30. – No. 6-7. – P.593-603.

13. *Свідоцтво* про реєстрацію авторського права на твір № 111479. Комп'ютерна програма «Експертна система «TransProp Expert» для прогнозування термодинамічних транспортних властивостей (коефіцієнтів в'язкості та теплопровідності) органічних сполук» (Експертна система «TransProp Expert») / В.Є. Кузьмін, А.Г. Артеменко, В.І. Недоступ (Україна). – Заявл. 31.01.2022; Опубл. 31.03.2022, Бюл. № 69.

Надійшла до редакції 09.07.2024

Таблиця 3

Статистичні показники 1D QSPR моделей, побудованих методом MLR*

Властивість	Кількість дескрипторів	R^2	Q^2	$R^2_{ts(five\text{-}fold)}$	RMSE	F	t(cr)
(VII) В'язкість	15	0,932	0,930	0,922	4,96	4257>1,66	2,62>2,58
(VIII) Теплопровідність	13	0,902	0,900	0,896	0,06	3307>1,79	2,70>2,58

Примітка: * – Q^2 – коефіцієнт детермінації в умовах ковзного контролю; F – значення критерію Фішера; t(cr) – значення критерію Стьюдента.

Таблиця 4

Нормовані коефіцієнти регресії для дескрипторів 1D MLR QSPR

В'язкість (модель VII)		Теплопровідність (модель VIII)	
Дескриптор	Beta	Дескриптор	Beta*
Cnk(Rep)/B,B	2,05	I.EN	–1,31
Cnk(Attr)/G	0,49	Cnk(Rf)/D	–0,39
Cnk(Rep)/D	0,41	Cnk(Rep)/B,B,B,C	–0,27
Cnk(Rf)/A,A,D	0,20	Cnk(Elm)/I,I	–0,17
Cnk(Attr)/B,B,C	0,12	Cnk(Elm)/F	–0,15
Cnk(Rep)/B,C	0,12	Cnk(Elm)/H,H,O	–0,14
Cnk(En)/A,C,C,C	0,05	Cnk(Rep)/B,E,E	–0,11
Cnk(En)/A	–0,14	Cnk(Elm)/C,I	–0,08
Cnk(Elm)/I,I	–0,16	Cnk(Elm)/F,H	0,05
Cnk(Elm)/O	–0,17	Cnk(Elm)/Al,C,H	0,07
Cnk(Elm)/F,H	–0,19	Cnk(Elm)/B,C,H	0,08
Cnk(Rf)/A,D	–0,37	Cnk(En)/C,C	0,25
I.EN	–0,44	I.AW	2,17
Cnk(En)/B,C,C,C	–0,61		
Cnk(Rf)/B	–1,78		

Примітка: * – Оскільки при побудові моделей значення показників теплопровідності були конвертовані у –log λ, то в даному випадку, навпаки, чим більш негативне значення мають коефіцієнти, тим більш позитивний внесок вносять такі структурні параметри у властивість.

PREDICTION OF TRANSPORT PROPERTIES OF ORGANIC COMPOUNDS IN THE GAS PHASE BASED ON QSPR ANALYSIS

A.G. Artemenko, L.M. Ognichenko, V.E. Kuz'min *, V.I. Nedostup

O.V. Bogatsky Physico-Chemical Institute NAS of Ukraine, Odesa, Ukraine

* **e-mail: theorchem@gmail.com**

This work demonstrates that the transport properties of various organic substances in the gas phase, such as viscosity and thermal conductivity coefficients, can be estimated with acceptable accuracy by using 1D-QSPR models, which allow for the prediction of the studied property based solely on the chemical composition of the molecule. Here, we studied the transport properties (viscosity coefficient and thermal conductivity coefficient) of organic compounds for sufficiently representative database including approximately 5,000 carbon-, halogen-, oxygen-, nitrogen-, and sulfur-containing compounds. Using a simplex approach for modeling molecular structure and machine learning methods, such as multiple linear regression (MLR) and random forest (RF), adequate 1D QSPR models for the transport properties of individual substances in the gas phase were developed for the formed databases. Analysis of the influence of certain structural and physicochemical factors on the studied transport properties of organic compounds was carried out. Based on the developed 1D RF QSPR models, a computer expert system for predicting the viscosity coefficients and thermal conductivity of new substances was created.

Keywords: 1D; QSPR; simplex representation of molecular structure (SiRMS); transport properties; viscosity coefficient; thermal conductivity coefficient.

REFERENCES

1. Yang S, Lu W, Chen N, Hu Q. Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes. *J Mol Struct Theochem*. 2005; 719: 119-127. doi: 10.1016/j.theochem.2004.10.060.

2. Godavarthy SS, Robinson RL, Gasem KAM. Improved structure–property relationship models for prediction of critical properties. *Fluid Phase Equilib*. 2008; 264: 122-136. doi: 10.1016/j.fluid.2007.11.003.

3. Kazakov A, Muzny CD, Diky V, Chirico RD, Frenkel M. Predictive correlations based on large experimental datasets: critical constants for pure compounds. *Fluid Phase Equilib*. 2010; 298: 131-142. doi: 10.1016/j.fluid.2010.07.014.

4. Yaws CL. *Yaws' handbook of thermodynamic and physical properties of chemical compounds (electronic edition): physical, thermodynamic and transport properties for 5,000 organic chemical compounds*. Knovel: Norvich; 2003. 2078 p.

5. Kuz'min V, Artemenko A, Ognichenko L, Hromov A, Kosinskaya A, Stelmakh S, et al. Simplex representation of molecular structure as universal QSAR/QSPR tool. *Struct Chem*. 2021; 32(4): 1365-1392. doi: 10.1007/s11224-021-01793-z.

6. Kuz'min VE, Artemenko AG, Muratov EN, Polischuk PG, Ognichenko LN, Liahovsky AV, et al. Virtual screening and molecular design based on hierarchical QSAR technology. In: Puzyn T, Leszczynski J, Cronin MTD, editors. *Recent advances in QSAR studies*. London: Springer; 2010. p. 127-176. doi: 10.1007/978-1-4020-9783-6_5.

7. Kuz'min VE, Artemenko AG, inventors; O.V. Bogatsky Physico-Chemical Institute NAS of Ukraine, assignee. Komp'yuterna programa «Programmnyi kompleks dlya resheniya zadach structura–svoistvo «HIT QSAR» [Computer program «Software complex for solving structure-property tasks «HIT QSAR»]. Ukraine copyright certificate UA 66633. 2016 Oct 28. (*in Ukrainian*).

8. Rappe AK, Casewit CJ, Colwell KS, Goddard III WA, Skiff WM. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J Am Chem Soc*. 1992; 114(25): 10024-10035. doi: 10.1021/ja00051a040.

9. Breiman L. Random forest. *Mach Learn*. 2001; 45: 5-32. doi: 10.1023/A:1010933404324.

10. Forster E, Ronz B. *Methoden der Korrelations- und Regressionsanalyse* [Methods of correlation and regression analysis]. Berlin, Verlag Die Wirtschaft; 1979. 324 p. (*in German*).

11. Report on the regulatory uses and applications in OECD member countries of (quantitative) structure-activity relationship [(Q)SAR] models in the assessment of new and existing chemicals. *OECD Papers*. 2006; 6: 79-157. doi: 10.1787/oecd_papers-v6-art37-en.

12. Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati SA. Interpretation of QSAR models based on random forest method. *Mol Inf*. 2011; 30(6-7): 593-603. doi: 10.1002/minf.201000173.

13. Artemenko AG, Kuz'min VE, Nedostup VI, inventors; O.V. Bogatsky Physico-Chemical Institute NAS of Ukraine, assignee. Komp'yuterna programa «Ekspertna systema «TransProp Expert» dlya prognozuvannya termodynamichnykh transportnykh vlastyvostei (koefitsiientiv viazkosti ta teploprovidnosti) organichnykh spoluk» [Computer program «Expert system «TransProp Expert» for predicting thermodynamic transport properties (viscosity and thermal conductivity coefficients) of organic compounds»]. Ukraine copyright certificate UA 111479. 2022 Mar 31. (*in Ukrainian*).