ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

4 • 2012

ХІМІЯ

УДК 544.169

Е. В. Варламова, Е. Н. Муратов, П. Г. Полищук, А. Г. Артеменко, В. Е. Кузьмин

QSPR-анализ температур кипения и конденсации двухкомпонентных смесей

(Представлено академиком НАН Украины Г. Л. Камаловым)

Основными целями данного исследования являлись: 1) модификация метода симплексного представления молекулярной структуры для QSPR-анализа смесей соединений; 2) применение разработанного метода для анализа температуры кипения и конденсации различных бинарных смесей. Независимая внешняя валидация показала, что полученные модели пригодны для прогноза свойств смесей (точность прогноза ±7 K) для заполнения пропущенных элементов в матрице смесей (более 2000 смесей), сформированной 67 индивидуальными соединениями из обучающей выборки.

Данные по равновесию системы жидкость — пар представляют собой важную информацию, необходимую для описания фазового поведения бинарных смесей жидкостей. Информация о фазовом равновесии является ключевой для проектирования различных химико-технологических процессов. Однако экспериментальные данные по равновесию системы жидкость — пар представлены в литературе недостаточно широко и ограничены хорошо известными растворителями. Поэтому создание эффективных инструментов для прогноза параметров равновесия жидкость — пар является важной задачей.

Таким образом, цель данного исследования — разработка метода дескрипторного представления смесей соединений различного состава для QSPR*-анализа; построение адекватных QSPR моделей для описания температур кипения/конденсации двухкомпонентных смесей; реализация и внешняя валидация полученных моделей (прогноз температур кипения ($T_{\text{кип}}$) / конденсации ($T_{\text{конд}}$) для новых смесей).

Экспериментальная информация из базы данных [1] была использована для формирования исследуемой выборки. Она включает данные о 67 индивидуальных соединениях и 167 их смесях. Следовательно, матрица всех возможных пар исследуемых соединений достаточно разряжена (из 2211 возможных смесей $T_{\text{кип}}$ и $T_{\text{конд}}$ могут быть спрогнозированы для 2044 новых смесей). Каждая пара соединений представлена набором различных составов (от 7

*QSPR — quantitative structure — property relationship.

до 57) и соответствующими $T_{\text{кип}}$ и $T_{\text{конд}}$. Вся изучаемая выборка описывается 3185 точками (составами). Соединения, входящие в выборку, представляют различные классы: углеводороды, галогенпроизводные углеводородов, спирты, простые и сложные эфиры и т. д. Исследуемые свойства — $T_{\text{кип}}$ и $T_{\text{конд}}$ — распределены в интервале 277–475 K.

Состав смесей описывается мольными долями соответствующих компонентов. Экспериментальные ошибки измерения составляют 0,06 K для температуры и 0,001 для состава, согласно данным N. Alpert, P. Elving (1951) и публикации [2].

В настоящем исследовании были использованы три разные стратегии прогноза изучаемых свойств.

1. **"Points out".** Прогноз $T_{\text{кип}}$ и $T_{\text{конд}}$ для новых составов известных пар соединений (167 смесей). Для этого все индивидуальные вещества всегда оставались в обучающей выборке, и некоторые составы для указанных 167 пар соединений произвольно отбирались в тестовую выборку в соответствии с методикой N-кратной внешней кросс-валидации (см. ниже).

2. **"Mixtures out".** Прогноз $T_{\text{кип}}$ и $T_{\text{конд}}$ для тех пар соединений, для которых отсутствует экспериментальная информация (2044 смесей). Для этого все 67 индивидуальных веществ всегда оставались в обучающей выборке, а некоторые их смеси отбирались полностью в тестовую выборку (в соответствии с методикой N-кратной внешней кросс-валидации), т. е. все составы для соответствующей пары компонентов.

3. **"Compounds out".** Прогноз $T_{\text{кип}}$ и $T_{\text{конд}}$ для смесей, образованных одним либо двумя новыми индивидуальными веществами. Для этого некоторые компоненты и все их смеси были отнесены к тестовой выборке (в соответствии с методикой N-кратной внешней кросс-валидации).

Очевидно, что сложность задач QSPR возрастает в ряду "Points out" > "Mixtures out" > > "Compounds out".

Для описания структуры исследуемых веществ и смесей использовалось симплексное представление молекулярной структуры (СПМС) [3]. В рамках СПМС любая молекула может быть представлена в виде системы различных симплексов (четырехатомных фрагментов фиксированного состава и структуры). Атомы в симплексе могут быть дифференцированы на основе различных характеристик:

метка (символ), характеризующая индивидуальность атома (природа атома или более детализированный тип);

метка, характеризующая частичный заряд на атоме [4, 5] (отражает электростатические свойства);

метка, характеризующая липофильность атома [6] (отражает гидрофобные свойства);

метка, характеризующая рефракцию атома (электронную поляризуемость). Последняя, в определенной степени, отражает способность атома участвовать в дисперсионных взаимодействиях;

метка, характеризующая возможность атома быть донором/акцептором водорода в потенциальной водородной связи (ВС). Атомы делятся на три группы: A — акцептор водорода в ВС, D — донор водорода в ВС, I — индифферентный атом.

Для атомных характеристик, имеющих численные значения (заряд, липофильность и т. п.), на предварительной стадии проводится разделение диапазона значений на определенное количество дискретных групп. Количество групп (G) является настроечным параметром и может варьироваться (как правило, $G = 3$–7). Использование различных вариантов дифференциации вершин симплексов (атомов) является принципиальной особен-

ностью предлагаемого подхода. Мы полагаем, что реализованная во многих QSAR*/QSPR методах конкретизация атомов только по их природе (например, C, N, O) ограничивает возможности выделения активных фрагментов. Например, если группа −NH− выбрана в качества фрагмента, определяющего активность и возможность образования ВС, является фактором, определяющим его активность, то мы можем упустить такие доноры ВС, как, например, OH-группа и т. п. Использование дифференциации атомов по свойству быть донором/акцептором ВС позволяет избежать описанной выше ситуации. Аналогичные примеры можно привести для других атомных свойств.

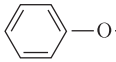
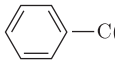
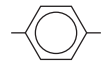
Дескрипторное представление смесей соединений не может быть выражено как соответствующая линейная комбинация дескрипторов компонентов смеси, так как в большинстве случаев зависимости "структура, состав — свойство" не являются аддитивными. Поэтому симплексный метод был дополнен и усовершенствован для QSAR/QSPR анализа бинарных смесей. Главное отличие от обычного симплексного подхода заключается в том, что используются симплексные дескрипторы, характеризующие непосредственно смесь соединений (см. ниже). Связные симплексы (5–11 типа, табл. 1) описывают только отдельные компоненты смеси, тогда как несвязные симплексы (3 и 4 типа, см. табл. 1) могут описывать и компоненты в отдельности, и смесь в целом. В связи с этим необходимо отметить следующее: какие несвязные симплексы будут принадлежать одной и той же молекуле, а какие разным. В последнем случае такие несвязные симплексы отражают структуру не одной молекулы, а характеризуют пару разных молекул. Фактически данные симплексы являются структурными дескрипторами смеси веществ (рис. 1). Чтобы различить такие симплексы, в процессе генерирования дескрипторов им присваивается специальная метка. В данном подходе учитывается состав смеси, т. е. дескрипторы отдельных компонентов (вещества A и B) взвешиваются, согласно мольным долям компонентов в смеси, а дескрипторы смеси умножаются на минимальную долю одного из компонентов.

Для построения моделей "структура–свойство" использовался метод Random Forest (RF). RF — сравнительно новый статистический метод анализа [7], который приобретает все большую популярность для построения QSAR/QSPR моделей [8, 9], благодаря таким своим преимуществам, как отсутствие проблемы переобучения моделей; отсутствие необходимости отбора (преселекции) переменных; наличие адекватной внутренней процедуры оценки качества и прогнозирующей способности моделей (параметр, R^2_{oob}); устойчивость моделей к наличию шума в исходной выборке; эффективная работа с большими базами данных; интерпретируемость получаемых моделей; возможность корректно анализировать выборки, включающие соединения с различным механизмом действия.

Предварительно, прогнозирующая способность QSPR моделей оценивалась методом N-кратной внешней кросс-валидации (N = 2–10). Суть которой заключается в том, что исследуемая выборка разбивается на N частей. Каждая из этих частей используются для верификации, при этом модели строятся из оставшихся $(N-1)$ частей выборки. Данный процесс повторяется N раз. Таким образом, все соединения исследуемой выборки используются для прогноза [10]. Полученные модели затем объединяются в консенсусные (усредненное значение прогноза по всем отобранным моделям). Было получено по две консенсусные модели (для $T_{кип}$ и $T_{конд}$) для каждой стратегии. Все построенные модели обладают хорошими статистическими характеристиками (R^2 = 0,99, R^2_{oob} = 0,95–0,99). Характеристики прогнозирующей способности консенсусных моделей представлены в табл. 2. Как видно из

*QSAR — quantitative structure — activity relationship.

Таблиця 1. Зміна частот на FTIR-спектрах композитів ЕП–(Me_2O_3 + ПАн) під дією фізичних полів

Характеристичні групи	Вихідні композити				Постійне магнітне поле		Постійне електричне поле	
	ЕП*	ЕП–ПАн	ЕП–(Fe_2O_3 + + ПАн)	ЕП–(Al_2O_3 + + ПАн)	ЕП–(Fe_2O_3 + + ПАн)	ЕП–(Al_2O_3 + + ПАн)	ЕП–(Fe_2O_3 + + ПАн)	ЕП–(Al_2O_3 + + ПАн)
γ(CH)$_{\text{аром}}$ 	830	826	826	826	826	826	826	826
Нереакційноздатні	916	918	—	—	—	918	—	918 плече
епоксидні цикли	930	933	—	930	933	930	—	930
ν(> C$_{\text{Ar}}$–O–C)	1039	1030	1039	1030	1034	1034	1034	1034
ν(C–O)	1084	1084	1086	1084	1088	1084	1088	1084
	1107	1107	1107	1107	1107	1107	1107	1107
Аліфатичні ν(C–N)	1180	1180	1182	1180	1180	1180	1180	1180
> C–O– у групі 	1246	1238	1248	1234	1238	1234	1238	1234
CH$_3$-групи у *біс*-фенолі А	1363	1362	1362	1362	1362	1362	1362	1362
Ar–CH(CH$_3$)$_2$–CH$_2$–CH	1385	1381	1383	1381	1385	1380 плече	1381	1381 плече
δ(CH$_2$, CH$_3$)	1460	1458	1460	1458	1458	1458	1458	1458
Ароматичне кільце	1508	1508	1508	1508	1508	1508	1508	1508
δNH-групи вторинних амінів,		—	1541	—	—	—	—	—
H-зв'язані ν(C–N)		—	1558	—	—	—	1551	—
		1586	1582	1581	1582	1581	1582	1581
Бензольне кільце νC + C;	1582	1582	1580	1577	1577	1582	1578	1582
cis –CH=CH–	1608	1605	160	16059	1605	1605	1609	1609
	1658	—	1654	1653	—	1655	—	
	1705	1706	1705	—	1705	—	1705	
	1744	1743	—	1740	1743	1744	—	
$\nu_{\text{асим}} + \nu_{\text{сим}}CH_2$	2852	2851	2852	2851	2855	2827/2866	2827/2855	2828/2870
	2920	2920	2922	2924	2924	2924	2924	2924
ν(CH$_3$) у фрагменті 	2955	2959	2960	2951 плече	2951 плече	2955	2951 плече	2963
ν(CH) 	3030	3031	3035	3044	—	3044	3043	3036
		3055	3055	3055	3055	3059	3055	3051
OH- і NH-групи	3320	—	3321 плече	3227	3217	—	—	—
	3420	3361	3414	—	3283	3302	3283	3337
			3495 плече		3356		—	3391

* За даними роботи [14].

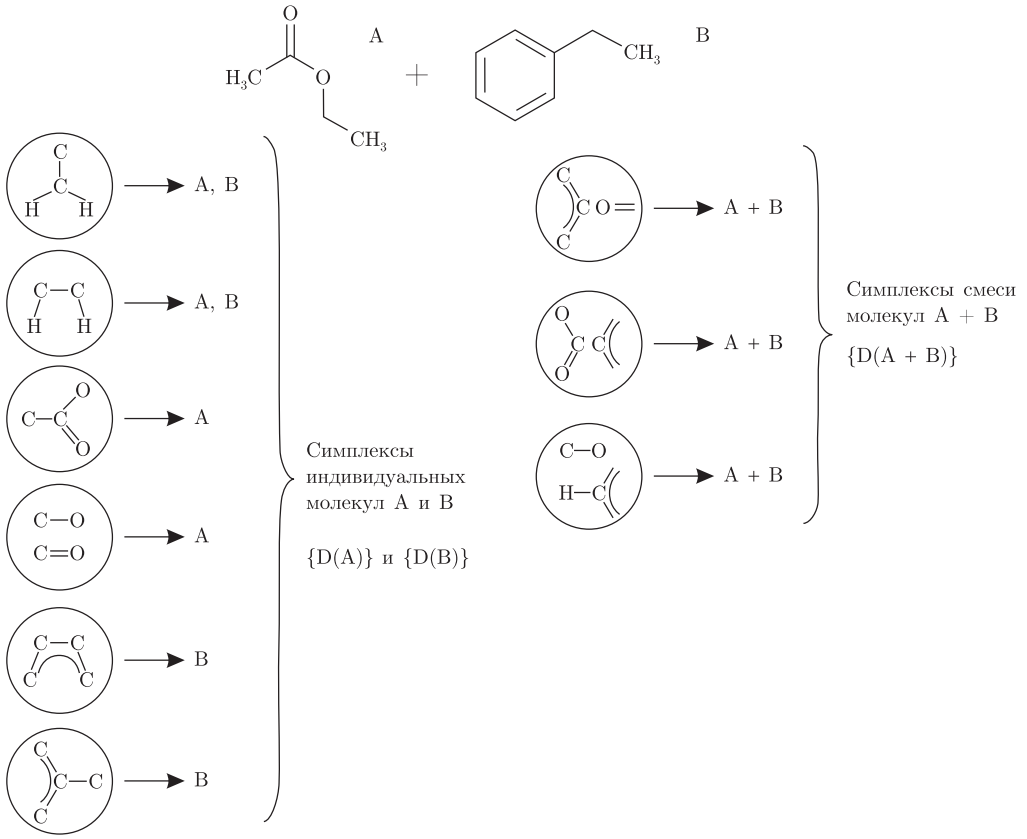


Рис. 1. Симплексные дескрипторы смесей соединений. Описание смеси: $\mathrm{D} = \begin{cases} [n\mathrm{A}\{\mathrm{D(A)}\} + n\mathrm{B}\{\mathrm{D(B)}\}], \\ n\mathrm{A}\{\mathrm{D(A+B)}\}, \end{cases}$
где nA и nB — мольные концентрации компонентов A и B в смеси; $n\mathrm{A} < n\mathrm{B}$, $n\mathrm{A} + n\mathrm{B} = 1$

таблицы, минимальная ошибка прогноза характерна для стратегии “Points out”, однако такие модели наименее интересны в практическом плане, они по существу являются апроксимационными для кривых кипения/конденсации. Больший прикладной интерес представляют модели стратегии “Mixtures out”, так как позволяют прогнозировать с приемлемой точностью температуры кипения/конденсации смесей соединений на основе экспериментальной информации об их компонентах из исходной матрицы данных (67×67). Наиболее сложная задача — прогноз температур кипения/конденсации, когда отсутствуют экспериментальные данные не только для смесей, но и для индивидуальных веществ (стратегия “Compounds out”) — решается наименее точно. Ошибку прогноза для этой стратегии можно немного понизить (до 8,5 K), если ввести информацию для соответствующих новых

Таблица 2. Характеристики прогнозирующей способности полученных QSPR моделей

Температура, K	Points out		Mixtures out		Compounds out	
	R^2_{ts}	$RMSE_{ts}$	R^2_{ts}	$RMSE_{ts}$	R^2_{ts}	$RMSE_{ts}$
$T_{кип}$	0,98	3,2	0,90	6,9	0,79	10,3
$T_{конд}$	0,97	3,9	0,88	7,6	0,78	10,7

Примечание. R^2_{ts} — коэффициент детерминации для N-кратной внешней кросс-валидации; $RMSE_{ts}$ — стандартная ошибка прогноза для N-кратной внешней кросс-валидации.

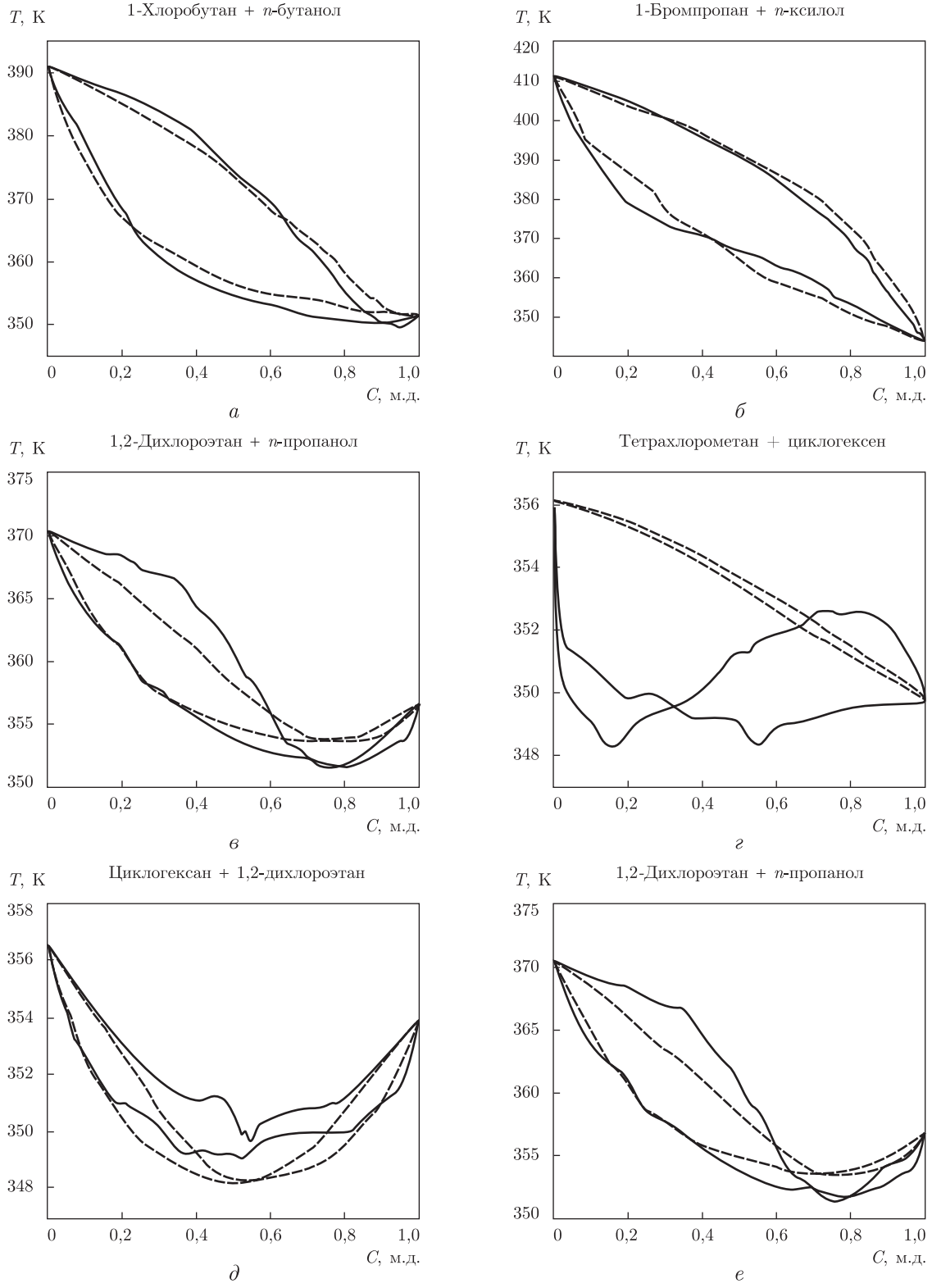


Рис. 2. Примеры экспериментальных (пунктирные линии) и предсказанных (сплошные линии) кривых равновесия жидкость — пар

индивидуальных веществ, отсутствующих в исходной матрице данных. Для того чтобы смоделировать такую ситуацию, были построены новые модели следующим образом: все индивидуальные соединения оставались всегда в обучающей выборке, а все смеси некоторых компонентов были отнесены к тестовой выборке.

Дополнительно предсказательная способность полученных моделей была проверена на внешней тестовой выборке. Для прогноза свойств 28 смесей (644 состава), которые не содержат новых соединений, использовались модели, полученные по стратегии "Mixtures out", а для 67 смесей (1366 составов), содержащих хотя бы одно новое индивидуальное вещество и 35 новых индивидуальных соединений использовались модели, полученные по стратегии "Compounds out". Для стратегии "Mixtures out" ошибка прогноза (7,2 K) для внешней тестовой выборки соизмерима с ошибкой (6,9 K) для N-кратной внешней кросс-валидации, тогда как для стратегии "Compounds out" ошибка составляет 18,5 K, что превышает ошибку кросс-валидации (10,3 K).

В заключительной части работы был проведен сравнительный анализ экспериментальных и расчетных кривых (модели "Mixtures out") конденсации и кипения (некоторые примеры приведены на рис. 2). Отметим, что указанные модели работают в том числе и для пар соединений, образующих азеотропные смеси (рис 2, *д*, *е*). В некоторых случаях, когда разница между температурами кипения индивидуальных веществ меньше ошибки прогноза, не удалось получить адекватные модели кривых конденсации/испарения соответствующих смесей (рис. 2, *г*).

Таким образом, метод симплексного представления молекулярной структуры был модифицирован и успешно применен для QSPR-анализа температур кипения/конденсации двухкомпонентных смесей. Показано, что модели "Mixtures out" могут быть успешно использованы для прогнозирования свойств смесей. Наши дальнейшие усилия будут направлены на совершенствование разработанного подхода для повышения точности моделей "Compounds out".

1. *Kang J. W., Yoo K. P., Kim H. Y. et al.* Development and current status of Korea Thermophysical Properties Databank(KDB) // Int. J. Thermophys. – 2001. – **22**. – P. 487–494.
2. *Miller K. J., Huang H. S.* Vapor-liquid equilibrium for binary systems 2-butanone with 2-butanol, 1-pentanol, and isoamyl alcohol // J. Chem. Eng. Data. – 1972. – **17**. – P. 77–78.
3. *Kuz'min V. E., Artemenko A. G., Muratov E. N.* Hierarchical QSAR technology on the base of Simplex representation of molecular structure // J. Comp. Aid. Mol. Des. – 2008. – **22**. – P. 403–421.
4. *Jolly W. L., Perry W. B.* Estimation of atomic charges by an electronegativity equalization procedure calibration with core binding energies // J. Am. Chem. Soc. – 1973. – **95**. – P. 5442–5450.
5. *Кузьмин В. Е., Берестецкая Е. Л.* Программа расчетов зарядов на атомах методом выравнивания орбитальных электроотрицательностей // Журн. структур. химии. – 1983. – **24**. – С. 187–188.
6. *Wang R., Fu Y., Lai L.* A new atom-additive method for calculating partition coefficients // J. Chem. Inf. Comp. Sci. – 1997. – **37**. – P. 615–621.
7. *Breiman L.* Random Forests // Mach. Learn. – 2001. – **45**. – P. 5–32.
8. *Svetnik V., Liaw A., Tong C. et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling // J. Chem. Inf. Comp. Sci. – 2003. – **43**. – P. 1947–1958.
9. *Polishchuk P. G., Muratov E. N., Artemenko A. G. et al.* Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity // J. Chem. Inf. Model. – 2009. – **49**. – P. 2481–2488.
10. *Tropsha A.* Recent advances in development, validation and exploitation of QSAR models // Burger's Medicinal Chemistry and Drug Discovery. Vol. 1 / Ed. D. Abraham. – New York: Wiley, 2010. – P. 505–533.

Физико-химический институт им. А. В. Богатского НАН Украины, Одесса

Поступило в редакцию 24.10.2011

К. В. Варламова, Є. Н. Муратов, П. Г. Поліщук, А. Г. Артеменко, В. Є. Кузьмін

QSPR-аналіз температур кипіння і конденсації двокомпонентних сумішей

Основними цілями даного дослідження були: 1) модифікація методу симплексного представлення молекулярної структури для QSPR-аналізу сумішей сполук; 2) застосування розробленого методу для аналізу температури кипіння і конденсації різних бінарних сумішей. Незалежна зовнішня валідация показала, що отримані моделі придатні для прогнозу властивостей сумішей (точність прогнозу ±7 K) для заповнення пропущених елементів у матриці сумішей (понад 2000 сумішей), сформованій 67 індивідуальними сполуками з навчальної вибірки.

E. V. Varlamova, E. N. Muratov, P. G. Polischuk, A. G. Artemenko, V. E. Kuz'min

QSPR-analysis of the temperatures of boiling and condensation of two-component mixtures

The main goals of our study were 1) modification and adaptation of the SiRMS-approach to QSPR-analysis of mixtures of compounds, 2) its application to the analysis of the boiling and condensation temperatures of various binary mixtures. The rigorous external validation shows that the obtained models are well-suitable (accuracy of ±7 K) for the gap-filling of missed data (more than 2000 mixtures) in the mixture matrix created by 67 pure liquids from the training set.