

А. Д. Лифанов, Е. Г. Лифанова

**ПОСТРОЕНИЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ЭНТАЛЬПИИ СУБЛИМАЦИИ
ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ
И ФРАГМЕНТНЫХ ДЕСКРИПТОРОВ**

Ключевые слова: большие данные, индустрия 4.0, энтальпия сублимации, машинное обучение, искусственный интеллект.

В настоящее время в химии накоплен большой массив экспериментальных данных. В связи с этим, возникает необходимость совершенствования вычислительных методов хранения, обработки экспериментальных данных. В данном подходе органические соединения представляются в виде набора дескрипторов, характеризующих особенности химической структуры молекул. На практике переход в газовую фазу термодинамически возможен не только для жидкостей, но и для веществ, находящихся в твердом состоянии. Для большого числа органических соединений возможен их переход из твердого состояния в газообразное минуя жидкую фазу. Данный процесс называется сублимацией. Количественной характеристикой такого перехода является энтальпия сублимации, которая обозначается как ΔH_{sub} . Данный параметр является важной термодинамической характеристикой и несомненно представляет практический интерес. Данные об октановых числах для органических соединений были взяты из литературных источников. В базу данных для нашей работы была включена информация об энтальпии сублимации для 845 органических веществ. Для упрощения анализа представления органических соединений, мы использовали 208 дескрипторов RDKit, поскольку они являются одними из лучших дескрипторов для прогнозирования свойств химических соединений. Данные дескрипторы создаются на основе общих ключей подструктуры. Кроме того, модели были рассчитаны с использованием молекулярных отпечатков Моргана, также известных как циркулярные отпечатки с радиусом 2. В рамках данной работы были реализованы гребневая регрессия, алгоритм случайного леса, метод ближайших соседей kNN, метод опорных векторов (SVM), искусственные нейронные сети. Для обучающей выборки полученная классификационная модель случайного леса показала безошибочную классификацию, ошибка прогноза для нее равна 0. Статистические характеристики построенной модели гребневой регрессии для выборки имеют следующие значения: $R^2 = 0.88$ и ошибкой предсказания $RMSE = 13.05$ кДж/моль.

A. D. Lifanov, E. G. Lifanova

**CONSTRUCTION OF A MODEL FOR PREDICTING THE ENTHALPY OF SUBLIMATION OF ORGANIC
COMPOUNDS USING MACHINE LEARNING METHODS AND FRAGMENTS DESCRIPTORS**

Keywords: big data, industry 4.0, enthalpy of sublimation, machine learning, artificial intelligence.

Currently, a large amount of experimental data has been accumulated in chemistry. In this regard, there is a need to improve computational methods for storing and processing experimental data. In this approach, organic compounds are represented as a set of descriptors that characterize the features of the chemical structure of molecules. In practice, the transition to the gas phase is thermodynamically possible not only for liquids, but also for substances in the solid state. For a large number of organic compounds, their transition from a solid state to a gaseous state is possible, bypassing the liquid phase. This process is called sublimation. The quantitative characteristic of such a transition is the enthalpy of sublimation, which is denoted as ΔH_{sub} . This parameter is an important thermodynamic characteristic and is undoubtedly of practical interest. Data on octane numbers for organic compounds were taken from literature data. Information on the enthalpy of sublimation for 845 organic substances was included in the database for this work. To simplify the analysis of the representation of organic compounds, we used 208 RDKit descriptors, as they are among the best descriptors for predicting the properties of chemical compounds. These descriptors are created based on the shared keys of the substructure. In addition, the models were calculated using Morgan's molecular fingerprints, also known as circular prints with a radius of 2. Within the framework of this work, ridge regression, the random forest algorithm, the kNN nearest neighbor method, the support vector machine (SVM) method, and artificial neural networks were implemented. For the training sample, the obtained random forest classification model showed an error-free classification, the prediction error for it is 0. The statistical characteristics of the constructed ridge regression model for the sample have the following values: $R^2 = 0.88$ and the prediction error $RMSE = 13.05$ kJ/mol.

Введение

В настоящее время в химии накоплен большой массив экспериментальных данных. В связи с этим, возникает необходимость совершенствования вычислительных методов хранения, обработки экспериментальных данных. Это, в свою очередь, позволяет разрабатывать эффективные модели и алгоритмы для создания новых функциональных материалов с заданными свойствами.

Некоторые свойства характеризуют определенные процессы, происходящие в материалах под воздействием различных факторов. Так, например, повышение температуры способствует испарению жидкостей. Количественной характеристикой, позволяющей описать условия данного процесса является температура кипения. Именно при этой температуре начинается переход органического вещества из жидкой фазы в газообразную [1].

На практике переход в газовую фазу термодинамически возможен не только для жидкостей, но и для веществ, находящихся в твердом состоянии. Для большого числа органических соединений возможен их переход из твердого состояния в газообразное минуя жидкую фазу. Данный процесс называется сублимацией.

Количественной характеристикой такого перехода является энтальпия сублимации, которая обозначается как ΔH_{sub} . Данный параметр является важной термодинамической характеристикой и несомненно представляет практический интерес. В более общем смысле под энтальпией сублимации ΔH_{sub} следует понимать минимальную энергию, которую необходимо приложить к молекуле для снижения сил межмолекулярного взаимодействия, которые удерживают ее в твердом состоянии.

С практической точки зрения, информация о величине ΔH_{sub} необходимо для решения проблемы распределения красителя на поверхности текстильных материалов, для предотвращения их выцветания под действием солнечного света [3].

Экспериментальное определение энтальпии сублимации органических соединений является довольно трудоемкой и дорогостоящей процедурой. Поэтому были предложены расчетные методики, позволяющие рассчитать величину ΔH_{sub} по химической структуре органического соединения. Одним из таких подходов является QSPR/QSAR. Подход Quantitative Structure – Property Relationships (QSPR) занимается исследованием закономерностей связи между химическим строением соединений и их свойствами. Модель «структура-свойство» устанавливает функциональную зависимость свойства от структуры [4].

Одним из основных требований к установлению закономерностей типа «структура – свойство» является обеспечение высокой прогностической способности моделей [2].

Методы QSPR получили широкое распространение для прогнозирования энтальпии сублимации органических соединений [5, 6]. В научной литературе имеется около 10 работ, посвященных прогнозированию энтальпии сублимации органических соединений [7-12]. Наборы данных, которые использовались для построения модели варьировались от 30 до 1800 записей. Анализ данных работ показывает, что при построении модели недостаточно внимания уделяется оценке достоверности экспериментальных данных. Кроме того, в качестве дескрипторов обычно выбираются квантово-химические [9], параметры электростатического потенциала на поверхности молекулы [11]. При этом, выбранные дескрипторы не в полной мере учитывают особенности структуры молекул.

В настоящее время прогнозирование величины ΔH_{sub} выполнено на относительно небольшой выборке химических соединений, что ограничивает область применимости данных моделей. Тем не менее, построенные модели линейной регрессии, искусственных нейронных сетей обладают существенными преимуществами перед расчетными

методами. Тем не менее, мы считаем возможным апробировать подход с использованием фрагментных дескрипторов.

Наиболее перспективным является установление закономерностей связи химического строения органических соединений с показателями энтальпии сублимации. В данном подходе органические соединения следует представлять в виде фрагментных дескрипторов, характеризующих особенности химического строения молекул.

Цель работы заключалась в исследовании закономерностей связи химического строения органических соединений с октановыми числами.

Материалы и методы

Традиционно, построение QSPR-модели проходит несколько этапов: подготовка данных, выбор модели, отбор дескрипторов, обучение модели и анализ результатов.

Данные об октановых числах для органических соединений были взяты из базы данных ChemSpider [5].

Собранный набор данных содержал информацию о свойствах 845 органических соединений с их значениями энтальпии сублимации в диапазоне от 13.87 до 243.24 кДж/моль.

В нашей работе в качестве линейных представлений выбрана система SMILES [13, 14]. Далее на основе SMILES генерируется дескриптор, отражающий особенности структуры органического соединения [15]. Информация о структуре химических соединений кодируется дескрипторами.

Описание структуры химических соединений в данном подходе может осуществляться по-разному, то есть разными наборами дескрипторов, выбор которых определяется степенью детализации структуры в решаемой задаче [16-23].

Данные дескрипторы рассчитаны с использованием библиотеки RDKit 2 [16]. RDKit представляет собой инструментальный для хемоинформатики с открытым исходным кодом [16]. Кроме того, в качестве дескрипторов рассматривались PAdEL (Pharmaceuticals Data Exploration Laboratory) [17], CDK (Chemistry Development Kit) [18], ChemoPy (Chemoinformatics in Python) [19].

В качестве дескрипторов, отражающих структуры органических соединений, – молекулярные отпечатки (Morgan's fingerprints) [23] с параметрами: радиус 2, длина битовой строки 2048 бит (чем больше радиус, тем большее окружение атомов учитывается при кодировании фрагментов; чем больше длина битовой строки, тем больше информации о структурных фрагментах можно закодировать и избежать потери информации о связях «структура-свойство»), а также структурные ключи (MACCS keys) [20, 21].

Полученные молекулярные отпечатки и структурные ключи используются в качестве входных данных для обучения модели, которая может быть реализована на основе любого метода машинного обучения.

В рамках данной работы были реализованы гребневая регрессия, алгоритм случайного леса [24], метод ближайших соседей kNN [25, 26], метод опорных векторов (SVM) [27], градиентный бустинг (XGBoost) [28], AdaBoostRegressor (AdaBoost) [29].

Оценка качества регрессионных моделей осуществлялась с использованием специальных метрик sklearn.metrics. Для этого из библиотеки sklearn импортировались r2_score (коэффициент детерминации) и mean_squared_error (среднеквадратичная ошибка). Рассмотрим формулы для их расчета более подробно.

Коэффициент детерминации (R^2) рассчитывается следующим образом:

$$R^2 = 1 - \frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{exp}})^2}{\sum_i (y_i^{\text{pred}} - \bar{y}^{\text{exp}})^2} \quad (1)$$

где y_i^{pred} и y_i^{exp} - предсказанные и истинные значения характеристик объектов, соответственно, $i = 1, 2, 3, \dots, N$, N - объем выборки ($N=845$).

Как и коэффициент детерминации R^2 (1), статистический показатель Q^2 (тот же R^2 , после процедуры кросс-валидации) не может принимать значения, превышающие 1, причем, чем ближе его значение к 1, тем выше средняя прогнозирующая способность моделей [30].

Следующим важным статистическим критерием является среднеквадратичная ошибка (RMSE) (2). Среднеквадратичная ошибка RMSE может принимать любое положительное значение, включая ноль, причем, чем ниже, тем выше средняя прогнозирующая способность моделей, оцененная посредством процедуры 5-кратного скользящего контроля. RMSE рассчитывается следующим образом:

$$RMSE = \sqrt{\frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{exp}})^2}{N - 1}} \quad (2)$$

Таким образом, показатели RMSE и Q^2 являются универсальными показателями прогнозирующей способности моделей [31].

Таким образом, показатели RMSE и Q^2 (тот же R^2 , но с процедурой n-кратного перекрестного скользящего контроля являются универсальными показателями прогнозирующей способности моделей [21].

Результаты и их обсуждение

В результате исследования мы получили адекватные модели «структура-свойство» с высокими статистическими показателями. Для прогноза и интерпретации использовались модели как результат пятикратного скользящего контроля (Рис. 1) [31].

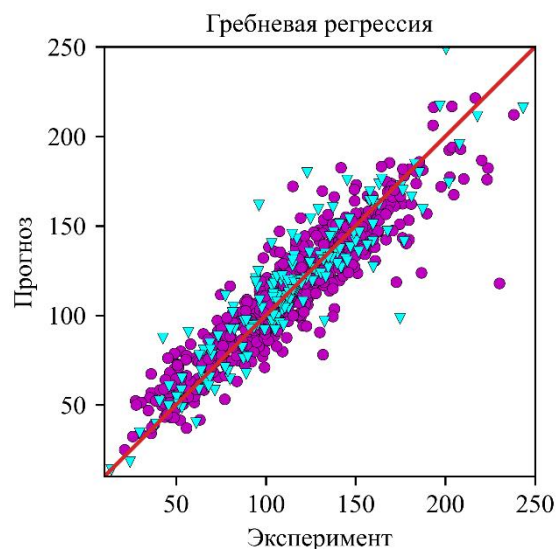


Рис. 1 - Регрессионная модель гребневой регрессии для прогнозирования энтальпии сублимации органических соединений

Fig. 1 - A ridge regression model for predicting the octane numbers of organic compounds

В работах [1, 2] показано, что прогностическая способность модели определяется числом дескрипторов, адекватно описывающих структуру химического соединения. В связи с этим, из 208 исходного набора дескрипторов RDKit, мы последовательно убирали из этапа обучения модели взаимно коррелирующие дескрипторы. Порог отсеивания взаимно коррелирующих дескрипторов находился в интервале от 0,9 до 0,1 с шагом 0,1.

Таблица 1 - Результаты фильтрации взаимно коррелирующих дескрипторов

Table 1 - Cross-correlation coefficients for selected descriptors

Порог	Число дескрипторов	Обучающая выборка		Тестовая выборка	
		R ²	RMSE	R ²	RMSE
Метод k-ближайших соседей (kNN)					
0,9	169	0,77	18,09	0,66	23,02
0,8	148	0,76	18,34	0,64	23,66
0,7	132	0,73	19,67	0,61	24,48
0,6	111	0,64	22,45	0,55	26,44
0,5	85	0,60	23,70	0,49	28,04
0,4	66	0,59	24,06	0,49	28,25
0,3	46	0,57	24,52	0,40	38,44
0,2	27	0,40	28,99	-0,04	40,12
0,1	17	-0,01	37,70	-0,00	39,66
Метод гребневой регрессии					
0,9	169	0,85	14,72	0,81	17,13
0,8	148	0,84	14,85	0,81	17,14
0,7	132	0,82	15,74	0,79	17,98
0,6	111	0,73	19,65	0,72	20,97
0,5	85	0,57	24,51	0,53	26,91
0,4	66	0,40	29,12	0,38	30,98
0,3	46	0,15	34,66	0,11	37,20
0,2	27	0,05	36,61	0,04	38,67
0,1	17	0,00	37,54	0,00	39,43

В результате моделирования мы построили адекватную модель гребневой регрессии с использованием в качестве дескрипторов RDKit с коэффициентом детерминации для обучающей выборки $R^2 = 0.85$ и ошибкой предсказания $RMSE = 14.48$ °C. Для тестовой выборки аналогичные показатели $R^2 = 0.82$ и $RMSE = 16.79$ °C. Из данных, представленных в табл. 1 можно видеть, что фильтрация взаимно коррелирующих дескрипторов не способствует улучшению качества моделей. Мы полагаем, что это связано со случайностью в отборе дескрипторов.

При выявлении сложных зависимостей между структурой и свойствами органических соединений, часто приходится сталкиваться с проблемой избыточности информации. Дескрипторы, описывающие структуру соединений, могут быть мультиколлинеарны. Это означает, что между дескрипторами может быть высокая корреляция. Так, дескриптор nC , описывающий количество атомов углерода в органическом соединении имеет высокую корреляцию с дескриптором $nAromaticRings$ (число ароматических циклов). Для уменьшения избыточности информации используют методы понижения размерности химического пространства. Одним из наиболее распространенных методов является метод главных компонент. Данный метод позволяет выявить латентные переменные, которые позволяют объяснить корреляцию между признаками (дескрипторами).

Поэтому, отбор дескрипторов нами проводился с использованием метода главных компонент (PCA – от англ. Principal component analysis). Гипотеза данного метода заключается в том, что главные компоненты являются лишь индикаторами дескрипторов, описывающих структуру химических соединений. Сами главные компоненты не измеряются экспериментально, а являются результатом объединения признаков.

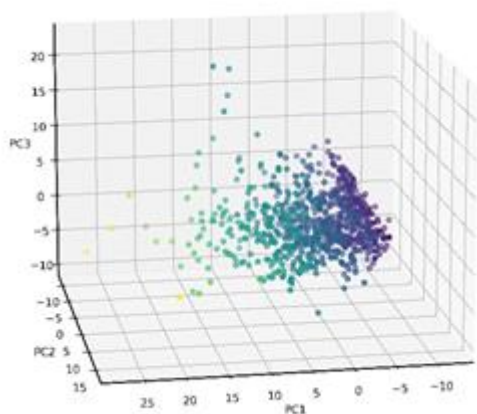


Рис. 2 - Визуализация главных компонент в трехмерном пространстве, характеризующих взаимную корреляцию дескрипторов

Fig. 2 - Visualization of the main components in three-dimensional space, characterizing the mutual correlation of descriptors

Окрашивание точек на рис. 2 выбрали согласно соответствующему значению молекулярной массы, но можно выбрать любой другой дескриптор.

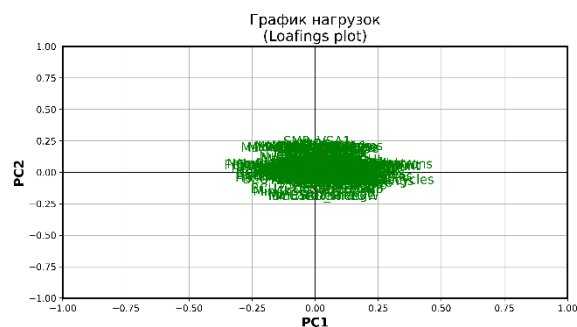


Рис. 3 - График нагрузок, характеризующих взаимную корреляцию дескрипторов

Fig. 3 - Graph of loads characterizing the cross-correlation of descriptors

Как видно из рис. 3, наблюдается взаимная корреляция дескрипторов, поэтому трудно отличить на этом графике дескрипторы. Для последовательного исключения взаимно коррелирующих фрагментных дескрипторов из QSPR-модели, мы использовали анализ главных компонент (PCA). Результаты отбора дескрипторов представлены на рис. 4.

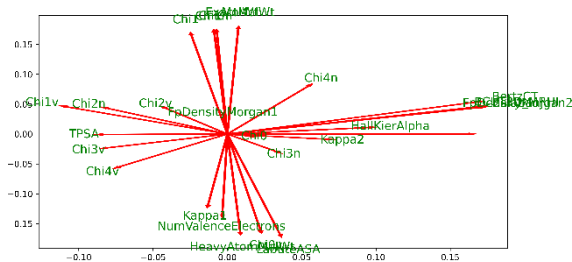


Рис. 4 - График нагрузок, характеризующих взаимную корреляцию для ограниченного набора дескрипторов

Fig. 4 - Graph of the loads characterizing the cross-correlation for a limited set of descriptors

При использовании в качестве дескрипторов главных компонент, качество моделей улучшилось (Рис. 5). На рис. 4 коррелирующие между собой дескрипторы объединяются в главные компоненты. Так, для гребневой регрессии, показатели качества регрессионной модели были следующие $R^2 = 0.88$, $RMSE = 13.05$ (на обучающей выборке) и $R^2 = 0.81$, $RMSE = 17.03$ (на тестовой выборке) (Рис. 4).

Еще одним действенным инструментом отбора дескрипторов является проведение процедуры n -кратного скользящего контроля (перекрестной проверки).

После процедуры 5-кратного перекрестного скользящего контроля качество модели осталось на прежнем уровне. В случае с методом ближайших соседей наблюдалось переобучение модели. Показатели качества регрессионной модели были следующие $R^2 = 1.00$, $RMSE = 0.63$ (на обучающей

выборке) и $R^2 = 0.39$, $RMSE = 30.79$ (на тестовой выборке).

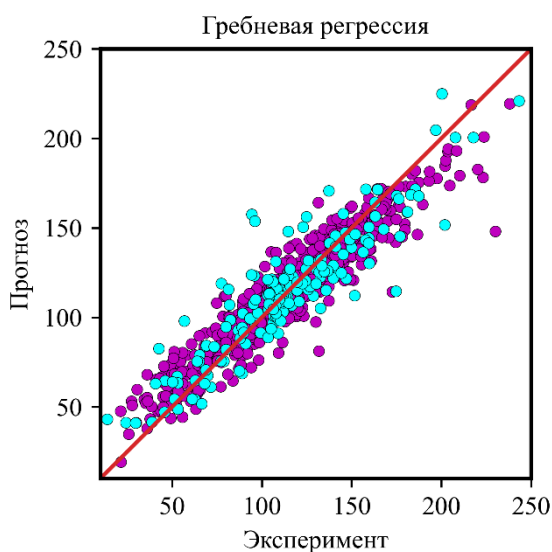


Рис. 5. - Регрессионная модель гребневой регрессии для прогнозирования энтальпии сублимации органических соединений (дескрипторы – главные компоненты, полученные методом PCA)

Fig. 5. - A regression model of ridge regression for predicting the enthalpy of sublimation of organic compounds (descriptors are the main components obtained by the PCA method)

Таким образом, мы приходим к выводу, что использование методов машинного обучения, которые не чувствительны к количеству дескрипторов (данные методы используют L1 и L2 – регуляризацию, т.е. вводят штраф за сложность модели) не требует предварительного отбора дескрипторов. К данным методам относятся алгоритм случайного леса, гребневая регрессия, метод опорных векторов, искусственные нейронные сети. Методы, которые не используют регуляризацию, например, метод ближайших соседей, осуществлять отбор наиболее значимых признаков не имеет смысла, поскольку это способствует снижению качества модели.

С целью определения влияния особенностей структуры органических соединений на их энтальпии сублимации, нами проведен корреляционный анализ, результаты которого представлены в таблице 2.

Из данных, представленных в табл. 2 можно видеть, что наибольшей корреляцией со свойством обладают молекулярная масса (MolWt), топологическая площадь полярной поверхности (TPSA), дескриптор, позволяющий получать стабильные результаты независимо от того, используются ли явные или неявные атомы водорода (LabuteASA), дескрипторы, характеризующие многоочлены матрицы смежности подавленного водородом графа молекулы (Kappa1, HallKierAlpha, Chi1n, Chi1, Chi0v, Chi0n, Chi0, BalabanJ, BertzCT), число валентных электронов (NumValenceElectrons),

точная молекулярная масса (Exact MolWt), молекулярная масса тяжелых атомов в структуре молекулы (HeavyAtomMolWt).

Таблица 2 - Результаты анализа корреляционных связей некоторых дескрипторов с энтальпией сублимации органических соединений

Table 2 - The results of the analysis of correlations of some descriptors with the enthalpy of sublimation of organic compounds

Класс	Наименование	Коэфф корр	Значимость дескриптора
Физико-химические	MolWt	0,60	значимый
	HeavyAtomMolWt	0,60	значимый
	Exact MolWt	0,58	мало значимый
	NumValenceElectrons	0,58	мало значимый
Графовые	BertzCT	0,53	мало значимый
	BalabanJ	-0,27	не значимый
	Chi0	0,59	мало значимый
	Chi0n	0,55	мало значимый
	ChiOv	0,57	мало значимый
	Chi1	0,58	мало значимый
	Chi1n	0,52	мало значимый
	HallKierAlpha	-0,53	мало значимый
	Kappa1	0,61	значимый
Молекулярные свойства	LabuteASA	0,59	мало значимый
	TPSA	0,43	мало значимый

Поскольку в работе использовалось большое количество дескрипторов (208 дескрипторов RDKit), мы оставили лишь те дескрипторы, которые оказывают существенное влияние на прогнозируемое свойство и имеют коэффициент корреляции 0,6 и выше. В том случае мы проводили отбор дескрипторов, оказывающих наибольшее влияние на энтальпию сублимации органических соединений. Однако это не означает, что все дескрипторы с коэффициентом корреляции с энтальпией сублимации необходимо включить в предложенную модель. Необходимо исключить взаимную корреляцию дескрипторов, поскольку они увеличивают способность модели к переобучению, а использование даже не чувствительных к избыточной информации методов машинного обучения, не позволяет учитывать все влияния, которые они оказывают на выходные данные модели (Табл. 2).

Из данных, представленных в табл. 2 можно видеть, что для большинства дескрипторов наблюдается значимая корреляция с энтальпией сублимации органических соединений. Поскольку

энтальпия сублимации – это энергия, необходимая для разрыва межмолекулярных связей, то наибольшей значимостью обладают физико-химические и молекулярные дескрипторы. Другие дескрипторы обладают гораздо меньшей корреляцией. При этом, графовый дескриптор BalabanJ показывает отрицательную корреляцию (-0,27). Поскольку нам необходимо было выяснить каким образом отбор дескрипторов сказывается на качестве регрессионных моделей, нами проведена оценка качества регрессионных моделей (Табл. 3).

Таблица 3 - Результаты прогнозирования энтальпии сублимации органических соединений с использованием регрессионных моделей

Table 3 - The results of forecasting the enthalpy of sublimation of organic compounds using regression models

Модель	Дескрипторы	Без отбора дескрипторов		С отбором дескрипторов	
		R ²	RMSE	R ²	RMSE
Ridge	MACCS	0,52	16,88	0,52	16,88
	ECPF4	0,83	9,96	0,83	9,96
AdaBoost	MACCS	0,42	18,54	0,33	19,88
	ECPF4	0,67	13,97	0,64	14,66
RF	ECPF4	0,80	10,88	0,47	17,68
	MACCS	0,45	18,13	0,43	18,41
XGBoost	MACCS	0,54	16,49	0,45	18,13
	ECPF4	0,78	11,38	0,79	11,29
kNN	MACCS	0,63	14,76	0,67	14,08
	ECPF4	0,66	14,23	0,68	13,74
SVR	MACCS	0,33	19,90	0,48	17,64
	ECPF4	0,11	22,94	0,80	10,77

В результате анализа фрагментов органических соединений не выявили четкой тенденции между вкладами фрагментов. Таким образом, мы получили адекватные регрессионные модели, на основе которых проведен прогноз определенных классов соединений для соответствующих свойств, а также провели сравнительный анализ фрагментов, способствующих повышению, или, наоборот, понижению энтальпии сублимации органических соединений [30-33].

Работа выполнена в рамках реализации Программы развития ФГБОУ ВО «КНИТУ» (ПСАЛ Приоритет 2030).

Литература

1. S. Fei, Y. Qi, W. Liu, Y. Wang, Z. Wang, H. Zhang, *Combustion Science and Technology*, **7**, 1-19 (2021).
2. K. Oyama, T. Sone, *Journal of the Japan Petroleum Institute*, **54**, 3, 201-207 (2011).
3. М.Ю. Доломатов, О.С. Коледин, Э.А. Ковалева, Р.А. Федина, Р.В. Гарипов, М.Р. Валеев, *Башкирский химический журнал*, **29**, 2, 65-70 (2022).
4. T.I. Madzhidov, A. Rakhimbekova, V.A. Afonina, *Mendeleev Communications*, **31**, 6, 769-780 (2021).
5. <http://www.chemspider.com/>.
6. W.J. Welsh, W. Tong, E.R. Collantes, J.S. Chickos, S.G. Gagarin, *Thermochim. Acta*, **290**, 55–64 (1997).

7. C. Ouyard, J.B.O. Mitchell, *Acta Cryst*, **B59**, 676-685 (2003).
8. M. Bagheri, M. Bagheri, A.H. Gandomi, A. Golbraik, *Thermochim. Acta*, **543**, 96–106 (2012).
9. F. Gharagheizi, *Thermochim. Acta*, **469**, 8-11 (2008).
10. G.L. Perlovich, O.A. Raevsky, *Cryst. Growth*, **10**, 2707-2712 (2010).
11. D. Mathieu, *Ind. Eng. Chem. Res.*, **51**, 2814–2819 (2012).
12. M. Salahinejad, T.C. Le, D.A. Winkler, *J. Chem. Inf. Model*, **53**, 223-229 (2013).
13. D. Weininger, *Journal of Chemical Information and Computer Sciences*, **28**, 1, 31-36 (1988).
14. D. Weininger, A. Weininger, J.L. Weininger, *Journal of Chemical Information and Computer Sciences*, **29**, 2, 97-101 (1989).
15. R. Todeschini, V. Consonni, New York: Wiley-VCH, 1-680 (2000).
16. <https://rdkit.org/>.
17. <http://padel.nus.edu.sg/sofware/padeldescriptor/>.
18. <http://sourceforge.net/projects/cdk>.
19. D. Cao, Q. Xu, Q. Hu, Y. Liang, *Bioinformatics*, **29**, 8, 1092–1094 (2013).
20. J. Gasteiger, T. Engel, Weinheim: Wiley-VCH (2003).
21. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, *J. Chem. Inf. Comput. Sci.*, **42**, 6, 1273-1280 (2002).
22. О.С. Коледин, М.Ю. Доломатов, Р.Ш. Япаев, А.Т. Гильмутдинов, М.Ф. Мухарметов, Р.В. Гарипов, М.Р. Валеев, *Журнал прикладной химии*, **95**, 5, 666-671 (2022).
23. H.L. Morgan, *Journal of Chemical Documentation*, **5**, 2, 107-113 (1965).
24. L. Breiman, *Machine Learning*, **45**, 1, 5-32 (2001).
25. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Consistency properties. Technical report. USAF School of Aviation Medicine (1951).
26. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Small sample performance. Technical report. USAF School of Aviation Medicine (1952).
27. Cortes, C., V. Vapnik, Support-vector networks. *Machine Learning*, **20**, 3, 273-297 (1995).
28. S.B. Gunturi, K. Archana, A. Khandelwal, R. Narayanan, *QSAR & Combinatorial Science*, **27**, 11-12, 1305–1317 (2008).
29. M. Budka, B. Gabrys, *Procedia Computer Science*, **1**, 1, 193–201 (2010).
30. Schein, A.I., L.H. Ungar, *Machine Learning*, **68**, 3, 235-265 (2007).
31. Gramatica, P. *QSAR & Comb Sci*, 694–701 (2007).
32. Meloun M, Militku J, Hill M. *Analyst*, **127**, 433–450 (2002).
33. Tharwat A. *Applied Computing and Informatics*, **17**, 1, 168–192 (2021).

References

1. S. Fei, Y. Qi, W. Liu, Y. Wang, Z. Wang, H. Zhang, *Combustion Science and Technology*, **7**, 1-19 (2021).
2. K. Oyama, T. Sone, *Journal of the Japan Petroleum Institute*, **54**, 3, 201-207 (2011).
3. M.Y. Dolomatov, O.S. Koledin, E.A. Kovaleva, R.A. Fedina, R.V. Garipov, M.R. Valeev, *Bashkir Chemical Journal*, **29**, 2, 65-70 (2022).
4. T.I. Madzhidov, A. Rakhimbekova, V.A. Afonina, *Mendeleev Communications*, **31**, 6, 769-780 (2021).
5. <http://www.chemspider.com/>.
6. W.J. Welsh, W. Tong, E.R. Collantes, J.S. Chickos, S.G. Gagarin, *Thermochim. Acta*, **290**, 55–64 (1997).
7. C. Ouyard, J.B.O. Mitchell, *Acta Cryst*, **B59**, 676-685 (2003).
8. M. Bagheri, M. Bagheri, A.H. Gandomi, A. Golbraik, *Thermochim. Acta*, **543**, 96–106 (2012).
9. F. Gharagheizi, *Thermochim. Acta*, **469**, 8-11 (2008).

10. G.L. Perlovich, O.A. Raevsky, Cryst. Growth, 10, 2707-2712 (2010).
11. D. Mathieu, Ind. Eng. Chem. Res., 51, 2814–2819 (2012).
12. M. Salahinejad, T.C. Le, D.A. Winkler, J. Chem. Inf. Model, 53, 223-229 (2013).
13. D. Weininger, Journal of Chemical Information and Computer Sciences, 28, 1, 31-36 (1988).
14. D. Weininger, A. Weininger, J.L. Weininger, Journal of Chemical Information and Computer Sciences, 29, 2, 97-101 (1989).
15. R. Todeschini, V. Consonni, New York: Wiley-VCH, 1-680 (2000).
16. <https://rdkit.org/>
17. <http://padel.nus.edu.sg/software/padeldescriptor/>
18. <http://sourceforge.net/projects/cdk>
19. D. Cao, Q. Xu, Q. Hu, Y. Liang, Bioinformatics 29, 8, 1092–1094 (2013).
20. J. Gasteiger, T. Engel, Weinheim: Wiley-VCH (2003).
21. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, J. Chem. Inf. Comput. Sci., 42, 6, 1273-1280 (2002).
22. O.S. Koledin, M.Yu. Dolomatov, R.Sh. Yapaev, A.T. Gil'mutdinov, M.F. Muharmtov, R.V. Garipov, M.R. Valeev, Journal of Applied Chemistry, 95, 5, 666-671 (2022).
23. H.L. Morgan, Journal of Chemical Documentation, 5, 2, 107-113 (1965).
24. L. Breiman, Machine Learning, 45, 1, 5-32 (2001).
25. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Consistency properties. Technical report. USAF School of Aviation Medicine (1951).
26. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Small sample performance. Technical report. USAF School of Aviation Medicine (1952).
27. Cortes, C., V. Vapnik, Support-vector networks. Machine Learning, 20, 3, 273-297 (1995).
28. S.B. Gunturi, K. Archana, A. Khandelwal, R. Narayanan, QSAR & Combinatorial Science, 27, 11-12, 1305–1317 (2008).
29. M. Budka, B. Gabrys, Procedia Computer Science, 1, 1, 193–201 (2010).
30. Schein, A.I., L.H. Ungar, Machine Learning, 68, 3, 235-265 (2007).
31. Gramatica, P. QSAR & Comb Sci, 694–701 (2007).
32. Meloun M, Militku J, Hill M. Analyst, 127, 433–450 (2002).
33. Tharwat A. Applied Computing and Informatics, 17, 1, 168–192 (2021).

© **А. Д. Лифанов** – к.х.н., доцент кафедры Общей химической технологии, Казанский национальный исследовательский технологический университет, Казань, Россия, lifanov84@mail.ru; **Е.Г. Лифанова** – ассистент кафедры Дизайна и национальных искусств, Казанский (Приволжский) федеральный университет, Россия, 678769@mail.ru.

© **A. D. Lifanov** – PhD (Chemical Sci.), Associate Professor of the Department of General Chemical Technology, Kazan National Research Technological University, Kazan, Russia, lifanov84@mail.ru; **E. G. Lifanova** – Assistant of the Department of Design and National Arts, Kazan (Volga Region) Federal University, Kazan, Russia, 678769@mail.ru.

Дата поступления рукописи в редакцию – 07.03.25.

Дата принятия рукописи в печать – 21.04.25.