Вестник ТвГУ. Серия "Химия". 2012. Выпуск 14. С. 160-172

УДК 541.6

КОЛИЧЕСТВЕННЫЕ МОДЕЛИ В КОРРЕЛЯЦИЯХ «СТРУКТУРА - СВОЙСТВО» ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

Ю.А. Федина, Ю.Г. Папулов, М.Г. Виноградова

Тверской государственный университет
кафедра физической химии

Представлен новый подход к построению количественных зависимостей “структура-свойство” для разнородных выборок органических соединений.
***Ключевые слова**: количественные модели, органические соединения, дескрипторы, QSPR-модели.*

Важной проблемой в современной теоретической химии является определение свойств химических соединений исходя из их молекулярной структуры. Повышенный интерес к проблеме «структура – свойство» обусловлен наличием большого количества синтезированных к настоящему времени веществ, а также широкого спектра возможностей их применения. Решение данной задачи, даже в рамках одного класса соединений, позволит прогнозировать свойства гипотетических молекул, вести синтез новых соединений с заданными свойствами. Поиск закономерностей в характере изменения различных свойств молекул в зависимости от их строения является одной из важнейших задач теоретической химии. Найденные закономерности можно использовать для прогнозирования свойств химических соединений, поиска новых соединений с заданным набором свойств, а также для систематизации молекул определенного класса. Несмотря на наличие в литературе достаточно большого числа различных количественных зависимостей «структура – свойство» (QSPR-моделей), проблема построения QSPR-моделей, описывающих свойства разнородной выборки органических веществ, по-прежнему представляется не вполне решенной. Четкие методологические правила построения подобных моделей в явном виде до сих пор не сформулированы, кроме того, в таких моделях часто используются параметры (дескрипторы), требующие достаточно сложных вычислений, но не дающие значительного улучшения статистических характеристик модели по сравнению с более простыми дескрипторами. Таким образом, исследование проблемы моделирования связи “«структура – свойство» и развитие методологии построения QSPR-моделей для разнородных выборок органических соединений представляется весьма важной и актуальной задачей. [1, 2].

Несмотря на наличие в литературе примеров построения QSPR-моделей для неоднородных выборок соединений, существует

устойчивое мнение, что добротные QSPR-модели могут быть построены только для выборки структурно и функционально схожих соединений. С целью исследования данной проблемы была рассмотрена база данных (БД) по нормальным температурам кипения (*T*кип) алканов (72 соединений) и по теплотам образования (H = ΔHº) алканов (47 соединений).

Для статистических расчетов использован програмный пакет Fathom Dynamic Data Software, Microsoft Excel и ActivStat. В работе использованы теоретические значения физико-химических характеристик структур найденные в технических публикациях. Построены корреляционные зависимости между структурными дескрипторами и физико-химическими параметрами для всего набора исследуемых соединений. Исследования показали, что топологические индексы в зависимости от способа расчета и их порядка по-разному коррелируют с физико-химическими характеристиками веществ. QSPR-модели построены с учетом индекса Рандича, индекса Винера и индекса Балабана.

Индекс Винера определен как полусумма топологических расстояний между всеми N атомами в молекулярном графе и рассчитывался по следующей формуле :

$$W = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij}$$

где D_{ij} —это *i*-й *j* -й элемент матрицы расстояний, который показывает наикратчайшее расстояние между вершинами *i* и *j* в графе *G*. Элементы матрицы определены по формулам:

$$d_{ij} = 1 - \frac{6}{z_i} \quad \text{и} \quad d_{ij} = \sum \frac{1}{b}\frac{36}{z_i z_j}$$

где z_i и z_j — заряд ядра (числа всех электронов) атомов *i* и *j*, соединенных данной связью; *b* — величина, характеризующая порядок (кратность) связи.

Индекс связности Рандича представляет собой математически закодированную информацию о числе атомов в молекуле, их связи между собой, о степени разветвления молекулы и может быть рассчитан для различных уровней связанности атомов молекулы между собой. Индекс связанности $\chi = \chi(G)$ графа *G* определяется как

$$\chi = \sum \left(\delta_i \cdot \delta_j\right)^{-1/2},$$

где δ_i и δ_j —валентности вершин *i* и *j* в графе *G*. Они соответствуют связям, соединяющим атомы *i* и *j*, и отображают состав графа. Суммирование проводится по всем ребрам графа G. Валентные вершины для молекул, содержащих ненасыщенные атомы углерода и гетероатомы (N, S , O и др.) δ^v, определяются по формуле

$$\delta^v = Z_i^v - h_i ,$$

где Z_i^v — число валентных электронов атома i, h_i — число связанных с ним атомов водорода. Данная величина несет информацию, относящуюся как к объемным, так и к электронным характеристикам.

Индекс Платта *F(G)* введен в 1947 г. и определен как сумма числа связей, смежных с каждой из связей в молекуле или в другой химической частице. Он равен сумме степеней каждого ребра в графе *G*, что отражается формулой:

$$F(G) = \sum_{f=1}^{f_{полн}} \deg e_f ,$$

где $\deg e_f$ - число ребер, смежных с ребром *f*, и $f_{полн}$ – полное число ребер в графе *G*.

Балабан предложил обладающий высокой дискриминирующей способностью топологический индекс, основанный на матрице расстояний молекулярного (химического) графа. Этот молекулярный дискриптор – индекс Балабана *J* – широко используется при исследованиях количественных корреляций структура – свойство или структура – активность. Индекс Балабана рассчитывается как средняя сумма расстояний связанности по формуле

$$J = \frac{b}{\mu + 1} \sum\nolimits_{i-j} (\delta_i \delta_j)^{-1/2}$$

где *b* представляет количество связей в графе *G*, μ – цикломатическое число графа *G*, которое указывает число независимых циклов в графе и равно минимальному числу разрезов (удалений ребер), необходимых для превращения графа *G* в дерево (ациклический граф). Оно может быть рассчитано с помощью соотношения

$$\mu = q - N + 1,$$

где N – число атомов в молекуле, за исключением водородных, тогда как q – число смежностей в молекуле. Индекс δ_i является суммой элементов строки в матрице расстояний. Этот индекс – мера связности атома *i* со всей остальной частью молекулы.

Коэффициент липофильности определяется при помощи стандартной системы 1-октанол – вода. Логарифм коэффициента распределения незаряженных форм субстрата стандартно обозначается logP и вычисляется по следующей формуле:

$$\log P_{oct/wat} = \log\left(\frac{[solut]_{octanol}}{[solute]_{water}}\right).$$

Для исследований физико-химических характеристик циклоалканов был использован индекс Гутмана, вычисляемый по следующей формуле:

$$Sz_{ij}(B) = \sum_{C} |C| n_1(C) \cdot n_2(C),$$

где $n_1(C)$ и $n_2(C)$ включает ребра с обоих сторон от атома углерода. Милан Рандич изучал корреляцию этого индекса с физико-химическими свойствами циклоалканов и предложил модификацию для колец с нечетным количеством атомов углерода, что значительно улучшает результаты. Индекс Гутмана, Szeged (Sz) индекс, учитывает количество атомов, находящихся ближе к рассматриваемой связи с каждой стороны. Рандич предложил для связей равноудаленных от группы атомов поровну разделить это количество атомов.

В работе были построены линейно-регрессионные модели «структура-свойство» вида *Ткип = aD + b* и *H = aD + b*. QSPR-модели, полученные для алканов, описываются следующими уравнениями и статистическими параметрами:

Ткип = 57,6χ – 97,303, R^2 = 0,9658, r = 0,98, s = 8,96 ºС;
Ткип = 1,2484W + 19,428, R^2 = 0,8698, r = 0,93, s = 21,4 ºС;
Ткип = 55,337ln(W) – 121,17, R^2 = 0,9531, r = 0,98, s = 8,8 ºС;
Ткип = 50,188J – 55,264, R^2 = 0,3338, r = 0,58, s = 32,88 ºС;
H = -40,882χ – 73,484, R^2 = 0,8415, r = 0,96, s = 16,5 кДж/моль;
$H = -80W^{0,2382}$, R^2 = 0,9106; r = 0,95, s = 9,8 кДж/моль;
H = -42,52J – 69,439, R^2 = 0,5473; r = 0,74, s = 19,68 кДж/моль

(R^2 – коэффициент детерминации, r – коэффициент корреляции, s – среднеквадратичное стандартное отклонение).

При учете индекса Рандича и логарифма индекса Винера статистические параметры имеют близкие значения. В случае использования индекса Балабана эти характеристики низки, что свидетельствует о значительных расхождениях в расчетных и экспериментальных величинах температуры кипения алканов.

Нельзя не отметить существенного вклада индекса Рандича в улучшение статистических параметров и общего вида корреляционной зависимости. Можно заключить, что индекс Рандича является наиболее перспективным дескриптором в расчете температуры кипения алканов.

Полученные зависимости (между экспериментальными и предсказанными значениями свойства) представлены на рис. 1 и 2.

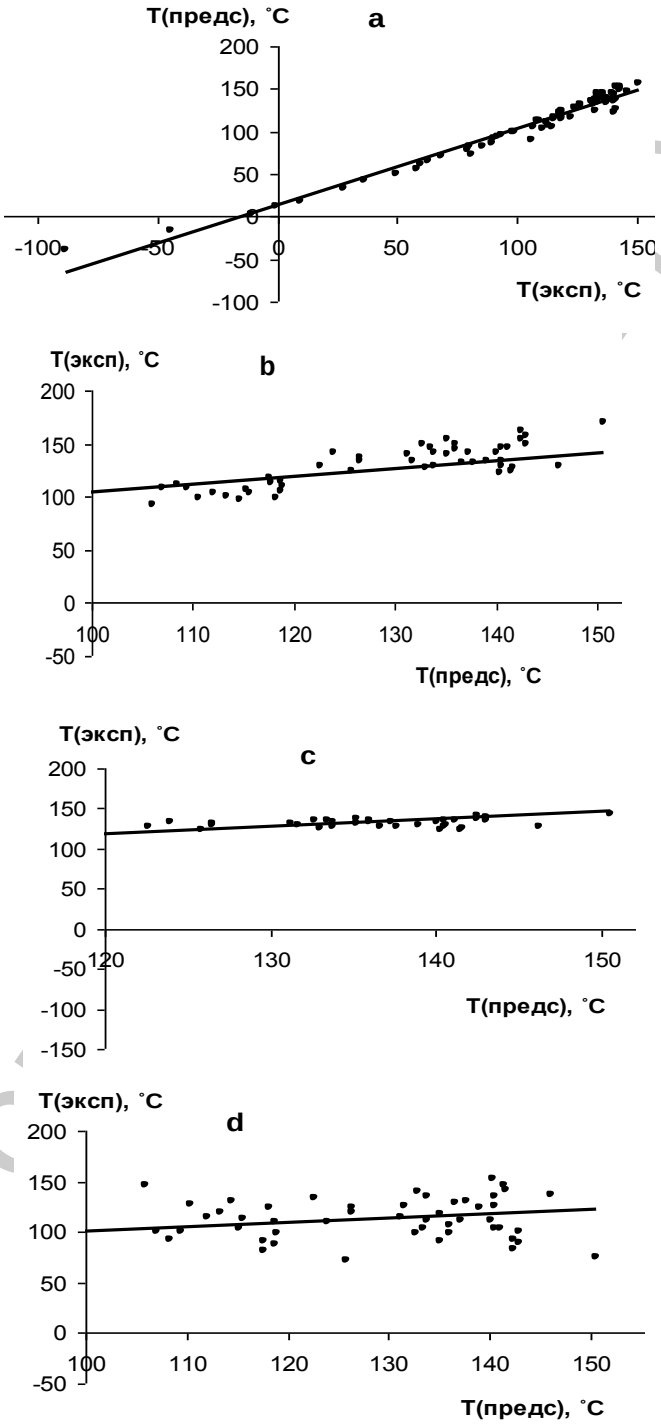
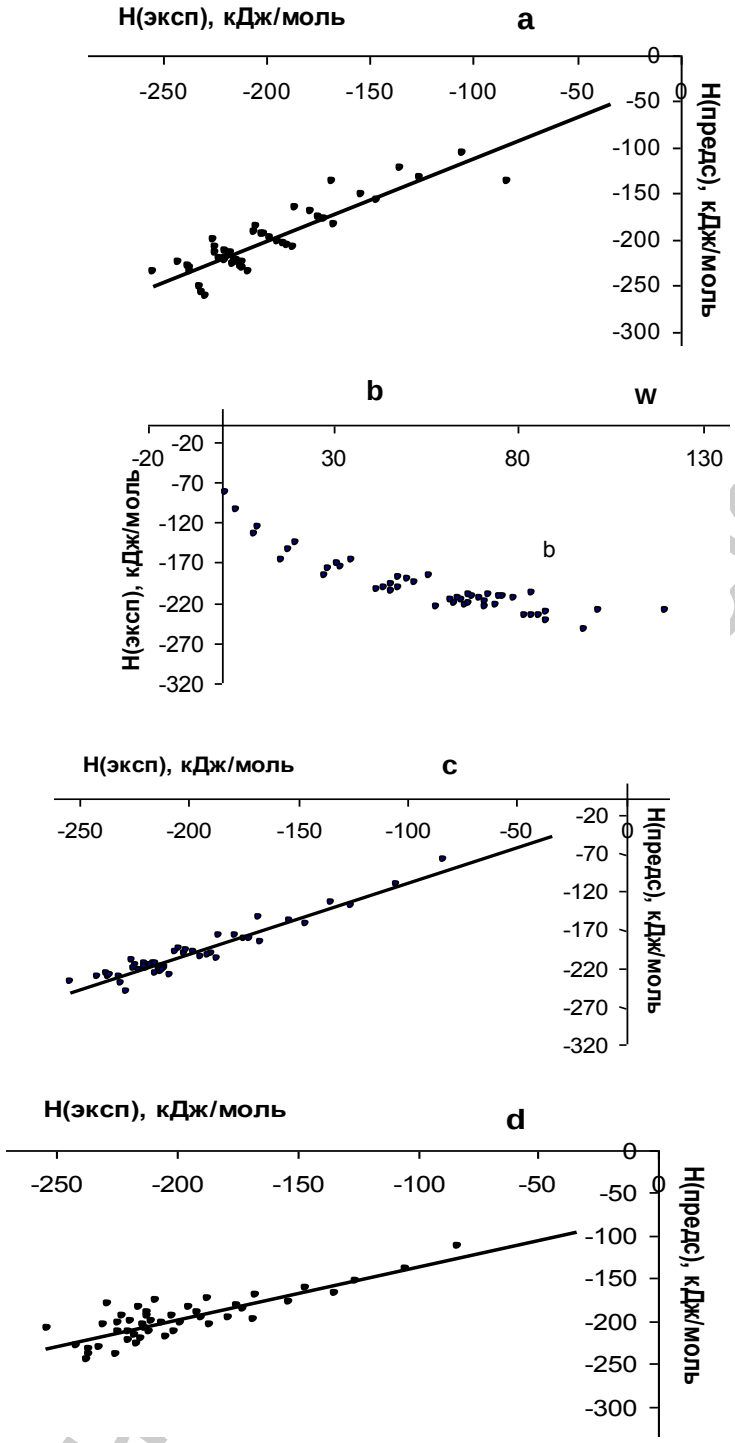


Рис. 1. QSPR-зависимости, построенные для *Ткип* алканов a) с учетом индекса Рандича; b) с учетом индекса Винера; c) с учетом логарифма индекса Винера; d) с учетом индекса Балабана



Р и с . 2. QSPR-зависимости, построенные для *H* алканов a) с учетом индекса Рандича; b) с учетом индекса Винера; c) как показательная функция с учетом индекса Винера; d) с учетом индекса Балабана

Наилучшие статистические характеристики демонстрируют QSPR-модели, построенные при использовании индекса Рандича и индекса Винера. Индекс Рандича ведет к более высокому значению коэффициента корреляции, а индекс Винера, выступающий параметром в показательной функции для вычисления теплоты образования алканов, уменьшает стандартное отклонение. Вторая модель прогнозирует величины теплоты образования ближе к их экспериментальным значениям.

Рассмотрим применимость предложенного подхода к построению количественных моделей «структура–свойство» для циклоалканов. Составленная база данных включает 24 соединения. Были построены следующие QSPR-модели для прогнозирования температуры кипения и теплоты образования циклоалканов:

Ткип = 62,661χ – 110,18, R^2 = 0,9716, r = 0,99, s = 9,36 ˚C;

Ткип = 0,8322RW + 22,448, R^2 = 0,9169, r = 0,96, s = 85,8 ˚C;

Ткип = 69,223ln(RW) – 190,6, R^2 = 0,9844, r = 0,99, s = 9,02 ˚C;

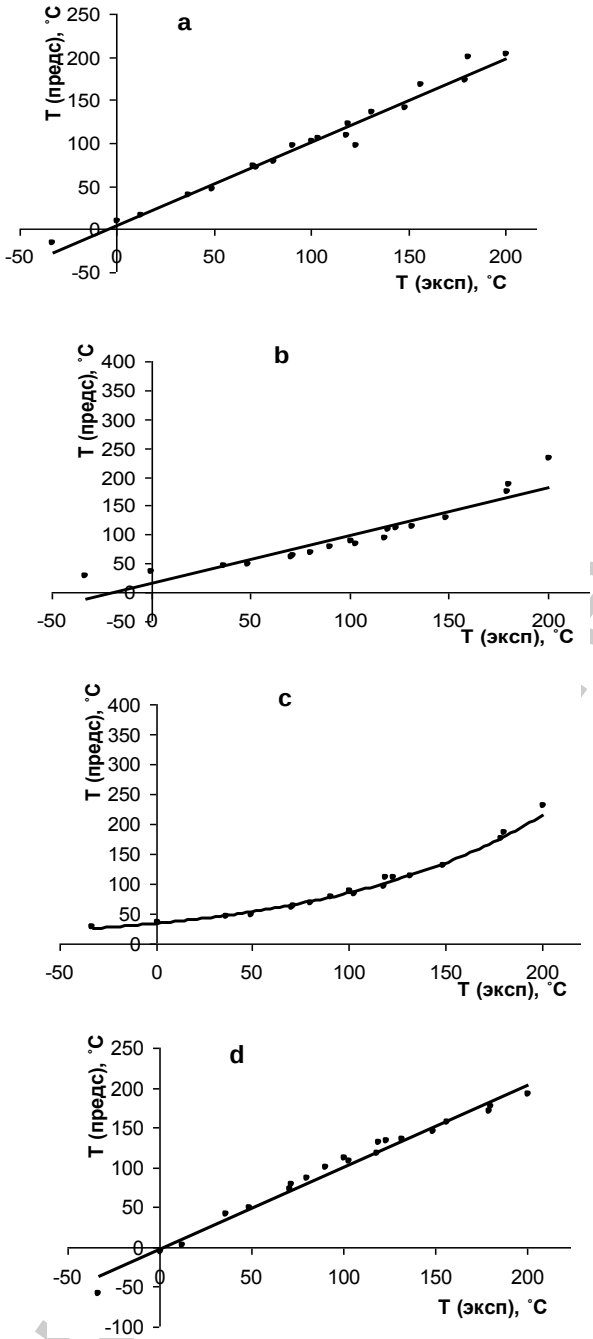
Н (экспер) = -57,508 χ + 82,462, R^2 = 0,7576, r = 0,87, s = 66,27 кДж/моль;

Н (экспер) = -0.5431RW - 72.882, R^2 = 0.6072, r = 0,78, s = 59,48 кДж/моль.

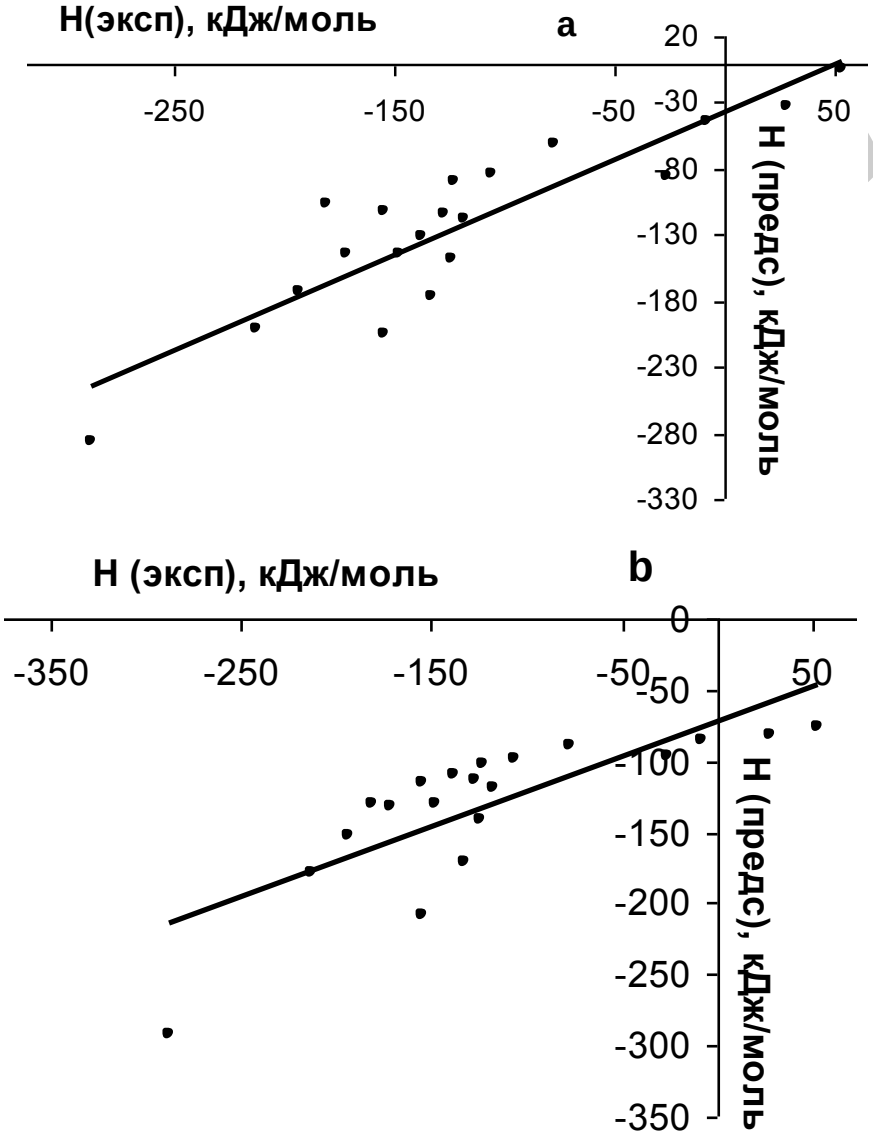
Модель «структура–свойство» для нормальной температуры кипения циклоалканов, построенная с применение индекса Рандича, имеет высокие статистические характеристики. Корреляция между температурой кипения соединений и модифицированным индексом Винера для нечетных колец и индексом Гутмана для четных колец (RW/Sz) значительно ниже первой. Наилучшая прогностическая способность принадлежит QSPR-модели, полученной с применением логарифма дескриптора RW/Sz.

Полученные QSPR-модели позволяют прогнозировать теплоту образования циклоалканов. Каждые примененный дескриптор имеет преимущество и ограничение. Индекс Рандича привел к более высокой корреляции. Рассчитанные величины теплоты образования циклоалканов с учетом модифицированного индекса Винера для нечетных колец и индекса Гутмана для четных колец более близки к экспериментальным данным. Об этом свидетельствует среднеквадратичное стандартное отклонение.

Полученные зависимости представлены на рис. 3 и 4.

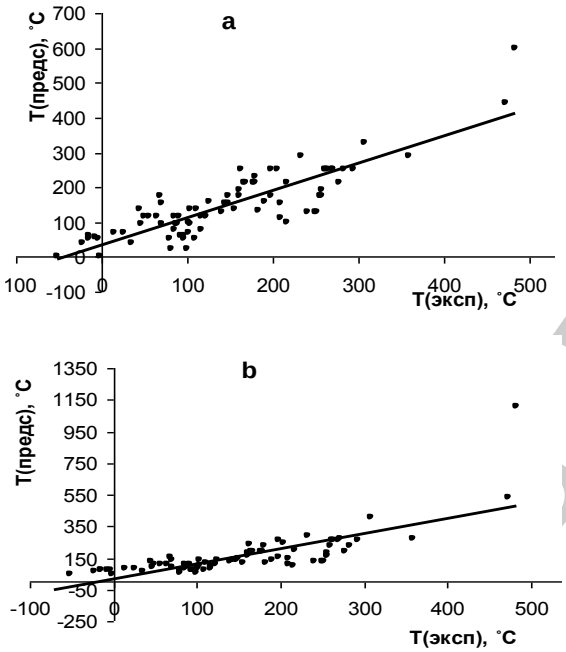


Р и с . 3. QSPR-зависимости, построенные для Ткип циклоалканов a) с учетом индекса Рандича; b) с учетом модифицированного индекса Винера для нечетных колец и индексом Гутмана для четных колец (RW/Sz) как линейная функция; c) с учетом дексриптора RW/Sz как показательная функция; d) с учетом логарифма дескриптора RW/Sz



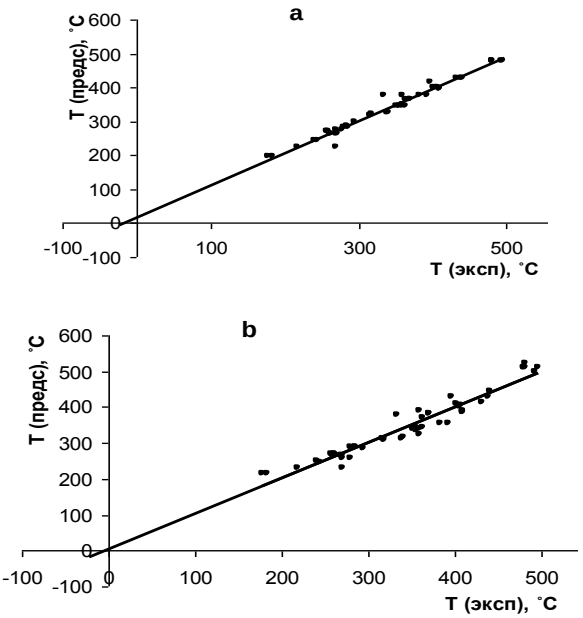
Р и с . 4. QSPR-зависимости, построенные для Н алканов а) с учетом индекса Рандича b) с учетом модифицированного индекса Винера для нечетных колец и индексом Гутмана для четных колец (RW/Sz)

Аналогичные модели структура–свойство для температуры кипения и температуры плавления построены для полициклических ароматических углеводородов с применением индекса Рандича и индекса Винера. Была создана база данных, включившая 82 соединения. Получены следующие модели (рис. 5).



Р и с. 5. QSPR-зависимости, построенные для Тпл полициклических ароматических углеводородов а) с учетом индекса Рандича в) с учетом индекса Винера

Результаты поиска зависимости теплоты кипения ПАУ отражены следующими диаграммами рис.6.



Р и с. 6. QSPR-зависимости, построенные для *Ткип* полициклических ароматических углеводородов а) с учетом индекса Рандича в) с учетом индекса Винера

Зависимости температуры плавления и температуры кипения ПАУ от коэффициента распределения представлены на рисунке 7.

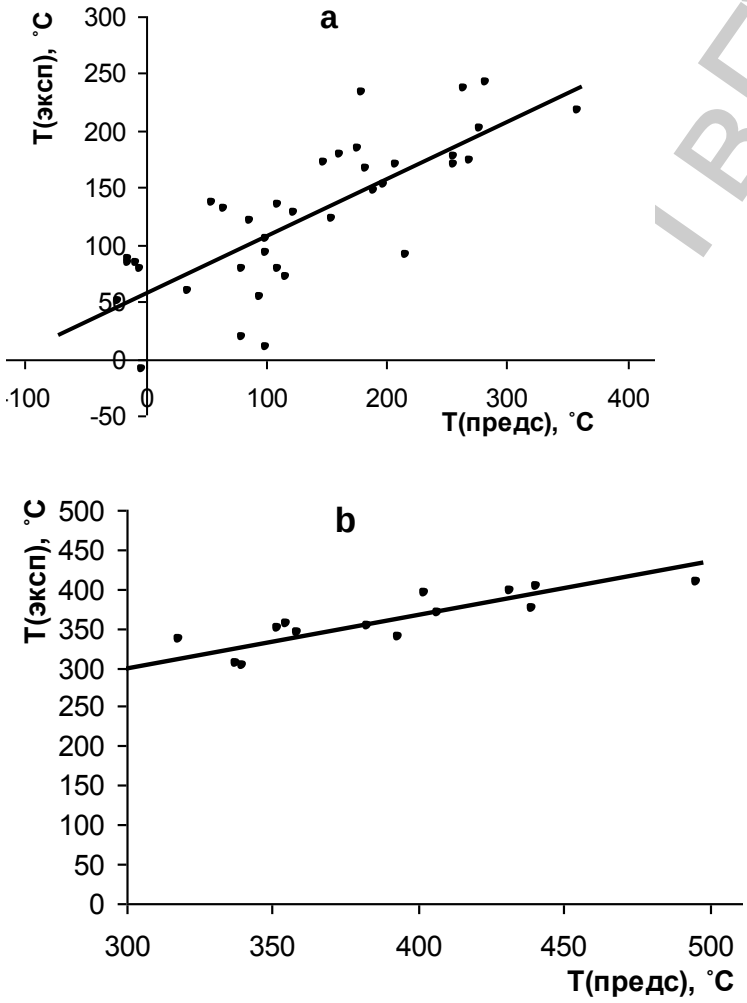


Рис. 7. QSPR-зависимости, построенные для а) *Тпл* и б) *Ткип* полициклических ароматических углеводородов с учетом коэффициента распределения в системе н-октанол – вода.

Таким образом, разработан новый подход к построению количественных зависимостей «структура–свойство» для разнородных выборок органических соединений. Он включает:

1) разделение общей выборки на группы структурно родственных соединений;

2) поиск «ведущего» дескриптора, позволяющего получить наилучшие однопараметровые QSPR-зависимости для каждой отдельной группы соединений, включая линейную, показательную и степенную;

3) фильтрацию незначимых и взаимно коррелирующих дескрипторов;

4) проверку прогностической способности модели.

Предложенный подход успешно применен для построения количественных линейно-регрессионных моделей «структура–свойство» для нормальной температуры кипения, нормальной температуры плавления и абсолютной энергии образования.

Построенные количественные зависимости «структура–свойство» позволяют достаточно быстро и с высокой степенью надежности оценивать значения рассматриваемых свойств органических соединений. Это представляется особенно полезным в тех случаях, когда необходимо оценить значения того или иного свойства сразу для достаточно большого числа соединений.

Предложенные в работе модели могут быть использованы для дизайна новых химических структур с заданными свойствами, что является актуальной проблемой современной химической науки.

Разработанные методологические основы построения QSPR-моделей для разнородных выборок органических соединений могут быть использованы при создании новых эффективных алгоритмов для исследования количественных соотношений «структура–свойство» и «структура–активность».

Список литературы

1. Руврэ Д.Г. Химию прогнозирует топология: пер. с англ. // В мире науки (Scientific American). 1986. №11. С. 14–22.
2. Виноградова М.Г., Папулов Ю.Г., Смоляков В.М. Количественные корреляции «структура–свойство» алканов. Аддитивные схемы расчета. Тверь:ТвГУ, 1999. 96 с.

QUANTITATIVE MODELS IN CORRELATION "STRUCTURE - PROPERTY" ORGANIC COMPOUNDS

Yu.A. Fedina, Yu.G. Papulov, M.G. Vinogradova

Tver State University
Department of physical chemistry

A new approach to the quantitative relationships "structure-property" for heterogeneous samples of organic compounds .
Keywords:. *quantitative models, organic compounds, descriptors, QSPR-model.*

Сведения об авторах :

ФЕДИНА Юлия Алексеевна – соискатель кафедры физической химии ТвГУ, e-mail: fedina_yuliya@yahoo.com

ПАПУЛОВ Юрий Григорьевич – профессор, доктор химических наук, зав. кафедры физической химии ТвГУ, e-mail: papulov_yu@mail.ru

ВИНОГРАДОВА Марина Геннадьевна – доктор химических наук, профессор кафедры физической химии ТвГУ, e-mail: mgvinog@mail.ru