УДК 004:519:544.16

МОДЕЛИ СВЯЗИ «СТРУКТУРА – СВОЙСТВО» ДЛЯ УГЛЕВОДОРОДОВ НА ОСНОВЕ СОБСТВЕННЫХ ЧИСЕЛ МОЛЕКУЛЯРНЫХ ГРАФОВ

Медведев Д.Ю., Скворцова М.И., Соломонова Е.В.

ФГБОУ «МИРЭА – Российский технологический университет» (Институт тонких химических технологий имени М.В. Ломоносова), Москва, e-mail: skvorivan@mail.ru

Показано, что для ряда физико-химических свойств алканов (температуры кипения, молярного объема, молярной рефракции, теплоты парообразования, критической температуры, критического давления, поверхностного натяжения) могут быть построены достаточно точные модели связи «структура – свойство», в которых в качестве молекулярных параметров используются исключительно инварианты спектрального типа, вычисляемые для соответствующих молекулярных графов. Эти инварианты задаются при помощи некоторых симметричных функций от собственных чисел определенных матриц графа. В качестве таких матриц рассмотрены, в частности, матрицы смежности, расстояний, Кирхгофа и др. Для матриц каждого типа построены как линейные, так и нелинейные модели. Установлено, что наилучший результат при таком моделировании дает матрица смежности. На основе статистического анализа частот встречаемости различных инвариантов в построенных корреляциях выявлены наиболее «популярные» параметры и предложено качественное объяснение этим фактам. Кроме того, показано, что использование для построения корреляций одновременно всех инвариантов всех рассмотренных матриц позволяет улучшить результаты, получаемые для каждой матрицы отдельно. Предлагаемый подход к построению корреляций «структура – свойство» допускает обобщение путем расширения перечня как используемых спектральных инвариантов, так и матриц молекулярных графов. Разработанная методика моделирования связи «структура – свойство» может быть применена к органическим соединениям любого класса и любым свойствам, измеряемым количественно.

Ключевые слова: молекулярный граф, инварианты графа, собственные числа графа, модели связи «структура-свойство», QSAR/QSPR, алканы

STRUCTURE-PROPERTY RELATIONSHIP MODELS FOR HYDROCARBONS BASED ON EIGENVALUES OF MOLECULAR GRAPHS

Medvedev D.Yu., Skvortsova M.I., Solomonova E.V.

MIREA – Russian Technological University (M.V. Lomonosov Institute of Fine Chemical Technologies), Moscow, e-mail: skvorivan@mail.ru

It has been shown that for a number of physicochemical properties of alkanes (boiling point, molar volume, molar refraction, heat of vaporization, critical temperature, critical pressure, surface tension), sufficiently accurate models of the "structure-property" relationship can be constructed, in which as molecular parameters, only spectral-type invariants are used, calculated for the corresponding molecular graphs. These invariants are defined by some symmetric functions of the eigenvalues of certain graph matrices. As such matrices, in particular, matrices of adjacency, distances, Kirchhoff, etc. are considered. Both linear and nonlinear models are constructed for matrices of each type. It has been established that the adjacency matrix gives the best result in such modeling. Based on a statistical analysis of the frequencies of occurrence of various invariants in the constructed correlations, the most "popular" parameters were identified and a qualitative explanation for these facts was proposed. In addition, it is shown that the use of all invariants of all considered matrices simultaneously for constructing correlations makes it possible to improve the results obtained for each matrix separately. The proposed approach to constructing "structure-property" correlations can be generalized by expanding the list of both spectral invariants used and molecular graph matrices. The developed technique for modeling the "structure-property" relationship can be applied to organic compounds of any class and any properties that can be measured quantitatively.

Keywords: molecular graph, graph invariants, graph eigenvalues, structure-property relationships, QSAR/QSPR, alkanes

Проблема моделирования связи между структурой и свойствами химических соединений – это важнейшая математическая задача современной теоретической и компьютерной химии [1-3]. Основная цель построения моделей связи «структура – свойство» – прогнозирование свойств химических соединений непосредственно по их структуре, при помощи соответствующих расчетов, минуя эксперимент. Результаты таких расчетов могут быть использованы для целенаправленного поиска соединений с заданными свойствами. Для моделей связи «структура – свойство», имеющих вид корреляционных уравнений, в литературе часто используется аббревиатура QSPR/QSAR (Quantitative Structure-Property/Activity Relationships), в зависимости от того, какое свойство соединений рассматривается: физико-химическое или какой-либо вид биологической активности.

Следует отметить, что экспериментальное определение различных свойств соединений с целью поиска соединений

с заданным набором свойств во многих случаях является технически сложным и, кроме того, требует определенных финансовых и временных затрат. В связи с этим разработка, обоснование и тестирование различных математических методов моделирования связи между структурой и свойствами химических соединений является актуальной задачей.

Одним из наиболее распространенных подходов к моделированию связи «структура – свойство», приводящим к моделям, имеющим вид корреляционных уравнений, является так называемый статистический подход. Он заключается в следующем. Имеется набор химических структур $\{S_i\}$ (структурных формул молекул) с известными значениями $\{y_i\}$ некоторого свойства (физико-химического или биологической активности), $i = 1, ..., N$. Требуется, анализируя эти данные, выявить зависимость свойства y от структуры S, т.е. найти функцию вида $y = f(S)$, определяющую модель связи «структура – свойство».

Как правило, в этих исследованиях для описания структуры молекул используются взвешенные молекулярные графы (МГ), которые строятся по структурным формулам соответствующих молекул. Вершины этих графов соответствуют атомам молекулы (возможно, группе атомов), а ребра – химическим связям между ними. Числовые веса вершин и ребер кодируют атомы и связи различных типов; кроме того, могут быть приписаны некоторые веса и парам не связанных вершин графа. МГ задается своей матрицей весов. Следует отметить, что по первоначально определенной матрице весов МГ можно также построить много других матриц, модифицируя определенным образом элементы этой исходной матрицы.

Для количественной характеристики структуры молекул используются инварианты $x_1, ..., x_n$ соответствующего МГ (т.е. числа, вычисляемые по графу или его матрице, не зависящие от способа нумерации его вершин). В этом случае модель связи «структура – свойство» приобретает вид уравнения: $y = f(x_1, ..., x_n)$.

Таким образом, каждой исходной химической структуре S ставится в соответствие по некоторому правилу граф (или его матрица M), а этому графу, в свою очередь, сопоставляется некоторый набор инвариантов $x_1, ..., x_n$, вычисляемых по определенным алгоритмам.

Далее предполагается, что функция многих переменных $f(x_1, ..., x_n)$ в уравнении связи «структура – свойство» имеет специальный вид (например, она является линейной или квадратичной), но зависит от ряда подгоночных параметров. Эти параметры определяются по исходным данным так, чтобы получаемое уравнение было бы как можно более точным на исходном наборе химических соединений.

Следует отметить, что при моделировании связи «структура – свойство» возникают проблемы выбора способа построения МГ, инвариантов МГ, а также аппроксимирующей функции f из бесконечного множества возможных вариантов. Однако этот выбор очень важен, так как от него зависит конечный результат моделирования.

Точность построенной модели можно оценивать разными способами (см., например, [4]). Например, можно использовать для этой цели такие статистические параметры соответствующего регрессионного уравнения, как коэффициент корреляции (R) и среднеквадратичное отклонение (s), максимальная и средняя относительные ошибки ($\delta_{макс}$ и $\delta_{ср}$) расчета свойств исходной выборки (в %) при помощи полученного уравнения и т.д. При этом необходимо задать пороговые значения этих параметров, на основе которых модель будет признана «приемлемой» или нет. Эти пороговые значения зависят от конкретной задачи. Например, для физико-химических свойств обычно считается, что модель «хорошая», если $R \geq 0{,}95$ или если $\delta_{макс} \leq 5\%$. Например, в [5] предложены следующие характеристики качества модели по величине коэффициента корреляции: $R \geq 0{,}990$ (outstanding), $R \geq 0{,}975$ (excellent), $R \geq 0{,}950$ (very good), $R \geq 0{,}925$ (good), $R \geq 0{,}900$ (fair).

Одним из важных видов инвариантов графов являются инварианты спектрального типа, т.е. построенные на основе каких-либо спектральных характеристик матрицы графа [6; 7]. Основными спектральными характеристиками графа с n вершинами являются его собственные числа $\lambda_1 \leq ... \leq \lambda_n$ (или коэффициенты характеристического полинома, которые выражаются через собственные числа), а также соответствующие собственные векторы.

Собственные числа графа несут в себе довольно большую информацию о структуре графа. Одно время даже считалось, что простой граф однозначно определяется по набору собственных чисел его матрицы смежности, но потом были найдены примеры, показывающие, что это не так. Как известно, матрица графа (как и любая матрица) однозначно восстанавливается по набору ее собственных чисел и соответствующих собственных векторов.

Инварианты спектрального типа широко используются в теории электронного строения сопряженных молекул. Примерами таких параметров являются: сумма положительных собственных чисел или их число, наименьшее положительное и наибольшее отрицательное собственное число и т.д. [6]. Кроме того, отдельные спектральные инварианты находят применение и в корреляциях «структура – свойство», совместно с молекулярными параметрами других видов [7; 8].

Цель исследования – построить и проанализировать регрессионные модели связи «структура – свойство» для разнообразных физико-химических свойств алканов, в которых используются инварианты МГ, основанные исключительно на собственных числах различных матриц этих МГ.

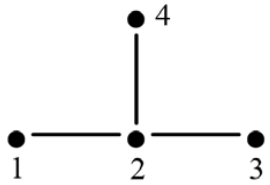
Исходные данные для построения корреляций

В качестве исходных данных для построения моделей связи «структура – свойство» была рассмотрена выборка структурных формул алканов с числом атомов углерода от 6 до 8 (всего $N = 32$ соединения), с известными значениями следующих физико-химических свойств: 1) температура кипения (bp – boiling point); 2) молярный объем (MV – molar volume at 20 °C); 3) молярная рефракция (MR – molar refraction at 20 °C); 4) теплота парообразования (HV – heat of vaporization at 25 °C); 5) критическая температура (TC – critical temperature); 6) критическое давление (PC – critical pressure); 7) поверхностное натяжение (ST – surface tension at 20 °C) [9].

Способы представления химических структур

Первоначально были построены простые МГ углеродных остовов алканов. Вершины этих МГ соответствуют атомам углерода, а ребра – химическим связям между ними. Затем для таких МГ были построены следующие простейшие матрицы: $M_1 = A$ (матрица смежности), $M_2 = D$ (матрица расстояний), $M_3 = V - M_1$ (матрица Кирхгофа), $M_4 = V+M_2$. Здесь V – матрица, диагональные элементы которой равны степеням соответствующих вершин МГ, а остальные элементы равны нулю. На рисунке приведен пример простого МГ углеродного остова 2-метилпропана и его матриц $M_1 – M_4$.

Заметим, что рассмотрение разных матриц одного простого МГ равносильно построению разных МГ, с разными весами вершин, ребер и пар несвязанных вершин.



Молекулярный граф 2-метилпропана

$$M_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix},$$

$$M_4 = \begin{pmatrix} 1 & 1 & 2 & 2 \\ 1 & 3 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ 2 & 1 & 2 & 1 \end{pmatrix}$$

Простой молекулярный граф 2-метилпропана и его матрицы M_1-M_4

Инварианты, используемые для построения моделей связи «структура – свойство»

Пусть n – число вершин МГ, $\lambda_1 \le ... \le \lambda_n$ – собственные числа какой-либо матрицы этого МГ (в данном случае – M_1-M_4). В работе рассматривались следующие инварианты спектрального типа, построенные при помощи некоторых простейших симметричных функций многих переменных и вычисленные для каждой из матриц M_1-M_4:

$$x_1 = f_1(\lambda_1, ..., \lambda_n) = max\ \lambda_i;$$

$$x_2 = f_2(\lambda_1, ..., \lambda_n) = min\ \lambda_i;$$

$$x_3 = f_3(\lambda_1, ..., \lambda_n) = \sum|\lambda_i|^2;$$

$$x_4 = f_4(\lambda_1, ..., \lambda_n) = \sum|\lambda_i|^3;$$

$$x_5 = f_5(\lambda_1, \ldots, \lambda_n) = \sum \lambda_i,\ \lambda_i > 0;$$

$$x_6 = f_6(\lambda_1, \ldots, \lambda_n) = \sum (1/\lambda_i),\ \lambda_i > 0;$$

$$x_7 = f_7(\lambda_1, \ldots, \lambda_n) = (1/n) \times \sum e^{\lambda i};$$

$$x_8 = f_8(\lambda_1, \ldots, \lambda_n) = \prod \lambda_i.$$

Отметим, что выбор этих симметричных функций был достаточно формальным и произвольным; он никак не связан ни с рассматриваемыми свойствами, ни с классом химических соединений. Кроме x_1-x_8, был рассмотрен также параметр $x_9 = n$.

Следует подчеркнуть, что существует бесконечно много инвариантов МГ, которые могут использоваться в моделях связи «структура – свойство», и эти инварианты могут быть разных типов (см., например, [6]). Заранее неизвестно, от каких именно инвариантов и каким образом зависит изучаемое свойство для данного класса соединений. В связи с этим возникает проблема выбора конечного числа инвариантов МГ из бесконечного числа возможных вариантов. В нашей работе мы рассматриваем инварианты только одного типа – спектрального, что позволяет уменьшить степень неопределенности при выборе молекулярных параметров и алгоритмизировать процесс построения моделей.

Алгоритм построения корреляций «структура – свойство»

Для всех семи физико-химических свойств и отдельно для каждой матрицы M_1-M_4 были построены как линейные, так и нелинейные корреляции специального вида. Опишем в общем виде алгоритм, использованный нами для этих целей.

Пусть имеется выборка, состоящая из N химических структур с известными численными значениями y_i (i=1,…,N) их некоторого свойства y (физико-химического или биологической активности) и известными численными значениями каких-либо их m структурных параметров $x_1,x_2,\ldots,x_m$. Задача заключается в построении корреляционного уравнения вида $y=f(x_1,x_2,\ldots,x_m)$ (где функция f – это некоторый многочлен от переменных $x_1,x_2,\ldots,x_m$), удовлетворяющего определенным заданным требованиям. При этом, возможно, некоторые переменные из набора $x_1,x_2,\ldots,x_m$ не будут входить в это уравнение; степень многочлена f может быть произвольной. Очевидно, что получаемое уравнение является линейным относительно выражений, являющихся произведением исходных переменных, взятых в некоторых целочисленных неотрицательных степенях, и поэтому оно определяет линейную регрессию, соответствующую исходной нелинейной регрессии.

Шаг 1. Назовем набор параметров $x_1,x_2,\ldots,x_m$ *исходным набором*. Выберем некоторое число k ($1 \le k \le m$). Из исходного набора параметров методом пошаговой линейной регрессии отберем k параметров $x_{i1},\ldots,x_{ik}$, дающих наилучшую линейную корреляцию для свойства y (т.е. с наибольшим коэффициентом корреляции), и построим ее.

Параметры $x_{i1},\ldots,x_{ik}$, вошедшие в полученную корреляцию, назовем *базовыми*.

Если качество полученной модели устраивает исследователя, то работа алгоритма на этом заканчивается. Если необходимо построить более точную модель, то переходим к Шагу 2.

Шаг 2. Образуем *новое множество параметров*, состоящее из базовых параметров (см. Шаг 1), их квадратов и всевозможных попарных произведений. Очевидно, что общее число параметров в новом множестве будет равно $m_1=k+0.5(k^2+k)=0.5k^2+1{,}5k$.

Шаг 3. Идем на Шаг 1 и выполняем все действия, указанные в этом пункте, взяв в качестве исходного набора параметров новое множество параметров, сформированное на Шаге 2. Теперь в качестве m используется m_1.

Результатом работы алгоритма являются регрессионные уравнения, полученные на Шаге 1, с такими их статистическими характеристиками, как коэффициент корреляции, среднеквадратичное отклонение, максимальная абсолютная ошибка расчетов свойств химических соединений исходной выборки по построенному уравнению (с указанием той структуры, на которой она реализуется), средняя относительная ошибка (в %). Кроме того, для каждой структуры заданной выборки приводятся расчетное значение свойства y, а также соответствующие абсолютная и относительная ошибки.

Отметим, что результат работы алгоритма (т.е. получаемая модель) существенно зависит от последовательности чисел k, выбираемых на Шаге 1 (при многократном повторении Шага 1). В связи с этим для получения лучшего результата или разных результатов целесообразно поварьировать цепочку чисел k. Из получаемых различных моделей можно выбрать одну или несколько наилучших (по какому-либо критерию). Таким образом, процедура построения моделей носит итерационный характер и в определенной степени управляется исследователем.

Таблица 1

Коэффициенты корреляции R нелинейных моделей для матриц M_1-M_4

Физ.-хим. св-во / Матрица	bp	MV	MR	HV	TC	PC	ST
M_1	0,9907	0,9963	0,9998	0,9914	0,9835	0,9355	0,9724
M_2	0,9855	0,9930	0,9997	0,9790	0,9588	0,9105	0,9291
M_3	0,9857	0,9940	0,9998	0,9896	0,9710	0,9397	0,9424
M_4	0,9888	0,9955	0,9998	0,9888	0,9438	0,9105	0,8435

Вышеописанный алгоритм реализован нами в виде компьютерной программы на языке Java; при этом пользователь на первом этапе работы программы может выбрать некоторое подмножество множества параметров $x_1,x_2,\ldots,x_m$, с которым будет работать дальше.

При построении корреляций в настоящей работе в соответствии с описанным выше алгоритмом в качестве исходного набора использовались параметры x_1-x_9 (m=9), на Шаге 1 первоначально было выбрано k=3, на Шаге 2 новое множество параметров состояло из m_1=9 параметров, при повторном выполнении Шага 1 было выбрано k=3. На этом процесс построения моделей заканчивался. Заметим, что похожая методика построения нелинейных корреляций использовалась нами ранее в работе [10].

Определение матрицы, дающей наилучшую модель

Далее была поставлена задача выбора *одной* матрицы графа из нескольких рассмотренных выше матриц, позволяющих получить наилучшую модель (как линейную, так и нелинейную). Наилучшей мы считаем модель с наибольшей величиной коэффициента корреляции R. При совпадении коэффициентов корреляции наилучшей считаем модель с наименьшим значением среднеквадратичного отклонения s. Соответствующую матрицу также назовем наилучшей.

В таблице 1 приведены значения коэффициентов корреляции (с четырьмя знаками после запятой) для нелинейных моделей для всех семи физико-химических свойств и всех матриц M_1-M_4. Такая точность в определении R необходима для максимальной дифференциации построенных моделей и выбора только одного варианта из нескольких возможных. При округлении R до трех знаков после запятой в ряде случаев различия между моделями по этому критерию исчезают.

На основе таблицы 1 можно сделать следующие выводы.

1. Наилучшую модель дает матрица M_1 (для свойств bp, MV, HV, TC, ST).

2. Для свойства MR матрицы M_1, M_3, M_4 дают одинаковый результат, но если учитывать s, то наилучшая – это M_4.

3. Наилучший результат для свойства PC дает M_3 (R = 0,9355). Однако в линейной корреляции наилучший результат дает M_1 (R = 0,9413). Отметим, что, используя матрицу M_1, в рамках предложенного алгоритма можно получить более точную модель, с большей степенью нелинейности относительно исходных молекулярных параметров, а именно:

$$PC = 61.3\text{-}6.38x_9\text{+}0.11(x_1)^2x_5x_9\text{-}0.02x_9\,(x_1)^2(x_8)^2$$

($\delta_{ср}$ = 1.42%, s = 0.634, R = 0.9515).

4. Все модели, отобранные как наилучшие, содержат 3 структурных параметра (если их рассматривать как линейные) и обладают очень высоким коэффициентом корреляции. Опираясь на классификацию моделей по коэффициенту корреляции, приведенную в [5], их можно охарактеризовать, по крайней мере, как «очень хорошие».

Частоты появления в корреляциях различных инвариантов

Весьма интересным представляется вопрос о том, какие из рассмотренных, формально определенных инвариантов наиболее часто встречаются в построенных корреляциях и с чем может быть связана их «популярность». Для ответа на этот вопрос были найдены суммарные частоты появления всех 9 рассмотренных инвариантов во всех 8 корреляциях (линейных и нелинейных, полученных отдельно для матриц M_1-M_4) для каждого свойства (табл. 2). Частота встречаемости инварианта в отдельной корреляции определялась как число его появлений в соответствующей формуле, с учетом степени этого инварианта. Например, для приведенной выше корреляции для PC частоты появления x_1, x_5, x_8, x_9 равны 4, 1, 2, 3 соответственно.

Таблица 2

Суммарные частоты появления параметров x_1-x_9 во всех 8 корреляциях

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
bp	1	4	4	1	5	7	1	0	6
MV	4	4	5	0	4	5	0	0	9
MR	3	4	2	2	2	5	0	0	12
HV	4	3	1	2	5	6	1	1	8
TC	5	3	3	2	3	6	0	2	7
PC	7	4	0	1	1	2	6	0	8
ST	2	2	1	3	3	10	5	2	6
Σ	26	24	16	11	23	41	13	5	56

Оказалось, что наиболее «популярные» 3 инварианта – это x_9, x_6, x_1: $x_9 = n$; $x_6 = \sum 1 / \lambda_i$, $\lambda_i > 0$; $x_1 = max\ \lambda_i$. Эти факты можно объяснить, например, следующим образом.

Параметр $x_9 = n$ может служить количественной характеристикой размера молекулы, а размер молекулы существенно влияет на ее свойства (например, чем больше размер молекул, тем больше энергии необходимо затратить, чтобы начался процесс кипения или испарения соответствующего вещества).

Параметр x_1 может служить мерой ветвления молекулярного графа-дерева (в случае матрицы смежности), так как при фиксированном n он принимает свои экстремальные значения на наиболее и наименее разветвленных графах (с интуитивной точки зрения) – графе-звезде и графе-цепи. Чем больше «ветвистость» МГ, тем сильнее молекулы «цепляются» друг за друга своими ветвями, тем больше энергия взаимодействия молекул и тем выше, например, температура кипения.

Нами установлено, что параметр $1/x_6$ (для матрицы смежности) совместно с инвариантом P_3, равным числу цепочек длины 3 в МГ, хорошо коррелирует с плотностью ρ соответствующих веществ (коэффициент корреляции $R = 0{,}965$). Итак, x_6 связан определенным образом с плотностью веществ, а она, в свою очередь, существенно влияет на многие другие свойства алканов. Например, для ρ можно получить довольно точную двухпараметрическую корреляцию с P_3 и bp или с P_3 и HV ($R = 0{,}97$); при добавлении в уравнение третьего параметра n коэффициент корреляции увеличивается: $R = 0{,}99$.

Нелинейные корреляции на основе спектров всех четырех матриц

Кроме того, для всех рассматриваемых свойств были построены нелинейные модели, для построения которых использовались параметры x_1-x_8 не для какой-либо одной матрицы, как выше, а для всех четырех, а также x_9 (т.е. всего 33 исходных параметра). Назовем такие модели *обобщенными.*

Цель построения обобщенных моделей – выяснить: 1) можно ли таким образом улучшить результаты, полученные ранее для отдельных матриц; 2) будут ли использоваться в них все рассмотренные матрицы или только наилучшая, выявленная ранее.

Для вышеуказанных обобщенных моделей была проведена проверка их адекватности при помощи так называемой процедуры скользящего контроля (cross validation). Суть этой процедуры заключается в следующем: из исходной выборки, состоящей из N структур, удаляют одну структуру, затем строят новую модель на оставшихся структурах, далее при помощи этой модели рассчитывают величину свойства для удаленной структуры; эти действия выполняют для всех N структур, а затем строят корреляцию между экспериментальными и рассчитанными значениями изучаемого свойства. По статистическим характеристикам этой корреляции (например, по коэффициенту корреляции и среднеквадратичному отклонению) делают вывод о качестве тестируемой модели.

В таблице 3 представлены результаты построения и тестирования нелинейных обобщенных моделей для всех семи физико-химических свойств. В ней указаны параметры, вошедшие в корреляцию, а также соответствующие матрицы, по которым они вычислялись; кроме того, приведены значения некоторых статистических характеристик полученных моделей: $\delta_{ср}$ – средняя относительная ошибка (в %), s – среднеквадратичное отклонение, R – коэффициент корреляции. При этом параметр R указан с четырьмя знаками после запятой, как и в таблице 1, для того чтобы можно было сравнивать модели из таблиц 1 и 3 по этому параметру.

Таблица 3

Результаты построения и тестирования нелинейных обобщенных моделей

	Инварианты МГ, вошедшие в корреляцию	$\delta_{ср}$ (%)	s	R	$\delta_{ср,cv}$ (%)	s_{cv}	R_{cv}
bp	$x_5(M_1)x_3(M_1)$, $(x_6(M_2))^2$, $x_5(M_4)x_6(M_2)$	2.18	2.65	0.9928	2.60	2.99	0.990
MV	$x_1(M_4)x_4(M_1)$, $x_4(M_4)x_8(M_2)$, $x_3(M_3)x_6(M_3)$	0.40	0.79	0.9979	0.73	1.39	0.994
MR	x_9, $x_2(M_4)x_4(M_3)$, $x_5(M_1)x_6(M_2)$	0.10	0.05	0.9999	0.10	0.04	0.999
HV	$x_5(M_1)x_5(M_4)$, $x_8(M_2)x_4(M_2)$, $x_8(M_1)$ $x_8(M_4)$	0.83	0.40	0.9948	1.37	0.61	0.989
TC	$x_5(M_1)x_1(M_1)$, $x_6(M_1)x_7(M_3)$, $x_8(M_4)x_7(M_3)$	1.14	4.14	0.9876	1.47	4.86	0.981
PC	$x_5(M_1)x_6(M_2)$,$x_8(M_4)x_7^4(M_3)$, $x_6(M_2)x_6(M_1)$	1.03	0.41	0.9802	1.64	0.69	0.931
ST	$x_5(M_1)x_7(M_1)$, $x_2(M_4)x_4(M_3)$, $x_6(M_1)x_1(M_3)$	1.21	0.33	0.9785	1.65	0.44	0.956

Кроме того, в таблице 3 приведены аналогичные результаты для процедуры скользящего контроля, выполненной для полученных моделей: $\delta_{cp,cv}$ – средняя относительная ошибка (в %), s_{cv} – среднеквадратичное отклонение, R_{cv} – коэффициент корреляции.

Сравнивая модели в таблицах 1 и 3 по величине коэффициента корреляции R, можно сделать вывод, что для каждого физико-химического свойства полученная обобщенная модель является более точной, чем наилучшая из предыдущих восьми, построенных для каждой из матриц M_1-M_4 отдельно. Результаты скользящего контроля свидетельствуют о хорошей стабильности полученных моделей.

Кроме того, установлено, что для четырех свойств (bp, HV, TC, ST) в корреляциях используются 3 матрицы, а для трех (MV, MR, PC) – все 4 матрицы (из четырех). Таким образом, практически все рассмотренные матрицы оказались полезными при моделировании связи «структура – свойство». Заметим, что при этом параметр $x_9 = n$ вошел в корреляцию только для свойства MR.

Выводы

1. В работе предложен новый подход к моделированию связи между физико-химическими свойствами алканов и их структурными характеристиками. Одна из особенностей разработанной методики – это использование в качестве молекулярных параметров только инвариантов спектрального типа соответствующих МГ. Вторая особенность – применение гибкой итерационной процедуры построения аппроксимирующей функции в моделях связи «структура – свойство», которая является многочленом специального вида от исходных молекулярных параметров.

2. Показано, что для ряда физико-химических свойств алканов разработанный подход позволяет строить достаточно точные модели связи «структура – свойство». При сравнении результатов, получаемых для разных четырех видов матриц МГ, установлено, что наилучший результат дает матрица смежности МГ. Использование для построения корреляций одновременно всех инвариантов всех рассмотренных четырех матриц позволяет улучшить результаты, получаемые для каждой матрицы отдельно. При этом полезными оказываются все матрицы.

3. При проведении статистического анализа частоты встречаемости различных молекулярных параметров в построенных корреляциях (по отдельным матрицам) выявлены 3 наиболее «популярных» параметра и предложено качественное объяснение этим фактам.

4. Предлагаемый подход к построению корреляций «структура – свойство» допускает обобщение путем расширения перечня используемых спектральных инвариантов или/и рассмотрения дополнительно каких-либо других матриц МГ, что может способствовать увеличению его эффективности. Получаемые корреляции могут быть также улучшены за счет увеличения степени их нелинейности. Отметим также, что разработанная методика моделирования связи «структура – свойство» может быть применена к органическим соединениям любого класса и любым свойствам, измеряемым количественно.

Список литературы

1. Begam B.F., Kumar J.S. Computer Assisted QSAR/QSPR Approaches – A Review. Indian Journal of Science and Technology. 2016. Vol. 9. No. 8. P. 1-8. DOI: 10.17485/ijst/2016/v9i8/87901.

2. Gramatica P. Principles of QSAR Modeling: Comments and Suggestions From Personal Experience. International Journal of Quantitative Structure-Property Relationships. 2020. Vol. 5. No. 3. P. 61-97. DOI: 10.4018/IJQSPR.20200701.oa1.

3. Шулаева Н.А., Скворцова М.И., Михайлова Н.А. Модели связи «структура-свойство» органических соединений на основе молекулярных графов с элемента-

ми пространственного строения молекул // Тонкие химические технологии. 2020. Т. XV. № 6. С.84-103. DOI: 10.32362/2410-6593-2020-15-6-84-103.

4. Gajewicz A. How to judge whether QSAR/read-across predictions can be trusted: a novel approach for establishing a model's applicabitility domain. Environmental Sciences: Nano. 2018. Vol. 5. No. 2. P. 408-421. DOI: 10.1039/C7EN00774D.

5. Randić M. Comparative Regression Analysis. Regressions Based on a Single Descriptor. Croatica Chemica Acta. 1993. Vol. 66. No. 2. P. 289-312.

6. Станкевич М.И., Станкевич И.В., Зефиров Н.С. Топологические индексы в органической химии // Успехи химии. 1988. Т. 57. С. 337-366.

7. Gligorijević A., Marković S., Redžepović I., Furtula B. Application of spectral graph theory on the enthalpy change of formation of acyclic saturated ketones. Journal of the Serbian Chemical Society. 2018. Vol. 83. No.12. P. 1339-1349. DOI: 10.2298/JSC180906086G.

8. Skvortsova M. Molecular Graphs and Molecular Hypergraphs of Organic Compounds: Comparative Analysis. Journal of Medicinal and Chemical Sciences. 2021. Vol. 4. No. 5. P. 452-465. DOI:10.26655/JMCHEMSCI.2021.5.6.

9. Needham D.E., Wei I.-C., Seybold P.G. Molecular Modeling of the Physical Properties of the Alkanes. Journal of American Chemical Society. 1988. Vol. 110. P. 4186-4194.

10. Скворцова М.И., Соломонова Е.В., Ратнов А.Г. О некоторых методах построения нелинейных уравнений, связывающих структуру и свойства органических соединений // Современные наукоемкие технологии. 2020. № 10. С. 85-92. DOI: 10.17513/snt.38260.