УДК 541.6

МОДЕЛИРОВАНИЕ СВЯЗИ МЕЖДУ СТРУКТУРОЙ И ТЕМПЕРАТУРОЙ ПЛАВЛЕНИЯ ПРОИЗВОДНЫХ ПИРАЗОЛОНА: ТОПОЛОГИЧЕСКИЙ ПОДХОД

М.И. Скворцова, доцент, [*]Н.С. Рукк, доцент

кафедра Высшей и прикладной математики

[*]кафедра Неорганической химии

МИТХТ им. М.В. Ломоносова

e-mail: skvorivan@mail.ru

В рамках статистического подхода построена нелинейная модель связи «структура – температура плавления» для производных пиразолона. Для количественного описания молекулярной структуры использован ряд топологических параметров, вычисляемых непосредственно из структурных формул молекул. Приведены результаты использования построенной модели для расчета температур плавления соединений рассматриваемого класса, не включенных в исходную выборку.

In the framework of the statistical approach the non-linear «structure – melting point» model for pyrazolone derivatives is constructed. Number of topological parameters calculated from the structural formula was used for the quantitative description of the molecular structure. The results of calculation of the melting points for a series of compounds of considered class are reported.

Ключевые слова: *температура плавления, модели связи «структура – свойство», производные пиразолона*

Key words: *melting point*, *«structure – property» model, pyrazolone derivatives*

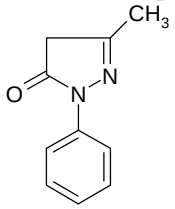
Введение. Проблема моделирования связи между структурой и свойствами органических соединений является одной из важнейших задач современной математической химии. Найденные закономерности позволяют прогнозировать свойства химических соединений непосредственно по их структуре, минуя эксперимент, и могут быть использованы для планирования целенаправленного поиска соединений с заданными свойствами.

Одним из наиболее распространенных подходов к моделированию связи «структура-свойство» является так называемый статистический подход, суть которого заключается в следующем. Имеется выборка соединений с известными численными значениями некоторого свойства y этих соединений. Структура соединений описывается при помощи набора молекулярных параметров $x_1,\ldots,x_n$, в качестве которых могут быть использованы топологические, электронные, геометрические характеристики молекул или значения каких-либо их физико-химических свойств. Как правило, математическая модель связи «структура-свойство» в рамках этого подхода имеет вид уравнения $y=f(x_1,\ldots,x_n)$, связывающего свойство y и параметры $x_1,\ldots,x_n$ при помощи некоторой функции f. Тип функции f предполагается известным; однако f зависит от ряда подгоночных параметров. Эти параметры подбираются по известным данным для заданной выборки соединений так, чтобы вышеуказанное соотношение выполнялось бы как можно более точно на этой выборке. Для оценки качества построенной модели могут быть использованы различные критерии. Например, можно рассматривать коэффициент корреляции и среднеквадратичное отклонение для корреляции между расчетными и экспериментальными значениями рассматриваемого свойства для соединений исходной выборки или среднюю относительную ошибку расчета свойств соединений этой выборки по построенной модели и т.д. Полученное уравнение затем используется для прогнозирования свойств других соединений, не включенных в исходную выборку. В процессе построения моделей в рамках статистического подхода возникают проблемы выбора молекулярных параметров $x_1,\ldots,x_n$ и функции f. Это связано с тем, что в задачах такого типа заранее неизвестно, от каких именно структурных особенностей зависит данное свойство и каким образом, а для выбора как параметров, так и аппроксимирующей функции имеется бесконечно много вариантов [1, 2].

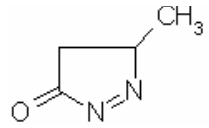
Одной из важнейших физико-химических характеристик органических соединений является их температура плавления. Как отмечено в [3], температура плавления веществ имеет многочисленные приложения в биохимии и экологии из-за ее тесной связи с растворимостью. Среди огромного количества работ, посвященных моделированию связи «структура – свойство» для различных свойств и классов соединений в рамках статистического подхода, работы, в которых рассматривается температура плавления, встречаются относительно редко. Это свойство считается «сложным» для подобного моделирования, особенно для выборок химических соединений, имеющих большое структурное разнообразие [3]. Следует отметить также, что многие вещества могут иметь различные кристаллические решетки, которым соответствуют различные температуры плавления. В работах [4, 5] рассматривается простейший

класс соединений – алканы, а в качестве молекулярных дескрипторов – различные инварианты соответствующих молекулярных графов, которые были успешно использованы в этих же работах при моделировании других свойств алканов. Однако даже в этом случае «хорошие» корреляции получаются лишь при достаточно большом числе параметров в модели, и только для нормальных алканов можно получить удовлетворительный результат. В работе [3] сообщается о моделировании температуры плавления пиридинов и пиперидинов, а также моно- и дизамещенных бензолов. Для поиска этих корреляций был использован очень большой набор разнообразных молекулярных дескрипторов – топологических, электронных, геометрических и др., из которого отбирались наилучшие параметры. Установлено, что наиболее важным в каждом случае является дескриптор, характеризующий способность молекул к образованию водородных связей и получаемый при помощи квантовохимических расчетов. Оценка качества модели проводилась по величинам R^2 и R^2_{cv}, где R – коэффициент корреляции уравнения регрессии, R^2_{cv} – коэффициент корреляции, получаемый в так называемой процедуре «cross-validation» (например, для пиридинов и пиперидинов R^2=0.831, R^2_{cv}=0.816 и эти результаты признаны авторами достаточно хорошими).

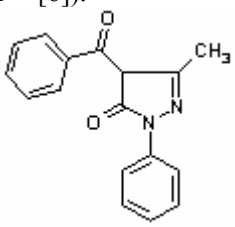
В настоящей работе рассматривается следующая задача: в рамках статистического подхода построить и исследовать модель связи «структура – температура плавления» для ряда производных пиразолона, используя для количественного описания этих соединений только топологические параметры. Исходные данные для решения этой задачи представлены на рис.1 (источник информации – [6]).



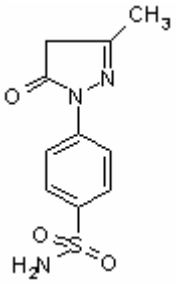
(1) $t_{эксп}$=126-128ºC



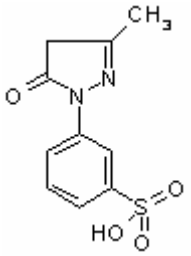
(2) $t_{эксп}$=217-223ºC



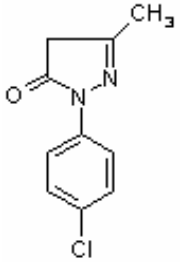
(3) $t_{эксп}$=90-92ºC



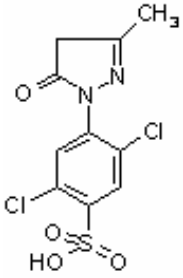
(4) $t_{эксп}$=246-250ºC



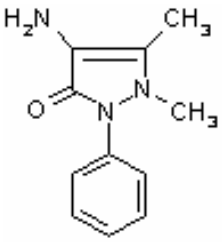
(5) $t_{эксп}$=335ºC



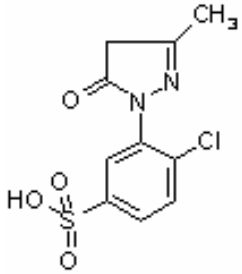
(6) $t_{эксп}$=167-171ºC



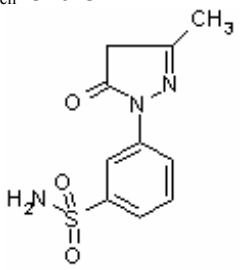
(7) $t_{эксп}$=346ºC



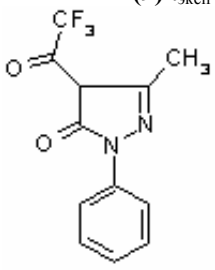
(8) $t_{эксп}$=107-109ºC



(9) $t_{эксп}$=344ºC



(10) $t_{эксп}$=202-206ºC



(11) $t_{эксп}$=145-148ºC

Рис. 1. Структуры производных пиразолона, использованные для построения модели.

Выбор молекулярных параметров. Как хорошо известно, на свойства органических соединений влияет наличие в их структуре определенных структурных фрагментов. Для учета этого фактора мы использовали параметры y_i (i=1,…,6), равные числам вхождения в

структуру следующих фрагментов:

1) Ph- ; 2) $-SO_2-$; 3) $-NH_2$;

4) -OH; 5) -Cl (или F); 6) O=.

Кроме того, для учета взаимного влияния этих фрагментов друг на друга были введены параметры

$$y_i y_j\ ,\quad y_i \oplus y_j \qquad (i,j=1,\ldots,6;\ i \neq j),$$

где $y_i \oplus y_j = 0$, если одновременно присутствуют (или отсутствуют) фрагменты i и j, $y_i \oplus y_j = 1$, если присутствует только один фрагмент из этих двух.

Естественно предположить, что на температуру плавления молекул влияет их компактность. Для учета этого фактора (на топологическом уровне) мы ввели следующие параметры:

$$W,\quad W/n,\quad W/n^2,\quad W/n^3,\quad 1/W$$

(W – индекс Винера, n – число неводородных атомов в молекуле). Индекс Винера определяется так: в графе, соответствующем структурной формуле молекулы со «стертыми» атомами водорода, рассчитываются расстояния (в ребрах) между всеми парами вершин и затем все эти числа суммируются.

Для учета влияния на температуру плавления водородных связей между молекулами мы рассмотрели следующие параметры:

$$X(H)/n,\quad X(H)/(n+X(H)),$$

где X(H) – число атомов водорода в молекуле.

Метод построения моделей. Пусть $X=\{x_1,\ldots,x_k\}$ – указанные выше молекулярные параметры. Сначала методом пошаговой линейной регрессии из множества X отбирается минимальное число параметров, дающих линейную модель заданной точности по критерию $\max\Delta_i(\%)\leq 5\%$. В данном случае этими параметрами оказались следующие:

1) X(-OH); 2) $X(-SO_2-)$; 3) W/n^3; 4) X(-Cl); 5) X(H)/(n+X(H)); 6) $X(-Ph)\oplus X(-NH_2)$; 7) X(=O).

Затем к этим семи параметрам добавляются их квадраты и попарные произведения. Из полученного нового набора опять отбирается минимальное число параметров, дающих линейную модель заданной точности. В результате построены две модели, нелинейные относительно исходных семи параметров. Первая из них, отобранная по критерию

$$\max\Delta_i(\%) \leq 5\ \%\ ,$$

содержит шесть параметров. Вторая модель, отобранная по критерию

$$\Delta_{ср}(\%) \leq 5\ \%\ ,$$

содержит пять параметров. Очевидно, что первый критерий является более жестким, чем второй, так как всегда $\Delta_{ср}(\%)\leq\max\Delta_i(\%)$. Однако при использовании метода наименьших квадратов для построения модели более естественно использовать критерий, связанный со средней ошибкой. Следует также отметить, что при моделировании связи «структура – свойство» довольно типичной является следующая ситуация: модели, содержащие меньшее число параметров, имеют лучшую прогнозирующую способность. Поэтому при прочих равных условиях предпочтение отдают моделям с меньшим числом параметров. В связи с этим для дальнейших исследований была выбрана модель с пятью параметрами.

Ниже приведено соответствующее уравнение, связывающее температуру плавления t соединений, представленных на рис.1, и пять отобранных параметров $\boldsymbol{X_1}$ – $\boldsymbol{X_5}$. Здесь через X(-OH), $X(-SO_2)$, X(-Cl) и т. д. обозначены числа соответствующих фрагментов в структурной формуле. Точность модели охарактеризована при помощи трех ошибок ($\Delta_{ср}(^0)$, $\Delta_{ср}(\%)$, $\max\delta_i$), а также среднеквадратичного отклонения S и коэффициента корреляции R.

$$t(^0) = -130.364 + 142.23\,\boldsymbol{X_1} + 836.96\,\boldsymbol{X_2} + 24086\,\boldsymbol{X_3} + 79.41\,\boldsymbol{X_4} - 165.67\,\boldsymbol{X_5}$$

$$\Delta_{ср}(^0)=5.7^0,\ \Delta_{ср}(\%)=3.37\%,\ \max\delta_i=11.5^0,\ S(^0)=9.7,\ R=0.997,$$

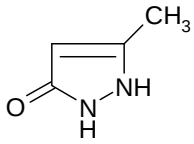
$$\boldsymbol{X_1}=X(-OH),\ \boldsymbol{X_2}=X(-SO_2)\cdot(W/n^3),\ \boldsymbol{X_3}=(W/n^3)^2,$$

$$\boldsymbol{X_4}=X(-Cl)\cdot(H/(n+H)),\ \boldsymbol{X_5}=(W/n^3)\cdot(X(-Ph)\oplus X(-NH_2))$$

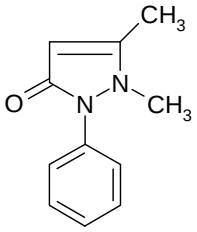
В таблице даны экспериментальные и расчетные значения температур плавления соединений исходной выборки, а также соответствующие абсолютные и относительные ошибки.

Тестирование модели. Полученное уравнение было использовано для расчета температуры плавления других соединений вышеуказанного класса, не включенных в исходную выборку [6]. На рис. 2 даны структурные формулы этих соединений, а также экспериментальные и расчетные значения их температур плавления. Во всех случаях относительные ошибки прогноза меньше 5%, что свидетельствует об удовлетворительной прогнозирующей способноси построенной модели на рассмотренных соединениях.

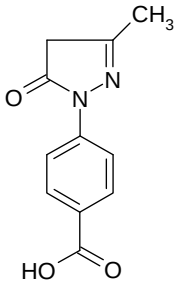
№	$t_{эксп.}$	$t_{расч.}$	δ_i	Δ_i
1	126	136.79	10.79	8.6%
2	217	213.59	3.41	1.6%
3	90	86.489	3.51	3.9%
4	246	241.68	4.32	1.8%
5	335	333.27	1.73	0.52%
6	167	170.43	3.43	2.1%
7	346	354.84	8.84	2.6%
8	107	110.98	3.98	3.7%
9	344	336.99	7.01	2%
10	202	206.63	4.63	2.3%
11	145	133.46	11.54	8%



$t_{эксп}$=109-112ºC
t_p=105ºC



$t_{эксп}$=285ºC
t_p=284ºC



$t_{эксп}$=223-225ºC
t_p=214ºC

Рис. 2. Структуры производных пиразолона, использованные для тестирования модели.

Заключение. Таким образом, в работе построена модель, связывающая температуру плавления производных пиразолона и их некоторые топологические структурные характеристики, Эти характеристики вычисляются довольно просто непосредственно по структурной формуле молекулы и не требуют установления пространственного и электронного строения молекулы. Выбор параметров основан на определенных рассуждениях качественного характера о тех факторах, которые должны оказывать существенное влияние на данное свойство. Для нахождения аппроксимирующей функции использован специальный подход, согласно которому модель строится в два этапа: сначала строится удовлетворительная линейная модель, а затем на ее основе – удовлетворительная нелинейная (второго порядка), однако являющаяся линейной относительно новых параметров. Этот прием позволяет уменьшить число переменных в модели и тем самым улучшить ее прогнозирующую способность.

На основе анализа проведенных исследований и построенной модели можно сделать следующие выводы.

1) Как было указано ранее, для многих соединений экспериментальные значения температур плавления лежат в некотором интервале. Если для построения модели для каждого соединения выбрать в качестве конкретного значения температуры плавления минимальное, среднее и максимальное значение из заданного интервала, то качество модели оказывается наилучшим для минимальных значений.

2) Наиболее важные факторы, влияющие на температуру плавления рассматриваемых соединений: а) компактность молекулы (наиболее адекватной количественной мерой компактности оказалась величина W/n^3); б) способность к образованию водородных связей (количественная мера – величина H/(n+H)); в) наличие определенных структурных фрагментов и их сочетаний.

4) Следует отметить, что при моделировании связи «структура – свойство» довольно типичной является следующая ситуация: модели, содержащие меньшее число параметров, имеют лучшую прогнозирующую способность. Поэтому при прочих равных условиях предпочтение отдают моделям с меньшим числом параметров. Предложенная нами процедура перехода от линейного уравнения к нелинейному (относительно исходных параметров) позволяет получить новую линейную модель, содержащую меньшее число параметров и имеющую примерно такую же точность, как и исходная модель. Окончательная модель с пятью параметрами прогнозирует (в среднем) лучше, чем нелинейная модель с шестью параметрами, но является менее точной на обучающей выборке. Эта модель с пятью параметрами и является итоговым результатом наших исследований.

ЛИТЕРАТУРА:

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

2. Raevsky, O. A. Molecular structure descriptors in the computer-aided design of biologically active compounds / O. A. Raevsky // Russ. Chem. Rev. – 1999. – Vol. 68. – P. 505–524.

3. QSPR and QSAR models derived using large molecular descriptor spaces. A review of CODESSA applications / M. Karelson [et al. // Collect. Czech. Chem. Commun. – 1999. – Vol. 64. – P. 1551–1571.

4. Needham, D. E. Modeling of the Physical Properties of the Alkanes / D. E. Needham, I. C. Wei, P. G. Seybold // J. Am. Chem. Soc. – 1988. – Vol. 110. – P. 4186–4194.

5. Моделирование связи между структурой и свойствами углеводородов на основе базисных топологических дескрипторов / М. И. Скворцова [и др.] // Изв. АН (сер. химическая). – 2004. – № 8. – С. 1527–1535.

6. http://www.sigmaaldrich.com