

ХИМИЧЕСКАЯ ТЕРМОДИНАМИКА  
И ТЕРМОХИМИЯ

УДК 544.022.544.3

ОЦЕНКА ТЕРМОДИНАМИЧЕСКИХ СВОЙСТВ ХИМИЧЕСКИХ  
СОЕДИНЕНИЙ НА ОСНОВЕ КОЛИЧЕСТВЕННЫХ  
СООТНОШЕНИЙ СТРУКТУРА–СВОЙСТВО

© 2019 г. О. В. Шульц<sup>а,\*</sup>

<sup>а</sup> Российский федеральный ядерный центр – Всероссийский научно-исследовательский институт технической физики им. академика Е.И. Забабахина, Снежинск, Челябинская обл., Россия

\* e-mail: O.V.Shults@VNIITF.ru

Поступила в редакцию 21.06.2018 г.

После доработки 05.12.2018 г.

Принята к публикации 15.12.2018 г.

Работа посвящена созданию математической модели, обобщающей информацию из базы данных по термодинамическим свойствам индивидуальных веществ. Описана математическая модель, предназначенная для оценки значений термодинамических функций (энтальпии образования, энтропии и изобарной теплоемкости) химических соединений по их структуре. Модель построена на основе линейной регрессии, отражающей связь между выбранным набором признаков и свойствами соединений. Описан способ пакетной обработки данных из базы для формирования исходной выборки. Описаны способ получения набора признаков и способ расчета параметров модели, приведены их численные значения. Выполнены тестовые расчеты, показано хорошее согласие расчетных данных со справочными.

*Ключевые слова:* QSPR, термодинамика, оценка свойств, математическая модель

DOI: 10.1134/S0044453719070264

ВВЕДЕНИЕ

При проектировании и моделировании хими-ко-технологических процессов всегда актуальной является проблема обеспечения исходными данными по свойствам веществ. Особенно остро эта проблема касается свойств сложных органических соединений. С другой стороны, в настоящее время существует большое количество справочников и баз данных по свойствам индивидуальных веществ. Логичным способом решения проблемы недостающих данных является построение моделей, обобщающих известную справочную информацию для прогнозирования свойств соединений, информация о которых отсутствует.

Одним из современных подходов к построению подобных моделей является QSPR/QSAR (Quantitative Structure-Property/Activity Relationship) – количественное соотношение структура–свойство/активность.

В рамках настоящей работы предлагается модель, использующая принципы QSPR, для оценки свойств различных соединений по данным об их структуре и агрегатном состоянии на примере расчета энтальпии образования, энтропии и изобарной теплоемкости соединений.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

*Связь свойств и признаков соединений*

В основе модели лежит предположение об аддитивности свойств фрагментов, из которых состоит химическое соединение. В общем виде это предположение может быть сформулировано следующим образом: любое химическое соединение может быть описано некоторым набором признаков, а значение некоторого свойства  $f$  этого соединения можно рассчитать как сумму

$$f = \sum_i n_i w_i, \quad (1)$$

где  $n_i$  – значение  $i$ -го признака у соединения,  $w_i$  – характеристика  $i$ -го признака. В такой формулировке это предположение приводится в ряде современных работ [1, 2]. В рамках теории аддитивности групп Бенсона [3, 4] в качестве таких признаков рассматриваются химические связи между различными атомами и группами атомов, а в качестве характеристики выступает энергия связи.

В настоящей работе в качестве признаков выступают количество фрагментов или функциональных групп  $i$ -го вида с учетом их связей, например  $-\text{CH}_3$ ,  $=\text{CH}-$ ,  $-\text{OH}$ ,  $-\text{COOH}$  и т.д., агрегатное состояние вещества, а также рассмотрен

вариант дополнения списка признаков признаками взаимного соседства фрагментов. Выражение (1) позволяет учитывать все признаки независимо от их вида, в связи с этим список признаков может быть расширен или сокращен для решения различных задач.

Для набора соединений с известными значениями свойств и известным набором признаков можно переписать уравнение (1) в виде:

$$\mathbf{F} = \mathbf{N} \cdot \mathbf{w}, \quad (2)$$

где  $\mathbf{F}$  – вектор значений выбранного свойства, соответствующих рассматриваемым соединениям;  $\mathbf{N}$  – матрица признаков, строки в которой соответствуют соединениям, а столбцы – признакам, на пересечениях стоят значения, соответствующие значениям признака в данном соединении;  $\mathbf{w}$  – вектор характеристик признаков.

Имея данные по значениям интересующего свойства для ряда соединений с известными значениями признаков, можно получить вектор термодинамических характеристик признаков как решение (или приближенное решение для переобусловленной системы) уравнения (2) и далее использовать его для прогнозирования свойств других соединений с известными признаками.

#### *Исходные данные для построения набора признаков*

Для построения набора признаков был проведен анализ базы данных (БД) термодинамических свойств индивидуальных веществ, включающей информацию из ряда источников [5–10]. БД содержит несколько десятков тысяч записей, содержащих брутто-формулу, систематическое название, описание фазового состояния, термодинамические данные и некоторую дополнительную информацию.

Так как для построения модели была необходима информация о структуре соединений, а в БД указание на структуру веществ содержалось только в виде систематического названия и брутто-формулы, было решено использовать компьютерную программу OPSIN [11], которая позволяет преобразовать систематические названия веществ в нотацию SMILES [12]. Дополнительно для проверки правильности интерпретации систематического названия, полученные строки SMILES проверялись на соответствие брутто-формулам соединений из БД.

Для интерпретации формата SMILES был разработан и реализован алгоритм, преобразующий такую запись в матричное представление молекулярного графа соединения. Построенный таким способом молекулярный граф далее разбивался на фрагменты. Количество фрагментов каждого вида использовалось в качестве значений признаков.

В качестве свойств соединений в рамках настоящей работы рассматривались энтальпия образования, стандартная энтропия и теплоемкость при 298 К.

Так как не все записи в БД содержат систематическое название и данные о термодинамических свойствах веществ в нужных диапазонах, а также не все систематические названия записаны в подходящем для распознавания программой OPSIN виде, для расчетов была использована только часть данных, удовлетворяющая всем названным условиям. Всего на этапе предварительного отбора для расчетов было выбрано порядка 20000 записей. Так как на месте недостающих данных в базе записи содержат нулевые значения и энтальпии простых веществ в стандартном состоянии также нулевые, все простые вещества в стандартном состоянии были исключены из выборки вместе с неполными записями.

#### *Построение молекулярного графа на основе нотации SMILES*

Построение молекулярного графа из строки SMILES производится рекурсивным образом. В первую очередь из исходной строки выделяется подстрока, не содержащая в себе подстрок в скобках. Если структура вещества простая и соответствующая строка SMILES не имеет скобок, первая итерация является единственной. Для выделенной строки строится простой неразветвленный граф. Место выделенной подстроки в строке, содержащей ее, помечается.

Процесс повторяется, пока есть подстроки в скобках плюс один раз. Последний полученный таким способом граф достраивается всеми подграфами в соответствии с отметками о вхождении подстрок. Вершины, помеченные как соседние в цикле, попарно соединяются соответствующими связями.

#### *Разбиение молекулярного графа на фрагменты*

Перед разбиением молекулярного графа на фрагменты для всех атомов проверяется соответствие количества связей наименьшей нормальной валентности. Если связей у атома меньше, чем его валентность, к атому присоединяются атомы водорода в необходимом количестве. Такая процедура необходима в связи с тем, что в нотации SMILES атомы водорода не указываются, когда их количество соответствует минимальной нормальной валентности атома, с которым они связаны.

В качестве фрагментов в рамках предлагаемого подхода предлагается использовать группы связанных атомов, соответствующие подграфам, полученным удалением из исходного графа всех ребер, не связанных с концевыми вершинами. Кон-

**Таблица 1.** Наиболее часто встречающиеся фрагменты и признаки агрегатного состояния

| Признак   | Кол-во | Признак            | Кол-во | Признак   | Кол-во | Признак                                 | Кол-во |
|-----------|--------|--------------------|--------|---|--------|---|--------|
| -C(-)(H)H | 46821  | [O <sup>2-</sup> ] | 1158   | -S-   | 400    | [Li <sup>+</sup> ]                      | 235    |
| -C(H)(H)H | 23701  | [Cl <sup>-</sup> ] | 1142   | -N(H)H  | 395    | -N(-)H                                  | 225    |
| =C(-)H    | 9500   | [F <sup>-</sup> ]  | 1062   | -C(=O)OH  | 385    | [Rb <sup>+</sup> ]                      | 221    |
| g         | 8682   | HOH                | 1059   | O=S(=O)([O <sup>-</sup> ])[O <sup>-</sup> ]           | 373    | -C#N                                    | 207    |
| -C(-)(-)H | 6098   | -C(-)=O            | 1019   | [Na <sup>+</sup> ]                                    | 369    | Cl(=O)(=O)(=O)[O <sup>-</sup> ]         | 199    |
| =C(-)-    | 5584   | [Br <sup>-</sup> ] | 672    | -C(F)(F)F   | 327    | -C(-)(F)F                               | 186    |
| l         | 2297   | [I <sup>-</sup> ]  | 659    | [K <sup>+</sup> ]                                     | 311    | [Se <sup>2-</sup> ]                     | 185    |
| -C(-)(-)- | 2045   | -C(=O)[O-]         | 597    | O=[N <sup>+</sup> ](O <sup>-</sup> )[O <sup>-</sup> ] | 301    | -[N <sup>+</sup> ](=O)[O <sup>-</sup> ] | 181    |
| -O-       | 2004   | =C(H)H             | 536    | [S <sup>2-</sup> ]                                    | 294    | I(=O)(=O)[O <sup>-</sup> ]              | 180    |
| -OH       | 1938   | [OH <sup>-</sup> ] | 507    | -C(Cl)(H)H  | 266    | -C(H)=O                                 | 179    |
| a         | 1719   | ia                 | 470    | [Cs <sup>+</sup> ]                                    | 260    | -N(-)-                                  | 176    |
| =C(-)Cl   | 1170   | #C-                | 423    | -SH   | 244    | [Ca <sup>2+</sup> ]                     | 176    |

Примечание. Строчными буквами обозначены признаки агрегатного состояния (g – газ, l – жидкость, a – водный раствор неэлектролита, ia – водный раствор электролита).

цевыми считаются вершины, имеющие только одну (возможно кратную) связь.

Группа –COOH разбивается указанным способом на две отдельные группы –C=O и –OH. Для выделения группы –COOH как единого фрагмента после прохода таким алгоритмом производится поиск групп –C(-)=O и –OH, соединенных вместе и их объединение в общий подграф. С другой стороны, в некоторых небольших молекулах, таких как молекулы метана, аммиака или воды есть только связи с концевыми атомами. Такие молекулы определяются алгоритмом как состоящие из единственного фрагмента.

Для удобства хранения и представления полученных фрагментов в читаемом для человека виде, все полученные фрагменты кодировались обратно в строку SMILES. Так как нотация SMILES допускает несколько вариантов записи для одной и той же структуры (обход графа при построении строки может быть начат с любой вершины), после разбиения молекулярного графа на подграфы, каждый полученный подграф упорядочивали таким образом, чтобы при формировании его строкового представления обход вершин начинался с первой по алфавиту. Эта дополнительная процедура была необходима, чтобы одинаковым фрагментам в разных соединениях всегда соответствовали одинаковые строковые представления SMILES.

#### Полученный набор признаков

Как было указано выше, в качестве значений признаков использовалось количество фрагментов каждого вида. Также в список признаков было включено агрегатное состояние вещества (такой признак может принимать только два значе-

ния “1” или “0”) и соседство фрагментов. По умолчанию все вещества без признака агрегатного состояния считаются твердыми.

В табл. 1 представлены наиболее часто встречающиеся фрагменты и признаки агрегатного состояния соединений, и их суммарное количество во всей рассмотренной выборке соединений. Полный список фрагментов и признаков агрегатного состояния включает 1501 пункт, из которых 376 встречаются не менее 10 раз в рассмотренной выборке.

Признаки соседства фрагментов были вынесены в отдельный список. Наиболее часто встречающиеся пары фрагментов-соседей приведены в табл. 2. Полный список рассмотренных пар соседних фрагментов содержит 706 пунктов, из которых 198 встречаются не менее 10 раз в рассмотренной выборке.

#### Расчет характеристик признаков

В первом приближении при построении модели рассматривались только признаки фрагментов и агрегатного состояния вещества. В соответствии с формулой (2) по справочным данным для выбранных признаков были рассчитаны термодинамические характеристики для оценки энталпии образования, стандартной энтропии и изобарной теплоемкости соединений.

В табл. 3 приведены полученные значения термодинамических характеристик наиболее частых признаков.

Так как не все записи в БД содержат полную информацию по значениям рассматриваемых термодинамических функций, для расчета характеристик признаков часть данных была отсеяна. Также для расчетов не использовались данные по

Таблица 2. Наиболее часто встречающиеся пары фрагментов-соседей

| Фрагмент 1 | Фрагмент 2 | Кол-во | Фрагмент 1 | Фрагмент 2 | Кол-во |
|------------|------------|--------|------------|------------|--------|
| -C(-)(H)H  | -C(-)(H)H  | 26151  | -C(-)(H)H  | -OH        | 456    |
| =C(-)H     | =C(-)-     | 6736   | -C(-)(-)H  | -C(-)(-)-  | 445    |
| -C(-)(H)H  | -C(H)(H)H  | 5722   | =C(-)H     | =C(H)H     | 436    |
| =C(-)H     | =C(-)H     | 5575   | -C(-)(-)H  | -OH        | 415    |
| -C(-)(H)H  | -C(-)(-)H  | 4254   | -O-        | -C(-)=O    | 414    |
| -C(H)(H)H  | -C(-)(-)H  | 4131   | -C(-)(H)H  | -C(-)=O    | 403    |
| =C(-)-     | =C(-)-     | 3520   | #C-        | #C-        | 357    |
| -C(H)(H)H  | -C(-)(-)-  | 2693   | #C-        | #CH        | 324    |
| =C(-)H     | =C(-)Cl    | 1343   | -C(-)(H)H  | -N(-)-     | 282    |
| -C(-)(H)H  | -C(-)(-)-  | 1318   | -C(-)(H)H  | -N(-)H     | 233    |
| -C(H)(H)H  | =C(-)-     | 1199   | -C(-)(H)H  | -S-        | 213    |
| -C(-)(-)H  | -C(-)(-)H  | 689    | =C(-)-     | -OH        | 212    |
| -C(-)(H)H  | -O-        | 664    | =C(-)-     | =C(H)H     | 190    |
| =C(-)Cl    | =C(-)Cl    | 626    | -C(H)(H)H  | -O-        | 189    |
| -C(-)(H)H  | =C(-)-     | 618    | -C(-)(-)H  | -O-        | 171    |
| =C(-)-     | -O-        | 588    | =C(-)H     | =N-        | 169    |
| -C(-)(H)H  | =C(-)H     | 532    | -C(-)(-)H  | =C(-)-     | 163    |
| =C(-)-     | =C(-)Cl    | 524    | -C(-)(-)-  | -OH        | 153    |

свойствам соединений, с единственным признаком со значением 1 (т.е. состоящие из единственного фрагмента и находящиеся в твердом агрегатном состоянии). В связи с этим при формировании выборки для расчета термодинамических характеристик некоторые признаки могли не встретиться ни разу.

С целью повышения точности расчета в модель была добавлена возможность учета влияния взаимного соседства фрагментов. В рамках модели такое влияние описывается дополнительным поправочным коэффициентом в формуле (1):

$$f = \sum_i n_i w_i + d,$$

где  $n_i$ ,  $w_i$  – в соответствии с обозначениями (1),  $d$  – поправочный коэффициент:

$$d = \sum_j m_j Z_j,$$

где  $m_j$  – количество пар  $j$ -го вида,  $Z_j$  – характеристика  $j$ -й пары.

Полученные значения термодинамических характеристик наиболее частых признаков, соответствующих соседству фрагментов, приведены в табл. 4.

#### Анализ результатов

Для проверки качества оценки термодинамических характеристик для всех рассмотренных

соединений из БД были рассчитаны значения энтальпии образования, стандартной энтропии и изобарной теплоемкости при 298 К двумя способами – с учетом и без учета соседства фрагментов.

Рассмотренная выборка данных по энтальпиям соединений содержала 5287 записей. Среднеквадратичное отклонение расчетных значений от справочных составило 93.51 кДж/моль без учета соседства фрагментов и 89.21 кДж/моль с учетом. Медианная относительная ошибка без учета знака составила 5.50% без учета соседства фрагментов и 3.74% с учетом.

Рассмотренная выборка данных по энтропиям соединений содержала 3629 записей. Среднеквадратичное отклонение расчетных значений от справочных составило 22.71 Дж моль<sup>-1</sup> К<sup>-1</sup> без учета соседства фрагментов и 20.36 Дж моль<sup>-1</sup> К<sup>-1</sup> с учетом. Медианная относительная ошибка без учета знака составила 1.98% без учета соседства фрагментов и 1.60% с учетом.

Рассмотренная выборка данных по изобарным теплоемкостям соединений содержала 1574 записи. Среднеквадратичное отклонение расчетных значений от справочных составило 28.22 Дж моль<sup>-1</sup> К<sup>-1</sup> без учета соседства фрагментов и 23.00 Дж моль<sup>-1</sup> К<sup>-1</sup> с учетом. Медианная относительная ошибка без учета знака составила 0.74% без учета соседства фрагментов и 0.66% с учетом.

На рис. 1–3 показано сравнение расчетных и соответствующих справочных данных с учетом соседства фрагментов.

Таблица 3. Термодинамические характеристики наиболее частых признаков

| №  | Признак            | $w_H$ , кДж/(моль) | $w_S$ , Дж/(моль К) | $w_{C_p}$ , Дж/(моль К) |
|----|--------------------|--------------------|---------------------|-------------------------|
| 1  | -C(-)(H)H          | -29.48             | 34.93               | 22.71                   |
| 2  | -C(H)(H)H          | -84.95             | 51.71               | 33.52                   |
| 3  | =C(-)H             | -10.48             | 23.07               | 15.92                   |
| 4  | g                  | 269.15             | 181.24              | -13.67                  |
| 5  | -C(-)(-)H          | -3.14              | 1.79                | 10.22                   |
| 6  | =C(-)-             | 25.25              | 1.39                | 9.02                    |
| 7  | l                  | 143.48             | 39.68               | 118.95                  |
| 8  | -C(-)(-)-          | 36.66              | -38.99              | 2.24                    |
| 9  | -O-                | -173.65            | 21.49               | 7.82                    |
| 10 | -OH                | -255.26            | 38.19               | 28.96                   |
| 11 | a                  | 11.37              | -                   | -                       |
| 12 | =C(-)Cl            | -33.90             | 36.80               | 23.76                   |
| 13 | [O2-]              | -310.73            | 7.20                | 16.35                   |
| 14 | [Cl-]              | -224.43            | 35.49               | 21.73                   |
| 15 | [F-]               | -394.21            | 24.04               | 18.74                   |
| 16 | HOH                | -293.02            | 40.19               | 45.64                   |
| 17 | -C(-)=O            | -223.65            | 58.29               | 25.28                   |
| 18 | [Br-]              | -188.69            | 48.36               | 23.51                   |
| 19 | [I-]               | -115.32            | 56.31               | 23.42                   |
| 20 | -C(=O)[O-]         | -497.09            | 31.67               | -                       |
| 21 | =C(H)H             | -27.21             | 38.16               | 29.45                   |
| 22 | [OH-]              | -321.98            | 37.55               | 30.42                   |
| 23 | ia                 | 1.48               | -                   | -                       |
| 24 | #C-                | 84.60              | 16.54               | 13.54                   |
| 25 | -S-                | -25.42             | 34.61               | 17.04                   |
| 26 | -N(H)H             | -61.72             | 16.87               | 29.98                   |
| 27 | -C(=O)OH           | -433.15            | 60.77               | 47.14                   |
| 28 | O=S(=O)([O-])[O-]  | -1072.74           | 65.25               | 55.78                   |
| 29 | [Na+]              | -161.14            | 38.41               | 25.41                   |
| 30 | -C(F)(F)F          | -743.77            | 101.85              | 64.09                   |
| 31 | [K+]               | -160.68            | 49.37               | 27.41                   |
| 32 | O=[N+](O-)[O-]     | -291.23            | 79.09               | -                       |
| 33 | [S2-]              | -93.82             | 17.99               | 21.42                   |
| 34 | -C(Cl)(H)H         | -196.58            | 75.27               | 44.05                   |
| 35 | [Cs+]              | -165.89            | 68.89               | 30.02                   |
| 36 | -SH                | -99.03             | 68.24               | 32.47                   |
| 37 | [Li+]              | -197.69            | 16.58               | 19.55                   |
| 38 | -N(-)H             | 11.57              | 8.62                | 17.18                   |
| 39 | [Rb+]              | -172.22            | 58.37               | 27.69                   |
| 40 | -C#N               | 78.45              | 56.98               | 40.00                   |
| 41 | Cl(=O)(=O)(=O)[O-] | -211.54            | 113.00              | -                       |
| 42 | -C(-)(F)F          | -440.13            | 62.54               | -                       |
| 43 | [Se2-]             | -32.86             | 31.88               | 24.33                   |
| 44 | -[N+](=O)[O-]      | -49.23             | 74.62               | 40.04                   |
| 45 | I(=O)(=O)[O-]      | -295.61            | 87.41               | -                       |
| 46 | -C(H)=O            | -291.03            | 69.96               | 28.74                   |
| 47 | -N(-)-             | 83.03              | -24.34              | 17.51                   |
| 48 | [Ca2+]             | -386.24            | 42.24               | 28.78                   |

Таблица 4. Термодинамические характеристики наиболее частых пар фрагментов-соседей

| №  | Пары фрагментов-соседей |            | $Z_H$ , кДж/(моль) | $Z_S$ , Дж/(моль К) | $Z_{C_p}$ , Дж/(моль К) |
|----|-------------------------|------------|--------------------|---------------------|-------------------------|
|    | Фрагмент 1              | Фрагмент 2 |                    |                     |                         |
| 1  | -C(-)(H)H               | -C(-)(H)H  | -1.22              | 0.43                | -0.34                   |
| 2  | =C(-)H                  | =C(-)-     | -0.55              | 0.29                | 1.04                    |
| 3  | -C(-)(H)H               | -C(H)(H)H  | 2.57               | -7.46               | -1.00                   |
| 4  | =C(-)H                  | =C(-)H     | 4.10               | 0.87                | -0.13                   |
| 5  | -C(-)(H)H               | -C(-)(-)H  | 7.69               | 5.42                | 1.13                    |
| 6  | -C(H)(H)H               | -C(-)(-)H  | -8.26              | -3.88               | -1.12                   |
| 7  | =C(-)-                  | =C(-)-     | -1.33              | -2.12               | -1.12                   |
| 8  | -C(H)(H)H               | -C(-)(-)-  | 2.83               | -2.79               | -0.33                   |
| 9  | =C(-)H                  | =C(-)Cl    | 13.52              | -2.86               | -1.17                   |
| 10 | -C(-)(H)H               | -C(-)(-)-  | 5.86               | 7.56                | 0.23                    |
| 11 | -C(H)(H)H               | =C(-)-     | -2.26              | 4.62                | 0.00                    |
| 12 | -C(-)(-)H               | -C(-)(-)H  | -8.51              | 0.89                | 1.13                    |
| 13 | -C(-)(H)H               | -O-        | 4.68               | -0.59               | 1.54                    |
| 14 | =C(-)Cl                 | =C(-)Cl    | 9.66               | -2.70               | -0.94                   |
| 15 | -C(-)(H)H               | =C(-)-     | 9.58               | 6.42                | -0.33                   |
| 16 | =C(-)-                  | -O-        | -18.41             | 6.32                | 0.70                    |
| 17 | -C(-)(H)H               | =C(-)H     | -42.63             | -0.19               | 1.24                    |
| 18 | =C(-)-                  | =C(-)Cl    | -6.70              | 0.60                | 2.33                    |
| 19 | -C(-)(H)H               | -OH        | -15.15             | 1.99                | 1.23                    |
| 20 | -C(-)(-)H               | -C(-)(-)-  | -10.19             | 7.89                | -0.47                   |
| 21 | =C(-)H                  | =C(H)H     | 4.44               | -6.63               | -1.12                   |
| 22 | -C(-)(-)H               | -OH        | -43.99             | -35.79              | 1.90                    |
| 23 | -O-                     | -C(-)=O    | 14.76              | 2.54                | -9.74                   |
| 24 | -C(-)(H)H               | -C(-)=O    | -16.84             | 0.21                | 3.59                    |
| 25 | #C-                     | #C-        | -2.84              | 1.17                | -2.61                   |
| 26 | #C-                     | #CH        | -12.87             | -2.03               | -3.95                   |
| 27 | -C(-)(H)H               | -N(-)-     | -3.06              | 7.80                | 0.00                    |
| 28 | -C(-)(H)H               | -N(-)H     | 23.66              | -4.09               | 0.72                    |
| 29 | -C(-)(H)H               | -S-        | 19.62              | 5.62                | 3.19                    |
| 30 | =C(-)-                  | -OH        | 13.12              | 5.80                | -6.30                   |
| 31 | =C(-)-                  | =C(H)H     | -18.76             | 9.76                | 6.08                    |
| 32 | -C(H)(H)H               | -O-        | 8.44               | 0.35                | -3.90                   |
| 33 | -C(-)(-)H               | -O-        | 30.90              | 0.73                | -1.69                   |
| 34 | =C(-)H                  | =N-        | 5.31               | -2.49               | -0.25                   |
| 35 | -C(-)(-)H               | =C(-)-     | 6.66               | -4.64               | -3.43                   |
| 36 | -C(-)(-)-               | -OH        | -75.70             | -6.66               | 10.51                   |

Как видно из рис. 1–3 применение описанной модели с использованием полученных значений характеристик признаков позволяет оценивать значения термодинамических функций (энтальпии образования, энтропии и изобарной теплоемкости) с довольно высокой точностью. Учет взаимного влияния соседних молекулярных

фрагментов позволяет дополнительно повысить точность расчетов.

Дополнительно следует отметить, что анализ погрешностей справочных данных из источников [5–10] не проводился. В связи с этим все погрешности полученных оценок были рассчитаны относительно использованных справочных данных.

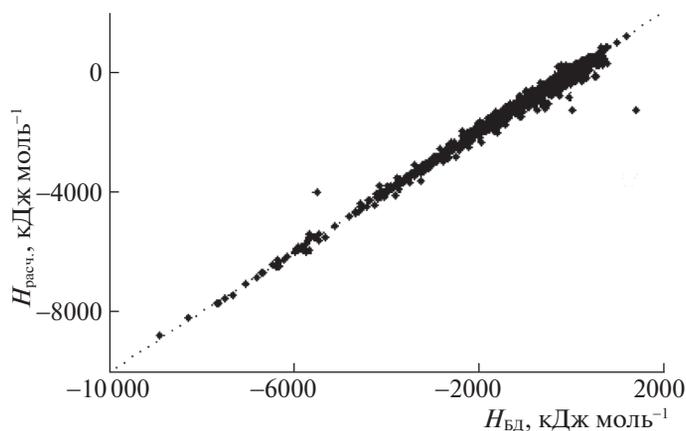


Рис. 1. Сопоставление расчетных значений энтальпии образования соединений при 298 К со справочными.

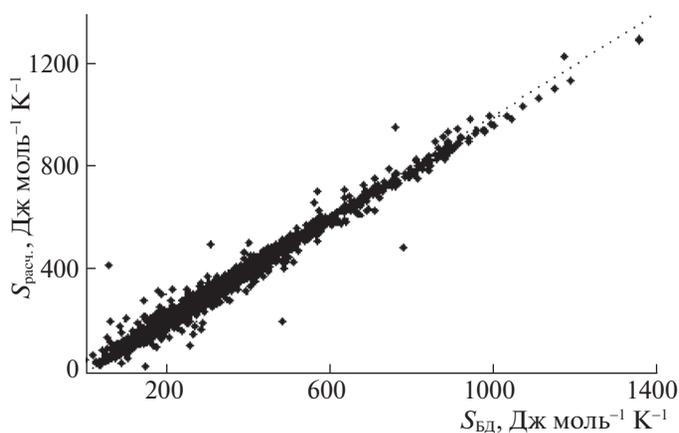


Рис. 2. Сопоставление расчетных значений стандартной энтропии соединений при 298 К со справочными.

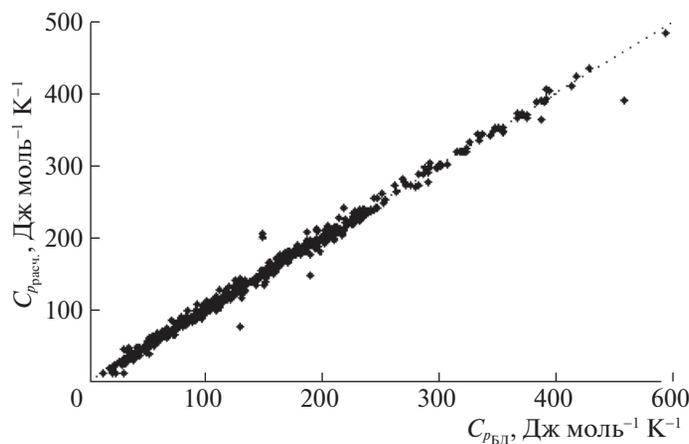


Рис. 3. Сопоставление расчетных значений изобарной теплоемкости соединений при 298 К со справочными.

В работе описана методика оценки свойств различных химических соединений по данным об их структуре и агрегатном состоянии с помощью

линейной регрессии и набора рассчитанных коэффициентов. В качестве параметров модели используется набор признаков, характеризующих

структуру соединения и его агрегатное состояние. Рассмотрено два варианта наборов признаков – с учетом и без учета соседства молекулярных фрагментов (функциональных групп). Описан вариант расчета энтальпии образования, энтропии и теплоемкости соединений при 298 К. Показано хорошее согласие данных, рассчитанных по предлагаемой методике, со справочными. Показано, что учет соседства молекулярных фрагментов позволяет повысить точность оценки. Для энтальпии среднеквадратичное отклонение рассчитанных данных от справочных с учетом соседства фрагментов меньше на 4.6%, чем без учета, для энтропии на 10.3%, для теплоемкости на 18.5%.

Описанная методика может быть применена для решения широкого круга задач. В первую очередь к таким задачам можно отнести оценку недостающих данных при моделировании химико-технологических процессов, в которых участвуют или могут образовываться различные соединения с недостаточно описанными свойствами.

Другим вариантом применения методики может быть проверка имеющихся справочных данных. В случае, когда БД содержит большое количество записей, проверка каждой из них вручную практически невозможна. Применение описанной методики позволяет в автоматическом режиме найти те записи, данные в которых сильно расходятся с оценочными и, таким образом, существенно сократить область поиска грубых ошибок.

Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 17-01-00873.

#### СПИСОК ЛИТЕРАТУРЫ

1. *Zefirov N.S., Palyulin V.A.* // J. Chem. Inf. Comput. Sci. 42 (5) 2002. P. 1112. doi: 10.1021/ci020010e
2. *Baskin I., Varnek A.* // Chemoinformatic Approaches to Virtual Screening, A. Varnek, A. Tropsha, eds., RCS. doi: 10.1002/chin.200920272
3. *Benson S.W.* // J. Chemical Education. V. 42. № 9, September 1965. Stanford Research Institute Menlo Park, California. doi: 10.1021/ed042p502
4. *Benson S.W., Buss J.H.* / J. of Chemical Physics. 29. № 3 september, 1958. Chemistry Department, University of Southern California, Los Angeles 7, California. doi: 10.1063/1.1744539
5. *Barin I.* Thermochemical Data of Pure Substances. Third Edition in collaboration with Gregor Platzki. Published jointly by VCH Verlagsgesellschaft mbH, Weinheim (Federal Republic of Germany) VCH Publishers, Inc., New York, NY (USA). doi: 10.1002/9783527619825
6. *Глушко В.П., Гурвич Л.В., Бергман Г.А. и др.* Термодинамические свойства индивидуальных веществ. Справочное издание в четырех томах. М.: Наука, 1982.
7. *Wagman D.D., Evans W.H., Parker V.B. et al.* Washington, DC Journal of Physical and Chemical Reference Data. V. 11. 1982. Supplement № 2. doi: 10.1063/1.555845
8. *Wilhoit R.C., Zwolinski B.J.* // J. Physical and Chemical Reference Data. V. 2. 1973. Supplement № 1. doi: 10.1021/es00107a726
9. *Wilhoit R.C., Chao J., Hall K.R.* Thermodynamic Properties of Key Organic Oxygen Compounds in the Carbon Range C1 to C4. Part 1. Properties of Condensed Phases. Thermodynamics Research Center, Texas A&M University, College Station, Texas 77843. doi: 10.1063/1.555747
10. *Chase M.W., Jr., Curnutt J.L., Downey J.R., Jr. et al.* JANAF Thermochemical Tables, 1982 Supplement. Physicochemical Studies, 1707 Building, The Dow Chemical Company, Midland, Michigan 48640. doi: 10.1063/1.555666
11. *Lowe D.M., Corbett P.T., Murray-Rust P., Glen R.C.* Chemical Name to Structure: OPSIN, an Open Source Solution Journal of Chemical Information and Modeling 2011 51 (3), 739–753. doi: 10.1021/ci100384d
12. [https://en.wikipedia.org/wiki/Simplified\\_molecular-input\\_line-entry\\_system](https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system) (доступ в апреле 2018).