ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ

2017. Том 58, № 4 *Май – июнь* *С. 685 – 691*

УДК 544.169:544.412.2

СООТНОШЕНИЕ "СТРУКТУРА—РЕАКЦИОННАЯ СПОСОБНОСТЬ" В РЕАКЦИЯХ ДИЛЬСА—АЛЬДЕРА С ИСПОЛЬЗОВАНИЕМ ПОДХОДА КОНДЕНСИРОВАННЫХ ГРАФОВ РЕАКЦИЙ

Т.И. Маджидов[1], Т.Р. Гимадиев[1,2], Д.А. Малахова[1], Р.И. Нугманов[1], И.И. Баскин[3], И.С. Антипин[1], А.А. Варнек[1,2]

[1]*Казанский федеральный университет, Россия*
E-mail: Timur.Madzhidov@kpfu.ru
[2]*Страсбургский университет, Франция*
[3]*Московский государственный университет им. М.В. Ломоносова, Россия*

Статья поступила 27 ноября 2016 г.

С использованием структурного представления химической реакции в виде конденсированного графа впервые была построена модель, позволяющая предсказывать константы скорости (lg*k*) реакций Дильса—Альдера, проводимых в различных растворителях и при различных температурах. Полученная модель показывает хорошее согласие предсказанных и экспериментальных значений lg*k*: среднеквадратичная ошибка расчетов составляет менее 0,75 логарифмических единиц. Проведен анализ ошибочных предсказаний, показывающий, что таковые соответствуют реакциям, реагенты которых содержат редко встречаемые структурные фрагменты. Модель доступна для пользователей на сервере https://cimm.kpfu.ru/predictor/.

DOI: 10.15372/JSC20170402

Ключевые слова: [4+2]π-циклоприсоединение, реакция Дильса—Альдера, константа скорости реакции, конденсированный граф реакции, химические реакции, хемоинформатика.

ВВЕДЕНИЕ

Реакции циклоприсоединения являются одними из наиболее распространенных и важных реакций в синтетической химии. Особый интерес к ним объясняется тем, что в ходе реакции образуются ароматические и ненасыщенные кольца, исключительно важные в медицинской химии [1]. Возросший интерес к ним в последнее время объясняется тем, что многие реакции клик-химии [2], особенно те, которые используются в биоортогональной химии [3], являются реакциями циклоприсоединения, например, азид—алкин [4], алкин—нитрон [5], тетразин—алкен [6] и др. Перспективы использования реакции в биоортогональной химии, как правило, определяются скоростью ее протекания с учетом весьма малых концентраций реагентов, которые необходимо использовать для обеспечения биосовместимости [7]. Кроме того, учитывая возможность образования в реакции циклоприсоединения региоизомеров, их соотношение можно оценить, зная константы скорости реакций. Таким образом, исключительно важно умение предсказывать скорости реакций циклоприсоединения. Вообще же, константы скорости реакций позволяют не только оценить динамику химических процессов, но и вычислить выход продуктов и их соотношение.

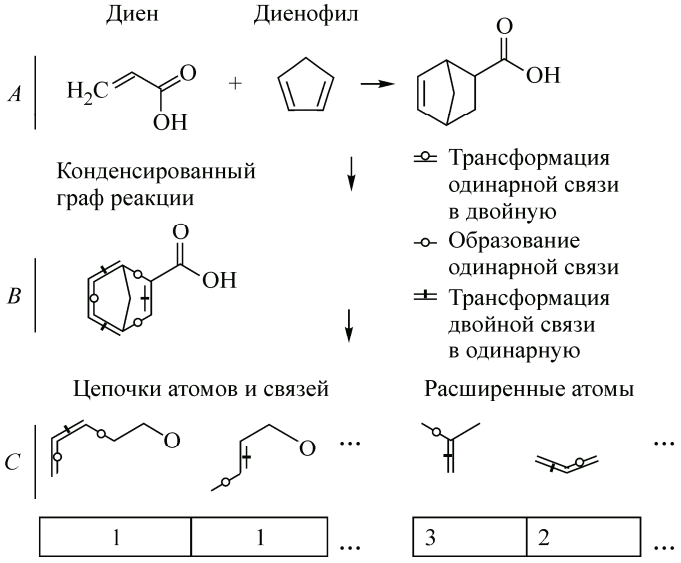

Рис. 1. Пример реакции (*A*), соответствующего ей конденсированного графа (*B*), а также дескрипторов ISIDA на основе цепочек атомов и расширенных атомов (*C*).
Цифры соответствуют встречаемости данного фрагмента

При этом не существует общих подходов, которые могут оценить скорость реакций циклоприсоединения в широком диапазоне растворителей. Использование методов квантовой химии для предсказания скорости и условий проведения реакций циклоприсоединения не являются эффективными по той причине, что проведение точных расчетов непрактично в силу их ресурсоемкости, причем оценка влияния растворителя и конформационно гибкие молекулы во много раз усложняют эту проблему. Относительно успешно с предсказанием констант скоростей справляются методы, которые базируются на использовании простых корреляционных уравнений, констант заместителей и растворителей [8, 9]. Эти закономерности строятся на ограниченном наборе структурно-однородных соединений либо для одного соединения в разных растворителях, что сильно ограничивает области их применения.

Использование средств хемоинформатики позволяет обойти оба указанных ограничения: химическая реакция и условия ее проведения кодируются в виде набора дескрипторов, и использование методов машинного обучения позволяет построить модель, способную предсказывать характеристики реакций в широком диапазоне структур и условий [10—15]. Основная проблема кодирования заключается в том, что реакция представляет собой сложный объект, иерархически связывающий несколько молекул. Подход конденсированного графа реакции, предложенный Г. Вледуцем [16] и развитый А. Варнеком [17, 18], является очень удобным способом представления реакций, поскольку позволяет закодировать всю реакцию в один псевдомолекулярный (конденсированный) граф. В конденсированном графе реакции (КГР) превращение представлено в сжатом виде, где помимо обычных химических связей, не изменяющихся в реакции, присутствуют так называемые динамические связи, кодирующие изменение порядка связи (рис. 1). Для получения конденсированного графа реакции необходимо установление атом-атомного отображения, т.е. соответствия между атомами реагентов и продуктов. Для него в дальнейшем можно использовать стандартные подходы хемоинформатики для расчета дескрипторов и далее строить регрессионные [10—12, 18], классификационные модели [19, 20] для химических реакций, идентифицировать ошибки атом-атомного отображения [21] или уточнять его, используя консенсус нескольких программ [22].

Одними из наиболее интересных и востребованных реакций циклоприсоединения являются реакции [4+2]π-циклоприсоединения, называемые также реакциями Дильса—Альдера. В данной работе впервые проводится моделирование констант скорости реакций Дильса—Альдера с использованием технологий машинного обучения. Пример реакции из базы данных приведен на рис. 1.

В Казанском федеральном университете долгое время велись исследования по изучению реакций циклоприсоединения в группе А.И. Коновалова, в результате чего был собран большой массив данных по скоростям и иным кинетическим параметрам (барьеры, энтропии, объемы активации реакции и иные) разнообразных реакций этого типа в различных условиях. Эти данные послужили основой для создания модели для предсказания скоростей интересующих нас реакций Дильса—Альдера.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

В качестве системы управления базой данных для хранения реакций использовали пакет Instant JChem [27] от компании ChemAxon. Структуры химических соединений, участвующих в реакции, стандартизовались с использованием инструмента Standardizer пакета JChem [28]. Процедура стандартизации включала: ароматизацию структур, удаление изотопов, стандартизацию нитрозогрупп, ароматических N-оксидов, азидов, нитрогрупп, изоцианатов, сульфонов, третичных N-оксидов, удаление явно указанных атомов водорода.

Для проведения атом-атомного отображения также использовали инструмент Standardizer. Ошибки атом-атомного отображения идентифицировали и исправляли вручную. Конденсированные графы реакций генерировали с использованием собственной программы CGR Condenser.

Фрагментные дескрипторы ISIDA для конденсированных графов были рассчитаны с помощью программы Fragmentor [17]. В качестве дескрипторов растворителя использовали константы Каталана (SPP [29], SA [30], SB [31]), константы Камлета—Тафта (α [32], β [33], π^* [34]), а также дескрипторы, характеризующие влияние полярности и поляризуемости растворителя: функция Борна $f_{\text{B}} = \dfrac{\varepsilon - 1}{\varepsilon}$, Кирквуда $f_{\text{K}} = \dfrac{\varepsilon - 1}{2\varepsilon + 1}$, $f_1 = \dfrac{\varepsilon - 1}{\varepsilon + 1}$ и $f_2 = \dfrac{\varepsilon - 1}{\varepsilon + 2}$ (ε — диэлектрическая проницаемость растворителя), $g_1 = \dfrac{n^2 - 1}{n^2 + 2}$, $g_2 = \dfrac{n^2 - 1}{2n^2 + 1}$, $h = \dfrac{(n^2 - 1)(\varepsilon - 1)}{(2n^2 + 1)(2\varepsilon + 1)}$ (n — показатель преломления n_D^{20} растворителя).

В качестве метода машинного обучения использовали регрессию на опорных векторах [24], требующую подбора оптимального ядра сходства, параметра ширины ″трубы″ ε и параметра штрафа C (гиперпараметров). Кроме того, модели, полученные с использованием различных фрагментных дескрипторов, отличаются по качеству. Выбор оптимальных типов дескрипторов и значений гиперпараметров метода SVR проводили с использованием программы SVMOptimizer [25].

В качестве показателей качества моделей использовали среднеквадратическое отклонение (RMSE) и коэффициент детерминации (Q^2) на контрольной выборке:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (x_i^{\text{пред}} - x_i^{\text{эксп}})^2}{N}},$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{N} (x_i^{\text{пред}} - x_i^{\text{эксп}})^2}{\sum_{i=1}^{N} (x_i^{\text{эксп}} - \langle x_i^{\text{эксп}} \rangle)^2},$$

где $x_i^{\text{пред}}$, $x_i^{\text{эксп}}$ — предсказанные и экспериментальные значения lg$k$ для i-й реакции; $\langle x_i^{\text{эксп}} \rangle$ — среднее значение логарифма константы скорости; N — число объектов в выборке. Контрольная выборка создавалась с использованием 10-кратно повторяющейся процедуры 5-кратного скользящего контроля.

Предсказание получалось усреднением предсказаний индивидуальных моделей, полученных с использованием разных поднаборов данных, созданных в ходе процедуры 5-кратного скользящего контроля.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Модели были построены с использованием набора данных по скоростям реакций и условиям их проведения, извлеченных вручную из кандидатских и докторских диссертаций группы академика А.И. Коновалова. Всего было извлечено более 880 реакций, проведенных в различных растворителях. Наиболее распространенными были толуол (347 реакции), хлорбензол (172) и дихлорэтан (170). Наиболее распространенными диенофилами в извлеченных реакциях были малеиновый ангидрид, тетрацианоэтилен и *пара*-бензохинон, среди диенов наиболее часто встречались замещенные фенциклоны, пентацены и тетрацены, а также 2,5-дикарбо-метокси-3,4-дифенилциклопентадиен. Значения констант скоростей реакций циклоприсоединения простира-

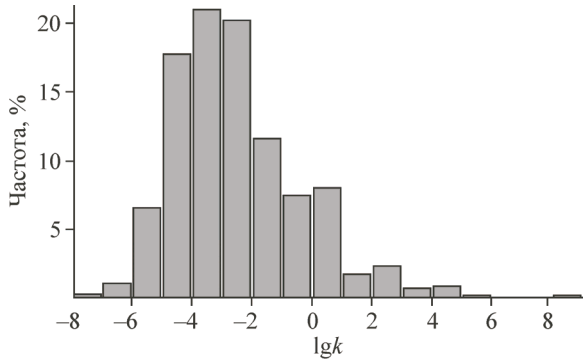

Рис. 2. Гистограмма распределения констант скоростей реакций Дильса—Альдера по частоте встречаемости

ются в широком диапазоне от –7,7 до +8,5 логарифмических единиц. Гистограмма распределения констант скоростей реакций циклоприсоединения по частоте встречаемости в базе данных приведена на рис. 2.

В качестве дескрипторов, характеризующих реакционное превращение, нами были использованы фрагментные дескрипторы ISIDA [17]. Значение каждого дескриптора равно частоте встречаемости заданного фрагмента в молекуле. Дескрипторы ISIDA подсчитывают число всех возможных фрагментов определенной топологии. Имеются две основные топологии фрагментов ISIDA: (i) цепочки атомов заданной длины или (ii) атомы с ближайшим окружением (расширенные атомы) — фрагменты содержат все атомы, удаленные от центрального атома на заданное топологическое расстояние (см. рис. 1). В качестве опций при фрагментации можно оставлять только фрагменты, содержащие хотя бы одну динамическую связь, оставлять только цепочки минимальной длины при наличии нескольких маршрутов, соединяющих данную пару атомов, а также включать или не включать в описание фрагмента информацию о типах атомов, связей, наличии или изменении формальных зарядов на атомах.

Таким образом, для использованного набора данных имелось 616 различных типов дескрипторных описаний, отличающихся только способом фрагментации конденсированного графа реакции Дильса—Альдера.

Из полученных наборов дескрипторов необходимо выбрать лучшие, позволяющие построить модель с наивысшей предсказательной способностью. Контроль прогнозирующей способности модели проводили с использованием процедуры 5-кратного скользящего контроля. Суть перекрестного контроля заключается в том, что вся выборка данных разбивается на 5 равных частей, одна из которых отбирается в так называемую контрольную выборку, остальные 80 % объектов образуют обучающую выборку. На обучающей выборке строится модель. С использованием полученной модели предсказывают характеристики объектов контрольной выборки. Процедура повторяется для каждой из пяти частей по очереди так, что каждый из объектов присутствует в контрольной выборке один раз. В конце предсказания объединяются, и проводится расчет ошибки предсказания. Оценку качества моделей проводили с использованием коэффициента детерминации (Q^2) и среднеквадратичного отклонения предсказанных значений от экспериментальных (RMSE). Для устранения влияния порядка соединений в выборке на показатель качества перед проведением скользящего контроля выборку случайным образом перемешивали. Данную процедуру повторяли 10 раз, показателем качества служило среднее значение по отдельным повторениям.

Для предсказания констант скорости для новых реакций все модели, генерируемые в ходе процедуры скользящего контроля, сохранялись и использовались для прогноза lg$k$. В результате для каждого объекта получаются 50 предсказаний, которые впоследствии усредняются. Этот подход, схожий с баггингом [23], позволяет снизить ошибку предсказания и увеличить стабильность модели за счет устранения влияния случайных флуктуаций, связанных с попаданием в обучающую выборку разных объектов. Качество модели можно оценить, усредняя предсказания для объектов из контрольных выборок. Сравнивая полученные предсказания с экспериментальными данными, рассчитывали RMSE и Q^2 согласно приведенным в экспериментальной части формулам.

В качестве метода машинного обучения использовали регрессию на опорных векторах (SVR) [24]. В рамках данного метода в дескрипторном пространстве строится многомерная ″труба″ с радиусом, задаваемым параметром ε, таким образом, чтобы объекты попадали внутрь нее.

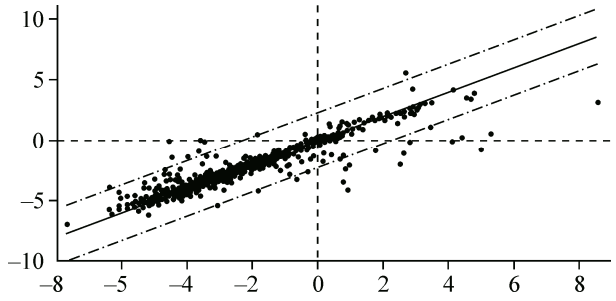

Рис. 3. Предсказанные значения констант скоростей реакций [4+2]π-циклоприсоединения в сравнении с экспериментальными.
Сплошная линия соответствует идеальному совпадению предсказанных и экспериментальных значений. Пунктиром обозначены линии, соответствующие отклонению предсказанных значений от наблюдаемых на 3·RMSE

Если объект оказывается вне ″трубы″, то модель штрафуется пропорционально удаленности объекта от поверхности ″трубы″ и параметра штрафа *C*. С использованием ядер сходства можно строить нелинейные зависимости, сложность которых определяется типом ядра и его параметрами. В работе использовали три типа ядра — линейное, полиномиальное и гауссово. Таким образом, имеется два гиперпараметра (коэффициент *C* и параметр ε) и вид ядра, которые подбираются так, чтобы обеспечить максимальную предсказательную способность модели. Для каждого типа дескриптора оптимальные гиперпараметры и ядра могут отличаться. Поскольку число комбинаций дескрипторов и значений гиперпараметров SVR-модели очень велико, они выбирались с использованием генетического алгоритма [25] таким образом, чтобы обеспечить оптимальную предсказательную способность модели на 5-кратном скользящем контроле.

Наиболее высокую предсказательную способность показали модели, построенные с использованием фрагментных дескрипторов на основе цепочек длиной от одного до шести атомов с учетом информации о зарядах на них. Среднеквадратичное отклонение предсказанных значений от экспериментальных для модели RMSE с использованием консенсусного подхода составляло 0,75 логарифмических единиц. Коэффициент детерминации Q^2 для модели составил 0,87.

График соответствия предсказанных значений экспериментальным представлен на рис. 3. Более качественную модель создать было невозможно, по-видимому, из-за наличия шума в данных, который возникает по нескольким причинам: расхождения в методиках измерений, пренебрежение некоторыми аспектами строения (например, в данной работе мы не принимали во внимание стереохимию присоединения), случайные ошибки. Очевидно, что отклонение предсказанных данных от экспериментальных в среднем не может быть меньше, чем шум в данных. Оценить этот шум сложно, однако полученная нами ошибка предсказания по порядку величины согласуется с полученными ранее моделями для реакций бимолекулярного нуклеофильного замещения [11]. В ней были показаны некоторые примеры расхождения значений констант скорости, полученных в разных статьях, в целом подтверждающие соответствие шума в данных неточности предсказаний. Кроме того, близкая по величине ошибка предсказания была получена в модели, предсказывающей скорость бимолекулярного элиминирования [12]. Таким образом, мы считаем, что качество предсказаний для предложенной модели сопоставимо с воспроизводимостью константы скорости в разных экспериментах.

Для оценки качества работы модели и ее ограничений был проведен анализ объектов, для которых обнаружились значительные (более чем на 3·RMSE) расхождения экспериментальных и предсказанных значений. Анализ таких ошибок прогноза показал, что они обусловлены наличием специфических, редко встречаемых структурных фрагментов (рис. 4). Если реакция, содержащая такие редкие фрагменты, оказывается в тестовой выборке, обучающая выборка, как правило, включает статистически незначимое количество реакций или вообще не включает реакций такого типа. По этой причине влияние таких редких фрагментов на скорость реакции не может быть предсказано, и выдаваемое значение скорости реакции содержит потенциально большую погрешность. Отметим, что такого рода ошибки предсказания могут быть успешно идентифицированы с помощью различных подходов оценки области применимости модели, например, метода контроля фрагментов [26]. В рамках данного подхода, если конденсированный граф некой тестовой реакции содержит фрагмент, который не встречался в графах для объектов обучающей выборки, то предсказание модели, полученной на данной обучающей выборке, считается недостоверным. Поскольку в финальной модели предсказание делается на основе

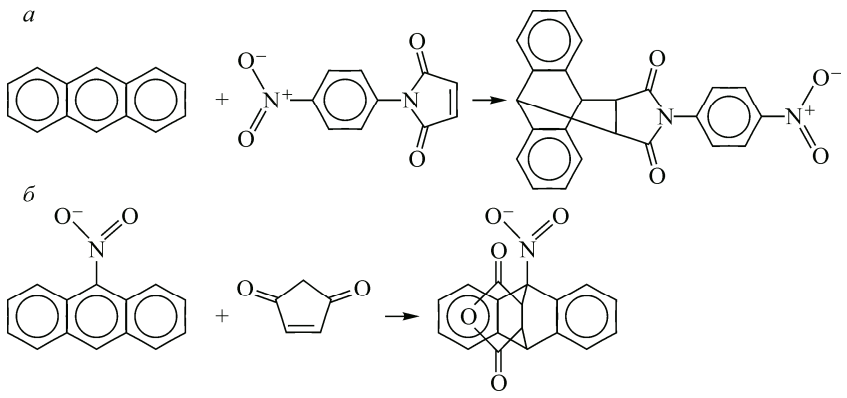

Рис. 4. Примеры ошибок прогноза. Условия реакции (растворитель, температура), предсказанное и экспериментальное значения: толуол, 60 °C, пред.: –3,16, эксп.: 0,76 (*а*); 1,4-диоксан, 130 °C, пред.: –1,92, эксп.: –4,23 (*б*). Обе реакции содержат нитрозаместитель, который отсутствует в других реакциях обучающей выборки

консенсуса множества моделей, полученных на различных наборах объектов, то домен применимости находится для каждой индивидуальной модели в отдельности. Если предсказание для более чем 70 % моделей, входящих в консенсус, оказывается недостоверным, то считается, что предсказание консенсусной модели также недостоверно. Данное пороговое значение было найдено перебором различных возможных значений (от 0 до 100 %) таким образом, чтобы минимизировалась ошибка предсказания в ходе процедуры скользящего контроля.

Полученная модель доступна для пользователей в он-лайн предикторе, доступном на сервере Лаборатории хемоинформатики и молекулярного моделирования КФУ https://cimm.kpfu.ru/predictor/. В редакторе имеется возможность нарисовать или загрузить интересующую реакцию, после чего будет автоматически создано атом-атомное отображение, имеется возможность указать интересующий растворитель и температуру. В результате возвращается предсказание константы скорости реакции, а также сведения об оценке надежности предсказаний (принадлежности предсказываемого объекта домену применимости модели).

ВЫВОДЫ

С использованием подхода конденсированного графа и сгенерированных с его помощью дескрипторов химической трансформации в сочетании с дескрипторами растворителя и температуры нами впервые была получена модель, предсказывающих константу скорости реакций Дильса—Альдера с участием различных реагентов, протекающих во множестве органических растворителей. Было показано, что точность предсказания сопоставима с уровнем экспериментального шума. Анализ ошибок прогноза также показал, что качество модели достаточно высоко для идентификации ошибок в данных и объектов с уникальной структурой по отношению к этому набору реакций. С использованием полученных знаний был реализован способ оценки принадлежности интересующей реакции домену применимости модели. Полученная модель доступна на сервере https://cimm.kpfu.ru/predictor/.

Исследование выполнено за счет гранта Российского научного фонда (проект № 14-43-00024).

Авторы благодарят компанию ChemAxon за предоставленное программное обеспечение.

СПИСОК ЛИТЕРАТУРЫ

1. *Hartenfeller M., Eberle M., Meier P., Nieto-Oberhuber C., Altmann K.-H., Schneider G., Jacoby E., Renner S.* // J. Chem. Inf. Model. – 2011. – **51**, N 12. – P. 3093.
2. *Kolb H.C., Finn M.G., Sharpless K.B.* // Angew. Chem. Int. Ed. Engl. – 2001. – **40**, N 11. – P. 2004.
3. *Sletten E.M., Bertozzi C.R.* // Acc. Chem. Res. – 2011. – **44**, N 9. – P. 666.

4. *Kolb H.C., Sharpless K.B.* // Drug Discov. Today. – 2003. – **8**, N 24. – P. 1128.
5. *MacKenzie D.A., Sherratt A.R., Chigrinova M., Cheung L.L., Pezacki J.P.* // Curr. Opin. Chem. Biol. – 2014. – **21**. – P. 81.
6. *Blackman M.L., Royzen M., Fox J.M.* // J. Am. Chem. Soc. – 2008. – **130**, N 41. – P. 13518.
7. *Gong Y., Pan L.* // Tetrahedron Lett. – 2015. – **56**, N 17. – P. 2123.
8. *Пальм В.А.* Основы количественной теории органических реакций. – Л.: Химия, 1977.
9. *Пальм В.А.* // Успехи химии. – 1961. – **30**, № 9. – С. 1069.
10. *Nugmanov R.I., Madzhidov T.I., Khaliullina G.R. et al.* // J. Struct. Chem. – 2014. – **55**, N 6. – P. 1026.
11. *Madzhidov T.I., Polishchuk P.G., Nugmanov R.I. et al.* // Russ. J. Org. Chem. – 2014. – **50**, N 4. – P. 459.
12. *Madzhidov T.I., Bodrov A.V., Gimadiev T.R. et al.* // J. Struct. Chem. – 2015. – **56**, N 7. – P. 1227.
13. *Kravtsov A.A., Karpov P.V., Baskin I.I. et al.* // Dokl. Chem. – 2011. – **441**, N 1. – P. 314.
14. *Kravtsov A.A., Karpov P.V., Baskin I.I. et al.* // Dokl. Chem. – 2011. – **440**, N 2. – P. 299.
15. *Halberstam N.M., Baskin I.I., Palyulin V.A., Zefirov N.S.* // Mendeleev Commun. – 2002. – **12**, N 5. – P. 185.
16. *Vladutz G.E.* // Inf. Storage Retr. – 1963. – **1**, N 2-3. – P. 117.
17. *Varnek A., Fourches D., Hoonakker F., Solov'ev V.P.* // J. Comput. Aided. Mol. Des. – 2005. – **19**, N 9-10. – P. 693.
18. *Hoonakker F., Lachiche N., Varnek A. et al.* // Int. J. Artif. Intell. Tools. – 2011. – **20**, N 2. – P. 253.
19. *Luca A. De, Horvath D., Marcou G. et al.* // J. Chem. Inf. Model. – 2012. – **52**, N 9. – P. 2325.
20. *Marcou G., Aires de Sousa J., Latino D.A.R.S. et al.* // J. Chem. Inf. Model. – 2015. – **55**, N 2. – P. 239.
21. *Muller C., Marcou G., Horvath D. et al.* // J. Chem. Inf. Model. – 2012. – **52**, N 12. – P. 3116.
22. *Маджидов Т.И., Нугманов Р.И., Гимадиев Т.Р., Лин А.И., Антипин И.С., Варнек А.А.* // Бутлеровские сообщения. – 2015. – **44**, № 12. – С. 170.
23. *Breiman L.* // Mach. Learn. – 1996. – **24**, N 2. – P. 123.
24. *Drucker H., Burges C.J.C., Kaufman L., Smola A., Vapnik V.* Support vector regression machines. Advances in Neural Information Processing Systems / M.C. Mozer, J.I. Jordan, J.I. Petsche, Ed.; MIT Press, 1997. – Vol. 9. – P. 155.
25. *Horvath D., Brown J., Marcou G. et al.* // Challenges. – 2014. – **5**, N 2. – P. 450.
26. *Tetko I.V., Sushko I., Pandey A.K. et al.* // J. Chem. Inf. Model. – 2008. – **48**, N 9. – P. 1733.
27. InstantJChem 15.7.27.0. ChemAxon, http://www.chemaxon.com, 2015.
28. Standardizer, JChem 15.8.3.0. ChemAxon, http://www.chemaxon.com, 2015.
29. *Catalán J., López V., Pérez P. et al.* // Liebigs Ann. – 1995. – **1995**, N 2. – P. 241.
30. *Catalán J., Díaz C.* // Liebigs Ann. – 1997. – **1997**, N 9. – P. 1941.
31. *Catalán J., Díaz C., López V. et al.* // Liebigs Ann. – 1996. – **1996**, N 11. – P. 1785.
32. *Taft R.W., Kamlet M.J.* // J. Am. Chem. Soc. – 1976. – **98**, N 10. – P. 2886.
33. *Kamlet M.J., Taft R.W.* // J. Am. Chem. Soc. – 1976. – **98**, N 2. – P. 377.
34. *Taft R.W., Kamlet M.J.* // J. Am. Chem. Soc. – 1976. – **98**, N 10. – P. 2886.