ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ

2015. Том 56, № 7 *Декабрь* *С. 1293 – 1300*

УДК 544.169:544.412.2

СООТНОШЕНИЕ ″СТРУКТУРА—РЕАКЦИОННАЯ СПОСОБНОСТЬ″ В РЕАКЦИЯХ БИМОЛЕКУЛЯРНОГО ЭЛИМИНИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ ПОДХОДА КОНДЕНСИРОВАННЫХ ГРАФОВ РЕАКЦИЙ

Т.И. Маджидов[1], А.В. Бодров[2], Т.Р. Гимадиев[1,3], Р.И. Нугманов[1], И.С. Антипин[1], А.А. Варнек[1,3]

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

Статья поступила 7 октября 2015 г.

С использованием структурного представления химической реакции в виде конденсированного графа впервые построена модель, позволяющая предсказывать константы скорости реакций бимолекулярного элиминирования. Предложенный подход позволяет предсказывать характеристики реакции в различных растворителях, водно-органических смесях и при различных температурах. Полученная модель показывает хорошее согласие предсказанных и экспериментальных значений, среднеквадратичное отклонение предсказаний от экспериментальных значений составляет менее 0,7 логарифмических единиц. Проведен анализ промахов предсказаний, который показал, что причиной ошибок является в основном несовершенство использованного набора данных, содержащего уникальные реакции. Модель доступна для пользователей на сервере arsole.u-strasbg.fr.

DOI: 10.15372/JSC20150701

Ключевые слова: бимолекулярное элиминирование, константа скорости реакции, конденсированный граф реакции, хемоинформатика, дескрипторы реакции, дескрипторы растворителя.

ВВЕДЕНИЕ

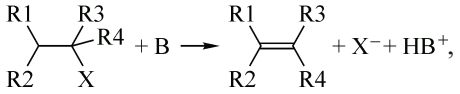
Константы скорости реакций являются исключительно важными характеристиками, которые позволяют не только оценить динамику химических процессов, но и вычислить выход продуктов, оценить селективность конкурирующих процессов и многое другое. В то же время современные химики пользуются во многом феноменологическими правилами для оценки на качественном уровне влияния тех или иных факторов на скорость и селективность реакции.

Несмотря на то, что аппарат моделирования в современной вычислительной химии достаточно развит, предсказание констант скоростей в конденсированной фазе представляет собой исключительно сложную задачу даже для одностадийных процессов, таких как реакции бимолекулярного нуклеофильного замещения и элиминирования. Решение данной проблемы с использованием строгих квантово-химических подходов совершенно непрактично, потому что скорость расчетов весьма невелика, а оценка влияния растворителя во много раз усложняет эту проблему. Достижение химической точности в предсказании требует проведения сложных и крайне ресурсоемких расчетов. Результаты ″слепого″ соревнования SAMPL2 Challenge [1], в котором различные, в основном квантово-химические подходы соревновались в качестве предсказаний

энергий сольватации и разности энергий таутомеров, показали, что достигаемая ими в настоящее время вычислительная точность (среднеквадратичная ошибка расчета энергии составляла ~2,5—3 ккал/моль) недостаточна для удовлетворительного описания энергии и сильно уступает экспериментальной. При этом почти все используемые методы имели построечные параметры, которые подбирались под предоставленный в ходе соревнования тренировочный набор. В расчете констант скоростей реакции следует ожидать еще менее точных предсказаний.

Относительно успешно с предсказанием констант скоростей справлялись методы, которые базируются на использовании простых корреляционных уравнений и констант заместителей и растворителей [2]. То есть практически проще извлекать закономерности из имеющихся данных, нежели использовать дедуктивные подходы как методы квантовой химии. Проблемой их является ограниченная применимость моделей: в основном они способны давать предсказания для одного класса соединений с различными заместителями либо для одного соединения в разных растворителях.

В последнее время к решению этой проблемы подошли с использованием методов хемоинформатики [3]. Закодировав химическую реакцию в виде строки численных дескрипторов, можно искать сложные нелинейные закономерности в данных. Основная проблема кодирования заключается в том, что реакция представляет собой сложный объект, иерархически связывающий несколько молекул. Подход конденсированного графа реакции [4, 5] является многообещающим способом решения проблемы предсказания характеристик химических реакций, поскольку позволяет очень органично использовать стандартные подходы хемоинформатики для моделирования связи структуры с реакционной способностью. До настоящего времени в литературе опубликованы только модели, которые позволяют предсказывать константы скорости реакций нуклеофильного замещения [6—10]. В данной работе впервые проводится моделирование констант скорости реакции бимолекулярного элиминирования (E2 реакций). Общая схема изученных реакций элиминирования имеет следующий вид:



где X — уходящая группа (чаще всего — атом галогена); B — основание (N-, O- или S-содержащее основание: анион или нейтральная молекула); R1—R4 — заместители.

Центральным элементом предложенного подхода является использование конденсированного графа реакции. Последний представляет собой обычный псевдомолекулярный граф (структурную диаграмму), на котором помечены образующиеся и разрывающиеся химические связи (рис. 1). Для получения конденсированного графа реакции необходимо установление атом-атомного отображения, т.е. установление соответствия между атомами реагентов и продуктов (см. рис. 1), после чего наложением атомов реагентов и продуктов с одинаковыми номерами идентифицируются динамические связи.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для построения модели химические превращения, соответствующие E2 реакциям, условия их проведения и значения логарифмов констант скоростей ($\lg k_2$) были вручную извлечены из справочника под редакцией Пальма [2]. Всего было извлечено 1043 реакции, 97 из которых проведены в воде, 213 — в водно-органических смесях и 733 — в органических растворителях.

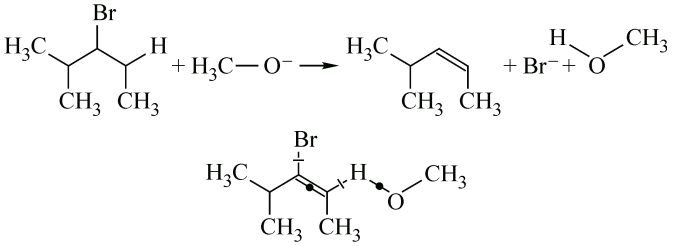


Рис. 1. Пример конденсированного графа (внизу), соответствующий реакции элиминирования (вверху).

Кружком • обозначены одинарная связь O—H, а также двойная связь C=C, образованная из одинарной. Зачеркнутые связи обозначают разрывающиеся одинарные связи C—H и C—Br

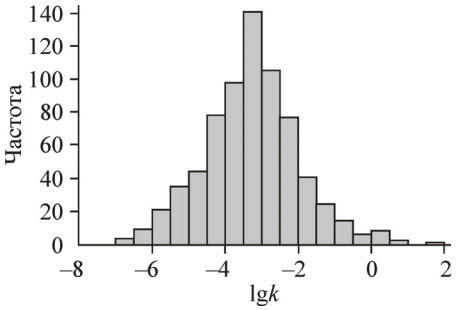


Рис. 2. Гистограмма распределения констант скоростей E2 реакций по частоте встречаемости

Значения констант скоростей E2 реакций простираются в диапазоне от –7,22 до 2,16 логарифмических единиц. Гистограмма распределения констант скоростей E2 реакций по частоте приведена на рис. 2. Обращает на себя внимание близкий к нормальному характер распределения этой характеристики с центром в районе –3 единиц.

В качестве дескрипторов, характеризующих реакционное превращение, нами были использованы фрагментные дескрипторы ISIDA [4]. Значение каждого дескриптора равно частоте встречаемости заданного фрагмента в молекуле. Дескрипторы ISIDA подсчитывают число всех возможных фрагментов определенной топологии (цепочек атомов определенной длины или расширенных атомов — атомов с ближайшим окружением, рис. 3). В качестве опций можно оставлять только фрагменты, содержащие динамическую связь, только кратчайшие цепочки, включать или не включать в описание фрагмента информацию о типах атомов, связей, наличии или изменении формальных зарядов на атомах. Всего с использованием всех возможных комбинаций опций было сгенерировано 260 наборов, включающих от нескольких десятков до нескольких тысяч дескрипторов, отвечающих за наличие того или иного фрагмента.

Поскольку скорость реакции зависит также от температуры и растворителя, в число дескрипторов для моделирования включали обратную температуру проведения реакции и 14 характеристик растворителя. В качестве последних использовали параметры полярности, поляризуемости, H-акцепторности и H-донорности растворителя (см. Экспериментальную часть). Для описания водно-органических смесей в качестве дескриптора была добавлена мольная доля органического растворителя в смеси.

Из полученных наборов дескрипторов необходимо выбрать лучшие, позволяющие построить модель с наивысшей предсказательной способностью. Прогнозирующую способность модели проверяли с использованием процедуры 5-кратного скользящего перекрестного контроля. Для этого всю выборку данных разбивали на 5 частей, одну из них отбирали в контрольную выборку, на остальных объектах строили модель. Полученную модель применяли на контрольной выборке. Процедуру повторяли для каждой из пяти частей по очереди, так что каждый из объектов один раз присутствовал в контрольной выборке. В конце предсказания объединяли и проводили расчет ошибки. Оценку качества моделей проводили с использованием коэффициента детерминации (Q^2) и среднеквадратичного отклонения предсказанных значений от экспериментальных (RMSE — Root-Mean Square Error). Всю процедуру 5-кратного скользящего контроля повторяли 30 раз. В качестве финальной использовали так называемую консенсусную модель, в которой все промежуточные модели, генерируемые в ходе скользящего контроля, сохраняются и могут использоваться для прогноза $\lg k_2$ для новых реакций. В результате для каждого объекта получается 150 предсказаний, которые усредняются, что позволяет снизить ошибку предсказания и увеличить стабильность модели за счет уменьшения флуктуаций, связанных с попаданием в обучающую выборку разных объектов. Качество консенсусной модели мы оценивали усреднением предсказаний объектов из контрольных выборок с последующим расчетом $\mathrm{RMSE_{cons}}$ и Q^2_{cons}, согласно обычным формулам. Поскольку всегда во внимание принимаются

Рис. 3. Примеры дескрипторов ISIDA на основе цепочек атомов и расширенных атомов, внизу приведена встречаемость фрагмента, образующая дескрипторную строку

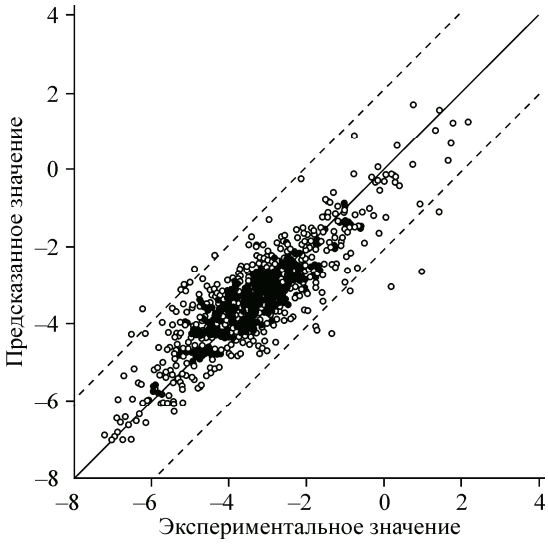


Рис. 4. Предсказанные значения констант скоростей E2 реакций в сравнении с экспериментальными.
Сплошная линия соответствует идеальному совпадению предсказанных и экспериментальных значений. Пунктиром обозначены линии, соответствующие отклонению предсказанных значений от наблюдаемых на $3RMSE_{cons}$

характеристики, прогнозируемые моделью на контрольной выборке, то показатели $RMSE_{cons}$ и Q^2_{cons} позволяют оценить качество предсказания новых данных.

В качестве метода машинного обучения использовали метод регрессии на опорных векторах (SVR) [11]. Данный метод пытается провести в дескрипторном пространстве гибкую многомерную трубу определенного радиуса таким образом, чтобы обеспечить максимальное попадание объектов внутрь трубы. У метода имеется три гиперпараметра (коэффициент *C*, характеризующий штраф за непопадание внутрь трубы, параметр ядра γ и параметр ε, характеризующий радиус трубы), значение которых подбирается так, чтобы обеспечить максимальную предсказательную способность модели. Таким образом, имеется несколько параметров, которые необходимо подобрать при моделировании: набор дескрипторов и гиперпараметры SVR. Оптимальные параметры подбирали с использованием эволюционного алгоритма [12] таким образом, чтобы обеспечить оптимальную предсказательную способность модели с использованием 5-кратного скользящего контроля.

Наиболее высокую предсказательную способность модели обеспечивали дескрипторы на основе цепочек атомов длиной от двух до шести атомов с учетом информации зарядов на атомах. Среднеквадратичное отклонение предсказанных значений от экспериментальных для консенсусной модели ($RMSE_{cons}$) составляло 0,69 логарифмических единиц, несколько хуже такового для описанной ранее модели для S_N2 (около 0,5) [6, 7]. Коэффициент детерминации (Q^2_{cons}) для консенсусной модели оставил 0,75. График соответствия предсказанных значений экспериментальным для консенсусных предсказаний представлен на рис. 4. Из приведенных данных и рис. 4 видно, что $\lg k_2$ предсказывается достаточно хорошо. В базе данных содержатся измерения константы скорости, проведенные различными методами и в различных лабораториях. Наличие расхождений в определении констант скоростей приводит к возникновению ″шума″ в данных. Поскольку использованные методы извлекают зависимости из экспериментальных данных, то неопределенность предсказания из-за шума данных возникает в самой модели. К сожалению, в базе данных не удалось обнаружить одинаковых реакций, проведенных в одних и тех же условиях, но в разных лабораториях, поэтому выполнить анализ межлабораторных ошибок измерения не удалось. Однако мы считаем, что ошибки измерения не могут быть меньше аналогичной величины для S_N2 реакций, которая достигала 0,5 и более логарифмических единиц [6]. Таким образом, мы можем считать, что качество предсказаний для предложенной модели сопоставимо с точностью измерений константы скорости.

Для анализа качества работы модели был проведен поиск объектов, для которых обнаруживается значительное отклонение прогнозируемых значений от наблюдаемых. Если разница между предсказанным и экспериментальным значением превышала три среднеквадратичных отклонения для консенсусного прогноза, то считалось, что имеет место ″промах″ модели. Всего было обнаружено 14 реакций такого типа (рис. 5). Анализ ″промахов″ позволил выявить следующие основные причины их появления:

- При наличии недостаточного числа фрагментов определенного типа модель не может построить обобщения относительно его вклада в общую константу скорости реакции. Очевидно,

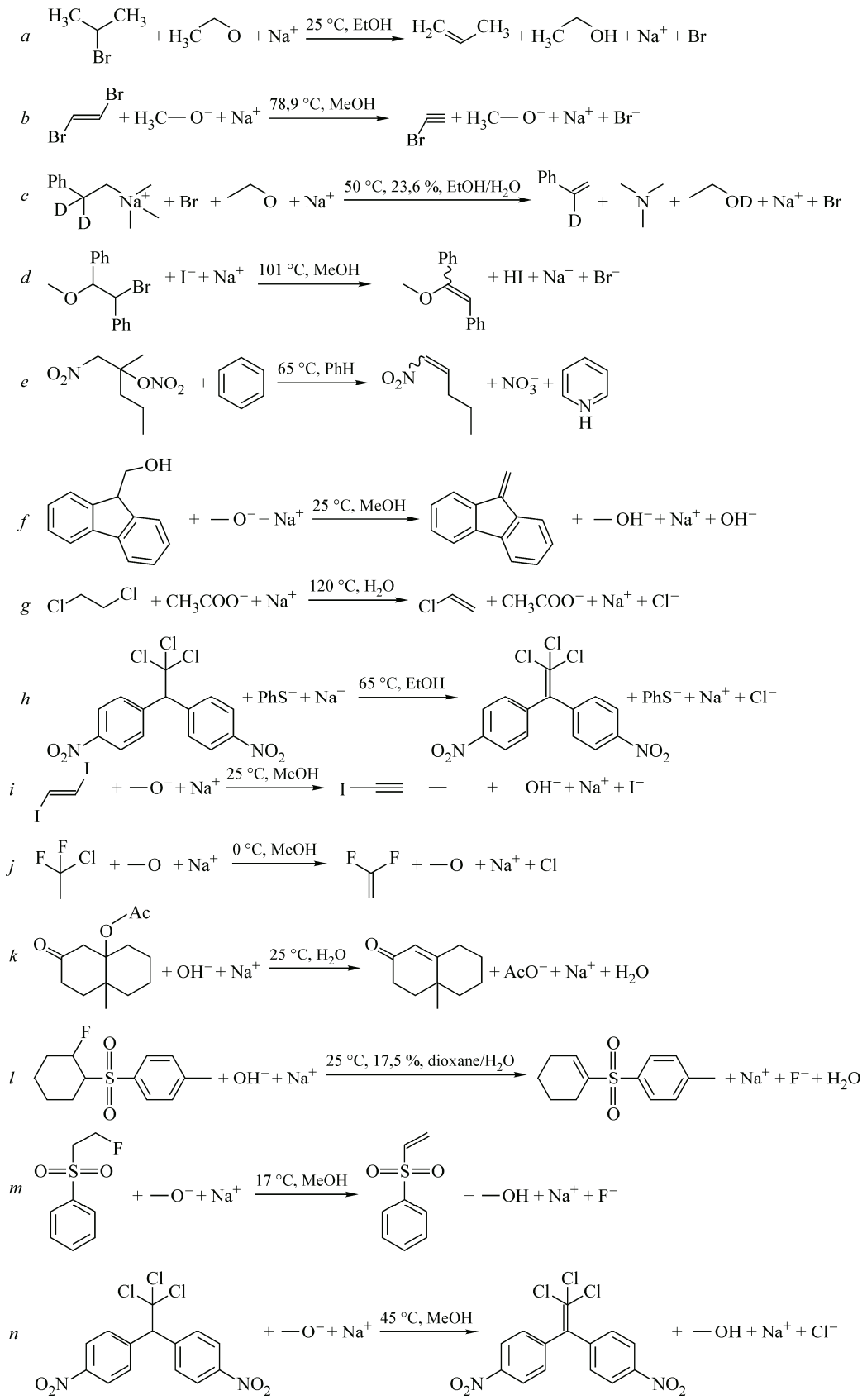


Рис. 5. Реакции, для которых получено сильное отклонение предсказанных значений константы скорости от экспериментальных

что если в реакции в контрольной выборке имелся фрагмент, который существенно влиял на значение константы, и при этом реакций с таким же фрагментом не имелось в обучающей выборке, то не стоит ожидать качественного предсказания искомой характеристики. По этой причине константа скорости предсказывалась некорректно для семи реакций (*c*, *d*, *e*, *h*, *j*, *m*, *n*). Обнаружили несколько причин возникновения уникальных фрагментов в дескрипторном представлении реакций. К примеру, в реакции *m* некорректное атом-атомное отображение привело к созданию неправильного конденсированного графа реакции, в результате чего дескрипторная строка сильно отличалась от таковой для структурно близких реакций. Несмотря на то, что атом-атомное отображение проверялось, имела место ошибка ручного анализа. Еще одна ошибка предсказания (реакция *n*) была вызвана проблемой в автоматической стандартизации представления нитрогрупп, в результате чего нитрогруппа оказалась представленной уникальным образом. В реакциях *a*, *c*, *d*, *e*, *h*, *j*, *k* присутствовали уникальные фрагменты ввиду специфики самой базы данных. Например, реакция *d* представляет собой единственный пример использования иодид-аниона в качестве основания, реакция *h* — дифенилтрихлорметилметана с акцепторными заместителями (заметим, что похожая реакция *n* была неправильно представлена), ситуация дополнительно осложняется использованием редкого тиофенолят-аниона в качестве основания. В реакции *j* в качестве субстрата используется дифторхлорзамещенное, тогда как в остальных случаях имеются только трифторзамещенные или монохлорзамещенные. Аналогичные проблемы можно отметить с другими реакциями *c* (имелся дейтерий), *e* (уходящая группа — нитрат).

- Если реакции с очень большими или очень малыми (по отношению к обучающей выборке) значениями $\lg k_2$ оказываются в контрольной выборке, предсказания потенциально подвержены ошибке вследствие проблем эктраполяции. Особенно предсказание осложняется для малых по размерам реагентов, содержащих небольшое число молекулярных фрагментов, используемых в качестве дескрипторов. Для реакции *a* имеет место ошибка подобного типа.
- Реакция *f*, по-видимому, была ошибочно отнесена автором справочника к типу E2 и после этого извлечена в базу. Эта реакция дегидратации едва ли проходит по одностадийному механизму E2. Всего в базе пять подобных реакций, но поскольку оставшиеся четыре имеют значения скорости, близкие к –3, то модель в их случае дает предсказания, близкие к истинным. Неизвестные примеры SVR зачастую предсказывает значением, близким к средней величине предсказываемой характеристики по выборке (около –3).
- Еще один пример, представленный реакцией *g*, попал в число плохо предсказываемых объектов, по-видимому, ввиду специфики реакционных условий. Данная реакция очевидно проводилась в условиях автоклавирования, тогда как все остальные реакции в базе — при обычном давлении.

Несовершенство используемой техники моделирования также являлось одной из причин возникновения значительных ошибок в предсказании. В предложенной модели не учитывалась стереохимия, поскольку ее учет представляет собой исключительно сложную задачу и потребует, по-видимому, разработки новых дескрипторов. Чаще всего стереохимия реакционного центра несущественно влияла на константу скорости, ее влияние было меньше, чем ошибка предсказания. Однако константа скорости реакций *b*, *i*, представляющих собой дегидрогалогенирование при двойной связи с *транс*-конфигурацией, была существенно выше, чем для аналогов с *цис*-конфигурацией. В результате этого значение константы скорости сильно недооценивалось. Другой проблемой было влияние очень маленьких фрагментов. Поскольку в результате отбора оптимальных наборов дескрипторов были выбраны цепочки длиной от двух атомов, то при наличии в реакционном уравнении структур, состоящих из одного тяжелого атома (атомы водорода явно не учитывались), информация об этой молекуле теряется. Так, в реакциях *l* и *k* в качестве основания выступал гидроксид-анион, который оказывался неучтенным при моделировании. В большинстве своем одноатомные структуры представлены близкими по активности хлорид- и бромид-ионами, поэтому отсутствие в дескрипторах признаков, соответствующих основанию, интерпретировалось моделью как наличие Br^- и Cl^-. Более высокая активность

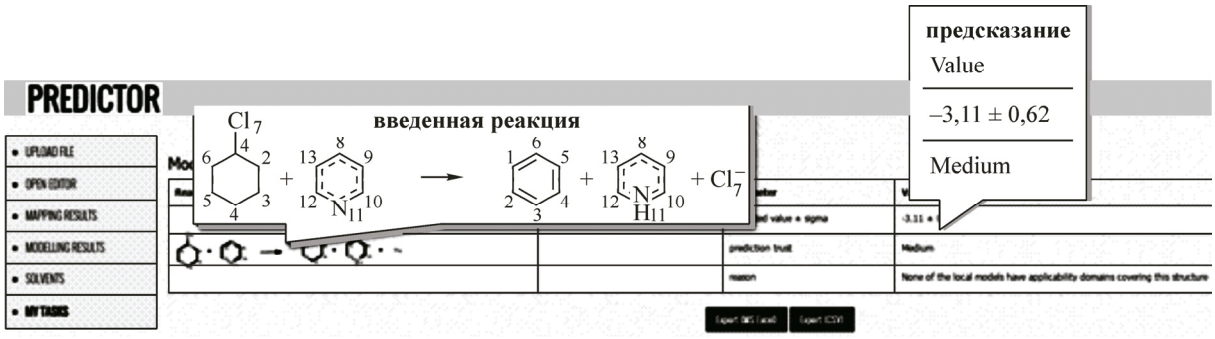


Рис. 6. Пример выдачи результатов предсказания на веб-сервере. В числе результатов выдается не только предсказываемая величина, но и оценка качества предсказания

гидроксид-иона приводила к возникновению ошибки. Этот эффект, по-видимому, также оказал влияние на появление в списке промахов реакции *d*.

Модель доступна для пользователей в онлайн предикторе на сервере arsole.u-strasbg.fr. В редакторе пользователь рисует интересующую реакцию, сервер автоматически создает атом-атомное отображение, после чего пользователь указывает интересующий растворитель и температуру (рис. 6). Программа выдает предсказанное значение логарифма константы скорости реакции с указанием оценки качества предсказаний (оптимальное, среднее или плохое).

ВЫВОДЫ

Таким образом, впервые с использованием подхода конденсированного графа и сгенерированных на его основе фрагментных дескрипторов нами были получены предсказательные модели связи структуры конденсированного графа, описывающего химическую реакцию, и условий, в которых она была проведена, с логарифмом константы скорости реакции бимолекулярного элиминирования. В отличие от других подходов предложенный метод позволяет проводить прогнозирование константы скорости реакций типа E2 для структурно разнородных субстратов и оснований в различных условиях: при различной температуре, различных органических растворителях и водно-органических смесях. Анализ ″промахов″ прогнозов показал, что качество модели достаточно высоко для идентификации ошибок во введенных данных и так называемых синглтонов, объектов с уникальной для данной базы данных структурой. Модель доступна для пользователей на сервере arsole.u-strasbg.fr.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

В качестве системы управления базой данных для хранения реакций использовался пакет Instant JChem [13] от компании ChemAxon. Структуры химических соединений, участвующих в реакции, стандартизовались с использованием инструмента Standardizer пакета JChem [14]. 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 использовался инструмент Standardizer. Ошибки атом-атомного отображения идентифицировались и исправлялись вручную. Конденсированные графы реакций генерировались с использованием собственной программы CGR Condenser.

Дескрипторы ISIDA для конденсированных графов были сгенерированы с использованием программы Fragmenter [15]. В качестве дескрипторов растворителя использовались: SPP [16], SA [17], SB [18] константы Каталана, α [19], β [20], π^* [21] константы Камлета—Тафта. Были взяты также следующие дескрипторы, характеризующие влияние полярности и поляризуемости растворителя: функция Борна $f_B = \frac{\varepsilon - 1}{\varepsilon}$, функция Кирквуда $f_K = \frac{\varepsilon - 1}{2\varepsilon + 1}$, функ-

ции $f_1 = \frac{\varepsilon - 1}{\varepsilon + 1}$, и $f_2 = \frac{\varepsilon - 1}{\varepsilon + 2}$ (ε — диэлектрическая проницаемость растворителя), $g_1 = \frac{n^2 - 1}{n^2 + 2}$, $g_2 = \frac{n^2 - 1}{2n^2 + 1}$, $h = \frac{(n^2 - 1)(\varepsilon - 1)}{(2n^2 + 1)(2\varepsilon + 1)}$ (как n был обозначен показатель преломления n_D^{20} растворителя). В качестве дескриптора смесей добавлялась мольная доля органического растворителя в водно-органической фазе. Он был равен 100 % для чистого растворителя.

Отбор дескрипторов и оптимальных значений гиперпараметров метода SVR проводили с использованием программы SVM Optimizer [12].

Расчет статистических характеристик качества предсказаний проводили с использованием следующих формул:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\chi_i^{\text{пред}} - \chi_i^{\text{эксп}})^2}{N}}, \qquad Q^2 = 1 - \frac{\sum_{i=1}^{N} (\chi_i^{\text{пред}} - \chi_i^{\text{эксп}})^2}{\sum_{i=1}^{N} (\chi_i^{\text{эксп}} - \overline{\chi_i^{\text{эксп}}})^2},$$

где $\chi_i^{\text{пред}}$, $\chi_i^{\text{эксп}}$ — предсказанные и экспериментальные значения $\lg k_2$ для i-й реакции; $\overline{\chi_i^{\text{эксп}}}$ — среднее значение логарифма константы скорости; N — число объектов в выборке.

Авторы выражают благодарность Российскому научному фонду (соглашение № 14-43-00024) за поддержку исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Geballe M., Skillman A.G., Nicholls A. et al.* // J. Comput. Aided Mol. Des. – 2010. – **24**, N 4. – P. 259.
2. *Пальм В.А.* // Успехи химии. – 1961. – **30**, № 9. – С. 1069.
3. *Varnek A., Baskin I.I.* // Mol. Inform. – 2011. – **30**, N 1. – P. 20.
4. *Varnek A., Fourches D., Hoonakker F. et al.* // J. Comput. Aided Mol. Des. – 2005. – **19**, N 9-10. – P. 693.
5. *Fujita S.* // J. Chem. Inf. Model. – 1986. – **26**, N 4. – P. 205.
6. *Madzhidov T.I., Polishchuk P.G., Nugmanov R.I. et al.* // Russ. J. Org. Chem. – 2014. – **50**, N 4. – P. 459.
7. *Nugmanov R.I., Madzhidov T.I., Khaliullina G.R. et al.* // J. Struct. Chem. – 2014. – **55**, N 6. – P. 1026.
8. *Hoonakker F., Lachiche N., Varnek A. et al.* // Int. J. Artif. Intell. Tools. – 2011. – **20**, N 2. – P. 253.
9. *Kravtsov A.A., Karpov P.V., Baskin I.I. et al.* // Dokl. Chem. – 2011. – **440**, N 2. – P. 299.
10. *Kravtsov A.A., Karpov P.V., Baskin I.I. et al.* // Dokl. Chem. – 2011. – **441**, N 1. – P. 314.
11. *Drucker H., Burges C.J.C., Kaufman L., Smola A., Vapnik V.* Support vector regression machines. In Advances in Neural Information Processing Systems / Ed. M.C. Mozer, J.I. Jordan, J.I. Petsche, MIT Press, 1997. – Vol. 9. – P. 155.
12. *Horvath D., Brown J., Marcou G. et al.* // Challenges. – 2014. – **5**, N 2. – P. 450.
13. InstantJChem 15.7.27.0. ChemAxon, http://www.chemaxon.com, 2015.
14. Standardizer, JChem 15.8.3.0. ChemAxon, http://www.chemaxon.com, 2015.
15. *Marcou G., Solov'ev V., Horvath D., Varnek A.* ISIDA Fragmentor2011-User Manual, 2012.
16. *Catalán J., López V., Pérez P. et al.* // Liebigs Ann. – 1995. – **1995**, N 2. – P. 241.
17. *Catalán J., Díaz C.* // Liebigs Ann. – 1997. – **1997**, N 9. – P. 1941.
18. *Catalán J., Díaz C., López V. et al.* // Liebigs Ann. – 1996. – **1996**, N 11. – P. 1785.
19. *Taft R.W., Kamlet M.J.* // J. Am. Chem. Soc. – 1976. – **98**, N 10. – P. 2886.
20. *Kamlet M.J., Taft R.W.* // J. Am. Chem. Soc. – 1976. – **98**, N 2. – P. 377.
21. *Kamlet M.J., Abboud J.L., Taft R.W.* // J. Am. Chem. Soc. – 1977. – **99**, N 18. – P. 6027.