ЖУРНАЛ ФИЗИЧЕСКОЙ ХИМИИ, 2025, том 99, № 1, с. 144–152

ФИЗИЧЕСКАЯ ХИМИЯ ПРОЦЕССОВ РАЗДЕЛЕНИЯ. ХРОМАТОГРАФИЯ

УДК 544.03, 543.544.3, 004.8

ПРИМЕНЕНИЕ МОЛЕКУЛЯРНОГО ПОДОБИЯ ДЛЯ ОЦЕНКИ ТОЧНОСТИ ПРЕДСКАЗАНИЯ ГАЗОХРОМАТОГРАФИЧЕСКИХ ИНДЕКСОВ УДЕРЖИВАНИЯ С ПОМОЩЬЮ ГЛУБОКОГО ОБУЧЕНИЯ

© 2025 г. Д. Д. Матюшин[a], А. Ю. Шолохова[a, *], М. Д. Хрисанфов[a, b], С. А. Боровикова[a]

[a]*ФГБУН Институт физической химии и электрохимии им. А. Н. Фрумкина РАН, 119071, Москва, Россия*

[b]*МГУ им. М. В. Ломоносова, Химический факультет, 119991, Москва, Россия*

**e-mail: shonastya@yandex.ru*

Поступила в редакцию 25.03.2024 г.
После доработки 22.05.2024 г.
Принята к публикации 24.05.2024 г.

При предсказании индексов удерживания с помощью глубокого обучения обычно нет способа оценить надежность предсказания для конкретной молекулы. В данной работе на примере неподвижных фаз на основе полиэтиленгликоля и базы данных NIST 17 показано, что в среднем предсказание тем точнее, чем более близкая по структуре к соединению, для которого выполняется предсказание, молекула находилась в обучающем наборе данных. Сходство по Танимото "молекулярных отпечатков пальцев" ECFP – наиболее подходящий для этой задачи алгоритм вычисления молекулярного подобия из четырех рассмотренных. Показано, что для ряда продуктов трансформации несимметричного диметилгидразина, структура которых была установлена с использованием такого предсказания, оно могло быть весьма ненадежным.

Ключевые слова: газовая хроматография, индексы удерживания, машинное обучение, глубокое обучение, молекулярное подобие

DOI: 10.31857/S0044453725010146, **EDN**: EHWTZH

ВВЕДЕНИЕ

Время удерживания в газовой хроматографии зависит от скорости потока газа-носителя, геометрических параметров хроматографической колонки, температурной программы и других факторов. В то же время индекс удерживания [1], характеризующий время удерживания вещества относительно времен удерживания *н*-алканов, зависит главным образом от структуры удерживаемого соединения и химической природы неподвижной фазы [1–3]. Таким образом, задача предсказания индекса удерживания для данной молекулы и данной неподвижной фазы – это задача предсказания одного единственного числа по структуре молекулы.

При хромато-масс-спектрометрическом анализе сложной смеси, содержащей неизвестные компоненты, предположение о структуре неизвестного соединения делается на основе масс-спектра, наиболее часто с помощью библиотечного поиска [4, 5]. Однако библиотечный поиск часто приводит к неверному результату, даже если рассматриваемое соединение содержится в базе данных [6]. В тех случаях, когда определяемые соединения в базах данных отсутствуют, задача становится еще более сложной [7]. Однако сопоставление наблюдаемого индекса удерживания с предсказанным с помощью машинного обучения позволяет отбросить неверных кандидатов [6, 8, 9] и подтвердить предварительную идентификацию [9–12]. Использование индексов удерживания существенно повышает надежность идентификации [6, 9]. Экспериментальные данные об индексах удерживания доступны лишь примерно для ста тысяч молекул [13], что в несколько раз меньше, чем количество молекул, для которых доступны экспериментальные масс-спектры, и на несколько порядков меньше, чем общее количество известных молекул. Таким образом, предсказание индексов удерживания – важная задача для современной химии.

Глубокое обучение, т.е. совокупность статистических методов, основанных на глубоких нейронных сетях, произвело революцию во многих областях науки и техники в последние годы. Глубокие нейронные сети используются для самых разных задач от аналитической химии [14] до задач машинного зрения и машинного перевода [15]. В частности, глубокое обучение применяется для предсказания газохроматографических индексов удерживания [13, 16–20] по структуре молекулы.

В последние годы был разработан целый ряд моделей такого типа [18]. Глубокое обучение существенно превосходит по точности ранее применявшиеся модели [16, 17]. В целом ряде работ [9–12] такие предсказанные индексы удерживания используются для уточнения идентификации.

Оценка точности моделей предсказания индексов удерживания проводится с использованием больших наборов данных и рассчитывается "средняя" метрика точности для всего набора данных [16–20] (например, среднеквадратичное или среднее абсолютное отклонение). Однако это совершенно не позволяет оценить, является ли точным предсказание для конкретной отдельно взятой молекулы. В некоторых работах точность рассчитывается для отдельных классов соединений [18, 19], однако и в этом случае классы (например, "ароматические соединения", "триметилсилильные производные") являются достаточно широкими и включают в себя самые разные молекулы. В связи с этим актуальна разработка способов, с помощью которых можно оценить, является ли надежным предсказание индекса удерживания для данной конкретной молекулы, т.е. способов оценить, можно ли доверять данному предсказанию. Использование предсказанных индексов удерживания может привести к неверным результатам, если именно для рассматриваемых молекул предсказание весьма ненадежно. Недавно для этой задачи был разработан подход, использующий сравнение между собой предсказаний, сделанных с помощью нескольких независимых моделей [21].

Существуют различные методы количественной оценки того, насколько структуры двух молекул близки между собой, т.е. оценки молекулярного подобия [22–25]. При этом, в частности, может использоваться сходство так называемых "молекулярных отпечатков пальцев" [25] (двоичных векторов, каждый бит которых показывает, содержится ли в молекуле тот или иной фрагмент), а также нахождение общего подграфа между двумя молекулами [22].

Целью данной работы является изучение того, как молекулярное подобие между молекулой, для которой выполняется предсказание индекса удерживания с помощью глубокого обучения, и молекулами, содержащимися в обучающем наборе данных, использованном для обучения модели, влияет на точность предсказания индекса удерживания. Данное исследование выполняется на примере индексов удерживания для полярных неподвижных фаз (тип "Standard polar" в базе данных NIST; полиэтиленгликоль и приблизительно эквивалентные по хроматографическому поведению полимеры на его основе) и ранее опубликованной модели глубокого обучения, встроенной в программное обеспечение SVEKLA [9, 16]. Также целью данной работы является предварительная оценка того, являются ли надежными предсказания индексов удерживания, сделанные в работе [9] и использованные для построения структуры новых продуктов трансформации несимметричного диметилгидразина.

МЕТОДЫ

Набор данных и модель глубокого обучения

В качестве набора данных использовалась база данных NIST 17. Процедура обработки и подготовки данных описана в предыдущей работе [16]. Набор данных был разбит на 5 наборов случайным образом. Обучение моделей было выполнено 5 раз, каждый раз 4 набора использовались в качестве обучающих, а пятый в качестве тестового. Результаты предсказаний для тестовых наборов (соединение, для которого выполняется предсказание, каждый раз отсутствует в обучающем наборе) были объединены и использованы для дальнейшей работы (5-fold кросс-валидация).

Были обучены две модели: одномерная сверточная нейронная сеть и глубокий многослойный перцептрон. Подробные описания моделей даны в работах [13, 16]. При этом использовалось трансферное обучение: сначала нейронные сети обучались для предсказания индексов удерживания для неполярных неподвижных фаз, а затем полученные веса нейронных сетей использовались в качестве начальных значений для обучения модели для предсказания индексов удерживания для неполярных неподвижных фаз. Молекулы, входящие в тестовый набор, каждый раз удалялись и из набора данных о индексах удерживания для неполярных неподвижных фаз, использованных для обучения. Таким образом, не происходило "утечки данных", то есть молекулы, использованные для тестирования, не использовались на обучении ни на каком этапе. Подробно процедура обучения описана в предыдущей работе [16].

База данных NIST 17 содержит по несколько записей данных для каждой из молекул. При обучении и тестировании использовались все эти записи (они отличаются тем, какая именно хроматографическая колонка использовалась, а также условиями измерения). После выполнения процедуры кросс-валидации для каждой записи имеется пара значений: экспериментальный индекс удерживания и предсказанный с помощью модели, которая "не видела" данную молекулу при обучении. При этом разбитие исходной базы данных на 5 наборов выполнялось так, что все записи для каждой из молекул помещались в один из наборов, выбранный случайным образом. Геометрические изомеры и стереоизомеры рассматривались как одна молекула. Более подробное описание процедур и алгоритмов содержится в ранее опубликованных работах [13, 16–17].

Расчет молекулярного подобия

Исходный набор данных содержал 89086 записей, каждая из которых содержала структуру молекулы, референсный и предсказанный индекс удерживания. Для каждой структуры было найдено медианное значение референсного индекса удерживания. Таким образом был получен набор данных, содержащий 9408 записей, состоящих из структуры молекулы, референсного и предсказанного значения. Каждая молекула встречается в данном наборе ровно один раз.

Для каждой молекулы были рассчитаны "молекулярные отпечатки пальцев" (векторы, показывающие наличие тех или иных фрагментов) с помощью алгоритма ECFP [25] (радиус 3, длина вектора 1024). Для каждой пары молекул вычислено сходство "молекулярных отпечатков пальцев" по Танимото:

$$S = \frac{N_{AB}}{N_A + N_B - N_{AB}}, \qquad (1)$$

где N_A, N_B – количество ненулевых битов в "молекулярных отпечатках пальцев" каждой из молекул, N_{AB} – количество битов, являющихся ненулевыми в каждом из двух "молекулярных отпечатков пальцев" одновременно. Для каждой молекулы было отобрано 100 наиболее близких структур (имеющих наибольшее значение молекулярного подобия S), входивших в обучающий набор данных при обучении модели, использованной для предсказания индекса удерживания для рассматриваемой молекулы. Затем было рассмотрено четыре способа вычисления молекулярного подобия. Для каждого из методов получено значение молекулярного подобия для молекулы, входившей в обучающий набор данных при обучении модели, использованной для предсказания индекса удерживания для рассматриваемой молекулы, и имеющей наибольшее значение молекулярного подобия с рассматриваемой. Это значение обозначается как $S_{\max}$. Так как эти методы более ресурсоемки, то поиск молекулы с наибольшим значением молекулярного подобия выполнялся только для 100 предварительно отобранных кандидатов.

Первым методом расчета молекулярного подобия, обозначенным MCS, было вычисление наибольшего общего фрагмента с помощью библиотеки RDKit, метод rdFMCS.FindMCS. После нахождения этого фрагмента подобие вычислялось по формуле, аналогичной уравнению (1):

$$S = \frac{M_{AB}}{M_A + M_B - M_{AB}}, \qquad (2)$$

где M_A, M_B – количество атомов каждой из молекул, M_{AB} – количество атомов в наибольшем общем фрагменте. При этом надо отметить, что учитывается только тип атомов и структура молекулярного графа. Атомы водорода не учитываются.

Вторым методом было вычисление сходства по методу Rascal. При этом также вычисляется наибольший общий фрагмент с помощью алгоритма Rascal [22] и рассчитывается количество связей и атомов в этом фрагменте. Сходство вычисляется по следующему уравнению:

$$S = \frac{\left(M_{AB} + B_{AB}\right)^2}{\left(M_A + B_A\right)\left(M_B + B_B\right)}, \qquad (3)$$

где M_A, M_B – количество атомов каждой из молекул, B_A, B_B – количество связей в каждой из молекул, M_{AB} и B_{AB} – количество атомов и связей в наибольшем общем фрагменте соответственно. В этом методе использовался модуль rdRascalMCES библиотеки RDKit.

Третий и четвертый методы были обозначены RDKitFP и ECFP. В них вычислялось сходство "молекулярных отпечатков пальцев" по формуле (1). Использовались молекулярные дескрипторы, рассчитанные с помощью классов GetRDKitFPGenerator и GetMorganGenerator соответственно. Длина вектора рассматривалась равной 4096, радиус (для ECFP) принимался равным 6. Метод ECFP соответствует "круговым молекулярным отпечаткам пальцев" [25]. Параметр maxPath для RDKitFP также принимался равным 6.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Молекулярное подобие и точность предсказания индексов удерживания

При осуществлении кросс-валидации исходный набор данных (база данных NIST 17) был разбит на 5 подмножеств. Для каждой молекулы из базы данных NIST 17, для которой доступно экспериментальное значение индекса удерживания на полярной неподвижной фазе, была найдена (четырьмя способами) наиболее близкая к ней, т.е. обладающая наибольшим значением меры молекулярного подобия, молекула, входящая в другое подмножество набора данных. Гипотеза, проверяемая в данной работе, состоит в том, что молекулярное подобие $S_{\max}$ между молекулой, для которой выполняется предсказание, и наиболее близкой молекулой из обучающего набора связано с точностью предсказания.

На рис. 1 показано распределение молекул (количество молекул в соответствующем интервале (бине) обозначено как N) из рассматриваемого набора данных по значению $S_{\max}$ для четырех методов расчета молекулярного подобия. Светло-серым показаны молекулы, для которых абсолютная ошибка предсказания с помощью рассматриваемого алгоритма [16] не больше 100, а темно-серым те, для

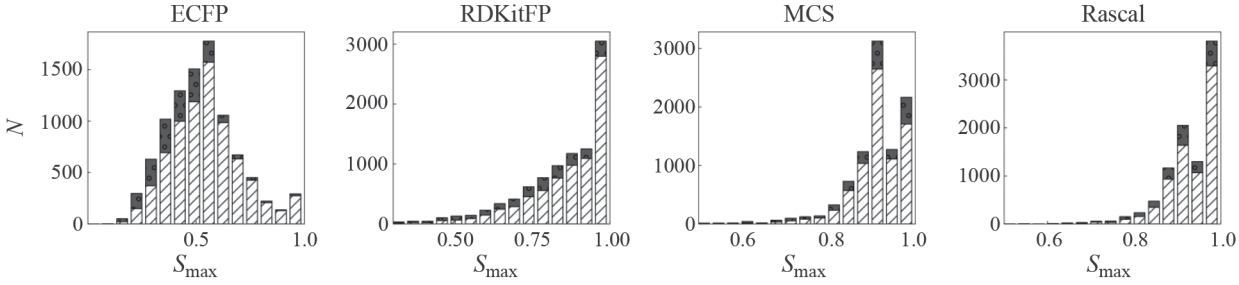

Рис. 1. Распределение количества молекул N в базе данных индексов удерживания NIST 17 (полярные неподвижные фазы) по значениям S_{max} (максимальное значение молекулярного подобия для всех пар, включающих в себя рассматриваемую молекулу и молекулы из обучающего набора) для четырех методов расчета молекулярного подобия. Темно-серым цветом обозначены "плохо предсказываемые молекулы" (абсолютная ошибка предсказания больше 100), светло-серым цветом обозначены остальные молекулы.

которых абсолютная ошибка предсказания больше 100. В дальнейшем мы называем такие молекулы "плохо предсказываемыми". Значение 100 было использовано в качестве порогового, так как в предыдущей работе [9] такое значение использовалось для отбрасывания ложных кандидатов при анализе сложной смеси. Таким образом, если рассматриваемая структура-кандидат является "плохо предсказываемой", то она может быть ложно отброшена (или наоборот не отброшена) на основании сравнения наблюдаемого и предсказанного индексов удерживания для полярной неподвижной фазы.

Как видно из рис. 1, при использовании метода расчета молекулярного подобия ECFP наибольшее число молекул имеет значение S_{max} около 0.5. Медианное значение S_{max} для всех молекул составляет 0.53 в этом случае. При этом для молекул со значениями S_{max} меньше ~0.5 доля "плохо предсказываемых" молекул существенно выше, чем для остальных. Для метода расчета молекулярного подобия RDKitFP медианное значение S_{max} для всех молекул существенно выше и составляет 0.89. Большинство молекул имеет достаточно высокие значения S_{max}, однако и в этом случае наблюдается аналогичная тенденция: количество "плохо предсказываемых" молекул сокращается с уменьшением S_{max} существенно медленнее по сравнению с количеством всех молекул. Для методов расчета молекулярного подобия MCS и Rascal, основанных на сравнении молекулярных графов, а не "молекулярных отпечатков пальцев", тенденция менее выражена. Для всех методов расчета молекулярного подобия в области самых малых значений S_{max} большинство молекул относится к "плохо предсказываемым".

Видно, что для всех методов, кроме RDKitFP, распределение молекул по S_{max} носит выраженный бимодальный характер. Для всех методов есть значительное количество молекул, для которых в обучающем наборе имеется очень похожая молекула, например, гомолог. В случае алгоритма MCS молекулярное подобие между, например, циклогексеном и циклогексаном равно 1.0: одна двойная связь в цикле игнорируется, так как общий подграф, включающий все связи между атомами углерода, кроме этой, включает в себя все неводородные атомы. Эта и другие особенности алгоритма приводят к тому, что для ряда весьма различных по химической природе молекул молекулярное подобие равно 1.0. Для алгоритма Rascal также возможно очень высокое значение молекулярного подобия для сильно различающихся по своей природе молекул. Так, например, 1-эйкозанол и эйкозановая кислота имеют значение молекулярного подобия 0.95, в то время как при использовании метода RDKitFP это значение равно 0.52 и при использовании ECFP равно 0.39. В то же время ECFP дает сходство, равное 1.0, для гомологов, содержащих длинную последовательность атомов углерода, например для эйкозанола и докозанола.

На рис. 2 наглядно показано то, как доля "плохо предсказываемых" (средняя абсолютная ошибка больше 100) молекул зависит от S_{max}. Для всех методов, кроме MCS, эта доля быстро растет с уменьшением S_{max}. Таким образом, маленькие значения S_{max} указывают на то, что, вполне вероятно, предсказание именно для рассматриваемой молекулы является весьма неточным. Для всех методов, кроме ECFP, общее количество молекул (также для удобства показанное на рис. 2) в соответствующем интервале быстро падает с падением значения S_{max}. В целом из рис. 1, 2 видно, что наилучшим алгоритмом вычисления молекулярного подобия для этой задачи является именно ECFP.

На рис. 1, 2 и в последующих разделах рассматривается, главным образом, доля "плохо предсказываемых" соединений, т.е. соединений, абсолютная ошибка предсказания для которых больше 100. Тем не менее интересно рассмотреть распределение ошибок для различных диапазонов S_{max}. Такие распределения абсолютной ошибки показаны на

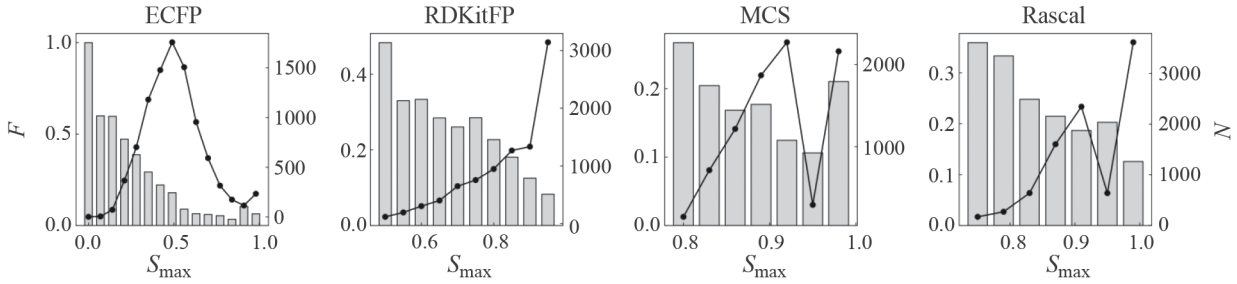

Рис. 2. Зависимость общего количества молекул N (сплошные круги и линии) и доли "плохо предсказываемых молекул" (абсолютная ошибка предсказания больше 100) F (прямоугольники) от значения S_{max} (максимальное значение молекулярного подобия для всех пар, включающих в себя рассматриваемую молекулу и молекулы из обучающего набора).

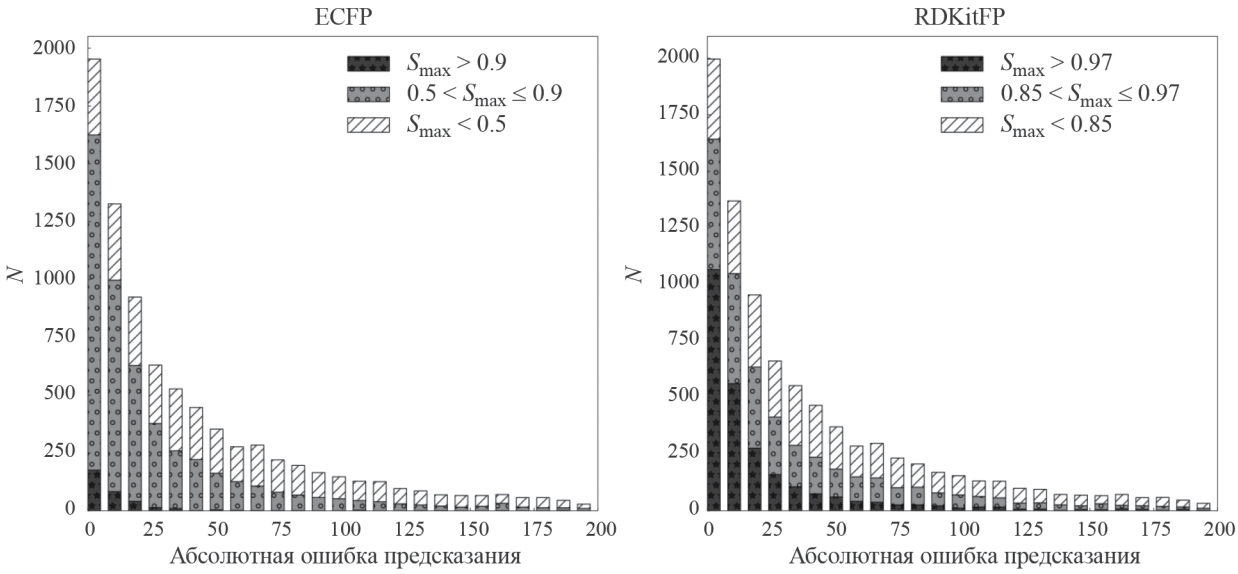

Рис. 3. Распределение количества молекул N по абсолютной ошибке предсказания для различных значений S_{max} (максимальное значение молекулярного подобия для всех пар, включающих в себя рассматриваемую молекулу и молекулы из обучающего набора) для двух методов расчета молекулярного подобия.

рис. 3 для алгоритмов ECFP и RDKitFP. В случае ECFP видно, что если $S_{max} > 0.9$, то подавляющее большинство значений абсолютной ошибки не превосходит 50, в то время как для значений абсолютных ошибок более 100 начинают доминировать молекулы с $S_{max} < 0.5$. Аналогичные закономерности есть и для алгоритма RDKitFP.

Количественное сравнение методов вычисления молекулярного подобия

Если использовать некоторое значение молекулярного подобия в качестве порогового, то молекулярное подобие можно использовать в качестве простейшего предиктора, характеризующего, является ли данная молекула "плохо предсказываемой". При изменении порогового значения от 0 до 1, чувствительность предсказания (доля выявленных "плохо предсказываемых" молекул среди всех "плохо предсказываемых" молекул) будет увеличиваться, а специфичность уменьшаться. Таким образом, можно построить кривую специфичность-чувствительность (Receiver Operator Characteristic, ROC-кривая) [26, 27], характеризующую надежность данной метрики молекулярного подобия при использовании в качестве предиктора. Площадь под данной кривой является [27] метрикой точности такого предиктора.

В табл. 1 показана площадь под кривой для различных алгоритмов расчета молекулярного подобия. При этом, в отличие от рис. 1, 2, в данном случае были рассмотрены алгоритмы RDKitFP

и ECFP с различными значениями параметров maxPath и radius. В табл. 1 приведены значения площади под кривой для различных значений этих параметров. Параметры maxPath (“молекулярные отпечатки пальцев” RDKitFP) и radius (“молекулярные отпечатки пальцев” ECFP) характеризуют размер субструктур, которым соответствуют биты “молекулярного отпечатка пальца”. Чем выше значения этих параметров, тем более крупные субструктуры рассматриваются.

Из табл. 1 видно, что лучше всего для рассматриваемой цели подходит алгоритм ECFP, причем зависимости от параметра radius практически нет. Алгоритм RDKitFP (при значениях параметра maxPath 6 и выше) дает худшие результаты. Остальные алгоритмы дают ненадежные результаты. Надо отметить, что значение площади под кривой 0.5 соответствует случайному классификатору, значение 1 – идеальному классификатору [26]. Значение 0.7 иногда рассматривается как наименьшее приемлемое [27]. ROC-кривые для алгоритмов расчета молекулярного подобия, основанных на “молекулярных отпечатках пальцев”, приведены на рис. 4. Видно, что RDKitFP со значением параметра maxPath = 3 не дает удовлетворительной точности, а ECFP превосходит RDKitFP.

Надежность идентификации ряда азотсодержащих соединений

Несимметричный диметилгидразин (НДМГ) – токсичное соединение, используемое в качестве ракетного горючего. При неконтролируемом хранении и попадании в окружающую среду это соединение образует множество продуктов трансформации [9, 12, 28], многие из которых не менее токсичны, чем сам НДМГ [12]. Исследование продуктов трансформации НДМГ – важная задача. Структуры большинства продуктов трансформации до сих пор неизвестны [9]. Различные методы хромато-масс-спектрометрии используются для предварительного определения структур продуктов трансформации НДМГ в сложных смесях. Недавно была опубликована работа [9], в которой для подтверждения структур неизвестных продуктов трансформации НДМГ использовалось в том числе предсказание индексов удерживания на полярной неподвижной фазе. Если различие между наблюдаемым и предсказанным индексом превосходило 100, то структура-кандидат отбрасывалась.

Суммарно 1754 из 9408 (19%) молекул в наборе данных являются “плохо предсказываемыми”. Однако, среди молекул, у которых $S_{max} < 0.5$ (алгоритм ECFP), 31% являются “плохо предсказываемыми”. В табл. 2 показано количество “плохо предсказываемых” молекул для различных диапазонов S_{max} и значения средних и медианных абсолютных ошибок для этих диапазонов. Значения

Таблица 1. Площадь под ROC-кривой при использовании различных метрик молекулярного подобия в качестве предиктора того, является ли молекула “плохо предсказываемой”

Метод	Площадь под кривой
RDKitFP (maxPath = 3)	0.62
RDKitFP (maxPath = 6)	0.69
RDKitFP (maxPath = 12)	0.70
RDKitFP (maxPath = 15)	0.69
ECFP (radius = 3)	0.72
ECFP (radius = 6)	0.72
ECFP (radius = 12)	0.72
MCS	0.55
Rascal	0.61

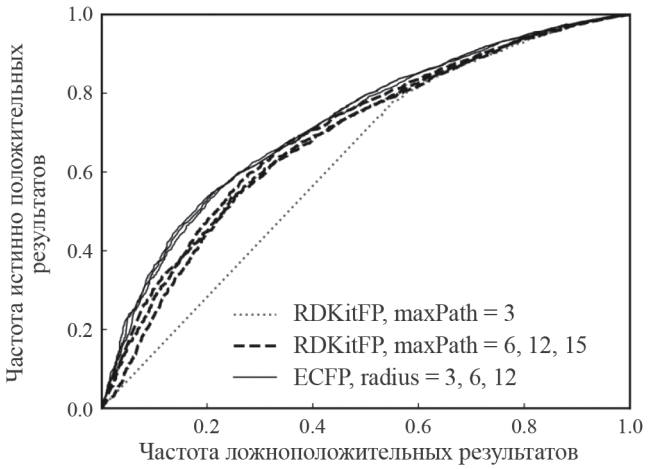

Рис. 4. ROC-кривые (кривые специфичность-чувствительность) для предсказания того, является ли молекула “плохо предсказываемой” (абсолютная ошибка предсказания больше 100) с помощью различных алгоритмов вычисления молекулярного подобия. Кривые для алгоритмов, для которых площадь под кривой отличается не более чем на 0.02, обозначены одним типом линии для читабельности.

ошибок в табл. 2 отличаются от таковых в предыдущей работе [16], где использовалась точно такая же модель, в связи с тем, что в прошлой работе [16] ошибки вычислялись для всех записей базы данных NIST, а в этой работе данные были предварительно усреднены по всем записям для каждого соединения. Таким образом, значительно уменьшился вклад в среднюю абсолютную ошибку тех соединений, для которых в базе данных NIST много записей, так как если для соединения имеется много записей, то для каждой записи вычисляется модуль ошибки и все эти значения (для каждой из записей

Таблица 2. Количество “плохо предсказываемых” молекул и метрики точности для различных диапазонов S_{max} (алгоритм ECFP)

Диапазон	“Плохо предсказываемые” молекулы	Всего молекул	Доля “плохо предсказываемых” молекул, %	Средняя абсолютная ошибка	Медианная абсолютная ошибка
Все молекулы	1754	9408	18.6	70.6	28.4
$S_{max} > 0.7$	83	1419	5.8	25.8	8.3
$S_{max} > 0.5$	512	5239	9.8	41.7	15.9
$S_{max} < 0.5$	1168	3820	30.6	109.9	57.0
$S_{max} < 0.3$	280	583	48.0	173.6	95.4
$S_{max} < 0.2$	50	30	60	311.0	181.7

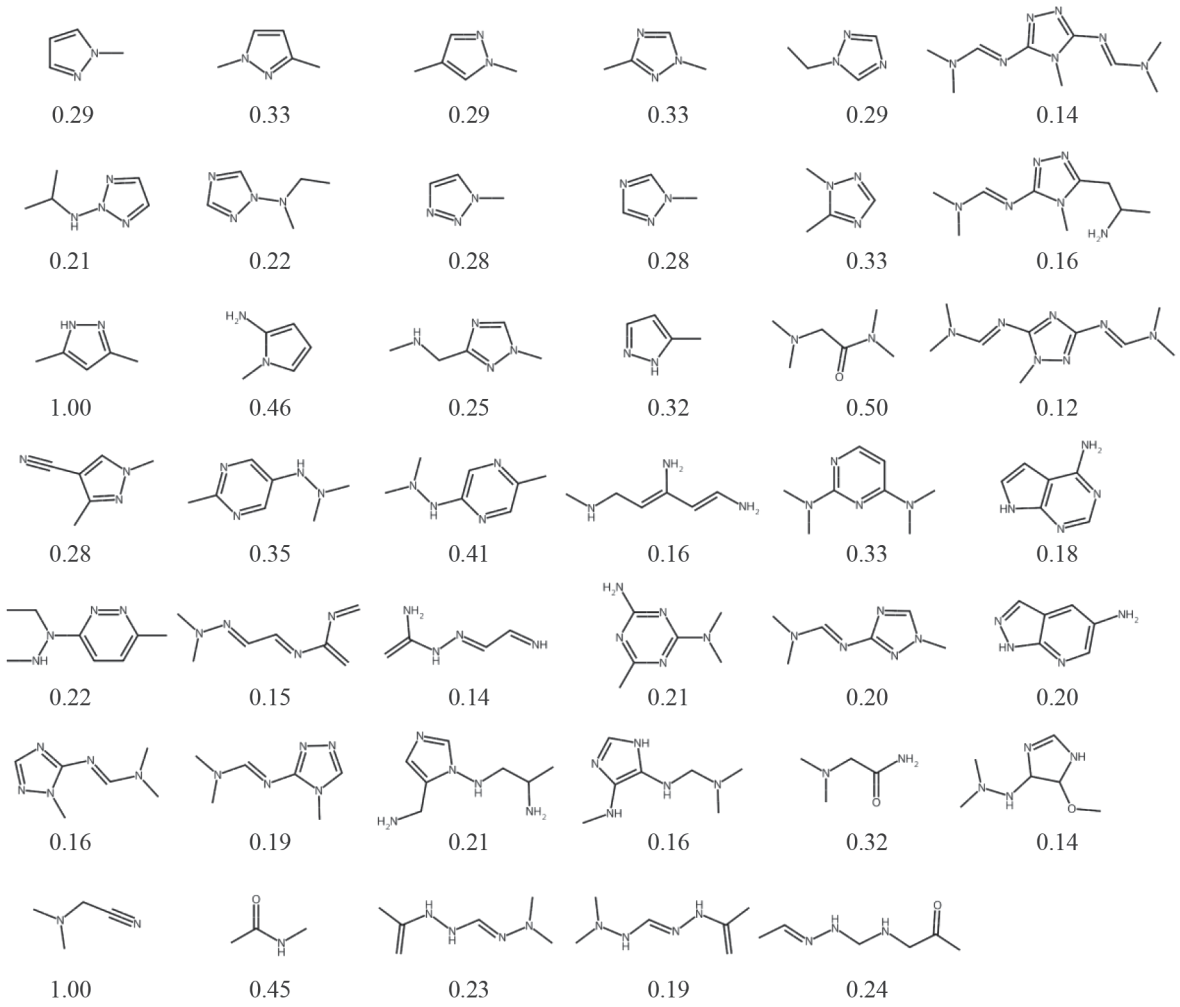

Рис. 5. Структуры продуктов трансформации несимметричного диметилгидразина, предложенные в работе [9], и значения S_{max} (величина молекулярного подобия между рассматриваемой молекулой и наиболее близкой молекулой из обучающего набора) для каждой из них. Метод расчета молекулярного подобия ECFP.

для всех соединений) усредняются. Средняя абсолютная ошибка это частное суммы модулей ошибок и количества записей или молекул. В данной работе каждому соединению соответствует одно слагаемое в сумме модулей ошибок, в отличие от работы [16]. Было отмечено, что большая часть

соединений, для каждого из которых база данных NIST содержит множество записей, имеют относительно простую структуру и в среднем предсказания для таких соединений точнее по сравнению со всеми соединениями. В базе данных NIST большинству структур соответствует ровно одна запись, однако для части структур (например такие молекулы как бензол и этанол) база данных содержит много записей. Разница в подходе к вычислению средней абсолютной ошибки приводит к разнице значений, приведенных в работе [16] и в табл. 2.

На рис. 5 показаны структуры, предложенные [9] в качестве структур продуктов трансформации НДМГ с использованием предсказания индексов удерживания на полярной неподвижной фазе. Для каждой из структур показано молекулярное подобие (алгоритм ECFP) с наиболее близкой структурой из базы данных NIST. Две структуры содержатся в базе данных NIST, для них это значение равно 1.0. Для остальных структур это значение не превосходит 0.5. Для ряда структур это значение даже меньше 0.2. Таким образом, нельзя быть уверенным в том, что такие предсказания приводят к верным результатам, и к представленным результатам следует относиться с осторожностью. Тем не менее структуры, подтвержденные с помощью нескольких методов хромато-масс-спектрометрии (газовой и жидкостной) и нескольких методов машинного обучения, могут быть рассмотрены как достаточно надежные [9, 12].

ВЫВОДЫ

Точность моделей, предсказывающих газохроматографические индексы удерживания, оценивается с помощью метрик, таких как средняя абсолютная ошибка, которые, однако, не позволяют оценить точность для конкретных молекул. В этой работе было показано, что факт наличия в обучающем наборе молекул, которые близки по структуре к молекуле, индекс удерживания которой предсказывается, очень сильно повышает вероятность того, что предсказание для этой молекулы будет точным. Наиболее подходящим для этой задачи способом оценки молекулярного подобия являются “молекулярные отпечатки пальцев” ECFP. В ситуациях, когда предсказание индексов удерживания используется при построении структур неизвестных химических соединений, необходимо оценивать точность предсказания тем или иным способом. Так, например, в одной из работ по изучению продуктов трансформации несимметричного диметилгидразина [9] для большинства рассмотренных структур в обучающем наборе данных не было молекул с высокими значениями меры молекулярного подобия. А значит выводы, сделанные с помощью предсказания индексов удерживания для этих структур, могут быть не вполне надежными. Исходный код сценариев, использованных для выполнения данной работы, доступен онлайн: https://github.com/mtshn/molsimwax

Работа выполнена при поддержке Российского Научного Фонда (проект № 22-73-10053), https://rscf.ru/project/22-73-10053/

СПИСОК ЛИТЕРАТУРЫ

1. *Tarján G., Nyiredy S., Györ M. et al.* // J. of Chromatography A. 1989. V. 472. P. 1. https://doi.org/10.1016/S0021-9673(00)94099-8
2. *Franke J.-P., Wijsbeek J., De Zeeuw R.A.* // J. of Forensic Sciences. 1990. V. 35. № 4. P. 813. https://doi.org/10.1520/JFS12893J
3. *Zellner B.A., Bicchi C., Dugo P. et al.* // Flavour and Fragrance J. 2008. V. 23. № 5. P. 297–314. https://doi.org/10.1002/ffj.1887
4. *Milman B.L., Zhurkovich I.K.* // TrAC Trends in Analytical Chemistry. 2016. V. 80. P. 636–640. https://doi.org/10.1016/j.trac.2016.04.024
5. *Vinaixa M., Schymanski E.L., Neumann S. et al.* // TrAC Trends in Analytical Chemistry. 2016. V. 78. P. 23. https://doi.org/10.1016/j.trac.2015.09.005
6. *Matyushin D.D., Sholokhova A.Yu., Karnaeva A.E. et al.* // Chemometrics and Intelligent Laboratory Systems. 2020. V. 202. P. 104042. https://doi.org/10.1016/j.chemolab.2020.104042
7. *Schymanski E.L., Meringer M., Brack W.* // Analytical Chemistry. 2011. V. 83. № 3. P. 903. https://doi.org/10.1021/ac102574h
8. *Dossin E., Martin E., Diana P. et al.* // Analytical Chemistry. 2016. V. 88. № 15. P. 7539–7547. https://doi.org/10.1021/acs.analchem.6b00868
9. *Sholokhova A.Yu., Matyushin D.D., Grinevich O.I. et al.* // Molecules. 2023. V. 28. № 8. P. 3409. https://doi.org/10.3390/molecules28083409
10. *Su Q.-Z., Vera P., Salafranca J. et al.* // Resources, Conservation and Recycling. 2021. V. 171. P. 105640. https://doi.org/10.1016/j.resconrec.2021.105640
11. *Su Q.-Z., Vera P., Nerín C. et al.* // Resources, Conservation and Recycling. 2021. V. 167. P. 105365. https://doi.org/10.1016/j.resconrec.2020.105365
12. *Sholokhova A.Yu., Grinevich O.I., Matyushin D.D. et al.* // Chemosphere. 2022. V. 307. P. 135764. https://doi.org/10.1016/j.chemosphere.2022.135764
13. *Matyushin D.D., Buryak A.K.* // IEEE Access. 2020. V. 8. P. 223140. https://doi.org/10.1109/ACCESS.2020.3045047
14. *Debus B., Parastar H., Harrington P. et al.* // TrAC Trends in Analytical Chemistry. 2021. V. 145. P. 116459. https://doi.org/10.1016/j.trac.2021.116459
15. *Dong S., Wang P., Abbas K.* // Computer Science Review. 2021. V. 40. P. 100379. https://doi.org/10.1016/j.cosrev.2021.100379

16. *Matyushin D.D., Sholokhova A.Yu., Buryak A.K.* // Intern. J. of Molecular Sciences. 2021. V. 22. № 17. P. 9194. https://doi.org/10.3390/ijms22179194
17. *Matyushin D.D., Sholokhova A.Yu., Buryak A.K.* // J. of Chromatography A. 2019. V. 1607. P. 460395. https://doi.org/10.1016/j.chroma.2019.460395
18. *Anjum A., Liigand J., Milford R. et al.* // Ibid. 2023. V. 1705. P. 464176. https://doi.org/10.1016/j.chroma.2023.464176
19. *Qu C., Schneider B.I., Kearsley A.J. et al.* // Ibid. 2021. V. 1646. P. 462100. https://doi.org/10.1016/j.chroma.2021.462100
20. *Vrzal T., Malečková M., Olšovská J.* // Analytica Chimica Acta. 2021. V. 1147. P. 64. https://doi.org/10.1016/j.aca.2020.12.043
21. *Geer L.Y., Stein S.E., Mallard W.G. et al.* // J. of Chemical Information and Modeling. 2024. V. 64. № 3. P. 690–696. https://doi.org/10.1021/acs.jcim.3c01758
22. *Raymond J.W., Gardiner E.J., Willett P.* // The Computer J. 2002. V. 45. № 6. P. 631–644. https://doi.org/10.1093/comjnl/45.6.631
23. *Bender A., Glen R.C.* // Organic & Biomolecular Chemistry. 2004. V. 2. № 22. P. 3204. https://doi.org/10.1039/B409813G
24. *Morehouse N.J., Clark T.N., McMann E.J. et al.* // Nature Communications. 2023. V. 14. № 1. P. 308. https://doi.org/10.1038/s41467-022-35734-z
25. *Rogers D., Hahn M.* // J. of Chem. Inform. and Modeling. 2010. V. 50. № 5. P. 742. https://doi.org/10.1021/ci100050t
26. *Hoo Z.H., Candlish J., Teare D.* // Emergency Medicine J. 2017. V. 34. № 6. P. 357. https://doi.org/10.1136/emermed-2017-206735
27. *Polo T.C.F., Miot H.A.* // J. Vascular Brasileiro. 2020. V. 19. P. e20200186. https://doi.org/10.1590/1677-5449.200186
28. *Popov M.S., Ul'yanovskii N.V., Kosyakov D.S.* // Microchemical J. 2024. V. 197. P. 109833. https://doi.org/10.1016/j.microc.2023.109833

Physical chemistry of separation processes. Chromatography

APPLYING MOLECULAR SIMILARITY USED FOR EVALUATING THE ACCURACY OF RETENTION INDEX PREDICTIONS IN GAS CHROMATOGRAPHY USING DEEP LEARNING

D. D. Matyushin[a], A. Yu. Sholokhova[a, *], M. D. Khrisanfov[a, b], and S. A. Borovikova[a]

[a] A. N. Frumkin Institute of Physical Chemistry and Electrochemistry of the Russian Academy of Sciences, Moscow, 119071 Russia

[b] M. V. Lomonosov Moscow State University, Department of Chemistry, Moscow, 119991 Russia

**e-mail: shonastya@yandex.ru*

Abstract. When predicting retention indices using deep learning, there is usually no way to assess the reliability of the prediction for a particular molecule. In this work, using stationary phases based on polyethylene glycol and the NIST 17 database as an example, it is shown that, on average, the closer the molecule in the training data set is to the compound being predicted, the more accurate the prediction. Tanimoto similarity of "molecular fingerprints" ECFP is the most appropriate molecular similarity calculation algorithm for this problem among the four considered. It is shown that for a number of transformation products of unsymmetrical dimethylhydrazine, whose structure was established using this prediction, it could be very unreliable.

Keywords: gas chromatography, retention indices, machine learning, deep learning, molecular similarity