Сорбционные и хроматографические процессы. 2024. Т. 24, № 4. С. 481-499.
Sorbtsionnye i khromatograficheskie protsessy. 2024. Vol. 24, No 4. pp. 481-499.

ISSN 1680-0613

ОРИГИНАЛЬНЫЕ СТАТЬИ

Научная статья
УДК 544.03:543.544.3:004.8
doi: 10.17308/sorpchrom.2024.24/12405

Эмпирические уравнения для прогнозирования газохроматографических индексов удерживания для неподвижной фазы DB-35MS

Дмитрий Дмитриевич Матюшин, Анастасия Юрьевна Шолохова✉
Институт физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва, Россия,
shonastya@yandex.ru✉

Аннотация. На данный момент большинство работ по прогнозированию индексов удерживания по структуре молекулы посвящены стандартным неподвижным фазам: полидиметилсилоксану, 5%-фенил-метилполисилоксану и полиэтиленгликолю. Информация об индексах удерживания на этих неподвижных фазах содержится в базе данных NIST, соответственно, доступен большой набор данных для обучения и возможно применение глубокого обучения. Это позволяет создать точные и универсальные модели для прогнозирования индексов удерживания. Однако другие неподвижные фазы также активно применяются в исследованиях при хромато-масс-спектрометрической идентификации компонентов сложных смесей. Создание алгоритмов прогнозирования индексов удерживания для этих неподвижных фаз также могло бы иметь большое значение. В данной работе рассматривается задача прогнозирования индексов удерживания для неподвижной фазы DB-35MS (35%-фенил-полидиметилсилоксан). Рассмотрен набор данных об удерживании 52 летучих органических соединений, содержащихся в бутонах сирени, для этой неподвижной фазы. Предложены эмпирические уравнения, в которые входят: вычисленный с помощью глубокого обучения индекс удерживания для неподвижной фазы DB-5 (5%-фенилметилполисилоксан) и ряд молекулярных дескрипторов, рассчитанных с помощью фреймворка RDKit. Показано, что использование сложных топологических молекулярных дескрипторов, а так же величин рассчитанных с помощью методов квантовой химии, не дает сильного повышения точности по сравнению с наиболее простыми целочисленными молекулярными дескрипторами, такими как число связей, подверженных внутреннему вращению. В то же время применение индексов удерживания для неподвижной фазы DB-5, спрогнозированных с помощью глубокого обучения, в качестве молекулярного дескриптора, приводит к сильнейшему уменьшению ошибки прогнозирования по сравнению с применением только обычных молекулярных дескрипторов. Если вместо индексов удерживания, спрогнозированных для неподвижной фазы DB-5, использовать индексы удерживания, спрогнозированные для неподвижной фазы DB-624, то также можно достигнуть относительно высокой точности прогноза. В работе представлены линейные уравнения, которые могут быть использованы на практике для вычисления индексов удерживания летучих соединений растительного происхождения, содержащих углерод, водород и кислород, для неподвижной фазы DB-35MS и аналогичных неподвижных фаз. Представлено также менее точное, но более универсальное уравнение, содержащее в себе только спрогнозированный с помощью глубокого обучения индекс удерживания для неподвижной фазы DB-5 в качестве молекулярного дескриптора. Достигнутые значения среднеквадратичной ошибки, средней абсолютной ошибки и медианной абсолютной ошибки составили 28.9, 19.3 и 11.8 единиц соответственно.

Ключевые слова: газовая хроматография, индекс удерживания, DB-35MS, нейронные сети, машинное обучение, количественное соотношение структура-удерживание

Благодарности: данная работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках госбюджетной темы № 124041900012-4.

Для цитирования: Матюшин Д.Д., Шолохова А.Ю. Эмпирические уравнения для прогнозирования газохроматографических индексов удерживания для неподвижной фазы DB-35MS // *Сорбционные и хроматографические процессы*. 2024. Т. 24, № 4. С. 481-499. https://doi.org/10.17308/sorpchrom.2024.24/12405

Сорбционные и хроматографические процессы. 2024. Т. 24, № 4. С. 481-499.
Sorbtsionnye i khromatograficheskie protsessy. 2024. Vol. 24, No 4. pp. 481-499.

ISSN 1680-0613

Original article

Empirical equations for the prediction of gas chromatographic retention indices for the DB-35MS stationary phase

Dmitriy D. Matyushin, Anastasia Yu. Sholokhova✉
A.N. Frumkin Institute of Physical Chemistry and Electrochemistry of Russian Academy of Sciences, Moscow, Russian Federation, shonastya@yandex.ru✉

Abstract. At the moment, most of the studies on the retention index prediction based on molecule structure are devoted to standard stationary phases: polydimethylsiloxane, 5%-phenyl-methylpolysiloxane, and polyethylene glycol. Retention index information for these stationary phases is contained in the NIST database, so a large training data set is available, and deep learning can be applied. This allows the creation of accurate and versatile retention index prediction models. However, other stationary phases are also actively used in research, for identification of components of complex mixtures using chromatography-mass spectrometry. The development of retention index prediction algorithms for these stationary phases could also be of great importance. In this paper, we consider the problem of predicting retention indices for the DB-35MS stationary phase (35%-phenyl-methylpolysiloxane). A data set of retention indices of 52 volatile organic compounds contained in lilac buds for this stationary phase is considered. Empirical equations are proposed that incorporate the retention index for the DB-5 stationary phase (5%-phenyl-methylpolysiloxane) predicted by deep learning and a number of molecular descriptors calculated using the RDKit framework. It was shown that the use of complex topological molecular descriptors, and features calculated using quantum chemistry does not provide a significant increase in accuracy compared to the simplest integer molecular descriptors, such as the number of bonds subject to internal rotation. At the same time, the use of the retention index for the DB-5 stationary phase predicted by deep learning as a molecular descriptor leads to a strong decrease in the prediction error compared to the use of only conventional molecular descriptors. When the retention indices predicted for the DB-624 stationary phase are used instead of the retention indices predicted for the DB-5 stationary phase, a relatively high prediction accuracy can also be achieved. Linear equations are presented that can be used in practice to calculate the retention indices of volatile plant compounds containing carbon, hydrogen, and oxygen for the DB-35MS stationary phase and similar stationary phases. A less accurate but more versatile equation is also presented that contains only the retention index predicted by deep learning for the DB-5 stationary phase as a molecular descriptor. The achieved values of the root-mean-square prediction error, the mean absolute prediction error, and the median absolute prediction error were 28.9, 19.3, and 11.8 units, respectively.
Keywords: gas chromatography, retention index, DB-35MS, neural networks, machine learning, quantitative structure-retention relationship**.**
Acknowledgments: This work was carried out with the support of the Ministry of Science and Higher Education of the Russian Federation within the framework of the state budget theme No. 124041900012-4.
For citation: Matyushin D.D., Sholokhova A.Yu. Empirical equations for the prediction of gas chromatographic retention indices for the DB-35MS stationary phase. *Sorbtsionnye i khromatograficheskie protsessy*. 2024. 24(4): 481-499. (In Russ.). https://doi.org/10.17308/sorpchrom.2024.24/12405

Введение

Индексы удерживания на основе *н*-алканов (т. е. значения, характеризующие удерживание молекулы относительно *н*-алканов), широко используются в газовой хроматографии [1] в качестве дополнительного фактора, позволяющего принять или отбросить кандидат при масс-спектрометрической идентификации [2]. Значение индекса удерживания главным образом зависит от структуры молекулы и неподвижной фазы и является переносимой характеристикой, не слишком сильно зависящей от других параметров системы [1]. Индексы удерживания на основе *н*-алканов были впервые описаны Ковачем [3] для изотермических условий, а затем адаптированы для условий программирования температуры [4]. К сожалению, для большинства соединений индексы удерживания отсутствуют в базах данных [5], а также базы данных содержат в себе ошибки и неточности [5-6].

Последние годы были разработаны методы прогнозирования индексов удерживания на основе глубокого обучения (глубоких нейронных сетей) [7-9]. Однако, чтобы обучить нейронную сеть требуется большой объем данных — экспериментальные данные об индексах удерживания для десятков тысяч молекул. В связи с этим глубокое обучение применяется главным образом для прогнозирования индексов удерживания для полидиметилсилоксана (DB-1 и аналогичные неподвижные фазы), 5%-фенил-метилполисилоксана (DB-5 и аналогичные неподвижные фазы) и полиэтиленгликоля (DB-WAX и аналогичные неподвижные фазы).

Для других неподвижных фаз также важно прогнозировать индексы удерживания, так как такие неподвижные фазы применяются для хромато-масс-спектрометрического анализа сложных смесей, а справочные данные для них недоступны. Ранее нами был разработан подход для применения глубокого обучения для прогнозирования индексов удерживания и в такой ситуации [10-11]. При этом используется «двухстадийная» модель машинного обучения — на первом этапе выполняется прогнозирование индексов удерживания с помощью глубоких нейронных сетей для наиболее распространенных неподвижных фаз (DB-1, DB-5, DB-WAX). Затем эти полученные *in silico* индексы удерживания используются в качестве исходных характеристик для моделей машинного обучения или эмпирических уравнений, прогнозирующих удерживание для других неподвижных фаз. Этот подход был применен для ряда неподвижных фаз промежуточной полярности (более полярных по сравнению с DB-5, но менее полярных, чем DB-WAX) [10], а также неподвижных фаз на основе ионных жидкостей [11].

Одной из важных неподвижных фаз, селективность которой отличается от селективности стандартных неподвижных фаз, является 35%-фенил-метилполисилоксан. Эти неподвижные фазы (например, DB-35, DB-35MS, OV-35) широко используются в аналитической практике [12-14], в том числе в новейших работах [15-16]. Работ по прогнозированию удерживания для них крайне мало, в частности, есть работы по прогнозированию удерживания замещенных полихлорированных бифенилов [17-18]. Есть публикация, в которой индексы удерживания для DB-35MS рассчитаны на основе индексов удерживания для HP-5MS, но в этом случае потребовалось знать экспериментальные индексы удерживания [19], модель была разработана только для конкретной пары условий, а также использовалось несвободное программное обеспечение Dragon.

В работе [20] рассматривалось прогнозирование времен удерживания (для одних конкретных хроматографических условий) ряда пестицидов и токсикантов для неподвижной фазы DB-35MS. Такой подход не является переносимым, и такие модели не могут быть использованы другими исследователями. Рассматривалось и прогнозирование индексов удерживания для гомологических рядов галогеналканов [21], однако такие уравнения не могут быть рассмотрены как универсальные и имеющие прогностическую силу. Также такие неподвижные фазы исследовались и с применением других методов [22-23]. Несмотря на большой интерес к таким неподвижным фазам и наличие ряда работ, посвященных прогнозированию удерживания для них, надежного и универсального метода, позволяющего прогнозировать индексы удерживания для 35%-фенил-метилполисилоксана и различных классов аналитов, исходя непосредственно из структуры аналита, на данный момент не существует. Целью данной работы является разработка такого метода, который позволит осуществить прогнозирование индексов удерживания для этих неподвижных фаз для летучих соединений растительного происхождения.

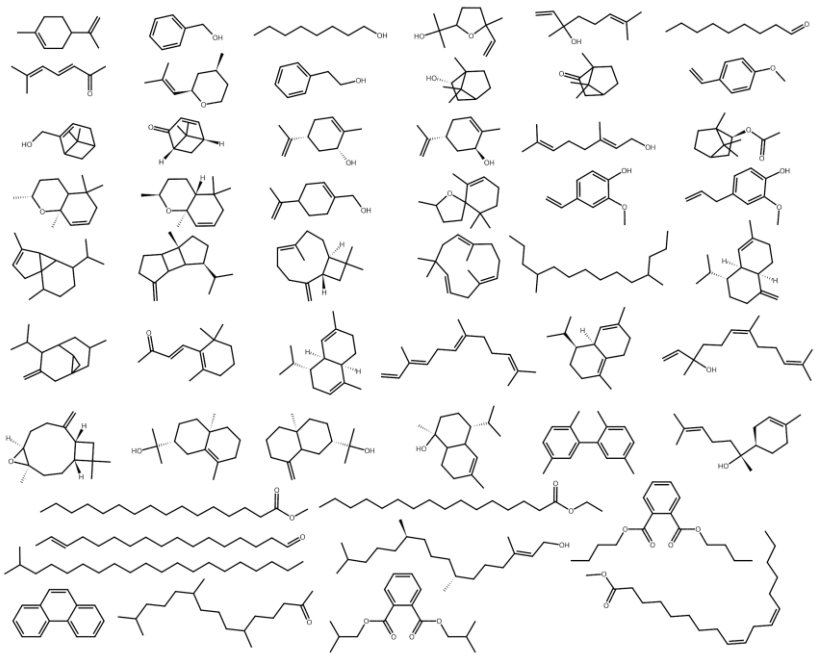


Рис. 1. Структурные формулы 52 летучих органических соединений из бутонов сирени, рассмотренных в данной работе.
Fig. 1. Structural formulas of 52 volatile organic compounds from lilac buds considered in this work.

Методы и набор данных

Для обучения моделей машинного обучения для прогнозирования индексов удерживания необходим набор экспериментальных данных. В данной работе набор данных был взят из статьи [14]. В этой статье выполнялся анализ компонентов эфирных масел, извлеченных из бутонов сирени. Авторами приведены данные об удерживании 95 компонентов эфирных масел на двух неподвижных фазах: HP-5MS или DB-35MS. Использовались капиллярные хроматографические колонки длиной 30 м, толщиной 0.25 мм, с толщиной слоя 0.25 мкм. Поскольку авторы не пользовались стандартными образцами и применяли только идентификацию с помощью масс-спектрометрии, то отнесение структурной формулы к хроматографическому пику может быть ненадежным. В связи с этим рассматривались только соединения, для которых есть информация об удерживании на обеих неподвижных фазах, а также не рассматривались соединения, структура которых была установлена с помощью более сложных эвристических методов. Предполагается, что таким образом были отобраны соединения, структура которых установлена наиболее надежно. Для части соединений не было возможно однозначно установить структурную формулу, исходя из данных, приведенных авторами [14]. В итоге был составлен набор данных об индексах удерживания (неподвижная фаза DB-35MS) для 52 молекул. Большая часть этих молекул содержит кислород, и ни одна из них не содержит азот и другие элементы, кроме кислорода, углерода, водорода. Структуры соединений весьма разнообразны. Структурные формулы рассмотренных 52 соединений изображены на рис. 1.

Для каждого соединения индекс удерживания зарегистрирован при трех скоростях нагрева (режим программирования температуры): 2, 4 или 6ºС/мин, начальная температура составляла 80ºС. Мы рассматривали среднее арифметическое значений, полученных при разных скоростях нагрева. Индекс удерживания

практически линейно зависит от скорости нагрева в режиме программирования температуры [24], таким образом, это среднее значение для большинства молекул близко к индексу удерживания при 4ºС/мин. Тем не менее такое усреднение может вносить ошибку. Для сравнения было так же рассмотрено прогнозирование индексов удерживания, полученных для одной конкретной скорости нагрева, а не «средних» индексов удерживания. Более подробно этот вопрос рассмотрен ниже. Молекулярная масса этих соединений находится в диапазоне от 108 до 296 Да. Индекс удерживания (неподвижная фаза DB-35MS) находится в диапазоне от 1095 до 2205. Более подробная информация об условиях хроматографирования, образцах бутонов сирени, а также «сырые» значения индексов удерживания содержатся в соответствующей работе [14].

Для построения линейных уравнений, связывающих молекулярные дескрипторы (легко рассчитываемые величины, характеризующие структуру молекулы, см. ниже) с индексами удерживания, было использовано ранее разработанное в нашей лаборатории программное обеспечение CHERESHNYA [11, 25] (версия 0.0.2-alpha1). Использовался пошаговый (stepwise) метод отбора молекулярных дескрипторов (МД): в этом методе МД отбираются один за другим. На каждом шаге отбирается МД, значение статистической значимости многомерной линейной регрессии (*f*-тест) для которого наибольшее с учетом ранее отобранных МД. Для оценки важности (значимости) МД процедура отбора выполнялась 200 раз, каждый раз из набора данных удалялась случайным образом одна молекула. Ранее [11] было показано, что процедура отбора МД является невоспроизводимой при незначительном изменении набора данных. За счет многократного повторения возможно добиться более надежной оценки значимости МД. Подробное описание этой процедуры дано в предыдущей работе [11].

Использовались (посредством программы CHERESHNYA) 2D молекулярные дескрипторы, генерируемые фреймворком RDKit (версия 2023.09.4) [26]. Полный набор МД включал в себя 208 МД. Половина из них (104) – это целочисленные величины с очевидным физическим смыслом, характеризующие наличие и количество тех или иных фрагментов в молекулах или, например, число валентных электронов. Другая половина (104) – это весьма разнообразные дескрипторы. Часть из них имеет простой физический смысл (в это число, например, входит молекулярная масса, длина самой длинной углеродной цепи), физический смысл других более сложный. Например, семейство МД BCUT представляет собой значения собственных значений матриц связности молекул (с модифицированными диагональными элементами), [27-28]. Несколько из рассмотренных МД составляют МД, характеризующие какие-то свойства молекул, например logP, молекулярную рефракцию. Они рассчитываются с помощью аддитивных схем и других моделей [29]. Все МД кратко описаны в документации RDKit [26], а более полное описание содержится в соответствующих публикациях, ссылки на которые приведены в этой документации. Все эти МД объединяет то, что они могут быть детерминистично и быстро (за доли секунды на одном ядре современного процессора) рассчитаны непосредственно на основе двумерной структуры молекулы без вычисления трёхмерных координат. Методы квантовой химии, молекулярной динамики при этом не используются.

Также рассматривался сокращенный набор МД: при этом рассматривались только 104 МД, в названии которых содержится фрагмент «Count», «Num» или «fr_». Все эти МД является целочисленными и имеют прозрачный физический смысл, очевидный из названия. Например, «NumRotatableBonds» – МД представляющий собой количество связей,

подверженных вращению, а «fr_benzene» – количество бензольных колец. МД, значение которых является константой для всех рассматриваемых молекул, а также МД с точностью до линейной зависимости, совпадающие с другими МД, в автоматическом режиме исключены с помощью программного обеспечения CHERESHNYA.

Помимо МД, вычисленных с помощью фреймворка RDKit, в качестве МД рассматривались индексы удерживания для неподвижных фаз DB-5, DB-WAX и DB-624 (6%-цианопропилфенил)-метилполисилоксан. Соответствующие МД обозначены как RI_DB-5, RI_DB-WAX, RI_DB-624. Индексы удерживания для этих неподвижных фаз в автоматическом режиме вычисляются с помощью ранее разработанного нами программного обеспечения. Модели для их прогнозирования подробно описаны в предыдущей работе [10]. Именно использование этих индексов удерживания в качестве МД позволяет достигнуть высокой точности прогноза. В данной работе мы рассматриваем традиционное для хемоинформатики определение МД [30], а именно мы называем термином МД численные величины, характеризующие (представляющие) структуру молекулы, которые могут быть легко и однозначно рассчитаны на основе структуры молекулы. В этом смысле спрогнозированные с помощью ранее опубликованных (и обученных на совершенно других наборах данных вне рамок этой работы) моделей глубокого обучения индексы удерживания могут быть рассмотрены в качестве МД. Для их вычисления используются глубокие нейронные сети [10].

Для сравнения эффективности различных типов МД в качестве МД так же были рассмотрены дипольный момент и поляризуемость, рассчитанные с помощью квантовохимического пакета nwchem (версия 7.2.3) [31]. Была использована теория функционала плотности, базисный набор 6-311G*, гибридный функционал PBE0 [32]. Начальное приближение трёхмерных координат атомов было получено с помощью пакета Open Babel 3.0.0 [33], минимизация потенциальной энергии (оптимизация структуры) была выполнена с помощью модели машинного обучения ANI-1ccx [34] и программного обеспечения TorchANI (версия 2.2.4) [35]. Эта модель позволяет оценивать потенциальную энергию молекулы с точностью, сравнимой с современными *ab initio* методами [34, 36]. Поляризуемость была оценена с помощью опции «response» программы nwchem, поле было задано как 1 (атомные единицы) и частота 0. В качестве значения поляризуемости рассматривалось среднее значение диагональных компонентов матрицы поляризуемости.

Все линейные уравнения для прогнозирования индексов удерживания построены с помощью метода наименьших квадратов, оценки точности выполнены с помощью «10-fold» кросс-валидации: набор данных случайно разбивался на 10 подмножеств, каждое из которых по очереди рассматривалось в качестве тестового набора. Кросс-валидация повторена 200 раз со случайным удалением одной молекулы из набора данных и случайным разбитием на 10 подмножеств. В качестве метрик для оценки точности использовались среднеквадратичная ошибка (СКО), средняя абсолютная ошибка (САО), медианная абсолютная ошибка (МАО) и коэффициент детерминации R^2.

Обсуждение результатов

Отбор молекулярных дескрипторов для построения моделей. На рис. 2 показана зависимость метрик точности (СКО и МАО) от количества отобранных МД для четырех различных наборов МД. Все четыре набора включают полный набор МД, рассчитанных при помощи пакета RDKit. Три набора также включают один из МД RI_DB-5, RI_DB-624 или RI_DB-

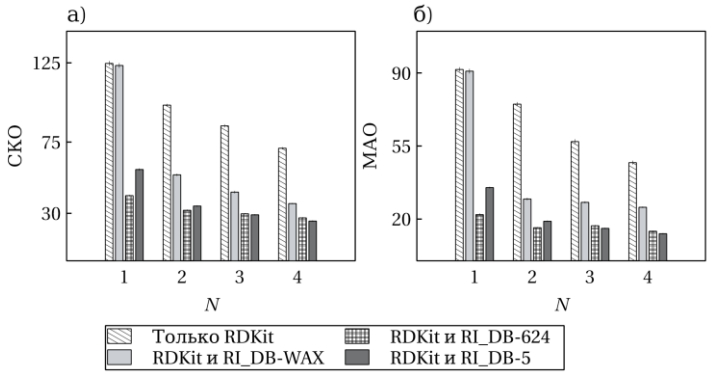


Рис. 2. Зависимость точности прогнозирования (среднеквадратичной (а) и медианной абсолютной (б) ошибки) индекса удерживания для неподвижной фазы DB-35MS от количества молекулярных дескрипторов (*N*), выбранных для четырех наборов молекулярных дескрипторов.

Fig. 2. Dependence of the prediction accuracy (root-mean-square (a) and median absolute (b) errors) of the retention index for the stationary phase DB-35MS on the number of molecular descriptors (*N*) selected for four sets of molecular descriptors.

WAX. Следует отметить, что использование индексов удерживания, полученных *in silico* с помощью глубокого обучения, приводит к чрезвычайно сильному повышению точности по сравнению с МД из пакета RDKit. Наилучшая точность достигается при использовании RI_DB-5, на втором месте идет набор, содержащий RI_DB-624. Для этих наборов при использовании пошагового метода отбора на первом шаге всегда отбираются именно эти МД. Для набора, содержащего RI_DB-WAX, этот МД отбирается на втором шаге пошагового алгоритма, поэтому, когда количество используемых МД равно 1, то точность с использованием RI_DB-WAX совпадает с точностью, достигаемой без его использования. Ошибка сильно уменьшается во всех случаях при переходе от 1 к 2 отбираемым МД. Это показывает, что RI_DB-5, RI_DB-624 и RI_DB-WAX в одиночку не способны обеспечить высокой точности прогноза, и именно использование их вместе с другими МД в качестве одного из МД позволяет добиться высокой точности прогноза.

При использовании большого числа МД (>5) точность (даже определяемая с помощью кросс-валидации) существенно улучшается. Однако это достигается за счет «запоминания» моделью особенностей набора данных, содержащего много однотипных соединений, и приводит к нефизичным уравнениям, не имеющим реальной прогностической силы. Чтобы избежать таких уравнений, мы ограничились уравнениями, содержащими четыре МД: индекс удерживания, полученный с помощью глубокого обучения, и еще три МД. Далее рассматриваются именно такие уравнения. Также ниже везде используются набор МД, содержащий МД, сгенерированные с помощью пакета RDKit, и МД RI_DB-5, дающий наилучшую точность.

На рис. 3 показаны результаты оценки важности МД, характеризуемой вероятностью *p* быть отобранным с помощью пошагового метода. Вероятность быть отобранным существенно отличается от 1. Это показывает, что использование однократной процедуры отбора не может быть надежным методом оценки важности МД, и из результатов такой процедуры нельзя делать физико-химических выводов. МД FpDensityMorgan2 характеризует количество ненулевых битов в молекулярном отпечатке пальцев, что, в свою очередь, характеризует количество и разнообразие различных фрагментов (субструктур) в молекуле. МД MolMR

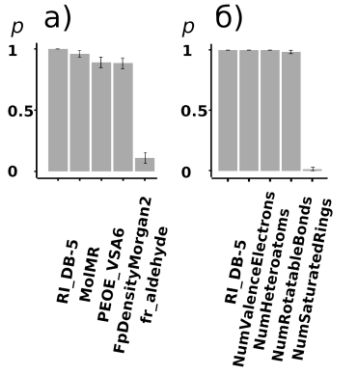


Рис. 3. Вероятность *p* выбора молекулярного дескриптора при прогнозировании индекса удерживания; использовался полный набор молекулярных дескрипторов из пакета RDKit (а) или сокращенный набор целочисленных молекулярных дескрипторов с четким физическим смыслом (б), а также молекулярный дескриптор RI_DB-5.

Fig. 3. Probability *p* of choosing a molecular descriptor when predicting the retention index; a complete set of molecular descriptors from the RDKit package (a) or an abbreviated set of integer molecular descriptors with a clear physical meaning (b), as well as the RI_DB-5 molecular descriptor were used.

означает оцененную с помощью аддитивной схемы величину молекулярной рефракции. МД PEOE_VSA6 характеризует доступную поверхность атомов, заряд по Гастейгеру для которых находится в диапазоне [− 0.1; − 0.05]. Несмотря на то, что эти МД не являются «черным ящиком» и имеют определенный физический смысл, вычислить их без использования специального программного обеспечения невозможно, а физический смысл весьма сложный для интерпретации.

В связи с этим был рассмотрен сокращенный набор МД, содержащий лишь целочисленные МД с прозрачным физическим смыслом, очевидным из названия. Так, МД NumRotatableBonds, NumHeteroatoms, NumValenceElectrons означают количество связей, подверженных внутреннему вращению, количество атомов, не являющихся атомами углерода и водорода, и количество валентных электронов. Необходимо уточнить, что NumHeteroatoms – число всех атомов, кроме углерода и водорода, в том числе не входящих в цикл и в главную цепь. Так как наш набор данных содержит в себе только углеводороды и кислородсодержащие молекулы, то фактически речь идет о числе атомов кислорода. Более подробное обсуждение границ применимости наших моделей дано ниже. Эти МД могут быть рассчитаны даже без применения специального программного обеспечения. Их использование является предпочтительным для получения простых и интерпретируемых уравнений для прогнозирования индексов удерживания.

Добавление к рассматриваемым наборам МД, рассчитанных с помощью теории функционала плотности: дипольного момента и поляризуемости, не влияет на описанные выше результаты. В данной работе МД для построения уравнений отбираются не «вручную» на основе априорной важности, а автоматически, с помощью пошагового алгоритма, на каждом шагу которого выбирается МД [11], наиболее сильно повышающий значимость линейной регрессии (см. выше). К сожалению, ни при использовании полного набора МД RDKit, ни при использовании сокращенного набора МД с понятным физическим смыслом, вне зависимости от того, используются или нет рассчитанные с помощью машинного обучения индексы удерживания в качестве МД, дипольный момент и поляризуемость не

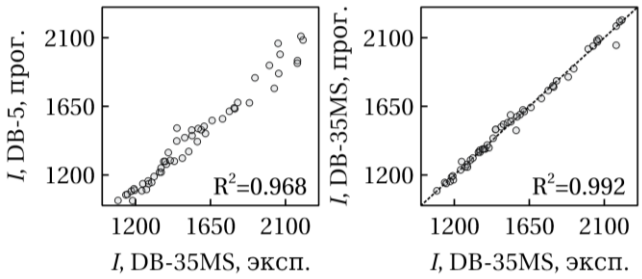
Таблица 1. Точность прогноза индексов удерживания для неподвижной фазы DB-35MS с помощью линейных моделей; приведены средние значения метрик точности для 200 повторов кросс-валидации и доверительные интервалы (N = 200; p = 0.95; отбирается не более 4 молекулярных дескрипторов; молекулярная рефракция, молекулярная масса, поляризуемость (атомные единицы) и дипольный момент (Дб) обозначены MolMR, *M,* α, μ, соответственно

Table 1. Accuracy of prediction of retention indices for the stationary phase DB-35MS using linear models; average values of accuracy metrics for 200 cross-validation repeats and confidence intervals are given (N = 200; p = 0.95; no more than 4 molecular descriptors are selected; molecular refraction, molecular weight, polarizability (atomic units) and the dipole moment (Db) are denoted by MolMR, *M*, α, μ, respectively

Набор молекулярных дескрипторов	СКО	САО	МАО
Полный набор RDKit	71.6 ± 0.6	57.4 ± 0.5	47.6 ± 0.8
Полный набор + RI_DB-624	27.4 ± 0.3	19.7 ± 0.2	14.0 ± 0.2
Полный набор + RI_DB-5	25.2 ± 0.3	18.3 ± 0.2	12.8 ± 0.2
Сокращенный набор RDKit	91.4 ± 1.0	72.6 ± 0.9	63.9 ± 1.6
Сокращенный набор + RI_DB-624	27.7 ± 0.3	20.5 ± 0.3	16.1 ± 0.3
Сокращенный набор + RI_DB-WAX	37.0 ± 0.6	27.7 ± 0.3	22.3 ± 0.5
Сокращенный набор + RI_DB-5	28.9 ± 0.5	19.3 ± 0.3	11.8 ± 0.3
Только RI_DB-624	41.2 ± 0.4	31.0 ± 0.3	22.5 ± 0.5
Только RI_DB-5	57.4 ± 0.6	44.5 ± 0.5	34.7 ± 0.3
RI_DB-WAX, RI_DB-624, RI_DB-5	33.0 ± 0.4	24.6 ± 0.3	21.4 ± 0.3
M, α, μ	144.2 ± 2.2	103.8 ± 1.3	75.6 ± 1.2
MolMR, *M,* α, μ	139.9 ± 1.2	102.8 ± 1.2	84.7 ± 1.1
RI_DB-5, MolMR, *M,* α, μ	34.1 ± 0.3	25.5 ± 0.3	17.2 ± 0.2

отбираются пошаговым алгоритмом в числе первых четырех МД.

Линейные уравнения для прогнозирования индексов удерживания. В таблице 1 показана точность прогнозирования индексов удерживания, достигаемая при использовании разных наборов МД. Хорошо видно, что при использовании RI_DB-624 или RI_DB-5 точность, достигаемая при применении полного и сокращенного наборов МД, полученных с помощью пакета RDKit, практически не отличается. Таким образом, использование именно сокращенного набора МД из пакета RDKit является наиболее целесообразным. Также видно, что относительно высокую точность можно получить с использованием RI_DB-WAX, RI_DB-624 и RI_DB-5 без применения МД из пакета RDKit. В целом достигнутая точность является весьма высокой примерно на том же уровне, что и точность прогноза для аналогичных соединений (эфирные масла и летучие соединения растительного происхождения) на других неподвижных фазах [10]. В работе [14], данные из которой использованы в настоящей работе, информация об индексах удерживания приведена для трех скоростей нагрева (режим программирования температуры). Среднеквадратичное, среднее абсолютное и медианное абсолютное отклонение между индексами удерживания, зарегистрированными при наименьшей и наибольшей скоростях нагрева, составляют 10.2, 6.7 и 4.5 соответственно. Это меньше, чем достигнутая ошибка прогнозирования, однако ошибка прогнозирования достаточно низкая,



а б

Рис. 4. Корреляция между экспериментальным (эксп.) индексом удерживания (*I*) для неподвижной фазы DB-35MS и спрогнозированными (прог.) индексами удерживания для неподвижных фаз DB-5 (а) и DB-35MS (б).

Fig. 4. Correlation between the experimental (exp.) retention index (I) for the stationary phase DB-35MS and the predicted (prog.) retention indices for the stationary phases DB-5 (a) and DB-35MS (b).

чтобы использовать такой прогноз при масс-спектральной идентификации.

На рис. 4 показаны корреляции между наблюдаемыми индексами удерживания для неподвижной фазы DB-35MS и спрогнозированными индексами удерживания для неподвижных фаз DB-5 и DB-35MS. Из рис. 4 и таблицы 1 видно, что уравнения, содержащие спрогнозированный с помощью глубокого обучения индекс удерживания и три МД из пакета RDKit, дают существенно лучшую точность прогноза, чем уравнения, содержащие только один полученный с помощью глубокого обучения индекс удерживания. Это показывает, что такие индексы удерживания должны рассматриваться именно как МД при прогнозировании индексов удерживания для нестандартных неподвижных фаз. Не давая высокой точности в качестве единственного значения-предиктора, они позволяют построить точное уравнение вместе с другими МД.

Для сравнения была оценена точность прогнозирования индексов удерживания с помощью линейных уравнений, не включающих МД из пакета RDKit, отбираемых с помощью пошагового алгоритма. При этом были рассмотрены следующие МД: дипольный момент и поляризуемость, вычисленные с помощью теории функционала плотности (программное обеспечение nwchem), молекулярная рефракция (пакет RDKit, линейная модель), молекулярная масса, индекс удерживания для неподвижной фазы DB-5 (программное обеспечение SVEKLA). Соответствующие метрики точности также приведены в таблице 1. Видно, что применение величин, рассчитанных с помощью методов квантовой химии, не позволяет достигнуть высокой точности прогнозирования индексов удерживания без использования полученного с помощью глубокого обучения индекса удерживания для неподвижной фазы DB-5.

В таблице 2 приведены примеры готовых линейных уравнений, которые можно использовать для оценки индекса удерживания для неподвижной фазы DB-35MS. Спрогнозированные с помощью глубокого обучения индексы удерживания можно рассчитать с помощью разработанного нашей группой интерактивного программного обеспечения с открытым исходным кодом SVEKLA [37] и CHERESHNYA [25]. Программное обеспечение является бесплатным и доступно онлайн. Значения молекулярных дескрипторов из пакета RDKit можно также рассчитать с помощью программного обеспечения CHERESHNYA [25] или непосредственно с помощью пакета RDKit [26]. МД из сокращенного набора

Таблица 2. Уравнения для прогнозирования индексов удерживания на неподвижной фазе DB-35MS ($RI_{DB-35MS}$), исходя из молекулярных дескрипторов и индекса удерживания для неподвижной фазы DB-5, полученного с помощью машинного обучения; молекулярная рефракция, молекулярная масса, поляризуемость (атомные единицы) и дипольный момент (Дб) обозначены MolMR, *M,* α, μ соответственно

Table 2. Equations for predicting retention indices in the non-mobile phase DB-35MS (RIDB-35MS) based on molecular descriptors and retention index for the stationary phase DB-5 obtained using machine learning; molecular refraction, molecular weight, polarizability (atomic units) and dipole moment (Db) are indicated by MolMR, *M*, α, μ, respectively

N	Equation	R^2
1	$RI_{DB-35MS}$ = -41.8 + 1.3444 * RI_DB-5 - 5.06 * NumRotatableBonds + 30.02 * NumHeteroatoms - 4.39 * NumValenceElectrons	0.992
2	$RI_{DB-35MS}$ = -74.2 + 0.1353 * RI_DB-WAX + 0.6263 * RI_DB-624 + 0.3017 * RI_DB-5	0.989
3	$RI_{DB-35MS}$ = 82.6 + 1.0225 * RI_DB-5	0.968
4	$RI_{DB-35MS}$ = 406.6 - 1.39 * α + 17.97 * μ + 9.09 * MolMR + 4.01 * *M*	0.798
5	$RI_{DB-35MS}$ = 7.4 + 1.4458 * RI_DB-5 + 0.33 * α - 4.01 * μ - 9.52 * MolMR	0.988

МД сравнительно легко рассчитать непосредственно из структуры даже без специального программного обеспечения. Таким образом, представленные в таблице 2 уравнения являются готовыми к использованию и могут быть применены на практике. Все многомерные корреляции, приведенные в таблице 2 являются статистически значимыми, значение *f*-теста [11] находится в диапазоне ~50-2000, вероятность того, что наблюдаемая корреляция является случайной (*p*-value), не превосходит 10^{-16} для всех уравнений.

Достигнутая в этой работе точность прогнозирования индексов удерживания находится на том же уровне, что и точность моделей, использующих глубокое обучение для других неподвижных фаз высокой и промежуточной полярности и аналогичных соединений (разнообразные летучие соединения растительного происхождения, эфирные масла). Модели, не использующие глубокое обучение для аналогичных соединений и неподвижных фаз, дают существенно худшую точность [38-42]. Соответствующие значения метрик точности приведены в таблице 3. Авторам неизвестны работы, в которых для каких-либо неподвижных фаз была достигнута существенно более высокая точность прогнозирования индексов удерживания для разнообразных соединений, принадлежащих к разным классам, в т.ч. полифункциональных. Соизмеримую точность демонстрируют лучшие и наиболее современные из прогностических моделей [43] для неполярных неподвижных фаз (например, DB-1, DB-5), однако для неподвижных фаз более высокой полярности (см. таблицу 3) точность существующих моделей намного ниже. Весьма точными могут быть и некоторые модели [44-45], обученные для очень узкого класса химических соединений, например, серии однотипных изомеров или нескольких гомологических рядов, похожих по своей структуре. Однако это существенно более простая задача, и такие модели не могут быть использованы на практике при анализе большинства реальных объектов.

Прогнозирование индексов удерживания для разных скоростей нагрева. Хорошо известно, что индексы удерживания зависят от температуры в изотермическом режиме [46]. В режиме программирования температуры индексы удерживания также зависят от скорости нагрева, при этом для одних веществ эта

Таблица 3. Сравнение приведенных в предшествующих работах метрик точности прогнозирования индексов удерживания компонентов эфирных масел и других летучих веществ растительного происхождения для различных неподвижных фаз и наборов данных
Table 3. Comparison of the accuracy metrics given in previous studies for predicting retention indices of components of essential oils and other volatile substances of plant origin for various non-mobile phases and data sets

Неподвижная фаза	СКО	САО	МАО	Количество молекул	Источник
DB-35MS	28.9	19.3	11.8	52	Эта работа
Полиэтиленгликоль	125.4	89.8	68.6	1169	[38]
Полиэтиленгликоль	177.3	139.0	123.5	427	[39]
Полиэтиленгликоль	86.1	46.6	26.1	1169	[10]*
Полиэтиленгликоль	58.8	40.4	26.4	427	[10]*
DB-624	54.2	32.5	19.2	545	[40]
DB-624	36.8	24.3	16.7	545	[10]*
DB-225	179.9	**	**	269	[41]
DB-1701	56.1	37.1	20.0	36	[10]*
OV-17	58.8	47.3	42.9	192	[42]
OV-17	43.8	30.7	22.5	192	[10]*

*работа научной группы авторов; **не указано в источнике

зависимость выражена сильно, а для других практически отсутствует [24]. Кроме того, индекс удерживания зависит и от других экспериментальных условий, например, от толщины слоя неподвижной фазы, начальной температуры [47]. В идеале для прогнозирования индексов удерживания должны учитываться не только структура молекулы и тип (структура) неподвижной фазы, но и все экспериментальные условия. К сожалению, на практике это невозможно из-за отсутствия соответствующих экспериментальных данных. Частичный же учет условий, например, обучение различных моделей для различных скоростей нагрева может не привести к значительному повышению точности, так как зависимость индекса удерживания от скорости нагрева неодинакова для различных хроматографических систем [24]. Так как учет всех экспериментальных условий невозможен, то разумно применить приближение, в котором индекс удерживания зависит только от структуры молекулы и типа неподвижной фазы, а влияние остальных факторов рассматривать как случайную ошибку. В связи с этим наиболее частым [7-10, 43] является подход (использованный и в этой работе), в котором разница в условиях, в которых получены индексы удерживания, игнорируется и осуществляется прогнозирование неких «усредненных» индексов удерживания.

Тем не менее, так как в нашем случае доступны индексы удерживания для трёх скоростей нагрева (режим программирования температуры) и интересно рассмотреть точность прогноза для каждой из скоростей нагрева в отдельности. Такие данные (аналогичные таблице 1) приведены в таблице 4. Все основные закономерности, касающиеся значимости тех или иных МД, в этом случае такие же, как и при использовании «усредненных» индексов удерживания. Точность прогноза (значения метрик точности) в этом случае так же почти не отличается от представленных в таблице 1. Это доказывает обос

Таблица 4. Точность прогноза индексов удерживания для неподвижной фазы DB-35MS с помощью линейных моделей; различные скорости нагрева *r* (°C/мин); обозначения аналогичные использованным в таблице 1.
Table 4. Accuracy of prediction of retention indices for a fixed phase DB-35MS using linear models; different heating rates r (°C/min); designations similar to those used in Table 1.

Набор молекулярных дескрипторов	*r* = 2			*r* = 4			*r* = 6		
	СКО	САО	МАО	СКО	САО	МАО	СКО	САО	МАО
Полный набор RDKit	70.1 ± 0.7	55.8 ± 0.7	45.9 ± 1.0	73.1 ± 0.6	59.1 ± 0.6	50.2 ± 0.8	72.5 ± 0.7	57.7 ± 0.6	48.2 ± 0.9
Полный набор + RI_DB-5	24.4 ± 0.3	17.7 ± 0.2	13.6 ± 0.2	27.2 ± 0.3	20.0 ± 0.2	14.3 ± 0.2	24.9 ± 0.3	18.4 ± 0.2	13.3 ± 0.2
Сокращенный набор + RI_DB-5	28.2 ± 0.5	19.2 ± 0.3	12.8 ± 0.2	30.5 ± 0.5	20.2 ± 0.3	12.5 ± 0.2	27.8 ± 0.5	18.5 ± 0.3	12.4 ± 0.3
Только RI_DB-5	55.9 ± 0.6	43.2 ± 0.4	32.9 ± 0.2	59.9 ± 0.6	45.9 ± 0.5	35.9 ± 0.4	57.9 ± 0.7	45.2 ± 0.4	35.8 ± 0.5
MolMR, *M,* α, μ	141.0 ± 2.0	102.8 ± 1.2	82.7 ± 1.2	143.2 ± 2.2	104.7 ± 1.2	88.8 ± 1.3	144.2 ± 2.3	104.8 ± 1.2	85.4 ± 1.4
RI_DB-5, MolMR, *M,* α, μ	32.8 ± 0.3	23.8 ± 0.3	16.2 ± 0.2	36.0 ± 0.4	27.5 ± 0.3	20.0 ± 0.3	34.2 ± 0.3	26.5 ± 0.3	20.6 ± 0.3

нованность подхода с усреднением индексов удерживания для различных скоростей нагрева.

Тестирование с применением внешнего набора данных и границы применимости. Обучающий набор, использованный в данной работе, не содержит в себе молекул, содержащих атомы азота, фосфора, серы, галогенов (рис. 1). Таким образом, маловероятно, что формулы из таблицы 2 позволят получить точные прогнозы для таких молекул. Область применимости этих формул ограничена лишь соединениями, аналогичными представленным на рис. 1, т. е. состоящим из углерода, водорода и кислорода, летучим соединениям растительного происхождения. В работе [20] имеются данные об временах удерживания ряда пестицидов и токсикантов на неподвижной фазе DB-35MS. Большинство этих соединений содержат серу, фосфор или хлор. Нами были выбраны 8 из них, которые содержат лишь углерод, водород, кислород и азот. Логарифмы времен удерживания коррелируют со спрогнозированными индексами удерживания для неподвижных фаз DB-5 и DB-35MS. На рис. 5 показаны корреляции между спрогнозированными индексами удерживания, а также спрогнозированными авторами работы [20] логарифмами времен удерживания и экспериментально наблюдаемыми логарифмами времен удерживания.

Наиболее сильная корреляция наблюдается со спрогнозированными индексами удерживания для неподвижной фазы DB-5, на втором месте идет прогноз, сделанный авторами [20] с помощью регрессии на опорных векторах. Самая слабая корреляция наблюдается с прогнозом, сделанным авторами [20] с помощью линейной модели. Все наблюдаемые значения весьма близки. Авторы работы [20] обучали свои модели для прогнозирования времен удерживания для одного конкретного хроматографического режима. В нашем случае наши универсальные модели показывают примерно такую же или даже немного луч-

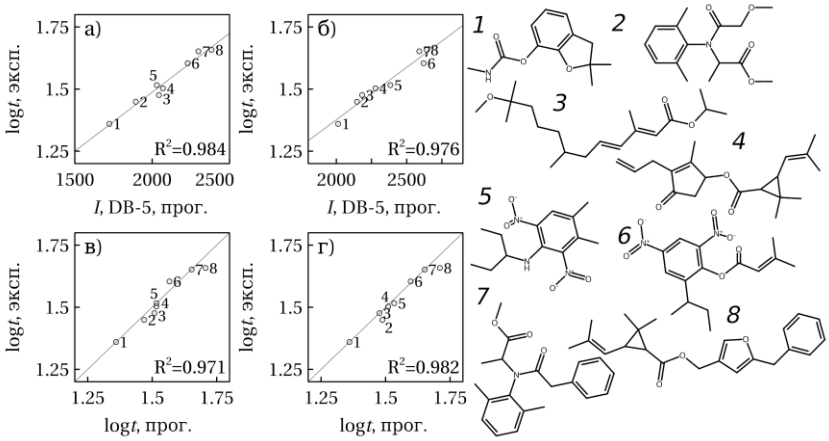


Рис. 5. Корреляция между наблюдаемым (эксп.) логарифмом времени удерживания (мин.) на неподвижной фазе DB-35MS для ряда соединений и спрогнозированными (прог.) величинами. В качестве спрогнозированных величин выступают: а) индекс удерживания на неподвижной фазе DB-5, спрогнозированный с помощью глубокого обучения; б) индекс удерживания на неподвижной фазе DB-35MS; в) логарифм времени удерживания, спрогнозированный с помощью линейной модели; г) логарифм времени удерживания, спрогнозированный с помощью метода опорных векторов.

Fig. 5. Correlation between the observed (exp.) logarithm of retention time (min.) in the stationary phase DB-35MS for a number of compounds and predicted (prog.) values. The predicted values are: a) the index of retention on the stationary phase DB-5, predicted using deep learning; b) the index of retention in the stationary phase DB-35MS; c) the logarithm of retention time, predicted using a linear model; d) the logarithm of retention time, predicted using the support vector regression method.

шую точность по сравнению с построенными ими моделями. Однако для соединений, содержащих атомы других элементов, помимо углерода, кислорода, водорода, можно порекомендовать использовать более универсальное уравнение 3 из таблицы 2, так как такие соединения находятся за границами применимости уравнения 1.

Модель, использованная для расчета RI_DB-5, была обучена с использованием базы данных NIST 17. Необходимо отметить, что для соединений, которые входят в данную базу данных, может наблюдаться существенно лучшая точность прогноза, чем для соединений, которые в нее не входят, это будет влиять и на точность прогнозирования индекса удерживания для неподвижной фазы DB-35MS. Однако, т. к. в базу данных NIST 17 входит на несколько порядков больше соединений, чем количество соединений, для которых есть информация об индексах удерживания для неподвижной фазы DB-35MS, уравнения из таблицы 2 могут быть использованы для оценки этих индексов удерживания.

Заключение

Прогнозирование индексов удерживания на основе структуры молекулы является важной задачей: так как базы данных содержат индексы удерживания далеко не для всех соединений и лишь для стандартных неподвижных фаз, то именно такие прогнозы могут быть использованы при хромато-масс-спектрометрической идентификации. Неподвижные фазы на основе 35%-фенил-метилполисилоксана широко применяются на практике, однако надежные и универсальные подходы к прогнозированию индексов удерживания для таких неподвижных фаз на данный момент отсутствуют. В ходе данной работы были разработаны модели для прогнозирования индексов удерживания для таких неподвижных фаз.

Добиться удовлетворительной точности прогноза для таких неподвижных фаз возможно только с применением «двухстадийного подхода». Модели, основанные на глубоких нейронных сетях, обученные с использованием базы данных NIST 17, используются для прогнозирования индекса удерживания для популярных неподвижных фаз. Затем строится линейное уравнение, включающее в себя эти спрогнозированные индексы удерживания и молекулярные дескрипторы, для вычисления требуемых индексов удерживания. Индексы удерживания, полученные с помощью глубокого обучения, с использованием ранее опубликованных моделей, могут быть рассмотрены именно как молекулярные дескрипторы, согласно определению – это численные величины, характеризующие молекулу, которые могут быть быстро и однозначно рассчитаны на основе структуры. Такой подход существенно превосходит по точности традиционный подход, не использующий глубокое обучение.

При построении моделей был использован набор данных об индексах удерживания 52 летучих органических веществ, содержащихся в бутонах сирени. Часть из этих веществ являются углеводородами, остальные состоят из кислорода, водорода и углерода. Соединения имеют довольно сложную и разнообразную структуру, многие являются полифункциональными. Соответственно, область применимости наиболее точной из полученных моделей ограничена таким классом молекул: разнообразные летучие органические соединения, не содержащие азот, серу и другие элементы, в частности растительного происхождения. Для оценки индексов удерживания других соединений, в частности азотсодержащих, может быть рекомендовано уравнение, включающее в себя только индекс удерживания для неподвижной фазы DB-5, полученный с помощью глубокого обучения.

Топологические молекулярные дескрипторы дают лишь незначительное увеличение точности по сравнению с простыми целочисленными молекулярными дескрипторами (такими как число валентных электронов и число связей, подверженных внутреннему вращению). Дипольный момент и поляризуемость, рассчитанные с помощью теории функционала плотности, также не дают никаких преимуществ в точности по сравнению с простейшими целочисленными молекулярными дескрипторами. Точность моделей для прогнозирования индексов удерживания характеризуется значениями среднеквадратичной ошибки, средней абсолютной ошибки и медианной абсолютной ошибки, которые составляют 28.9, 19.3 и 11.8 единиц соответственно. Точность полученной модели выше, чем точность известных авторам моделей для прогнозирования индексов удерживания аналогичных соединений: органических молекул растительного происхождения (в том числе компонентов эфирных масел), разнообразной структуры. Все расчеты могут быть выполнены с помощью разработанного нами свободного программного обеспечения с открытым исходным кодом CHERESHNYA [25].

Конфликт интересов

Авторы заявляют, что у них нет известных финансовых конфликтов интересов или личных отношений, которые могли бы повлиять на работу, представленную в этой статье.

Список литературы/References

1. Tarjan G., Nyiredy S., Györ M., Lombosi E. R., Lombosi T. S., Budahegyi M. V., Mészáros S.Y., Takács J. M., Thirtieth anniversary of the retention index according to Kováts in gas-liquid chromatography, *Journal of Chromatography A*, 1989; 472: 1-92. https://doi.org/10.1016/S0021-9673(00)94099-8

2. Sholokhova A.Yu., Borovikova S.A., Matyushin D.D., Buryak A.K., Identifikatsiya komponentov ozonirovannoi piroliznoi zhidkosti s ispol'zovaniem gazovoi khromato-mass-spektrometrii, ionnoi zhidkosti v kachestve nepodvizhnoi fazy i mashinnogo obucheniya, Sorbtsionnye i khromatograficheskie protsessy, 2022; 22(4): 413-420. https://doi.org/10.17308/sorpchrom.2022.22/10570 (In Russ.)

3. Kováts E., Gas-chromatographische charakterisierung organischer verbindungen. Teil 1: retentions indices aliphatischer halogenide, alkohole, aldehyde und ketone, *Helvetica Chimica Acta,* 1958; 41(7): 1915-1932. https://doi.org/10.1002/hlca.19580410703

4. Van Den Dool H., Dec. Kratz P., A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography, *Journal of Chromatography A,* 1963; 11: 463-471. https://doi.org/10.1016/S0021-9673(01)80947-X

5. Karnaeva A.E., Sholokhova A.Yu., Validation of the identification reliability of known and assumed UDMH transformation products using gas chromatographic retention indices and machine learning, *Chemosphere,* 2024; 362: 142679. https://doi.org/10.1016/j.chemosphere.2024.142679

6. Khrisanfov M.D., Matyushin D.D., Samokhin A.S., A general procedure for finding potentially er-roneous entries in the database of retention indices, *Analytica Chimica Acta,* 2024; 1297: 342375. https://doi.org/10.1016/j.aca.2024.342375

7. Qu C., Schneider B. I., Kearsley A. J., Keyrouz W., Allison T. C., Predicting Kováts Retention Indices Using Graph Neural Networks, *Journal of Chromatography A,* 2021; 1646: 462100. https://doi.org/10.1016/j.chroma.2021.462100

8. Anjum A., Liigand J., Milford R., Gautam V., Wishart D. S., Accurate prediction of isothermal gas chromatographic Kováts retention indices, *Journal of Chromatography A,* 2023; 1705: 464176. https://doi.org/10.1016/j.chroma.2023.464176

9. Matyushin D.D., Sholokhova A.Yu., Buryak A.K., A deep convolutional neural network for the estimation of gas chromatographic retention indices, *Journal of Chromatography A,* 2019; 1607: 460395. https://doi.org/10.1016/j.chroma.2019.460395

10. Matyushin D.D., Sholokhova A.Yu., Buryak A.K., Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases, *International journal of molecular sciences,* 2021: 22(17): 9194. https://doi.org/10.3390/ijms22179194

11. Sholokhova A.Yu., Matyushin D.D., Shashkov M.V., Quantitative structure-retention relationships for pyridinium-based ionic liquids used as gas chromatographic stationary phases: convenient software and assessment of reliability of the results, *Journal of Chromatography A,* 2024; 1730: 465144. https://doi.org/10.1016/j.chroma.2024.465144

12. Li M., Li R., Wang Z., Zhang Q., Bai H., Lv Q., Optimization of headspace for GC-MS analysis of fragrance allergens in wooden children's products using response surface methodology, *Separation Science Plus*, 2019; 2(1): 26-37. https://doi.org/10.1002/sscp.201800125

13. Evdokimova M.A., Onuchak L.A., Kuraeva Yu.G., Platonov V.I., Termodinamicheskie aspekty sorbtsii i razdeleniya enantiomerov nekotorykh monoterpenov na kapillyarnoi kolonke β-DEX 120, Sorbtsionnye i khromatograficheskie protsessy, 2015; 15(2): 288-300. (In Russ.)

14. Zhao C. X., Liang Y. Z., Fang H. Z., Li X. N., Temperature-programmed retention indices for gas chromatography–mass spectroscopy analysis of plant essential oils, *Journal of Chromatography A,* 2005; 1096(1-2): 76-85. https://doi.org/10.1016/j.chroma.2005.09.067

15) Volkova G.I., Zubarev D.A., Kadychagov P.B., Effect of Ultrasonic treatment on the properties and composition of High-

Wax crude oil and its precipitates, *Petroleum Chemistry*, 2024; 1-8. https://doi.org/10.1134/S0965544124020026

16. Yu P., Banh R., Sohn A., Martis S., Biancur D., Yamamoto K., Lin E., Kimmelman A., Topographical investigation of metabolites in excised squares (TIMES2): Comprehensive cross-sectional metabolite quantification of pancreatic cancer in vivo, *Cancer Research*, 2024; 84(6_Supplement): 4440-4440. https://doi.org/10.1158/1538-7445.AM2024-4440

17. Hassanzadeh Z., Ebrahimi P., Kompany-Zareh M., Ghavami R., Radial basis function neural networks based on projection pursuit approach and solvatochromic descriptors: single and full column prediction of gas chromatography retention behavior of polychlorinated biphenyls, *Journal of Chemometrics*, 2016; 30 (10): 589-601. https://doi.org/10.1002/cem.2822

18. Ghavami R., Sadeghi F., QSRR-based evaluating and predicting of the relative retention time of polychlorinated biphenyl congeners on 18 different high resolution GC columns, *Chroma*, 2009; 70(5-6): 851-868. https://doi.org/10.1365/s10337-009-1233-6

19. Zhao C. X., Zhang T., Liang Y. Z., Yuan D. L., Zeng Y. X., Xu Q. S., Conversion of programmed-temperature retention indices from one set of conditions to another, *Journal of Chromatography A*, 2007; 1144 (2): 245-254. https://doi.org/10.1016/j.chroma.2007.01.040

20. Li X., Luan F., Si H., Hu Z., Liu M., Prediction of retention times for a large set of pesticides or toxicants based on support vector machine and the heuristic method, *Toxicology letters*, 2007; 175(1-3): 136-144. https://doi.org/10.1016/j.toxlet.2007.10.005

21. Arruda A. C., Ampliação e aplicação do método semi-empírico topológico (IET) em modelos QSRR/QSPR/QSAR para compostos alifáticos halogenados e cicloalcanos, 2008. https://repositorio.ufsc.br/xmlui/handle/123456789/91111 (дата обращения: 27.07.2024)

22. Poole C. F., Qian J., Kiridena W., DeKay C., Koziol W. W., Evaluation of the separation characteristics of application-specific (volatile organic compounds) open-tubular columns for gas chromatography,*Journal of Chromatography A*, 2006; 1134(1-2): 284-290. https://doi.org/10.1016/j.chroma.2006.08.092

23. Zaitseva E. A., Obzor metodov klassifikatsii nepodvizhnykh faz v gazovoi khromatografii, Sorbtsionnye i khromatograficheskie protsessy, 2020; 20(2): 175-196. https://doi.org/10.17308/sorpchrom.2020.20/2772 (In Russ.)

24. Matyushin D. D., Sholokhova A. Y., Large-scale statistical study of the dependence of retention index on heating rate in temperature-programmed gas chromatography, *Journal of Chromatography A*, 2024; 1732: 465223. https://doi.org/10.1016/j.chroma.2024.465223

25. https://github.com/mtshn/chereshnya (дата обращения: 27.07.2024)

26. https://www.rdkit.org (дата обращения: 27.07.2024)

27. Pearlman R. S., Smith K. M., Metric validation and the receptor-relevant subspace concept, *Journal of Chemical Information and Computer Sciences*, 1999; 39(1): 28-35. https://doi.org/10.1021/ci980137x

28. Pearlman R. S., Smith K. M. Novel software tools for chemical diversity // 3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity. Dordrecht: Springer Netherlands. 2002: 339-353. https://doi.org/10.1007/0-306-46857-3_18

29. Wildman S. A., Crippen G. M., Prediction of physicochemical parameters by atomic contributions, *Journal of chemical information and computer sciences*, 1999; 39(5): 868-873. https://doi.org/10.1021/ci990307l

30. Consonni V., Todeschini R., Molecular descriptors, *Recent advances in QSAR studies: methods and applications*, 2010; 29-102. https://doi.org/10.1007/978-1-4020-9783-6_3

31. Valiev M., Bylaska E.J., Govind N. Kowalski K., Straatsma T.P., Van Dam H.J.J., Wang D., Nieplocha J., Apra E., Windus T.L., de JongW.A., NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations, *Computer Physics Communications,* 2010; 181(9): 1477-1489. https://doi.org/10.1016/j.cpc.2010.04.018

32. Adamo C., Barone V., Toward reliable density functional methods without adjustable parameters: The PBE0 model, *The Journal of chemical physics,* 1999; 110(13): 6158-6170. https://doi.org/10.1063/1.478522

33. Yoshikawa N., Hutchison G. R., Fast, efficient fragment-based coordinate generation for Open Babel, *Journal of cheminformatics,* 2019; 11(1): 49. https://doi.org/10.1186/s13321-019-0372-5

34. Smith J. S., Nebgen B. T., Zubatyuk R., Lubbers N., Devereux C., Barros K., Roitberg A. E., Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, *Nature communications,* 2019, 10(1): 2903. https://doi.org/10.1038/s41467-019-10827-4

35. Gao X., Ramezanghorbani F., Isayev O., Smith J. S., Roitberg A. E., TorchANI: a free and open source PyTorch-based deep learning implementation of the ANI neural network potentials, *Journal of chemical information and modeling,*2020; 60(7): 3408-3415. https://doi.org/10.1021/acs.jcim.0c00451

36. Zheng P., Yang W., Wu W., Isayev O., Dral P. O., Toward chemical accuracy in predicting enthalpies of formation with general-purpose data-driven methods, *The Journal of Physical Chemistry Letters,* 2022; 13(15): 3479-3491. https://doi.org/10.1021/acs.jpclett.2c00734

37. https://github.com/mtshn/svekla (дата обращения: 27.07.2024)

38. Rojas Villa C. X., Duchowicz P. R., Tripaldi P., Pis Diez R., Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase, *Journal of Chromatography A,* 2015; 1422: 277-288. https://doi.org/10.1016/j.chroma.2015.10.028

39. Qin L. T., Liu S. S., Chen F., Wu Q. S., Development of validated quantitative structure–retention relationship models for retention indices of plant essential oils, *Journal of separation science,* 2013; 36(9-10): 1553-1560. https://doi.org/10.1002/jssc.201300069

40. Dossin E., Martin E., Diana P., Castellon A., Monge A., Pospisil P., Guy P. A., Prediction models of retention indices for increased confidence in structural elucidation during complex matrix analysis: application to gas chromatography coupled with high-resolution mass spectrometry, *Analytical chemistry,* 2016; 88(15): 7539-7547. https://doi.org/10.1021/acs.analchem.6b00868

41. Rojas Villa C. X., Duchowicz P. R., Tripaldi P., Pis Diez R., Quantitative Structure-Property Relationships for Predicting the Retention Indices of Fragrances on Stationary Phases of Different Polarity, A*nales de la Asociación Química Argentina,* 2017; 104(2): 173-193. https://www.aqa.org.ar/images/anales/pdf104-2/104-2-abstracts.pdf

42. Yan J., Cao D. S., Guo F. Q., Zhang L. X., He M., Huang J. H., Liang Y. Z., Comparison of quantitative structure–retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds, *Journal of Chromatography A,* 2012; 1223: 118-125. https://doi.org/10.1016/j.chroma.2011.12.020

43. Geer L. Y., Stein S. E., Mallard W. G., Slotta D. J.AIRI: Predicting Retention Indices and Their Uncertainties Using Artificial Intelligence, *Journal of Chemical Information and Modeling,* 2024; 64(3): 690-696. https://doi.org/10.1021/acs.jcim.3c01758

44. Zenkevich I.G., Eliseenkov E.V., Kasatochkin A.N., Chromatographic identification of cyclohexane chlorination products by an additive scheme for the prediction of retention indices, *Chromatographia,* 2009; 70:

839-849. https://doi.org/10.1365/s10337-009-1213-x

45. Farkas O., Zenkevich I. G., Stout F., Kalivas J. H., Héberger K., Prediction of retention indices for identification of fatty acid methyl esters, *Journal of Chromatography A,* 2008; 1198: 188-195. https://doi.org/10.1016/j.chroma.2008.05.019

46. Zenkevich I.G., Pavlovskii A.A. Temperature dependence of gas chromatography retention indices as one of the main factors determining their interlaboratory reproducibility, *Protection of metals and physical chemistry of surfaces,* 2015; 51: 1058-1064. https://doi.org/10.1134/S2070205115060258

47. Wu L., Cho I. K., Li Y., Zhang G., Li Q. X., Evaluation of sources of irreproducibility of retention indices under programmed temperature gas chromatography conditions, *Journal of Chromatography A,* 2017; 1495: 57-63. https://doi.org/10.1016/j.chroma.2017.03.009

Информация об авторах / Information about the authors

Д.Д. Матюшин – н.с. лаборатории физико-химических основ хроматографии и хромато-масс-спектрометрии, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва, Россия

А.Ю. Шолохова – в.н.с. лаборатории «умных» методов химического анализа, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва, Россия

D.D. Matyushin – researcher, laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry; Institute of Physical chemistry and electrochemistry, Moscow, Russian Federation, email: dm.matiushin@mail.ru, ORCID: 0000-0003-0978-7666

A.Yu. Sholokhova – leading researcher, laboratory of "smart" methods of chemical analysis; Institute of Physical chemistry and electrochemistry, Moscow, Russian Federation, email: shonastya@yandex.ru, ORCID: 0000-0003-4192-1677

Статья поступила в редакцию 02.08.2024; одобрена после рецензирования 02.09.2024; принята к публикации 04.09.2024.
The article was submitted 02.08.2024; approved after reviewing 02.09.2024; accepted for publication 04.09.2024.