цМинистерство образования и науки Российской Федерации
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(государственный университет)
ФАКУЛЬТЕТ РАДИОТЕХНИКИ И КИБЕРНЕТИКИ
КАФЕДРА ПРОБЛЕМ УПРАВЛЕНИЯ

На правах рукописи
УДК 544.131

МИЛОСЕРДОВ ОЛЕГ АЛЕКСАНДРОВИЧ

ПРЕДСКАЗАНИЕ ИНДЕКСА УДЕРЖИВАНИЯ КОМПОНЕНТОВ БЕНЗИНОВ С ПОМОЩЬЮ ТОПОЛОГИЧЕСКИХ ИНДЕКСОВ

Выпускная квалификационная работа бакалавра

Направление подготовки 010900 «Прикладные математика и физика»

Заведующий кафедрой	_____	Д.А. Новиков
Научный руководитель	_____	М.В. Губко
Студент	_____	О.А. Милосердов

г. Долгопрудный
2014

Оглавление

1. Введение

1.1. Хроматографические методы анализа состава смесей

В современном мире часто возникает потребность определения состава различных смесей. С этой задачей успешно справляется хроматография – метод разделения смесей веществ или частиц, основанный на различиях в скоростях их перемещения в системе несмешивающихся и движущихся друг относительно друга фаз – подвижной и неподвижной. Неподвижной (стационарной) фазой служит твердое пористое вещество (часто его называют сорбентом) или пленка жидкости, нанесенная на твердое вещество. Подвижная фаза представляет собой жидкость или газ, протекающий через неподвижную фазу, иногда под давлением [1].

Разнообразные варианты хроматографии укладываются в относительно простую схему классификации в зависимости от используемой подвижной фазы и характера межмолекулярных взаимодействий. Классификация хроматографических методов по агрегатным состояниям подвижной и неподвижной фазе приведена в Таблице 1.1 [2].

Таблица 1.1. Классификация методов хроматографии по фазовым состояниям [2]

Подвижная фаза	Неподвижная фаза	Название варианта	
		частичное	общее
Газ	Адсорбент	Газоадсорбционная	Газовая хроматография
	Жидкость	Газожидкостная	
Жидкость	Адсорбент	Жидкостно-адсорбционная	Жидкостная хроматография
	Жидкость	Жидкостно-жидкостная	
Газ или пар в сверхкритическом состоянии	Адсорбент	Флюидно-адсорбционная	Флюидная хроматография
	Жидкость	Флюидно-жидкостная	
Коллоидная система	Сложная композиция твердых и жидких компонентов		Полифазная хроматография

В настоящее время создано множество аналитических приборов для разделения смеси веществ хроматографическими методами. Классификация приборов может быть также проведена на основании используемой подвижной фазы:

- газовые хроматографы (подвижной фазой является газ);
- жидкостные хроматографы (подвижной фазой является жидкость);
- флюидные системы (промежуточные между газовыми и жидкостными хроматографами, подвижной фазой является вещество при температуре и давлении выше критических – сверхкритический флюид).

Дополнительная информация о составе компонента смеси может быть получена при сочетании хроматографического разделения со спектральными детекторами, наиболее информативными среди которых являются масс-спектрометрические (МС) детекторы различного принципа действия (квадрупольные, времяпролетные и др.). Однако системы с использованием МС детекторов недешевы, а также требуют высокой квалификации оператора. В связи с этим непрерывно ведется поиск дополнительных возможностей идентификации соединений в составе сложной смеси по результатам хроматографического анализа без использования спектральных детекторов.

Для понимания цели данного исследования необходимо рассмотреть принцип работы хроматографических приборов. Газовая хроматография является самым распространенным методом для анализа сложных смесей летучих термостабильных соединений, к которым, например, относятся нефти и нефтепродукты. Рассмотрим схему устройства газового хроматографа.

Пример 1.1. Блок-схема газового хроматографа.

Рисунок 1.1. Блок - схема газового хроматографа [3]

Элементы газового хроматографа

1 — источник газа-носителя (подвижной фазы);

2 — регулятор расхода газа носителя;

3 — устройство ввода пробы;

4 — хроматографическая колонка в термостате;

5 — детектор;

6 — электронный усилитель;

7 — регистрирующий прибор;

8 — расходомер.

Основным конструктивным элементом хроматографов являются колонки — трубки, заполненные неподвижной фазой, по которым во время выполнения анализа движется подвижная фаза и исследуемый образец. Именно в колонке происходит разделение компонентов исследуемой смеси. После выхода из колонки смесь попадает в детектор. Детекторы предназначены для непрерывного измерения концентрации веществ на выходе из хроматографической колонки. Принцип действия детектора должен быть основан на измерении такого свойства аналитического компонента, которым не обладает подвижная фаза.

Результатом регистрации зависимости концентрации компонентов на выходе из колонки от времени является хроматограмма, которая состоит из ряда пиков, каждый из которых при полном разделении соответствует одному компоненту анализируемой пробы. Ниже приведен пример хроматограммы автомобильного бензина, полученной на газовом хроматографе с пламенно-ионизационным детектором (см. Рисунок 1.2).

После получения хроматограммы встает вопрос об идентификации пиков. Разделение пробы на индивидуальные компоненты происходит в соответствии со временем удерживания каждого компонента в хроматографической колонке.

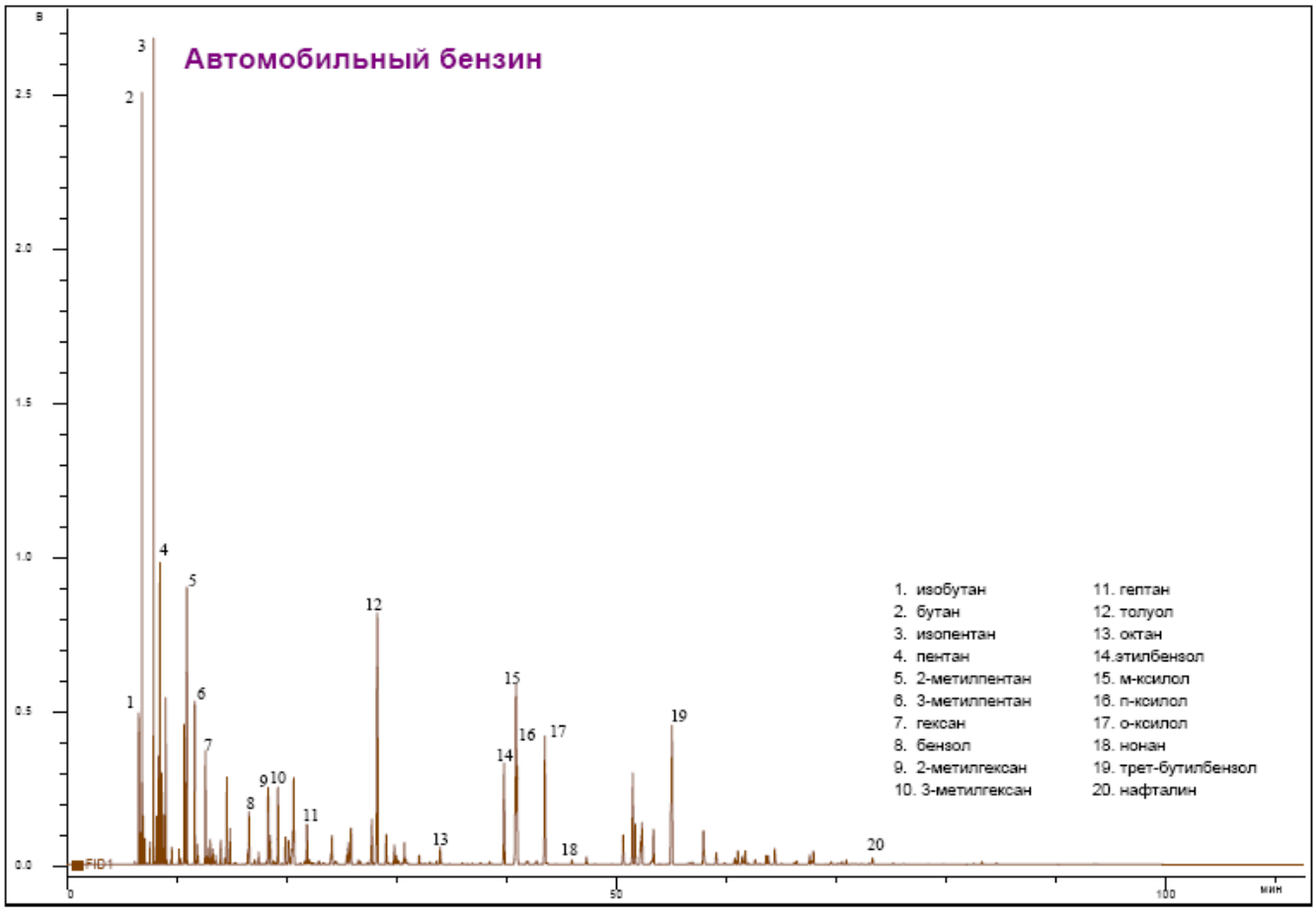


Рисунок 1.2. Пример хроматограммы автомобильного бензина [4]

Время, необходимое для элюирования компонента смеси из колонки, называется абсолютным временем удерживания (t_R) и определяется по времени выхода максимума его хроматографического пика (см. Рисунок 1.3). В процессе хроматографического разделения происходит распределение компонента пробы между подвижной и неподвижной фа-

зами. Время нахождения компонента в подвижной фазе (t_M) постоянно для всех составляющих анализируемой смеси. Величину t_M называют временем удерживания несорбирующегося вещества. Разность абсолютного времени удерживания и времени удерживания несорбирующегося вещества называют истинным (приведенным) временем удерживания:

$$t_R{}' = t_R - t_M.$$

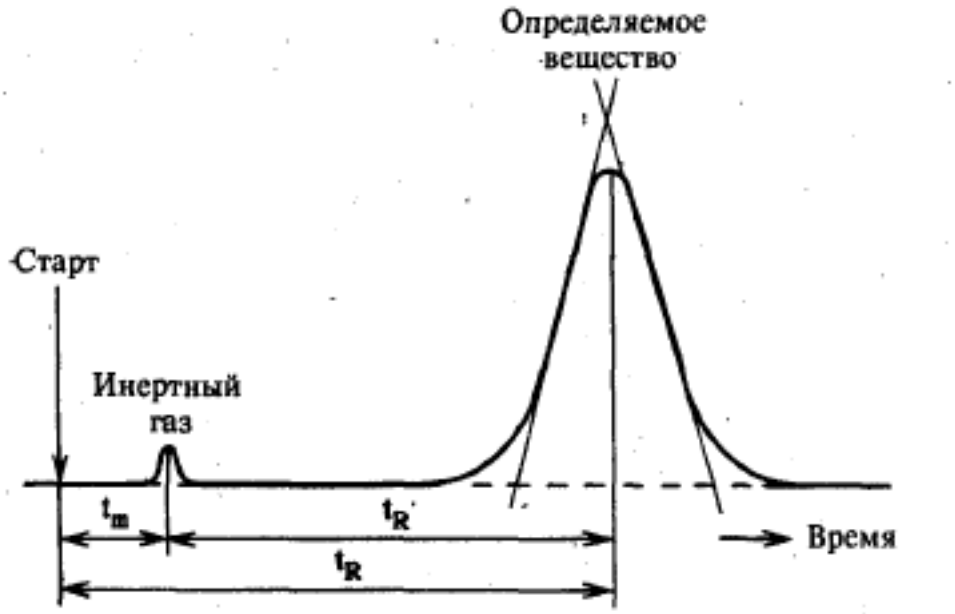


Рисунок 1.3 Связь между абсолютным и приведенным временем удерживания [5]

Для представления величин удерживания в газовой хроматографии используется индекс удерживания Ковача. По определению Ковача индекс удерживания — это мера относительного удерживания веществ, причем в качестве стандартного вещества сравнения, как правило, используются нормальные углеводороды [6]. Каждому нормальному углеводороду присвоен индекс удерживания, равный числу атомов углерода в его молекуле, умноженному на 100. Например, для метана – 100, пропана – 300, декана – 1000 и т. д. Водороду приписывают значение индекса, равное нулю. Эти числа образуют в шкале индексов удерживания серию фиксированных точек. Индекс удерживания Ковача задается следующей формулой:

$$I_i = 100 \cdot \left[m + \frac{\ln \frac{K_i}{K_m}}{\ln \frac{K_{m+1}}{K_m}} \right]$$

где:

$K \equiv K[T(t)]$ — функция, косвенно зависящая от времени выхода компонента t через температуру колонки T;

m — индекс нормального алкана, содержащего m атомов углерода;

I_i — индекс Ковача рассматриваемого вещества i, выходящего между алканами m и m+1.

При изотермическом режиме работы хроматографа функция $K[T(t)]$ пропорциональна приведенному времени удерживания t' [7]. Полезной особенностью индекса Ковача является то, что он слабо зависит от параметров режима, в частности, от температуры. Это свойство позволяет оценивать порядок хроматографического удерживания разных веществ, что играет ключевую роль в их идентификации.

Логарифмические индексы удерживания являются весьма информативной и удобной формой представления данных по относительному удерживанию органических соединений различных классов, и в настоящее время широко используются в качественном анализе для решения сложных задач, как, например, комплексная идентификация компонентов нефти или исследование запаха пищевых продуктов [7,8].

1.2. Актуальность решения обратной задачи структура – свойство

На данный момент составлена огромная база данных индексов удерживания Ковача, полученных экспериментально, но, тем не менее, на хроматограмме часто появляются пики, которые отсутствуют в базе, а исследователям нужно знать, что это за вещество. Поэтому актуальной задачей на сегодняшний день является прогнозирование индекса удерживания Ковача различных классов веществ.

Одной из важных и актуальных задач является детальный углеводородный анализ бензина. Технологам на нефтеперерабатывающих заводах важно знать, какие вещества появляются в бензиновых продуктах, выходящих с их установок, в зависимости от изме-

нений режима. В частности, повышение количества изоалканов увеличивает октановое число, тогда как увеличение количества бензола крайне нежелательно, потому что он является сильным канцерогеном и его содержание в товарном продукте строго лимитируется. Исчерпывающая информация о компонентном составе бензина позволяет вычислить такие важные его свойства, как октановое число, давление насыщенных паров, температура вспышки и плотность. Таким образом, точный хроматографический анализ бензина способен заменить множество лабораторных методов испытаний, некоторые из которых весьма трудоемки. Этот анализ проводится с использованием газовой хроматографии на неполярной капиллярной колонке длиной 100 м для обеспечения наиболее полного разделения всех соединений. Идентификация веществ в составе бензина производится на основании их логарифмических индексов удерживания. Однако до сих пор невозможно идентифицировать все компоненты смеси в связи с неполнотой базы данных индексов удерживания. Зачастую «идентификация» заключается лишь в отнесении компонента к тому или иному классу органических соединений по результатам масс-спектрометрического анализа.

База данных компонентов бензина, составленная на основе многочисленных проведенных хроматографических исследований бензинов с использованием масс-спектрометрического анализа, была предоставлена сотрудником компании PAC[1] Владимиром Чупиным. Данная база веществ была использована нами для выделения основных компонентов бензина.

Все вещества были разбиты на классы, затем составлена сводная таблица.

Таблица 1.2. Перечень основных классов веществ, содержащихся в бензине

№	Класс соединения	Кол-во в-в в смеси
1.	Алкадиены	18
2.	Алканы	160
3.	Алкены	74
4.	Алкины	6

[1] http://www.paclp.com/

№	Класс соединения	Кол-во в-в в смеси
5.	арен	62
6.	арен/циклоалканы	16
7.	арен/циклоалкены	10
9.	диароматические углеводороды (ДАУ)	4
11.	оксигенаты	16
13.	циклоалкадиены	2
14.	циклоалканы	88
15.	циклоалкены	8
	Всего в-в в базе	**464**

Поскольку численное значение индексов Ковача определяется лишь физико-химическими свойствами анализируемого вещества, природой неподвижной фазы и температурным режимом колонки, индекс удерживания вещества той или иной неподвижной фазой, отнесенный к определенной температуре, можно поставить в ряд с такими известными константами, как температура кипения (плавления), плотность или показатель преломления, которые, в свою очередь, определяются, в том числе, параметрами, зависящими от строения молекулы.

Итак, целью настоящего исследования является составление базы данных индексов удерживания компонентов бензина.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Перечислить все изомеры всех компонентов бензина.
2. В связи с отсутствием требуемого объема экспериментальных данных необходимо предсказать значения индексов удерживания всех веществ, классы которых содержаться в бензине.
3. Составить базу данных компонентов бензина, с предсказанными индексами удерживания.
4. Обеспечить удобное рабочее место пользователю базы данных.

1.3. Методы исследования

Предсказывать значения индекса удерживания позволяют методы QSRR[2] (количественный анализ “структура – индекс удерживания”), который является разделом QSPR[3] (структурного анализа “структура-свойство”). QSPR позволяет предсказывать свойства (физические, химические, биологическую активность и так далее) веществ по их молекулярной структуре [9]. При прогнозировании свойств на качественном уровне говорят о решении классификационной задачи, тогда как при прогнозировании числовых значений свойств говорят о решении регрессионной задачи. Нам необходимо спрогнозировать числовые значения индекса удерживания, а, следовательно, нужно построить новую или подобрать уже известную регрессию.

Один из способов прогнозирования свойств веществ осуществляется с помощью молекулярных дескрипторов, которые, в свою очередь, делятся на несколько категорий:

- Топологические индексы — инвариант молекулярного графа, некоторое числовое значение (или набор значений), характеризующее структуру молекулы. Обычно атомы водорода не учитываются. К наиболее известным топологическим индексам относятся индекс Винера [10], индекс Рандича [11], индекс Балабана [12] и другие.
- Физико-химические дескрипторы — это числовые характеристики, получаемые в результате моделирования физико-химических свойств химических соединений, либо величины, имеющие четкую физико-химическую интерпретацию. Наиболее часто используются в качестве дескрипторов: молекулярный вес (MW), молекулярные объемы и площади поверхностей.
- Квантово-химические дескрипторы — это числовые величины, получаемые в результате квантово-химических расчетов. Наиболее часто в качестве дескрипторов используются: энергии граничных молекулярных орбиталей, частичные заряды на атомах и частичные порядки связей, другие дескрипторы.

[2] Quantitative Structure-Retention Relationship
[3] Quantitative Structure-Property Relationship

В данной работе используются топологические индексы, хотя они являются не самыми точными. Топологические индексы используются нами по двум причинам. Во-первых, они легко вычислимы. Во-вторых, при их использовании возможно решение важной задачи поиска веществ с заданными свойствами – задачи оптимизации одних топологических индексов при ограничениях на другие. Решением этой задачи мы планируем заниматься в перспективе исследований.

Топологические индексы основываются на представлении молекул ненаправленными (возможно, помеченными) графами и классифицируются следующим образом [12]:

- Индексы, основанные множестве степеней вершин
    - Пример: Индекс Рандича [11]

$$r = \sum_{v_i, v_j} \frac{1}{\sqrt{d(v_i)d(v_j)}}$$

      где v_i и v_j — вершины, образующие ребро, $d(v_k)$— степень вершины v_k.

- Индексы, связанные с матрицей расстояния *D*
    - Пример: Индекс Винера [10] графа *G*

$$W = \sum_{\{v_i, v_j\} V(G)} d(v_i, v_j)$$

      где $d_G(v_i, v_j)$ — кратчайший путь между вершинами v_i u v_j, а $V(G)$ — множество вершин графа *G*.

- Индексы, зависящие от спектральных характеристик графа
    - Пример: Спектральный радиус матрицы смежности.
- Информационные топологические индексы
    - Пример: Фрагментные индексы.

      В качестве фрагментного индекса может быть использован любой фрагмент структуры молекулы. Например на Рисунке 1.4 приведен пример фрагмента,

представляющего из себя атом углерода, принадлежащий бензольному кольцу и с прикрепленным радикалом изопропилом.

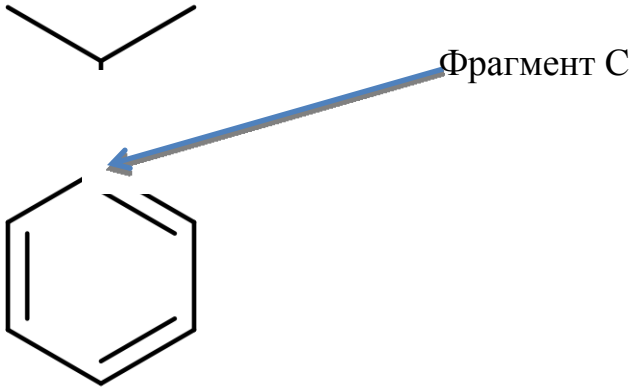


Рисунок 1.4. Пример фрагментного индекса

Используя топологические индексы, мы можем применить регрессионный анализ для построения регрессии, задающей индекс удерживания. Регрессионный анализ — статистический метод исследования влияния одной или нескольких независимых переменных на зависимую переменную. В нашем случае независимыми переменными будут топологические индексы, а зависимой – индекс удерживания Ковача.

2. Границы исследования

2.1. Классы веществ

В силу ограниченности времени исследования, в работе были рассмотрены три самых многочисленных класса веществ из 15 присутствующих в Таблице 1.2: алканы, алкены и арены.

Алканы — ациклические углеводороды линейного или разветвленного строения, содержащие только одинарные связи и образующие гомологический ряд с общей формулой $\mathbf{C_nH_{2n+2}}$.

Пример 2.1. Один из типичных представителей алканов - 3-этил-2,3,4-триметилгепта

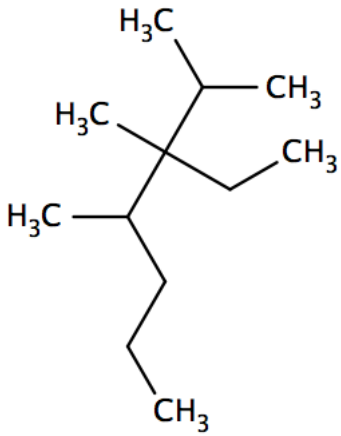


Рисунок 2.1. 3-этил-2,3,4-триметилгептан

Алкены — ациклические непредельные углеводороды, содержащие одну двойную связь между атомами углерода, образующие гомологический ряд с общей формулой $\mathbf{C_nH_{2n}}$. Для алкенов характерна цис-/транс- изомерия. В цис- изомерах заместители находятся по одну от плоскости двойной связи, а в транс- изомерах — по разные.

Пример 2.2. Цис- (Рисунок 2.2)/транс-(Рисунок 2.3) изомерия

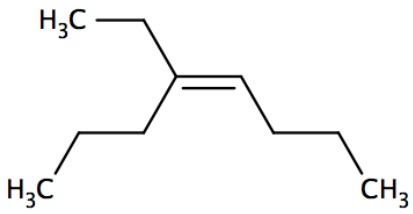


Рисунок 2.2. (4Z)-4-этил-4-октен

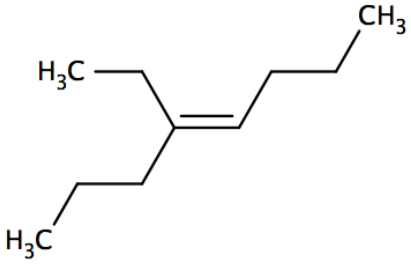


Рисунок 2.3. (4E)-4-этил-4-октен

Арены — карбоциклические соединения (бензол и алкилбензолы), полученные в результате замены атомов водорода в бензольном кольце на радикалы насыщенного углеводорода. Общая формула ароматических углеводородов $\mathbf{C_nH_{2n-6}}$.

Пример 2.3. 2-изопропил-1,4-диметилбензол

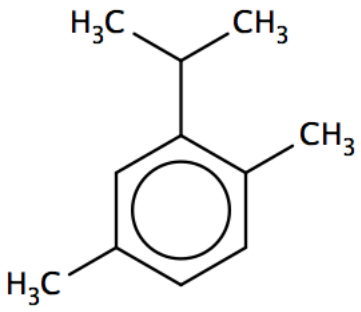


Рисунок 2.4. 2-изопропил-1,4-диметилбензол

Также из исследований известно, что важные для анализа компоненты бензина имеют от 4-х до 12-ти значимых атомов углерода.

2.2. Подбор регрессий

После определения границ исследования необходимо подобрать регрессию для каждого из трех классов веществ.

Для поиска подходящих исследований использовались два обзора К. Хебергера по современному состоянию QSRR за 2007 [13] и 2012 [14] годы. В этих обзорах проанализировано более 500 статей с работами по QSRR. К сожалению, во многих статьях пренебрегают основными рекомендациями, позволяющими вырабатывать надежные и корректные предсказания индекса удерживания, о которых пишет Хебергер.

Во-первых, он говорит о том, что необходимо хорошо изучить уже существующие исследования и не повторять их. Также Хебергер предлагает разделить выборку веществ на три равные части. Первая будет использоваться как обучающий сет, на котором будет строиться регрессия. Вторая часть необходима для корректной проверки на точность регрессии и сопоставимость ее с реальностью. И последняя часть используется для тестов. Разбиения должны быть случайными.

Из пяти сотен статей после тщательного изучения нами было отобрано пятнадцать. Также мы выяснили, что обучающей выборкой в большей части статей являлась база индексов удерживания веществ, полученных на сквалане при температуре 100 °C.

Сквалан – неполярная неподвижная фаза, применяемая в газожидкостной хроматографии. Также он используется как высококачественное смазочное масло и как компонент некоторых фармацевтических и косметических препаратов. Структурная формула сквалана приведена на Рисунке 2.5.

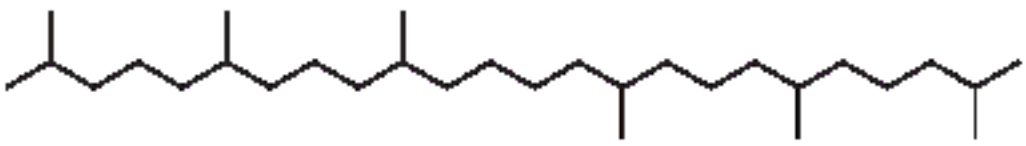


Рисунок 2.5. Структурная формула сквалана

Нами были определены следующие критерии отбора подходящих статей:

1. База значений индексов удерживания должна быть получена на сквалане при температуре 100 °C.
2. Высокая точность предсказательной способности: $R^2 > 0,99$.
3. Большая обучающая база данных: более 50 веществ.

Всего было подробно изучено 7 статей по аренам и 8 статей по алканам и алкенам (которые обычно исследуются вместе). Большая часть статей была отброшена в связи с тем, что авторы пренебрегали правилами, о которых писал Хебергер.

В качестве искомых регрессий были выбраны результаты [15] для алкенов, [16] для алканов и [17] для аренов. Выбор определялся точностью, удобством расчета и наличием учета различий между цис- и транс- изомерами алкенов.

2.3. Регрессия для предсказания индекса удерживания алканов и алкенов

Построение регрессии в [15] основано на следующих идеях. Известно, что индекс удерживания связан с количеством атомов углерода. Атом углерода вносит линейный вклад в индекс удерживания, но из-за стерических эффектов этот вклад уменьшается, поэтому для каждого фрагмента необходимо определить его вес, который он вносит в значение полуэмпирического индекса.

В качестве примера можно взять 2,3-диметилбутан, значение индекса удерживания для которого составляет 568,1. Но если мы попытаемся линеаризовать индекс удерживания путем присвоения каждому атому вклада, равного 100 единицам, повторив тем самым подход Ковача для линейных алканов, то получим 600 единиц. Следовательно, нетрудно догадаться, что атомы углерода в ответвлениях дают вклад в индекс удерживания, меньший 100. Из экспериментов следует, что невозможно посчитать какой либо постоянный вклад в индекс удерживания для конкретного атома углерода, что видно, например, из величины индекса удерживания 2,2-диметилпентана (I = 626,2) и 3,3-диметилпентана

(I = 660,2), а также других алканов нелинейного строения. Чтобы решить эту проблему, авторы [15] прибегли к следующему методу: назначили приблизительные (полученные экспериментально) значения вкладов для каждого из типов атомов: 100- для атомов метильной группы (CH_3–), 90 для вторичных (–CH_2–) атомов, 80 для третичных (–CH<) и 70 для четвертичных (>C<), разделив их на 100 для нормировки.

Определение степени стерических эффектов, присутствующих в углеводороде, зависит также от размера замещающей группы, а не только от местоположения конкретного атома, поэтому авторы [15] добавляют еще одно слагаемое в виде произведения степени фрагмента на его вес. В результате, полуэмпирический топологический индекс (I_{ET}) выражается в виде:

(2.1) $$I_{\mathrm{ET}} = \sum n_i \, (C_i + d_i \, lg \, C_i)$$

где C_i – вес фрагмента i-ого типа, n_i – количество фрагмента i-ого типа, d_i – степень i-ого фрагмента.

Таблицы 2.1 и 2.2 являются важной составляющей дальнейшего исследования.

Таблица 2.1 Вклады различных фрагментов в составе молекулы алкана в индекс удерживания [15].

Фрагмент	Номер места в цепи	Значение	*Ci*
CH_3–	–	100	1
–CH_2–	–	90	0,9
–CH<	–	80	0,8
>C<	–	70	0,7

Таблица 2.2 Вклады различных фрагментов в составе молекулы алкена в индекс удерживания [15].

Фрагмент	Номер места в цепи	Значение	*Ci*
CH_3–	–	100	1
–CH_2–	–	90	0,9
–CH<	–	80	0,8
>C<	–	70	0,7
CH2=; –CH=	1C	89,75	0,8975
–CH=*trans*	2C	89,5	0,895

Фрагмент	Номер места в цепи	Значение	*Ci*
cis		91	0,91
–CH=*trans*[a]	3C	872	8,72
cis [a]		88,5	0,885
–CH=*trans*[a]	4C	86,5	0,865
cis [a]		87	0,87
–CH=*trans*	5C	86,5	0,865
cis		85,5	0,855
–CH=*trans*	6C	86	0,86
Cis		85	0,85
–CH=*trans*	7C	85,75	0,8575
Cis		84,5	0,845

[a-] *если атомов углерода в молекуле больше 10, то значения для цис- и транс- изомеров должны быть поменяны местами.*

Стоит отметить, что данный подход к определению индекса удерживания алканов привлекателен тем, что он придуман не на пустом месте и недалек от физики. Он получен на основе предположения, что удерживание атома углерода в составе молекулы вызвано в первую очередь силами дисперсионного взаимодействия с неподвижной фазой, которое уменьшается благодаря соседним стерическим эффектам. А для алкенов в процессе удерживания участвуют еще и электростатические силы. Взаимодействие кратных связей с другими структурными факторами делает трудным предсказание их эффекта с учетом цис – и транс – изомерии. По этой причине значения для атомов в алкенах подбираются опытным путем на основании экспериментальных индексов удерживания алкенов в зависимости от того, являются ли они цис- или транс- алкенами.

Лучшая регрессия для алкенов, полученная основе базы из 79 алкенов имеет вид:

(2.2) $$I_{CALC} = 122{,}8446\, I_{ET} - 41{,}7054$$

и регрессия имеет точность r = 0,99996 и стандартное отклонение SD = 2,3541.

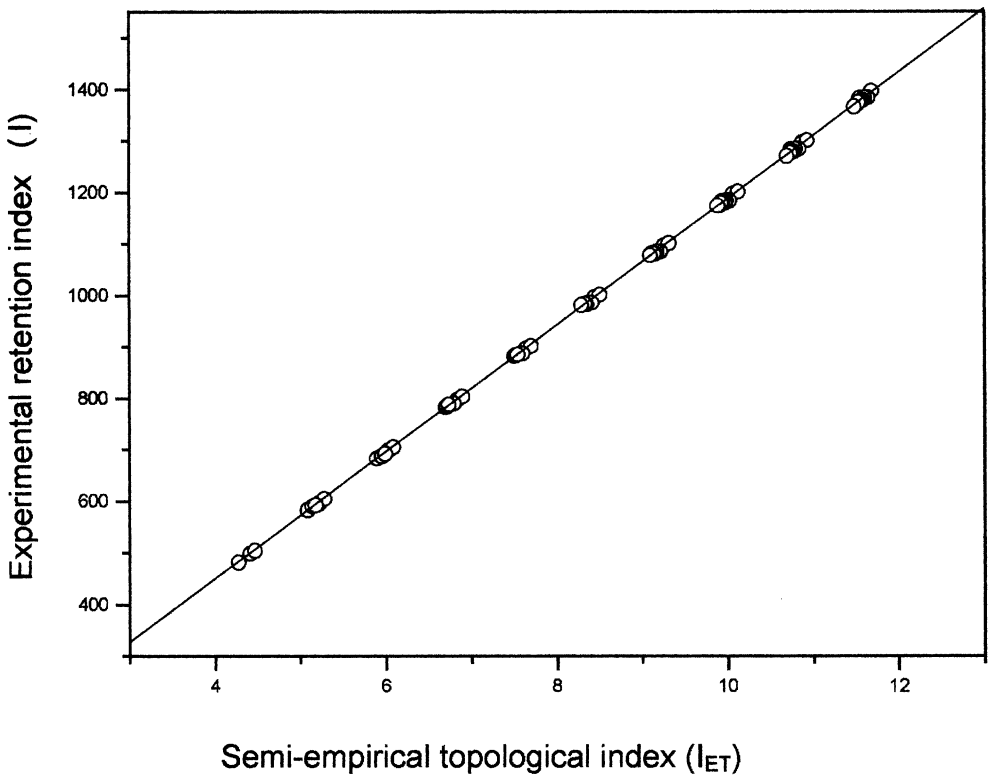


Рисунок 2.6 График зависимости экспериментального индекса удерживания алкенов от предсказанного [15]

Для алканов на основе базы из 157 веществ:

(2.3) $$I_{CALC} = 116{,}8\, I_{ET} - 19{,}05$$

точность r = 0,9901, а SD = 26.

Регрессия для алкенов, предложенная в этой работе, имеет высокую предсказательную способность, а значит может быть использована в качестве инструмента для прогнозирования хроматографического удерживания алкенов с цис-/транс- изомерной структурой. Однако в случае с алканами регрессия 2.3, не смотря на ее точность, указанную в статье [15], очень плохо прогнозирует индекс удерживания даже для таких простых соединений, как нормальные алканы. В связи с этим в данной работе была использована регрессия из статьи [16].

Для предсказания индекса удерживания авторы [16] используют количество определенных фрагментов молекулы, как топологические индексы. Все используемые фрагменты представлены на Рисунке 2.7.

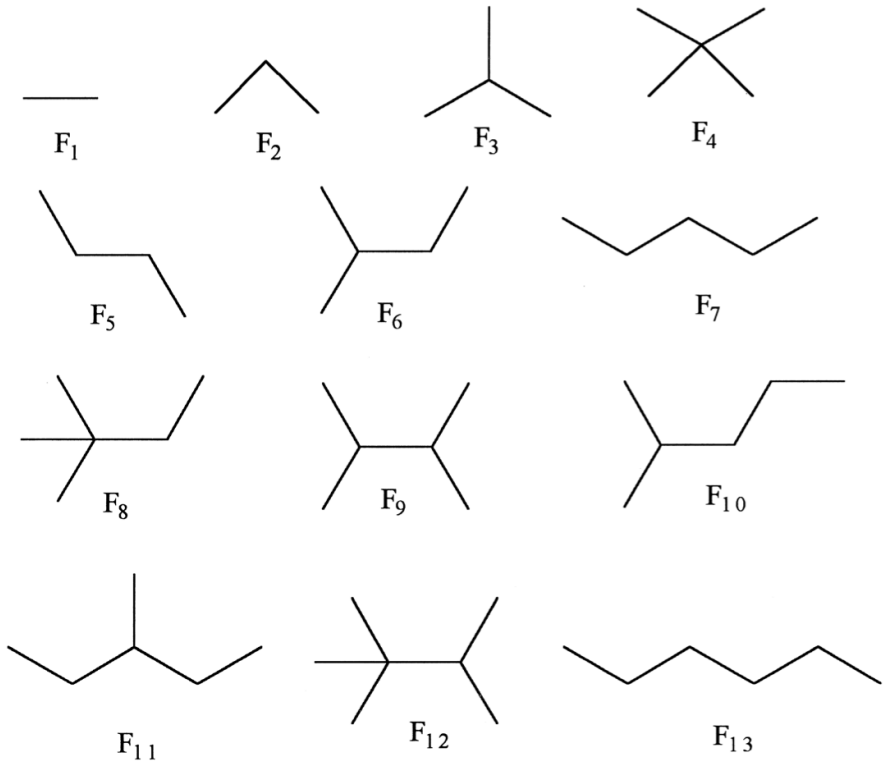


Рисунок 2.7 Структурные фрагменты молекул алканов [16]

Основное преимущество данного подхода состоит в том, что количественный вклад структурных фрагментов молекул в предсказываемый индекс удерживания может быть получены в явном виде и представлен в виде линейной регрессии 2.4.

(2.4) $$I = a_0F_0 + a_1F_1 + a_2F_2 + a_3F_3 + \cdots + a_kF_k + b$$

При использовании этого подхода в статье [16] был исследован набор данных состоящий из 156 алканов и получена линейная регрессия:

(2.5) $$I_{CALC} = 137{,}94 * F_1 - 3{,}66 * F_2 - 62{,}94 * F_3 + 118{,}80 * F_4 - 24{,}72 * F_5 + 29{,}98 * F_6 - 12{,}36 * F_7 - 25{,}86 * F_8 — 23{,}68 * F_9 + 3{,}22 * F_{10} + 3{,}22 * F_{11} + 19{,}32 * F_{12} + 31{,}15$$

точность r=0,9986, SD=9,44.

Точность регрессии 2.5 для алканов довольно высока, что иллюстрируется на Рисунке 2.8, а значит данный способ может быть использован для предсказания индекса удерживания алканов наряду с моделью для алкенов, предложенной в статье [15].

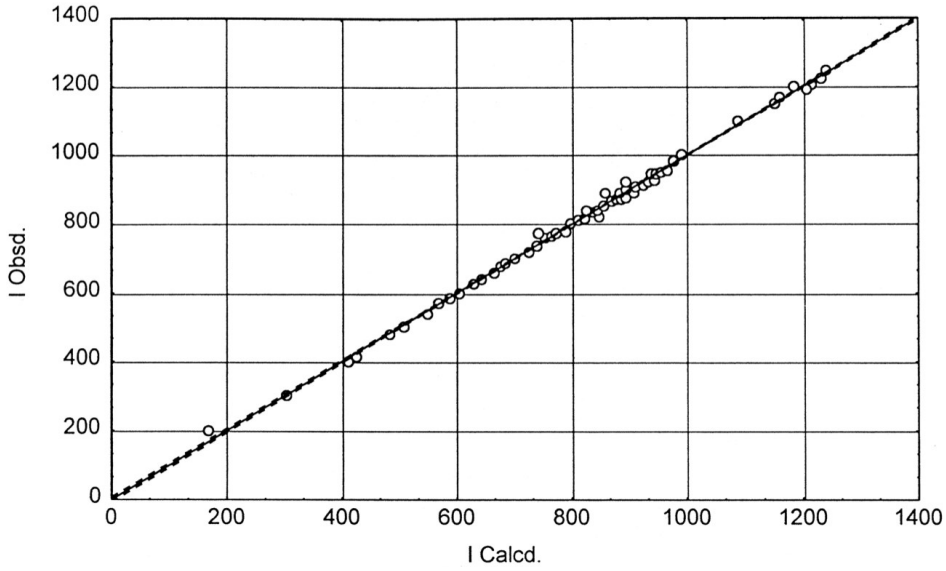


Рисунок 2.8 *График зависимости экспериментального индекса удерживания алканов от предсказанного* [16]

2.4. Регрессия для предсказания индекса удерживания аренов

Предложенный в [15] метод дал хорошие результаты, поэтому авторы решили использовать его и для других классов веществ. Например, для прогнозирования индекса удерживания аренов [17]. Основная цель этой статьи заключается в расширении полуэмпирического топологического метода для того чтобы предсказать индекс удерживания аренов и проверить его предсказательную силу на стационарных фазах с различной полярностью.

Построение регрессии производится аналогично рассмотренному выше. Оно основано на присвоении конкретных вкладов в индекс удерживания Ковача различным видам атомов углеводорода в зависимости от их расположения в структуре молекулы.

Арены были разделены на следующие группы: линейные; разветвленные; с замещением в орто-, мета- и пара-положениях; три-замещенные и тетра-замещенные.

Приставки орто-, пара- и мета- употребляются в органической химии для обозначения положения двух одинаковых или различных друг относительно друга заместителей в бензольном кольце:

- Орто-изомер — с соседним положением заместителей;

- Мета-изомер — заместители разделены одним атомом углерода;
- Пара-изомер — заместители находятся на максимальном удалении друг от друга.

Иллюстрация орто-, мета- и пара-положений приведена на Рисунке 2.6.

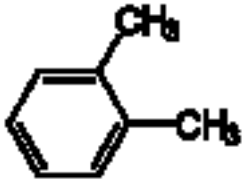
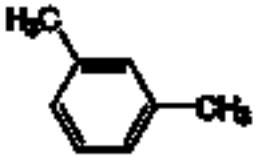
орто-ксилол	пара-ксилол	мета-ксилол
		

Рисунок 2.9. Орто-, мета-, пара- приставки

Таблица 2.3. Вклады различных фрагментов в составе арена в индекс удерживания [17]

Фрагмент	Позиция	C_i
CH_3	–	1
$–CH_2–$	–	0,9
–CH<	–	0,8
>C<	–	0,7
–CH=	–	0,9683
=C<[a]	mono	0,9263
	ortho	0,9535
	meta	0,9176
	para	0,9173
=C<[b]	mono	0,8927
Фрагмент	Позиция	C_i
	ortho	0,8767
	meta	0,8564
	para	0,8774
=C<[c]	mono	0,8401
	ortho	0,77
	meta	0,8062
	para	0,8237
=C<[d]	mono	0,8045
=C<[e]	mono	0,7576
=C<[f]	2	0,6275
	3	0,5407
	4	0,4959
	5	0,4738
	6	0,4564
=C<[g]	mono	0,7788
	meta	0,7181
	para	0,7624
=C<[h]	mono	0,8019
	ortho	0,7357
	meta	0,7304

Фрагмент	Позиция	C_i
	para	0,7836

[a] метил.
[b] этил.
[c] 3-10 атомов углерода в углеродной цепи.
[d]11-13 атомов углерода в углеродной цепи.
[e]α, β и γ разветвления позиции (до 6 атомов углерода).
[f] α ветвления позиция (10-13 атомов углерода).
[g] четвертичный атом углерода.
[h] изопропил.

На основе базы из 122-ти аренов была получена лучшая регрессия:

(2.6) $$I_{CALC} = -39{,}7381 + 123{,}0824\, I_{ET}$$

со следующими параметрами точности: R = 0,9998; SD = 5,5.

Точность этой регрессии достаточно велика, подтверждением чему является график зависимости вычисленного индекса удерживания от экспериментального (см. Рисунок 2.10). Как можно видеть, значения точно ложатся на прямую.

Таким образом, работы [15], [16] и [17] показывают, что фрагментные топологические индексы могут эффективно применяться для предсказания индекса удерживания Ковача, что позволяет их использовать для решения задачи данного исследования.

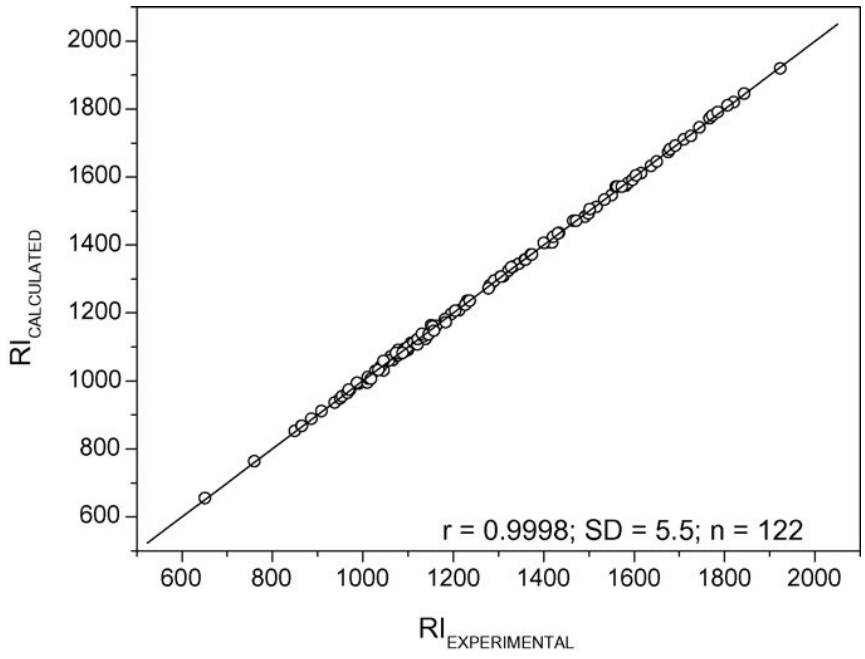


Рисунок 2.10. График зависимости экспериментального индекса удерживания аренов от предсказанного [17]

3. Перечисление химических структур

3.1. Перечисление двоичных деревьев

После выбора подходящих регрессий строим базу данных веществ. Мы рассматриваем структуру вещества как граф с вершинами в атомах углерода (водород не рассматриваем, так как он почти не вносит вклада). Вспомним об ограничениях нашего исследования:

1. Рассматриваются только три класса веществ: алканы, алкены и арены.
2. Исходя из базы данных состава бензина, максимальное количество атомов углерода – 14, а минимальное – 4.
3. У атома углерода не может быть больше 4 связей (связи с водородом не учитываются).

Таким образом, задача состоит в том, чтобы перечислить всевозможные химические графы с количеством вершин от 4 до 14 для трех случаев:

1. Для алканов. В этом случае нам просто нужно перебрать всевозможные деревья с количеством связей у вершины, не превышающим четырех.
2. Для алкенов. Задача аналогична алканам, но при этом одна из связей между атомами углерода является двойной, и необходимо учитывать наличие цис- и транс-изомеров.
3. Для аренов. В арене содержится одно бензольное кольцо, состоящее из шести атомов углерода, причем к каждому из этих атомов может крепиться атом водорода или цепь из $n \leq 8$ углеродов, причем она может быть разветвленной.

Определение 3.1. Дерево — это связный ациклический граф. Связность означает наличие путей между любой парой вершин, ацикличность — отсутствие циклов и то, что между парами вершин имеется только по одному пути [18].

Определение 3.2. Двоичное дерево — это такое дерево, у которого любая вершина может иметь не более двух потомков (детей).

Определение 3.3. Корневое дерево — дерево с выделенной вершиной.

Для решения этих задач нужно представить наши структуры как дерево. Поэтому построим алгоритм, перечисляющий все возможные деревья с заданными условиями и ограничениями.

Проще всего сначала перечислить все корневые двоичные деревья. Для этого используется алгоритм для генерации всех корневых двоичных деревьев с n вершинами, каждой из которых назначена его левая связь l_i и правая связь r_i, $i = 1, \ldots, n$ с дочерними вершинами [19]. Вершины перечисляются в прямом порядке. Таким образом, например, вершина 1 всегда корневая, а l_k всегда равно либо $k + 1$, либо 0; если $l_1 = 0$ и $n > 1$, то $r_1 = 2$.

Алгоритм перечисления двоичных деревьев с заданным количеством вершин [19]:

1. [Инициализация.] Установить $l_k \leftarrow k+1$ и $r_k \leftarrow 0$ для $1 \leq k < n$, установить также $l_n \leftarrow r_n \leftarrow 0$, и $l_{n+1} \leftarrow 1$.
2. [Посещение.] Посетить двоичное дерево, представленное связями l_1, $l_2 \ldots l_n$.
3. [Поиск j.] Установить $j \leftarrow 1$. Пока $l_j = 0$, устанавливать $r_j \leftarrow 0$, $l_j \leftarrow j+1$ и $j \leftarrow j+1$. Затем прекратить выполнение алгоритма, если $j > n$.
4. [Поиск k и y.] Установить $y \leftarrow l_j$ и $k \leftarrow 0$. Пока $r_y > 0$, $l_j \leftarrow j+1$ и
   $y \leftarrow r_y$.
5. [Продвижение y.] Если $k > 0$, установить $r_k \leftarrow 0$; в противном случае установить $l_j \leftarrow 0$. Затем установить $r_y \leftarrow r_j$, $r_j \leftarrow y$ и вернуться к шагу 2.

3.2. Преобразование двоичного дерева в лес

После того, как были перечислены все двоичные деревья, нужно было найти механизм, переводящий двоичные деревья в произвольные. Такой механизм был найден в [19]. Существует взаимно однозначное соответствие между двоичными деревьями размера n и

произвольными деревьями размера $n + 1$. И в самом деле, в существовании соответствия легко убедиться, если посмотреть на Рисунок 3.1. Слева находится изображение пяти первых двоичных деревьев размера 4, а рядом – пяти соответствующих произвольных деревьев размера 5.

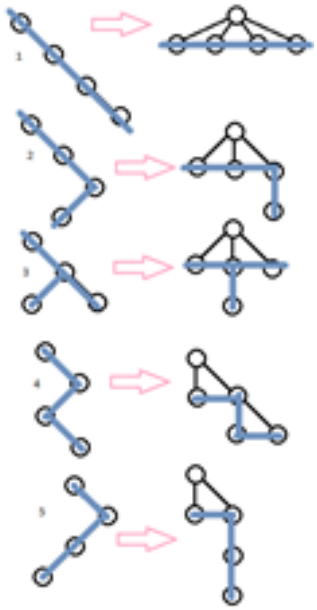


Рисунок 3.1. Поворот на 45 градусов и добавление корневой вершины

Для превращения двоичного дерева в соответствующее ненаправленное дерево нужно повернуть двоичное дерево на 45 градусов против часовой стрелки, добавить сверху корень, удалить горизонтальные связи, затем каждую вершину в ненаправленном дереве связать со следующим братом первого ребенка. Получилось в точности двоичное дерево, только вместо надписей на вершинах «первый ребенок» и «следующий брат» нужно написать «левый потомок» и «правый потомок». Есть только одно отличие между двоичным деревом и ненаправленным деревом в этой системе – правый потомок корня всегда пуст, т.к. корень не имеет братьев. Теперь ясно, что отличие между двоичным деревом и произвольным деревом состоит в именах полей в структуре данных, а значит, мы показали взаимно однозначное соответствие между ними (Рисунок 3.2).

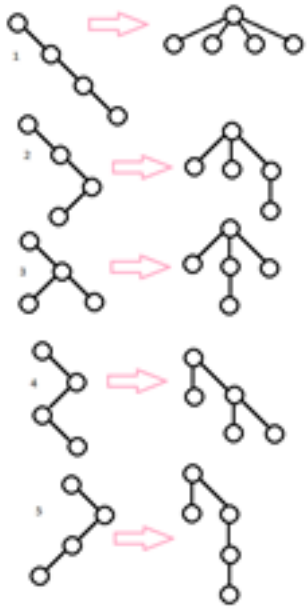


Рисунок 3.2. Соответствие двоичных деревьев произвольным.

Определение 3.4. Лес – это множество (обычно упорядоченное), не содержащее ни одного непересекающегося дерева или содержащее несколько непересекающихся деревьев. Вершины дерева при условии исключения корня образуют лес [20].

Важно отметить, что если не добавлять корневую вершину к построенному дереву, то мы получим лес с n вершинами. Поэтому наша задача по генерации всех деревьев сводится к задаче генерации леса.

Теперь можно снова обратиться к нашим подзадачам. Введем следующие условия:

- если количество компонент леса равно 1, то используем это дерево в качестве одной из структур перечисляемых алканов,
- если количество компонент леса равно 2, то используем это дерево в качестве одной из структур перечисляемых алкенов,
- если количество компонент леса равно 6, то используем это дерево в качестве одной из структур перечисляемых аренов.

Определение 3.5. Матрица смежности $– A = \|a_{ij}\|$ помеченного графа G с p вершинами называется $(□ \times □) –$ матрица, в которой $a_{ij} = 1$, если вершина $□_i$ смежна с $□_j$, и $a_{ij} = 0$ в противном случае. Таким образом существует взаимно однозначное соответ-

ствие между помеченными графами с p вершинами и симметрическими бинарными ($\square \times \square$) – матрицами с нулями на диагонали [18].

Для хранения структур графов будем использовать матрицу смежности. В нашем случае она будет заполнена нулями и единицами, поскольку мы рассматриваем простые графы.

3.3. Генерация алканов и алкенов

Для алканов, помимо проверки единственности компоненты леса, достаточно выполнить проверку на то, что граф является химическим, то есть у вершины имеется не более четырех связей.

Перечислить алкены труднее, так как необходимо учесть двойную связь и цис-/транс- изомерию. Для учета двойной связи был введен специальный вектор, в которой записывались корневые вершины леса i и j, а также добавлялись единицы в матрицу смежности в ячейки a_{ij} и a_{ji}.

Чтобы учесть цис-/транс – изомерию мы считали, что любой алкен может иметь цис- и транс- изомеры, поэтому дублировали полученные структуры. В случае если алкен не может иметь стереоизомерию, мы получаем две одинаковых структуры. Впоследствии на этапе удаления дублей все одинаковые структуры сливаются в одну.

3.4. Генерация аренов

Переходим к третьей подзадаче – перечислению аренов. В нашем случае, арены – это корневые леса с шестью компонентами, соответствующими радикалам, прикрепленным к шести атомам бензольного кольца, причем компонента из одной вершины соответствует незамещенному атому кольца. Для фильтрации неподходящих графов необходимо ввести проверку на то, что вершина, принадлежащая кольцу, может иметь присоединенный радикал и тогда степень соответствующей компоненты леса будет равна единице, в ином случае степень равна нулю.

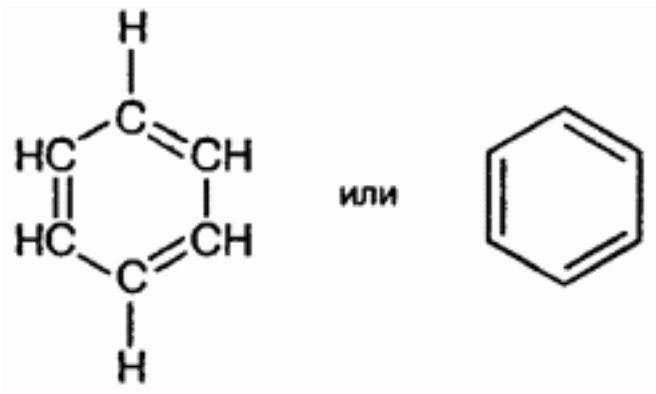


Рисунок 3.3. Бензольное кольцо

После того, как мы учли все ограничения, мы перебираем все леса, количество компонент которых равно 6, тем самым перебирая все возможные ответвления кольца (радикалы).

4. Работа с базой данных химических структур.

4.1. Общие сведения о химической СУБД компании ChemAxon

Результатом генерации всех нужных нам структур веществ является база данных, в которой содержится более 1 250 000 структур. Поэтому возникает вопрос обработки подобной базы данных. Решением этой проблемы стала программа Instant JChem [21], которая, помимо реализации работы с базой данных химических веществ, решает задачи визуализации, вычисления топологических индексов, и многие другие. Подробнее о ней будет рассказано ниже.

Итак, Instant JChem является инструментом, который позволяет исследователям создавать и анализировать базы данных химических структур. В этой программе есть своя база данных для многих веществ, но она больше расчитана для работы с импортированными базами данных. Импортировать данные можно различнми способами, как из существующийх баз данных (Microsoft SQL Server, MySQL, Oracle и другие) так и из различных файлов формата *.xls, *.txt, *.pdf и многих других, включая специализированные (*.inchi, *.smiles, *.mol и т.д.).

Для идентификации веществ в JChem необходимо было представить их в одном из стандартных форматов.

4.2. Построение SMILES

SMILES[4] – это общепринятая строковая нотация структурных формул.

В терминах теории графов SMILES представляет собой строку, полученную путем вывода символов вершин молекулярного графа в порядке, соответствующем их обходу в глубину, то есть для каждой не пройденной вершины необходимо найти все не пройденные смежные вершины и повторить поиск для них. Первоначальная обработка графа включает в себя удаление атомов водорода и разбивку циклов таким образом, чтобы по-

[4] Simplified Molecular Input Line Entry Specification, спецификация упрощенного представления молекул в строке ввода.

лучившийся граф представлял собой остовный лес. Местам разбиения графа ставятся в соответствие числа, показывающие наличие связи в исходной молекуле. Для указания точек ветвления молекулы используются скобки.

В нашей задаче были использованы следующие правила построения SMILES [22]:

1. Атомы обозначаются символами химических элементов в квадратных скобках. Для сокращения записи скобки часто опускают, поэтому атом углерода может быть обозначен как [C], так и просто как C.
2. Связи обозначаются символом «-», а ароматическая «(:)», но они чаще всего опускаются. Двойная связь обозначается знаком равенства «=», поэтому, например, двуокись углерода записывается как O=C=O.
3. Атомы в составе ароматических циклов обычно записываются строчными буквами вместо прописных, поэтому, например, для бензола получим c1ccccc1, где c1 – атомы, находящиеся на концах разорванной при построении остовного леса связи.
4. Разветвления цепи задаются с помощью скобок. Боковые цепи молекулы заключаются в круглые скобки.

   Пример 3.1. 2,2,4-триметилоктан задается довольно громоздко C(C(C(C(C(C(C)(C)(C)))(C))))), однако такая запись неудобна для чтения из-за своей перегруженности скобками, поэтому ту же молекулу допускается записывать в неканонической форме как CC(C)(CC(CCCC)C)C.

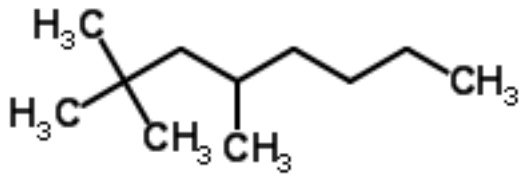


Рисунок 4.1. 2,2,4-триметилоктан

5. Также для нас важна цис-/транс- изомерия. Конфигурация относительно двойной связи записывается при помощи символов «/» и «\».

Пример 3.2. (3Z)-3-Гексен задается в виде CC/C=C\CC и является цис- изомером 3-Гексена, а (3E)-3-Гексен по правилам SMILES имеет вид CC/C=C/CC и соответствует транс-изомеру 3-Гексена (см. Рисунок 4.2 и 4.3).

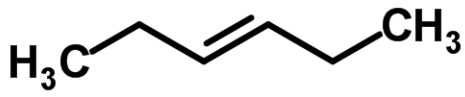


Рисунок 4.2. (3Z) -3-Гексан

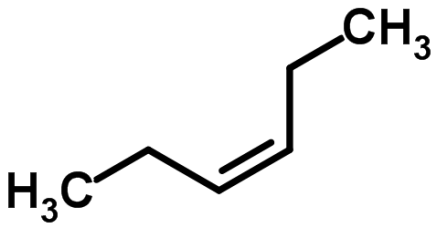


Рисунок 4.3. (3E) -3-Гексан

В случае алканов и аренов при построении SMILES не имеет смысла выбирать стартовую вершину обхода, так как в любом случае будет построена верная строка SMILES, которую примет на вход любая программа обработки данных химических веществ.

Однако при построении SMILES для дальнейшей идентификации алкенов необходимо учесть ряд особенностей построения названия алкенов.

При построении алкенов наиболее длинная углеродная цепь, содержащая двойную связь, получает название соответствующего алкана, в котором суффикс -ан заменен на -ен. Эта цепь нумеруется таким образом, чтобы углеродные атомы, участвующие в образовании двойной связи, получили номера, наименьшие из возможных. Затем радикалы называются и нумеруются, как и в случае алканов.

Поэтому требуется алгоритм, способный определить кратчайшие расстояния между всеми вершинами графа. На роль такого алгоритма подходит алгоритм Флойда — Уоршелла — динамический алгоритм для нахождения кратчайших расстояний между всеми вершинами взвешенного ориентированного графа.

Псевдокод алгоритма Флойда — Уоршелла [23]:

```
D^0=M            //M-матрица смежности
n=rows[M]     //размер матрицы
for k=1 to n do
      for i=1 to n do
            for j=1 to n do
                  d^k_ij=min(d^k-1_ij, d^k-1_ik+ d^k-1_kj)
return D^n
```

Сначала происходит инициализация матрицы кратчайших расстояний D^0, изначально она совпадает с матрицей смежности, в цикле увеличивается значение k и пересчитывается матрица расстояний, из D^0 получаем D^1, из D^1 — D^2 и так далее до $k = n$.

После применения этого алгоритма достаточно лишь выбрать самую длинную углеродную цепь, содержащую двойную связь.

4.3. Процедура очистки данных в Instant JChem

Для импорта мы использовали файл формата *.xls. После импортирования данных и некоторых преобразований таблицы к удобному виду мы получили со следующими атрибутами:

- Класс веществ (ALKANE, ALKENE, ARENE)
- Количество атомов углерода N
- Цис-/транс- изомерия (для алкенов)
- Строка SMILES
- Значения предсказанного индекса удерживания I

Одной из удобных функций Instant JChem являются автоматически вычисляемые поля, такие как название вещества, его формула, молекулярный вес, SMILES и множество других. Помимо этого, Instant JChem способен отображать структурные формулы молекул. В нашем случае все вычисления проводились программой по предоставленным нами строкам SMILES веществ. То есть, импорировав только построенные нами SMILES

веществ, мы получили полные названия всех веществ, соответствующих IUPAC[5], а также структурные формулы, тем самым решив задачу идентификации сгенерированных структур веществ. В итоге нами создана полная база данных химических веществ трех вышеуказанных классов (см. Рисунок 4.4).

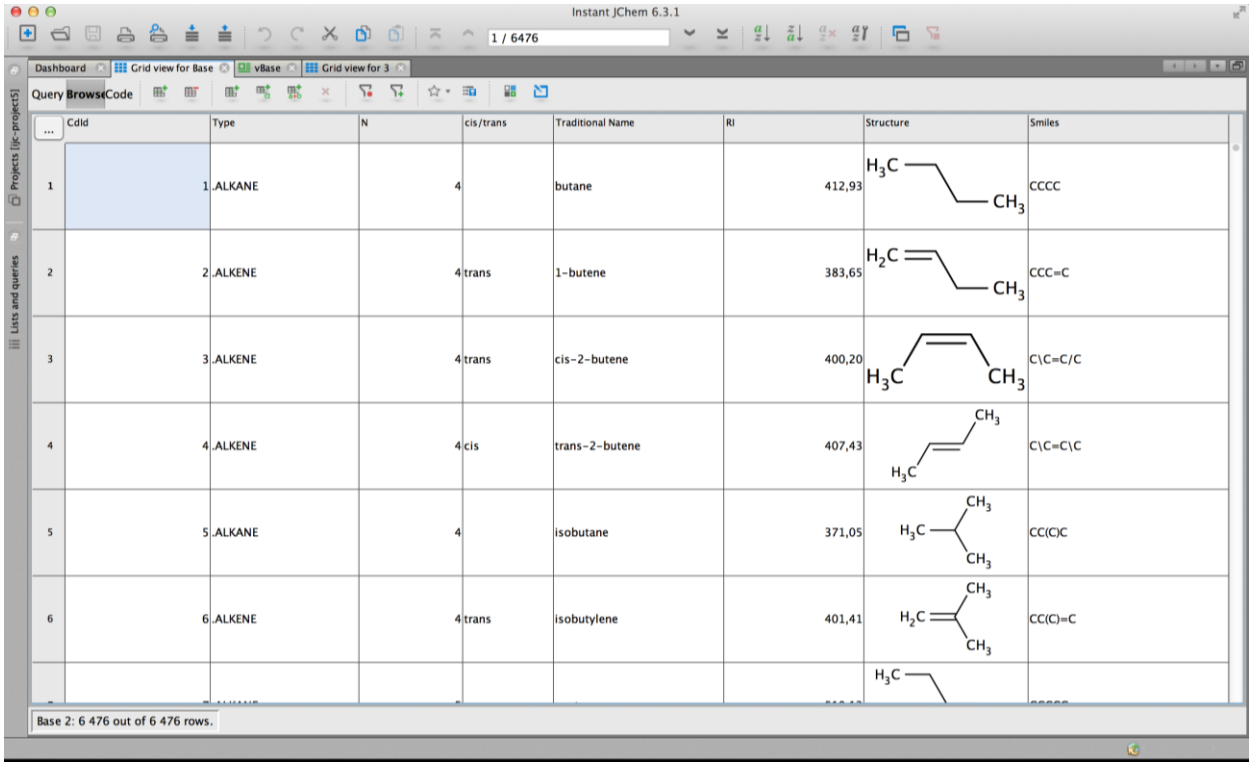


Рисунок 4.4 Представление базы данных в Instant JChem

В построенной базе данных присутствовало большое количество изоморфных графов (множество дублей). На уровне кода было бы очень затруднительно фильтровать дубли. Поэтому было необходимо придумать инвариант для матрицы смежности, то есть некое число, строку или вектор, однозначно характеризующий молекулярный граф. В качестве такого инварианта использовалось стандартное название вещества, построенное средвами JChem после загрузки и идентификации вещества по его SMILES-представлению структурной формулы.

После обработки данных в Instant JChem и удаления дублирующихся веществ по их

[5] International Union of Pure and Applied Chemistry – авторитетная международная структура, занимающаяся разработкой и распространением стандартов в области наименований химических соединений через межрегиональную комиссию по номенклатуре и обозначениям.

названию мы получили базу данных всех веществ трех классов алканов, алкенов и аренов с длиной цепочки от 4 до 12 атомов углерода. Полученная сводная таблица приведена в Таблице 4.1.

Таблица 4.1. Сводная таблица, демонстрирующая количество сгенерированных структур трех классов веществ, в зависимости от числа атомов углерода (N)

N	Алканы	Алкены	Алкилбензолы	N
4	2	4		6
5	3	6		9
6	5	17	1	23
7	9	36	1	46
8	18	91	4	113
9	35	215	8	258
10	75	542	22	639
11	159	1327	51	1537
12	355	3354	136	3845
Всего	**661**	**5592**	**223**	**6476**

4.4. Рабочее место для идентификации структур по индексу удерживания

Instant JChem предоставляет пользователям возможность самим создавать поля запросов, поиска, сортировки и представления данных. Например, на Рисунке 4.5 приведена форма поиска и отображения результатов запроса к базе, состоящая из слудующих полей:

1. Structure (в этом поле находится удобный редактор для изображения структур веществ и дальнейшего поиска по ним).
2. Type (в этом поле указывается класс веществ).
3. RI, где можно задать промежуток или какое-то конкретное значение значений индекса удерживания Ковача.
4. Traditional Name, где можно осуществлять поиск по названию вещества или же по части его названия.
5. Древовидная таблица Table, где создан удобный иерархический вид представления данных. Можно выбрать класс вещества, количество атомов углерода в нем и получить информацию по этому веществу в виде его названия, индекса удерживания

и структурной формулы.

6. Гистограмма, показывающая, какие вещества приходятся на какое значение индекса удерживания.

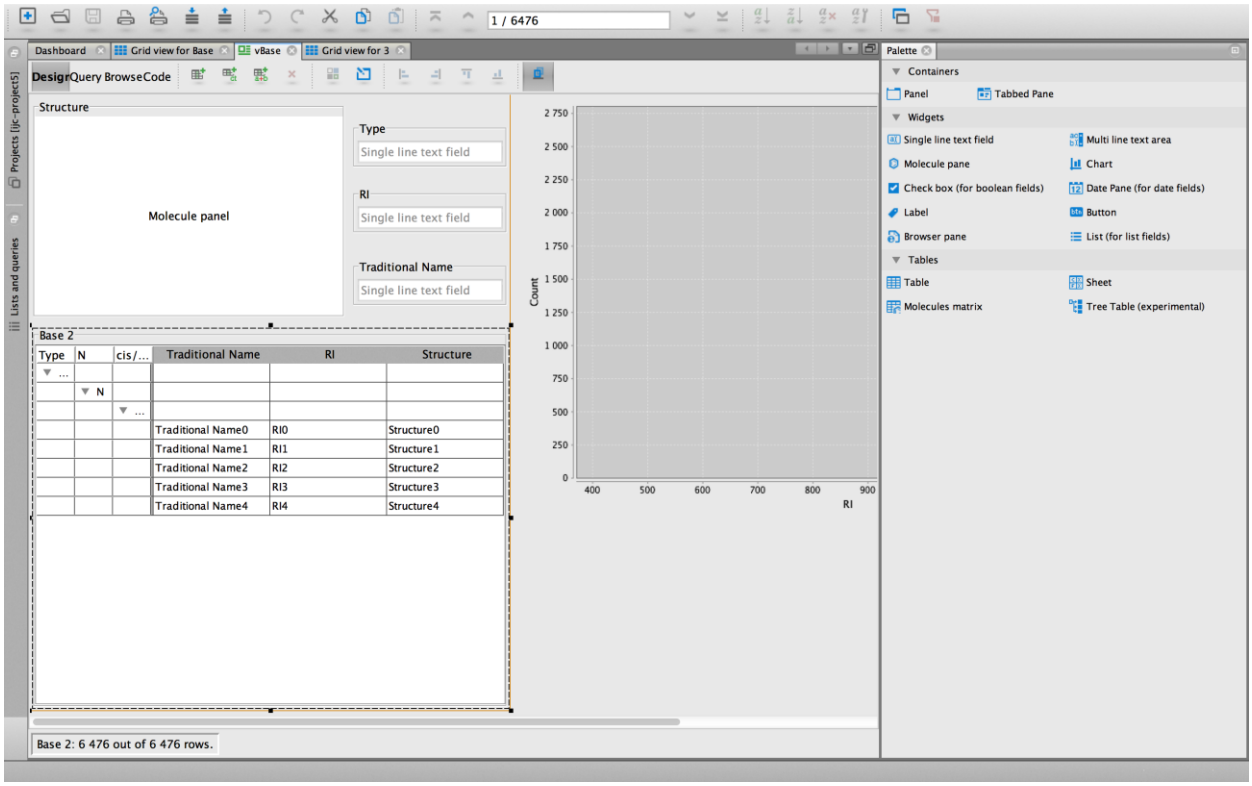


Рисунок 4.5. Создание полей для поиска и представления данных

Пример 4.1 Поиск данных в Instant JChem по индексу удерживания.

Чтобы найти всевозможные вещества с определенным индексом удерживания, необходимо ввести диапазон значений индекса удерживания в строке поиска RI. В качестве примера произведем поиск алканов с индексом удерживания в промежутке от 720 до 780 единиц (см. Рисунок 4.6).

В итоге получим девять алканов с индексом удерживания, лежащим в заданном промежутке (см. Рисунок 4.7).

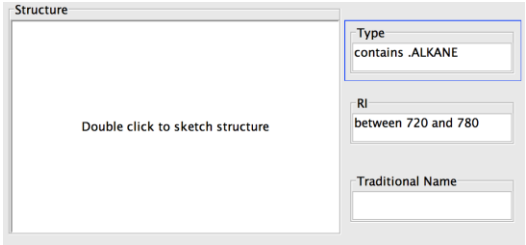


Рисунок 4.6. Пример поискового запроса

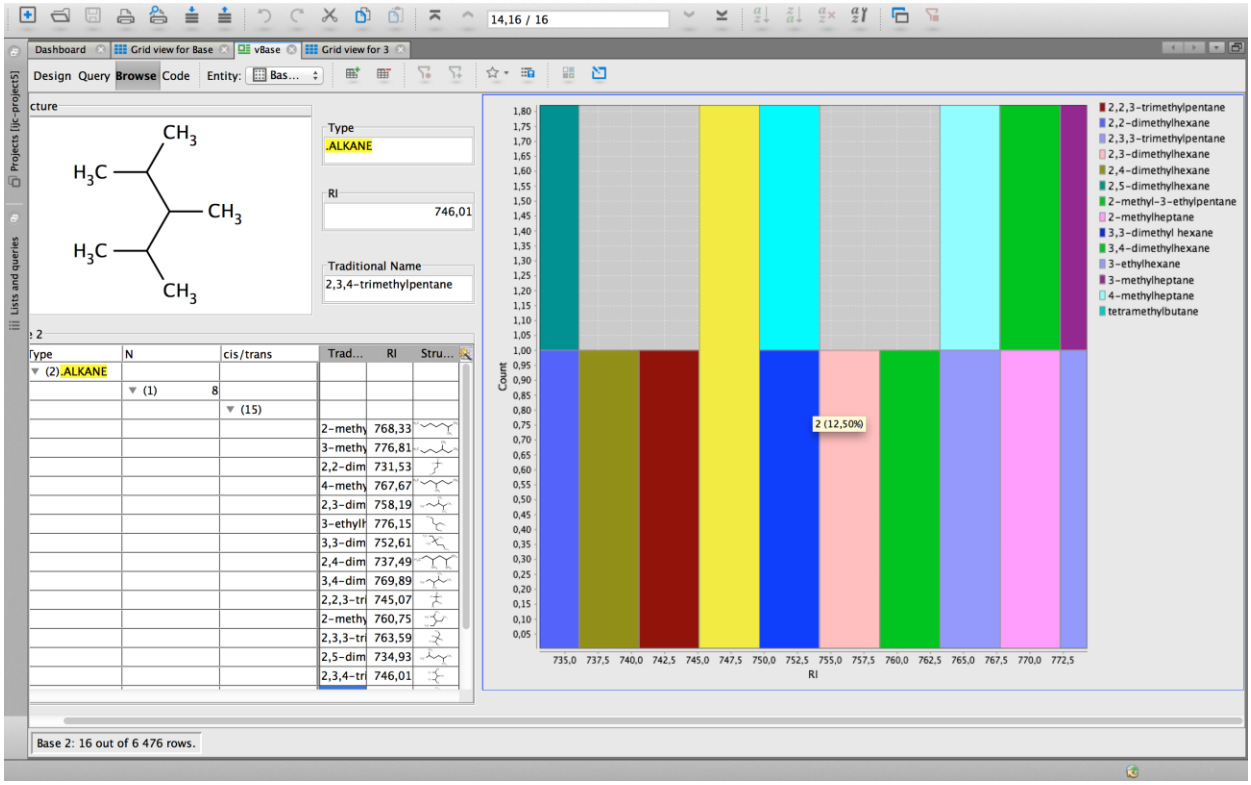


Рисунок 4.7. Пример результата поиска веществ

Пример 4.2 Поиск данных в Instant JChem по фрагменту молекулы.

Допустим, мы хотим осуществить поиск вещества по принадлежащему ему фрагменту. Для этого в окне Query открываем редактор структур молекул и в нем рисуем нужный нам фрагмент (Рисунок 4.8).

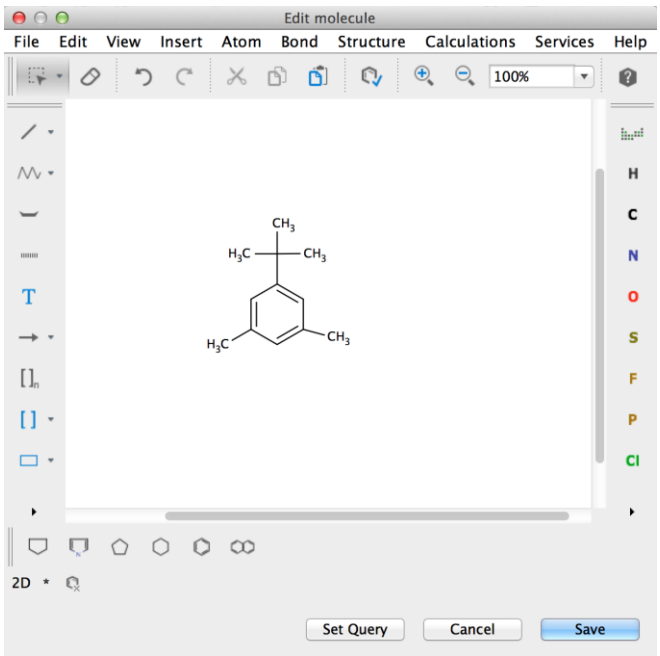


Рисунок 4.8 Фрагмент молекулы

Тогда, запустив поиск и перейдя в окно просмотра (Browse), мы увидим, что поисковому запросу соответствует двадцать одно вещество, все они принадлежат классу аренов, и количество атомов углеродов в этих молекулах лежит в диапазоне от 12 до 14 штук (см. Рисунок 4.9).

Также в качестве одного из вариантов визуализации результатов поиска на этом рисунке представлена гистограмма, показывающая соответствие индекса удерживания различным веществам. По оси абсцисс указан индекс удерживания Ковача (I), а справа от гистограммы расположен список веществ, выделенных разными цветами в соответствии с цветами на гистограмме.

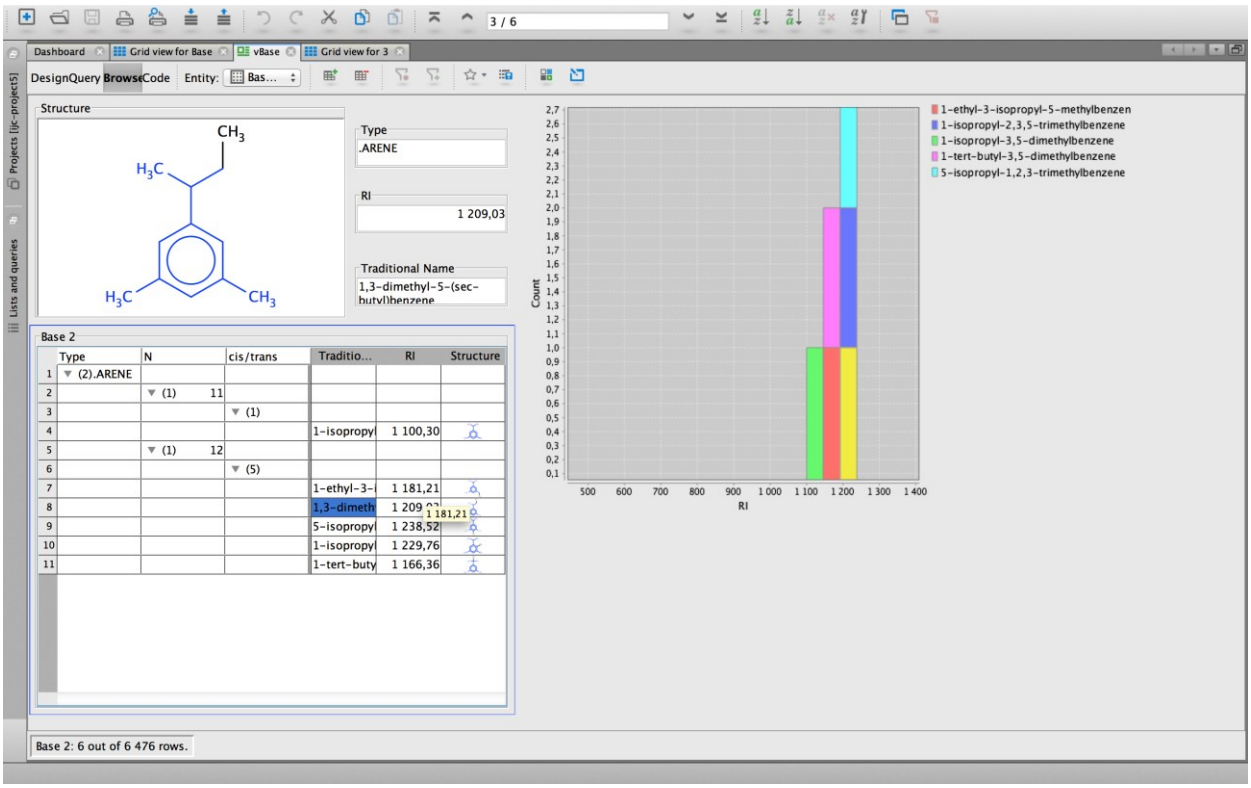


Рисунок 4.9. Пример результата поиска веществ по фрагменту молекулы

Пример 4.3 Построение графика.

Вместо гистограммы в качестве примера можно построить график зависимости индекса удерживания (RI) от количества атомов (N) в молекуле (см. Рисунок 4.10).

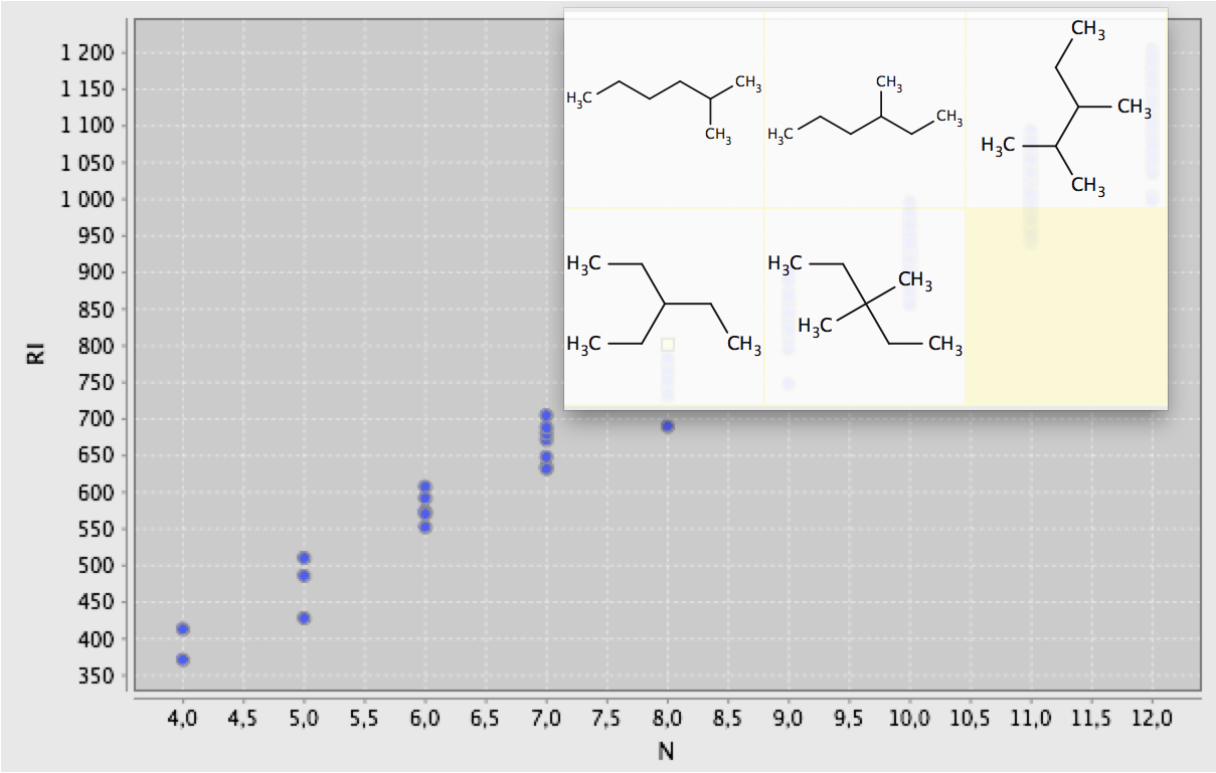


Рисунок 4.10 График зависимости RI от N

Чтобы узнать, какие вещества соответствуют данной точке, достаточно кликнуть мышью на эту точку, после чего появляется панель со структурами веществ (см. Рисунок 4.10).

Реализованный интерфейс позволяет успешно решать поставленную задачу идентификации пиков. Поисковые запросы можно задавать в различных формах (по индексу удерживания, по фрагментам, по названию). Также возможна визуализация данных в удобном для пользователя виде (график, гистограмма и так далее). Время отклика программы на запрос не превышает трех секунд, что позволяет оперативно получить необходимую информацию по необходимым веществам.

5. Выводы и перспективы

Таким образом, нами была решена обратная задача QSRR – определения веществ с заданным индексом удерживания. Сгенерированы все структуры алканов, алкенов, аренов – основных компонентов бензина. Высокая предсказательная способность у используемого полуэмпирического индекса позволяет точно рассчитать индекс удерживания для каждой структуры. Также в химической СУБД компании ChemAxon реализовано удобное рабочее место для идентификации структур по индексу удерживания.

Ввиду того, что были рассмотрены лишь три класса веществ, задача решена не полностью. Одной из перспектив данного исследования является рассмотрение остальных классов веществ, содержащихся в бензине.

6. Список литературы

1. ШАПОВАЛОВА Е.Н., ПИРОГОВ А.В. *Хроматографические методы анализа*. // – М: Кафедра аналитической химии Химического Факультета МГУ. 2007.
2. САКОДЫНСКИЙ К.И., БРАЖНИКОВ В.В., ВОЛКОВ С.А.и др. *Аналитическая хроматография*. – М.: Химия, 1993.
3. *Схема газового хроматографа*. [Электронный ресурс] // Википедия. – Режим доступа: http://ru.wikipedia.org/wiki/Газовая_хроматография
4. *Хроматограмма автомобильного бензина.* [Электронный ресурс] // Библиотека хроматограмм. – Режим доступа: http://www.unichrom.com/lib/gcr/a01.pdf
5. ХАЙВЕР К., НЬЮТОН Б., САНДРА П., и др. *Высокоэффективная газовая хроматография.* Пер. с англ. /Под ред. К. Хайвера. – М.: Мир, 1993.
6. KOVATS E., GIDDINGS J.C., KEUER R.A. *Advances in Chromatography, Volume 1, Chapter 7.* – New York: Marcel Dekker. 1965.
7. ПРУДКОВСКИЙ А.Г., ДОЛГОНОСОВ А.М. *Инструмент для оценки индекса Ковача по времени удерживания вещества в газовой хроматографии.* // Журнал аналитической химии 2008, Т.63, № 9. – С. 935–940.
8. ЦАРЕВ Н.И., ЦАРЕВ В.И., КАТРАКОВ И.Б. *Практическая газовая хроматография*. – Барнаул: Издательство Алтайского государственного университета, 2000.
9. NANTASENAMAT C., ISARANKURA-NA-AYUDHYA C., NAENNA T. et al. *A practical overview of quantitative structure-activity relationship*. // EXCLI Journal. – V. 8. – 2009. – P. 74–88.
10. WIENER H. *Structural determination of paraffin boiling points* // Journal of American Chemical Society – № 69 (1). 1947. – P. 17-20.
11. RANDIĆ M. *Characterization of molecular branching,* // Journal of the American Chemical Society – V. 97 (23). – 1975. P. 6609–6615.

12. СТАНКЕВИЧ М.И., СТАНКЕВИЧ И.В., ЗЕФИРОВ Н.С. *Топологические индексы в органической химии* // Успехи химии. Март 1988. – С. 337–350.
13. HE´BERGER K. *Review. Quantitative structure–(chromatographic) retention relationships* // Journal of Chromatography A, V. 1158. 2007. – P. 273–305.
14. HE´BERGER K. *Quantitative Structure-Retention Relationships* // Gas Chromatography. 2012. – P. 451–475.
15. HEINZEN V.E., SOARES M.F., YUNES R.A. *Semi-empirical topological method for the prediction of the chromatographic retention of cis- and trans-alkene isomers and alkanes*. // Journal of Chromatography A. V. 849, I. 2, 23 July 1999, P. 495–506.
16. Алканы ESTRADA E., GUTİERREZ Y. *Modeling chromatographic parameters by a novel graph theoretical sub-structural approach. //* Journal of Chromatography A, 858, July 15, 1999. P. 187–199.
17. PORTO L.C., SOUZA E.S., JUNKES B.S., et al. *Semi-empirical topological index: Development of QSPR/QSRR and optimization for alkylbenzenes* // Talanta V. 76, I. 2, July 15, 2008. P. 407-412.
18. ХАРАРИ Ф. *Теория графов*. // Изд. 2-е. – М.:Едиториал УРСС, 2003.
19. КНУТ Э.Д., Искусство программирования. //, Т. 4. Генерация всех деревьев. История комбинаторной генерации. выпуск 4. 2007.
20. КНУТ Э.Д., *Искусство программирования,* // Т. 1. Основные алгоритмы. — 3-е изд. — М.: Вильямс, 2006.
21. Instant JChem [Electronic resource] / ChemAxon website. – Режим доступа: https://www.chemaxon.com/products/instant-jchem-suite/instant-jchem/
22. SMILES - A Simplified Chemical Language. [Electronic resource] // Daylight Chemical Information Systems, Inc. – Режим доступа: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
23. ЛЕВИТИН А.В. *Динамическое программирование: Алгоритм Флойда поиска крат-*

чайших путей между всеми парами вершин // Алгоритмы: введение в разработку и анализ. — М.: Вильямс, 2006.