УДК 544.131

ПРЕДСКАЗАНИЕ ИНДЕКСА УДЕРЖИВАНИЯ КОМПОНЕНТОВ БЕНЗИНОВ С ПОМОЩЬЮ ТОПОЛОГИЧЕСКИХ ИНДЕКСОВ[1]

Губко М.В.[2], Милосердов О.А.[3]
(*Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Решается актуальная задача идентификации компонентов бензинов по значению их индекса удерживания Ковача. Для этого дается краткое введение в хроматографию и методы решения задач «структура-индекс удерживания» (QSRR), с помощью топологических индексов строятся точные регрессии, перечисляются всевозможные изомеры трех основных классов компонентов бензина, а также реализуется удобное рабочее место для идентификации структур по индексу удерживания в СУБД компании ChemAxon.

Ключевые слова: хроматография, индекс удерживания Ковача, предсказание индекса удерживания по структуре, фрагментные топологические индексы, перечисление деревьев, химическая СУБД Instant JChem, генерация SMILES.

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 относи-

[1] *Работа выполнена при поддержке РФФИ* (13-07-00389).
[2] *Михаил Владимирович Губко, кандидат технических наук, с.н.с.* (*mgoubko@mail.ru*).
[3] *Олег Александрович Милосердов, бакалавр, техник* (*oleg_milos@mail.ru*).

тельно 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.

Для понимания цели данного исследования необходимо рассмотреть принцип работы хроматографических приборов. В качестве примера рассмотрим схему устройства газового хроматографа – одного из самых популярных аналитических устройств, используемых исследователями (см. рис. 1).

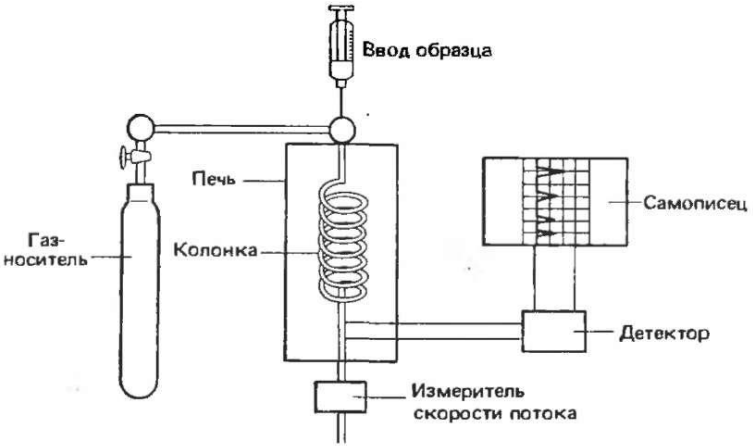

Рисунок 1. Блок-схема газового хроматографа

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. рис. 2).

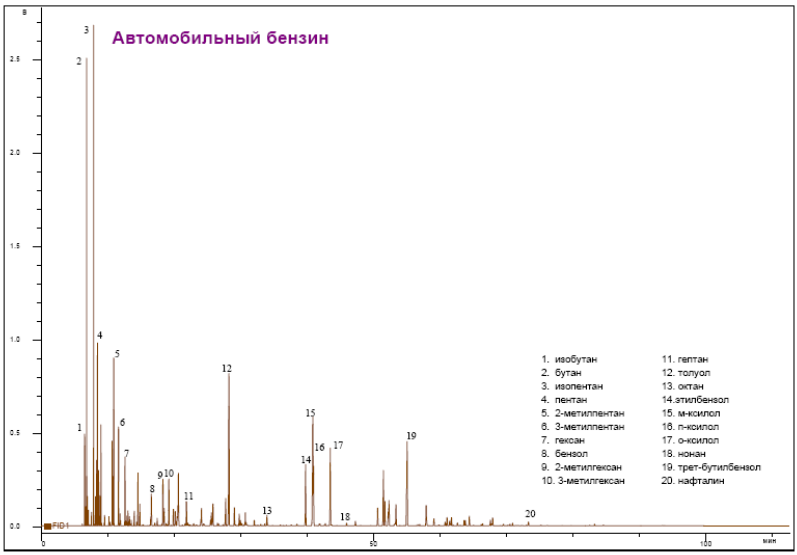

Рисунок 2. Пример хроматограммы автомобильного бензина

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. Индекс удерживания задается следующей формулой [2]:

$$(1) \qquad I_i = 100\left(m + \frac{\ln(K_i / K_m)}{\ln(K_{m+1} / K_m)}\right),$$

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

m – номер нормального алкана, содержащего m атомов углерода;
I_i – индекс Ковача рассматриваемого вещества i, выходящего между алканами m и $m + 1$.

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.

2. Методы QSRR и постановка задачи

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. В частности, в данном исследовании необходимо идентифицировать пики на хроматограмме бензина.

Делать это позволяют методы QSRR[4] (количественный анализ «структура – индекс удерживания»). Один из способов прогнозирования свойств веществ осуществляется с помощью молекулярных дескрипторов, которые, в свою очередь, делятся на несколько категорий: топологические индексы, физико-химические дескрипторы и квантово-химические дескрипторы. В данном исследовании были использованы топологические индексы, т.к. они легко вычислимы и при их использовании возможно решение важной задачи поиска веществ с заданными свойствами – задачи оптимизации одних топологических индексов при ограничениях на другие.

Существует несколько видов топологических индексов: основанные множестве степеней вершин (индекс Рандича), связанные с матрицей расстояния D (индекса Винера), зависящие от спектральных характеристик графа, а так же информационные топологические индексы, к которым относятся фрагмент-

[4] *Quantitative Structure-Retention Relationship*

ные индексы [3]. В качестве фрагментного индекса может быть использован любой фрагмент структуры молекулы. Например, атом углерода, принадлежащий бензольному кольцу с прикрепленным радикалом изопропилом. В настоящей работе мы использовали именно фрагментные индексы.

Итак, целью настоящего исследования является предсказание индекса удерживания компонентов бензинов с помощью топологических индексов.

Для достижения поставленной цели необходимо решить следующие задачи:

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

Нам была предоставлена база компонент бензина, в которой содержится более 450 компонентов. Всего в данной базе было выделено 15 классов веществ, но для первоначального рассмотрения мы выбрали три наиболее многочисленных: алканы, алкены и арены.

3. Подбор регрессий

Для поиска подходящих исследований использовались два обзора К. Хебергера по современному состоянию QSRR за 2007 [4] и 2012 [5] годы. В этих обзорах проанализировано более 500 статей с работами по QSRR. Критериями отбора подходящих регрессий была высокая точность предсказательной способности регрессии ($R^2 > 0{,}99$), большая обучающая база данных (не менее 50 веществ), а также тот факт, что база значений индексов

удерживания была получена на сквалане (это самая популярная неполярная фаза) при температуре 100 °C.

В итоге после тщательного анализа были выбраны две статьи: [6] для алканов и алкенов и [8] для аренов. Построение регрессии в [6] и [8] 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.

Из экспериментов следует, что невозможно посчитать какой либо постоянный вклад в индекс удерживания для конкретного атома углерода. Чтобы решить эту проблему, авторы [6] прибегли к следующему методу: назначили приблизительные (полученные экспериментально) значения вкладов для каждого из типов атомов: 100- для атомов метильной группы (CH3–), 90 для вторичных (–CH2–) атомов, 80 для третичных (–CH<) и 70 для четвертичных (>C<), разделив их на 100 для нормировки. Аналогичные веса введены для фрагментов алкенов, например 87,2 для фрагмента вида –CH=trans с номером 4 места в цепи, причем в статье учитывается цис-/транс-изомерия путем присвоения различных весов атомов в цепочках цис- и транс- изомеров. Определение степени стерических эффектов, присутствующих в углеводороде, зависит также от размера замещающей группы, а не только от местоположения конкретного атома, поэтому авторы [6] добавляют еще одно слагаемое в виде произведения степени фрагмента на его вес. В результате, полуэмпирический топологический индекс (I_{ET}) выражается в виде:

$$I_{ET} = \Sigma n_i(C_i + d_i \lg C_i), \quad (2)$$

где C_i – вес фрагмента i-ого типа, n_i – количество фрагмента i-го типа, d_i – степень центрального атома углерода i-ого фрагмента.

Аналогичная идея использовалась в статье [8] для предсказания индекса удерживания аренов. В итоге авторы [6] и [8] получили следующие регрессии:

- для алкенов – $I_{CALC} = 122{,}8446\ I_{ET} - 41{,}7054$ для которых регрессия была построена на обучающей базе из $N = 79$ ве-

ществ, а точность полученной регрессии составила $R = 0{,}99996$.

- для алканов – $I_{CALC} = 116{,}8\, I_{ET} - 19{,}05$; $N = 157$; $R = 0{,}9901$,
- для аренов – $I_{CALC} = 23{,}0824\ I_{ET} - 39{,}7381$; $N = 122$; $R = 0{,}9998$.

Точность данных регрессий достаточно велика, подтверждением чему является графики зависимости вычисленного индекса удерживания от экспериментального (см. рис. 3 и 4). Как можно видеть, предсказанные значения индекса точно ложатся на прямую.

Таким образом, работы [6] и [8] показывают, что фрагментные топологические индексы могут эффективно применяться для предсказания индекса удерживания Ковача, что позволяет их использовать для решения задачи данного исследования.

4. Перечисление химических структур и обработка полученной базы данных

После выбора подходящих регрессий перебираем все возможные структуры алкенов, алканов и аренов.

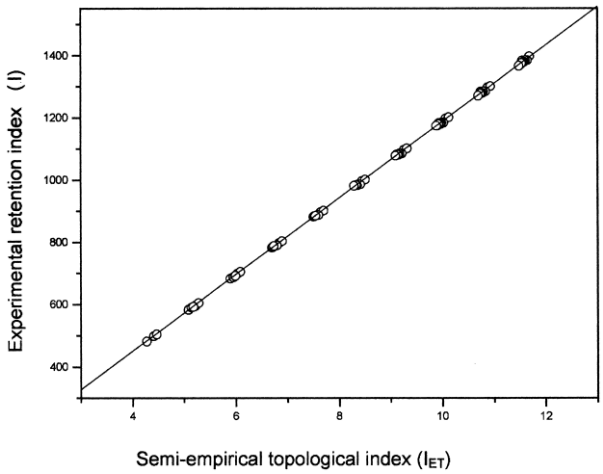

Рисунок 3. График зависимости экспериментального индекса удерживания алкенов от предсказанного [6]

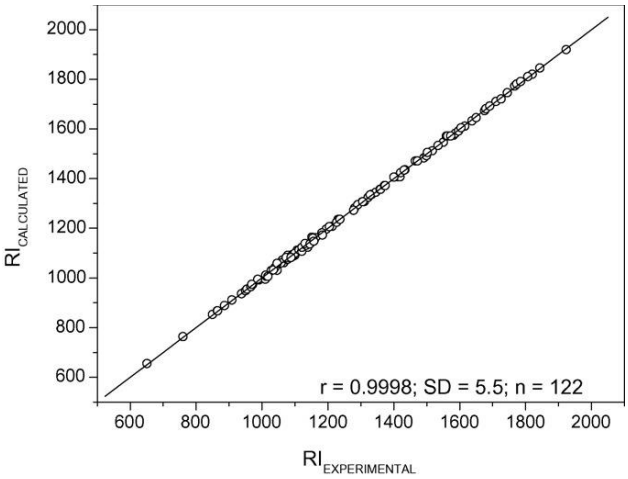

Рисунок 4. График зависимости экспериментального индекса удерживания аренов от предсказанного [8]

Используя методы QSRR, при представляем структуры веществ как графы с вершинами в атомах углерода и дугами вместо связей между молекулами. В качестве первого шага мы перечисляем все структуры алканов. Можно заметить, что граф алканов представляет собой ненаправленное дерево, поэтому, мы вначале перечисляли все бинарные деревья по алгоритму, взятому из [1], затем через взаимно однозначное соответствие из бинарных деревьев получим ненаправленные деревья [1]. Далее, убирая одну корневую вершину у полученных графов, мы получим корневой лес.

Получается, что алканы – это леса с одним компонентом, алкены – с двумя (т.к. мы учитываем двойную связь в молекулах алкенов), а у аренов – 6 компонентов (учитываем наличие бензольного кольца, состоящего из 6 атомов углерода).

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. В начале необходимо найти удобный вид представле-

ния структур молекул. Таким видом стал SMILES[5] – это общепринятая строковая нотация структурных формул. В качестве примера приведем SMILES для 2,2,4-триметилоктана – CC(C)(CC(CCCC)C)C.

Мы сгенерировали SMILES для каждой структуры по правилам, взятым из [9], и использовали их для загрузки полученной базы данных в программный продукт InstantJChem [7], который, помимо реализации работы с базой данных химических веществ, решает задачи визуализации, вычисления топологических индексов, и многие другие.

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[6], 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. рис. 5).

В построенной базе данных присутствовало большое количество изоморфных графов (множество дублей). 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

[5]SimplifiedMolecularInputLineEntrySpecification, спецификация упрощенного представления молекул в строке ввода.

[6]InternationalUnionofPureandAppliedChemistry –международная структура, занимающаяся разработкой и распространением наименований химических соединений через межрегиональную комиссию по номенклатуре и обозначениям.

дублирующихся веществ по их названию мы получили базу данных всех веществ трех классов алканов, алкенов и аренов с длиной цепочки от 4 до 14 атомов углерода. Полученная сводная таблица приведена в Таблице 1.

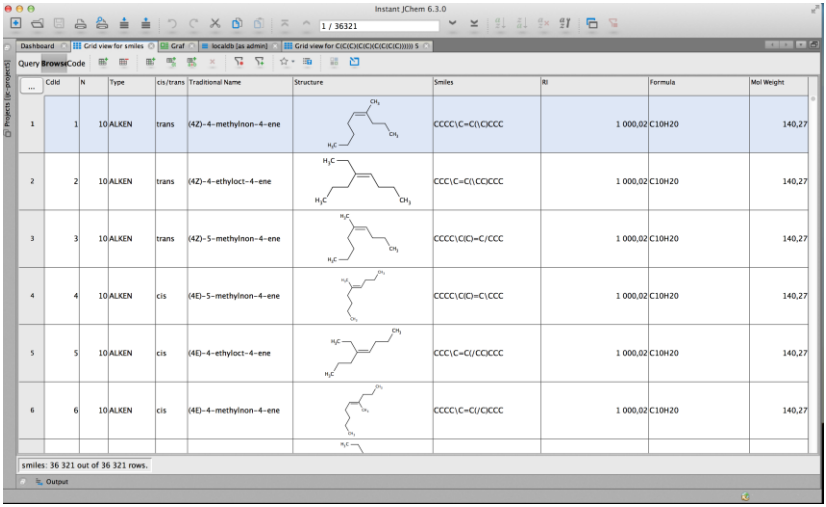

Рисунок 5. Представление базы данных в Instant JChem

Таблица 1. Сводная таблица, демонстрирующая количество сгенерированных структур трех классов веществ, в зависимости от числа атомов углерода (*N*)

N	Алканы	Алкены	Алкилбензолы
4	2	6	-
5	3	10	-
6	5	26	1
7	9	50	1
8	18	116	4
9	33	246	8
10	73	592	22
11	144	1314	51
12	323	3182	136
13	708	7562	334
14	1663	18810	869
Всего	**2981**	**31914**	**1426**

Instant JChem предоставляет пользователям возможность самим создавать поля запросов, поиска, сортировки и представления данных. Например, мы создали удобное рабоче место для поиска всевозможных веществ с определенным индексом удерживания. Для этого необходимо ввести диапазон значений индекса удерживания в строке поиска RI. В качестве примера произведем поиск алканов с индексом удерживания в промежутке от 1259 до 1261 единиц (см. рис. 6). В итоге получим семь алканов и 2 арена с индексом удерживания, лежащим в заданном промежутке (см. рис. 7).

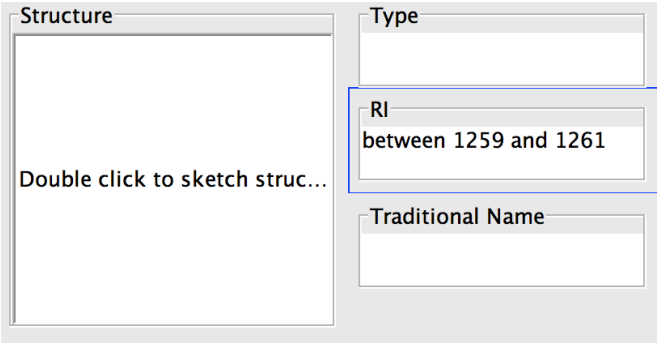

Рисунок 6. Пример поискового запроса

Вместо гистограммы, показанной на рис. 7, в качестве примера можно построить график зависимости индекса удерживания (I) от количества атомов (N) в молекуле. Чтобы узнать, какие вещества соответствуют данной точке, достаточно кликнуть мышью на эту точку, после чего появляется панель со структурами веществ (см. рис. 8).

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.

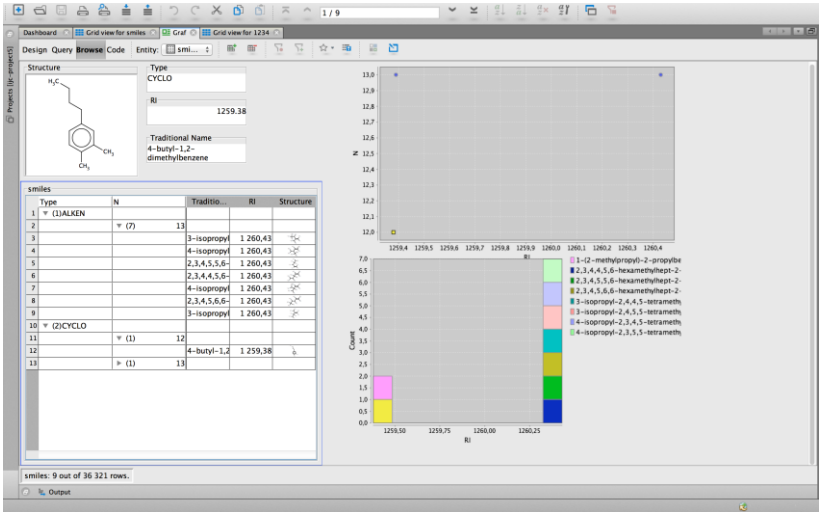

Рисунок 7. Пример результата поиска веществ

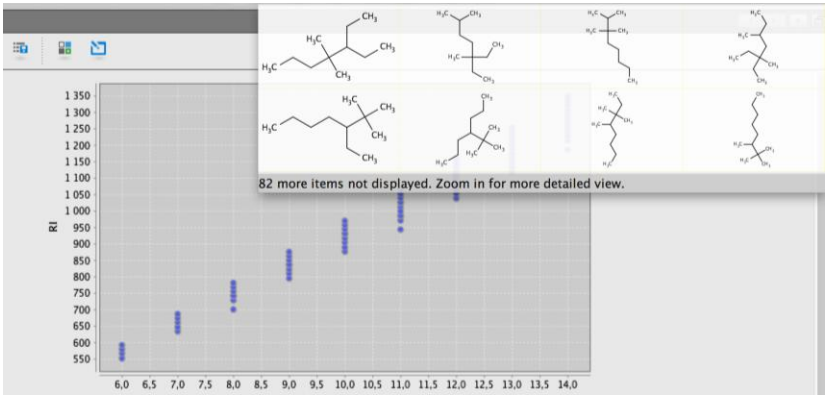

Рисунок 8. График зависимости I от N

5. Заключение

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.

Литература

1. КНУТ Э.Д. Искусство программирования // Т. 4. Генерация всех деревьев. История комбинаторной генерации. выпуск 4. – 2007.
2. 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.
3. СТАНКЕВИЧ М.И., СТАНКЕВИЧ И.В., ЗЕФИРОВ Н.С. *Топологические индексы в органической химии* // Успехи химии. Март 1988.– С. 337–350.
4. HÉBERGER K. *Review. Quantitative structure–*(*chromatographic*) *retention relationships //* Journal of Chromatography A, V. 1158. 2007. – P. 273–305.
5. HÉBERGER K. *Quantitative Structure-Retention Relationships*. in “Gas Chromatography” [ed. by C.F.Poole]. – Oxford: Elsevier, 2012. – P. 451–475.
6. 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.
7. Instant JChem [Electronic resource] / ChemAxon website. – Режим доступа: https://www.chemaxon.com/products/instant-jchem-suite/instant-jchem/
8. 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.
9. SMILES - A Simplified Chemical Language. [Electronic resource] // Daylight Chemical Information Systems, Inc. – Режим доступа: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

Abstract: *After a short introduction to the chromatographic analysis and quantitative structure-retention relation* (*QSRR*) *analysis we put the topical problem of petrol component identification basing on its retention index value. We use fragmental topological indices from the literature to build precise regressions. Then we enumerate all isomers of the three major petrol component classes and accompany them with the predicted values of retention index. Finally, we design a useful workplace using the chemical database management system Instant JChem from ChemAxon.*

Keywords: chromatography, Kovats retention index, quantitative structure-retention relation, fragmental topological indices, tree generation, Instant JChem, SMILES generation.