УДК 544.354.081.7:004.021

Оценка физических и физико-химических свойств молекул дендримеров и олигомеров на основе анализа количественной взаимосвязи «структура–свойство»

***Мансурова М.Г.**, студент*

кафедра «Теоретическая информатика и компьютерные технологии»

Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана

Научный руководитель: Дубанов А.В.,

Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана

qrcs@mail.ru

Введение. Дендримеры являются синтетическими макромолекулами с регулярной древовидной структурой. Они синтезируются в результате серии контролируемых реакций, каждая из которых приводит к экспоненциальному увеличению количества мономеров в составе полимера. Дендримеры широко используются как при изготовлении электронных устройств и покрытий [1], рентгеноконтрастных веществ [2], так и в качестве носителей для доставки лекарственных соединений [2]. Область применения того или иного дендримера определяется его физическими, физико-химическими и химическими свойствами. Эти свойства зависят, в том числе, от функциональных групп на поверхности молекул - концевых групп ветвей дендримера. Изменение таких групп может влиять на свойства самой молекулы, например на растворимость или токсичность [3].

Олигомерами называются полимеры, состоящие из небольшого числа мономеров. Некоторые олигомеры представляют собой разветвленные молекулы. Свойства этих веществ меняются при изменении числа составных звеньев [4] и их типов.

Построение моделей количественной взаимосвязи "структура-свойство" типов QSAR (Quantitative structure–activity relationship - модели "структура-свойство") или QSPR (Quantitative

Structure-Property Relationship - модели "структура-свойство", прогнозирующие физические и физико-химические свойства органических соединений) имеет широкое практическое применение. QSAR/QSPR-модели позволяют прогнозировать свойства вещества по его структуре до его синтеза и измерения его свойств. Для построения моделей "структура-свойство" применяются методы машинного обучения и математической статистики [5]. При описании молекулярной структуры химических соединений рассматриваемой выборки используются молекулярные дескрипторы, представляющие собой функционал, который отображает структуру рассматриваемой молекулы на множество вещественных чисел. Дескрипторы бывают структурные (топологические, фрагментные), дескрипторы межмолекулярных взаимодействий, молекулярной формы и другие [6, 7, 8].

Цель данного исследования заключалась в поиске таких наборов молекулярных дескрипторов, использование которых в QSPR позволило бы получать прогноз различных физических, физико-химических и химических свойств дендримеров и олигомеров с точностью, приемлемой для практического применения.

В качестве прогнозируемых свойств были рассмотрены следующие: плотность и индекс преломления (для смешанной выборки из олигомеров и дендримеров), температура кипения (для олигомеров). Такой выбор прогнозируемых показателей был обусловлен, с одной стороны - доступностью экспериментальных данных в литературе и открытых базах данных, с другой - важностью этих показателей для практического применения [9, 10, 11].

Объекты и методы. Построение QSPR-моделей прогноза выполнялось для соединений дендримеров, олигомеров, а также их смешанной выборки. В качестве данных для обучения и прогноза были взяты вещества из баз данных ChemSpider (для олигомеров) и Sigma-Aldrich (для дендримеров) [12, 13].

Вычисления дескрипторов проводились с помощью библиотеки молекулярных дескрипторов RDKit [14] для языка Python в среде исполнения PyCharm и пакета дескрипторов Rcpi [15, 16] для языка R в среде исполнения RStudio. Поиск наилучшей QSPR-модели был выполнен с помощью алгоритма пошаговой регрессии, реализованного в языке R [17, 18]. С помощью библиотеки, содержащей большое количество алгоритмов машинного обучения, sklearn [19, 20] для языка Python был построен качественный прогноз свойств рассматриваемых соединений.

Все вычисления были выполнены на ПК с процессором Intel Core i7, 2,40 ГГц, 5 Гб RAM под управлением операционной системы Ubuntu Linux (64 bit) версии 12.

Характеристика выборок. Выбор соединений дендримеров и олигомеров в качестве изучаемых веществ обуславливается следующим. Дендримеры стали активно исследоваться сравнительно недавно, их синтез был впервые произведен всего лишь около десяти лет назад. При этом, область применения этих веществ обширна и разнообразна. Исследование свойств дендримеров является актуальным для нескольких предметных областей [21]. Однако открытые базы данных практически не содержат информацию об их структуре и свойствах.

В результате выборка дендримеров была дополнена выборкой олигомеров. В отличие от дендримеров, олигомеры изучаются уже давно - таким образом, их физические и химические свойства определены экспериментально и доступны для широкого круга соединений. В то же время, в виду полимерной структуры, дендримеры и олигомеры можно рассматривать как соединения родственных классов.

Выборка дендримеров представляет собой различные соединения классов PAMAM и PPI от нулевого до седьмого поколения [22, 23]. В выборке также присутствуют вещества нецелых поколений.

В табл. 1 и 2 представлены величины, характеризующие выборку олигомеров A, дендримеров B и объединенную выборку этих соединений AB: минимум, максимум, нижний (Q1), средний (Q2 (медиана)) и верхний (Q3) квартили, а также среднеквадратическое (стандартное) отклонение σ, количество неизвестных значений в выборке U и общее число веществ N. В каждой таблице содержатся значения для определенного свойства веществ: плотности и индекса преломления - для обеих выборок, температуры кипения - для олигомеров. Выборка для температуры кипения представлена в табл. 3.

Стоит обратить внимание на то, что для плотности и индекса преломления разница между минимальными и максимальными значениями в выборках очень мала (~1). Среднеквадратическое отклонение, при этом, находится в интервалах 0,08 - 0,32 для плотности и 0,07 - 0,15 для преломления. Таким образом, для этих свойств необходимо построить такую модель, ошибка прогноза которой будет не большей, чем 0,1 - 0,2 единиц. Такая точность может считаться приемлемой для прогноза плотности и индекса преломления рассматриваемых соединений. Здесь и далее под стандартной ошибкой подразумевается остаточная стандартная ошибка модели (среднеквадратическое/стандартное отклонение или *RMSE*).

Для температуры кипения, однако, нет необходимости в такой точности, так как велико значение стандартного отклонения, а также разница между минимальными и максимальными значениями в выборке.

Таблица 1

Характеристика выборок A и B для плотности (г/см^3)

Выборка	Min	Max	Q1	Q2 (медиана)	Q3	σ	U	N
A	1,10	2,20	1,10	1,20	1,50	0,30	7	29
B	0,79	1,01	0,81	0,86	0,95	0,08	0	20
AB	0,79	2,20	0,86	1,10	1,20	0,32	7	49

Таблица 2

Характеристика выборок A и B для индекса преломления (n)

Выборка	Min	Max	Q1	Q2 (медиана)	Q3	σ	U	N
A	1,30	2,02	1,47	1,50	1,61	0,15	1	29
B	1,34	1,52	1,35	1,36	1,37	0,07	1	20
AB	1,30	2,02	1,37	1,47	1,54	0,15	2	49

Таблица 3

Характеристика выборки A для температуры кипения (°C)

Выборка	Min	Max	Q1	Q2 (медиана)	Q3	σ	U	N
A	50,00	1418,50	375,80	628,80	776,90	368,75	9	29

Множество двумерных дескрипторов. Библиотека RDKit языка Python предоставляет набор молекулярных дескрипторов в составе: фрагментные (fr_Al_COO, fr_Al_OH, fr_Al_OH_noTert и др.), физико-химические (MolLogP, MolMR, MolWt, PEOE_VSA1, SMR_VSA1, SlogP_VSA1, EState_VSA1, VSA_EState1, TPSA и др.), квантово-химические (NumValenceElectrons) и другие (HeavyAtomCount, NHOHCount, NOCount, NumHAcceptors,

NumHDonors, NumHeteroatoms, NumRotatableBonds, RingCount). Названия дескрипторов приведены согласно документации к пакету RDKit.

Набор молекулярных дескрипторов пакета Rcpi языка R содержит в себе большое число топологических дескрипторов, а также квантово-химических, фрагментных, физико-химических и ряда других.

Поиск значимых дескрипторов и методы прогноза. Для поиска дескрипторов, демонстрирующих наибольшую взаимосвязь с прогнозируемым значением, было выполнено обучение моделей (средствами языка R) на наборе из всех дескрипторов библиотеки RDKit языка Python и всех доступных дескрипторах Rcpi языка R. Дескрипторы выбирались методом регрессионного анализа по принципу допустимого уровня значимости статистического теста ($p << 0,05$) и наибольшего значения статистики *t*-критерия Стьюдента [24, 25]. Для точного *t*-теста необходимо, чтобы выборки были распределены нормально. Было найдено, что выборки для плотности и индекса преломления действительно являются нормальными [26]. При поиске коррелирующих значений в данных исследованиях применялся коэффициент корреляции Пирсона.

Дальнейшая работа с выборками была проведена с помощью классификатора DecisionTreeRegressor (при глубине дерева, равной 15) библиотеки sklearn. При прогнозе использовалась кросс-валидация LeaveOneOut языка Python. Перед разделением выборки на обучающую и контрольную индексы веществ перемешивались случайным образом - что необходимо для того, чтобы в каждой из подвыборок были молекулы как дендримеров, так и олигомеров (так как входные данные могут содержать вещества сначала одного класса, потом - другого).

Чтобы не выполнять поиск среди всех возможных дескрипторов для всех веществ выборки, поиск дескрипторов был выполнен сначала для каждого класса этих соединений, а затем для объединенной выборки. При этом множества дескрипторов из RDKit и Rcpi оценивались независимо друг от друга.

Алгоритм можно описать следующим образом: находим значимые дескрипторы для выборки A из библиотеки RDKit (группа дескрипторов G_1) и пакета Rcpi (G_2), аналогично - для выборки B (G_3, G_4). Затем для смешанной их выборки выполняем поиск дескрипторов по следующему правилу:

$$G = ((G_1 \cup G_3) \cup (G_2 \cup G_4))$$

Далее выбираем наиболее значимые дескрипторы из группы G, а затем удаляем коррелирующие дескрипторы, негативно влияющие на качество модели прогноза.

Поиск дескрипторов для прогноза плотности.

В табл. 4 представлена характеристика модели прогноза, полученной по значимым дескрипторам группы G для плотности: R^2 - скорректированный коэффициент детерминации модели, p - вероятность ошибки при отклонении нулевой гипотезы.

Таблица 4: Характеристика модели прогноза плотности смешанной выборки

Выборка	Стандартная ошибка ($г/см^3$)	Значение R^2	Значение p
AB	0,13	0,99	< 2,20e-16

Для группы из 15 значимых и не коррелирующих между собой дескрипторов коэффициент детерминации r^2 оказался равным 0,58 при ошибке *RMSE*, равной 0,19. Такие значения являются приемлемыми для практического применения.

Соответствующий набор дескрипторов: fr_SH, fr_Al_COO, fr_furan, fr_nitroso, fr_urea, fr_benzene, fr_phos_acid, fr_N_O (фрагментные дескрипторы), VSA_EState10 (гибридный электро-топологический дескриптор, учитывающий площадь поверхности Ван-дер-Ваальса), PEOE_VSA14 (гибридный дескриптор частичного заряда, учитывающий площадь поверхности Ван-дер-Ваальса), BalabanJ (топологический индекс Балабана), WPOL (топологический индекс Винера), FMF (дескриптор, характеризующий сложность структуры молекулы с помощью "каркаса Мурко"), C1SP2 (топологический дескриптор, характеризующий углеродные связи молекулы в условиях гибридизации), BCUTw_1h (дескриптор собственных значений, основанный на матрице Бердена). Удаленные из модели (коррелирующие) дескрипторы: Chi4v, Chi4n, MolWt, fr_sulfone, TPSA, nAtom, AMR, apol, ECCEN, MW, Zagreb.

Поиск дескрипторов для прогноза индекса преломления. Характеристика модели представлена в табл. 5.

Таблица 5: Характеристика модели прогноза индекса преломления смешанной выборки

Выборка	Стандартная ошибка (n)	Значение R^2	Значение p
AB	0,03	1,00	1,78e-14

Для группы из 20 значимых, не коррелирующих между собой дескрипторов коэффициент детерминации r^2 оказался равным 0,49 при ошибке *RMSE*, равной 0,11. Значения являются приемлемыми.

Соответствующий набор дескрипторов: fr_SH, fr_halogen, fr_furan, fr_nitroso, fr_urea, fr_phos_acid, fr_aniline, fr_N_O (фрагментные дескрипторы), SMR_VSA10, SMR_VSA7 (гибридные дескрипторы молекулярной рефракции и площади поверхности Ван-дер-Ваальса), PEOE_VSA14, PEOE_VSA12, PEOE_VSA11 (гибридные дескрипторы частичного заряда, учитывающие площадь поверхности Ван-дер-Ваальса), EState_VSA10, EState_VSA6 (гибридные электро-топологические дескрипторы, учитывающие площадь поверхности Ван-дер-Ваальса), MolLogP (молекулярная липофильность), BertzCT (топологический графовый дескриптор), Kappa3 (топологический каппа-индекс), ALogp2 (дескриптор квадрата липофильности), BCUTw_1l (дескриптор собственных значений, основанный на матрице Бердена). Удаленные из модели (коррелирующие) дескрипторы: fr_C_O_noCOO, Chi4v, fr_Al_COO, MolWt, fr_benzene, fr_sulfone, TPSA, EState_VSA8, PEOE_VSA13, PEOE_VSA10, BalabanJ, HeavyAtomMolWt, fr_hdrzine, Chi0, Chi1, fr_Al_OH, fr_C_O, NumValenceElectrons, Kappa2, nAtom.

Поиск дескрипторов для прогноза температуры кипения. В данном исследовании поиск дескрипторов проводился только для выборки олигомеров, так как известных значений температуры кипения для выборки дендримеров оказалось недостаточно для качественного прогноза. Группа G является объединением значимых дескрипторов библиотеки RDKit и пакета Rcpi. Характеристика модели представлена в табл. 6.

Таблица 6: Характеристика модели прогноза температуры кипения выборки олигомеров

Выборка	Стандартная ошибка (°C)	Значение R^2	Значение p

AB	181,10	0,92	5,66e-05

Для группы из 20 значимых, не коррелирующих между собой дескрипторов коэффициент детерминации r^2 оказался равным 0,65 при ошибке *RMSE*, равной 158,59. Значения являются приемлемыми, так как среднеквадратическое отклонение исследуемой выборки олигомеров было велико: 368,75 °C.

Соответствующий набор дескрипторов: fr_C_O_noCOO, fr_SH, fr_halogen, fr_Al_COO, fr_furan, fr_urea, fr_benzene, fr_phos_acid, fr_aniline, fr_N_O (фрагментные дескрипторы), Chi4v, Chi4n (индексы связности), SMR_VSA7, SMR_VSA10 (гибридные дескрипторы молекулярной рефракции и площади поверхности Ван-дер-Ваальса), VSA_EState10, EState_VSA8 (гибридные электро-топологические дескрипторы, учитывающие площадь поверхности Ван-дер-Ваальса), PEOE_VSA14 (гибридный дескриптор частичного заряда, учитывающий площадь поверхности Ван-дер-Ваальса), BalabanJ (топологический индекс Балабана), TPSA (топологическая площадь полярной поверхности молекулы), MolWt (молекулярный вес).

Физико-химические свойства веществ как молекулярные дескрипторы. В качестве молекулярных физико-химических дескрипторов были рассмотрены следующие свойства веществ: плотность, индекс преломления и температура кипения.

Прогноз выполнялся для каждого из представленных свойств с помощью модели, использующей в качестве дескрипторов другие два свойства. Таким образом было получено прогнозирование плотности и индекса преломления для дендримеров и олигомеров (AB), а также температуры кипения для выборки олигомеров (A).

Для прогноза плотности коэффициент детерминации r^2 оказался равным 0,77 при ошибке *RMSE*, равной 0,14. Для прогноза индекса преломления r^2 был равен 0,78 при ошибке *RMSE*, равной 0,07. Соответствующая модель для прогноза температуры кипения оказалась неприемлемой, так как коэффициент детерминации r^2 оказался равен всего 0,04.

Для выборки AB представлены следующие графики: зависимость предсказанных значений от экспериментальных данных для плотности - на рис. 1, для индекса преломления - на рис. 2.

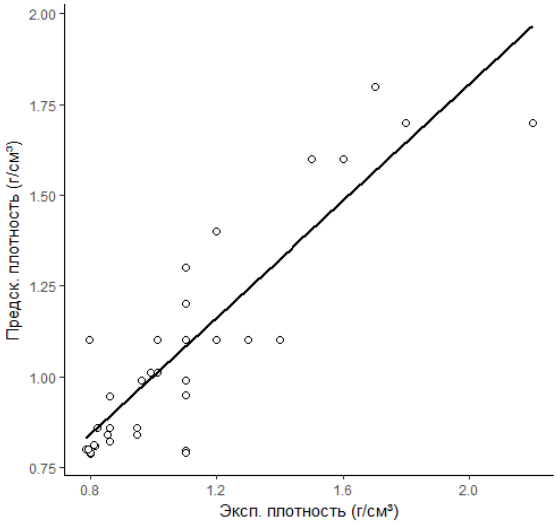

Рисунок 1. *График зависимости для плотности*

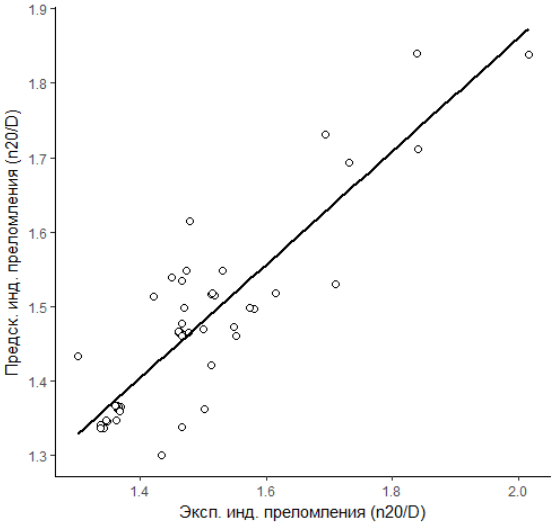

Рисунок 2. *График зависимости для индекса преломления*

В результате можно сделать вывод, что плотность дендримеров и олигомеров можно прогнозировать по индексу преломления и температуре кипения, а индекс преломления - по температуре кипения и плотности. Однако температуру кипения олигомеров лучше предсказывать с помощью химических дескрипторов библиотек Rcpi и RDKit.

Заключение. В результате исследования были найдены молекулярные дескрипторы для прогнозирования плотности и индекса преломления смешанной выборки из олигомеров и дендримеров, а также температуры кипения выборки олигомеров. Соответствующие результаты прогнозов представлены в табл. 7 и 8. По данным таблиц можно видеть, что большая точность прогноза достигается при использовании в качестве дескрипторов известных физико-химических свойств веществ. Однако стоит учитывать, что данные об этих свойствах не всегда доступны в открытых базах данных. Таким образом, наибольший интерес представляют дескрипторы библиотеки RDKit и пакета Rcpi.

Таблица 7: Результаты прогнозирования при использовании дескрипторов RDKit и Rcpi

Прогнозируемое свойство	Уравнение прямой	*RMSE*	Значение r^2

Плотность (г/см^3)	y = 0,60x + 0,41	0,19	0,58
Индекс преломления (n)	y = 0,62x + 0,56	0,11	0,49
Температура кипения (°C)	y = 0,59x + 243	158,59	0,65

Таблица 8: Результаты прогнозирования при использовании физико-химических свойств веществ как дескрипторов

Прогнозируемое свойство	Уравнение прямой	*RMSE*	Значение r^2
Плотность (г/см^3)	y = 0,80x + 0,20	0,14	0,77
Индекс преломления (n)	y = 0,76x + 0,34	0,07	0,78

Для выборки AB представлены графики зависимости предсказанных значений от экспериментальных данных при использовании дескрипторов RDKit и Rcpi в модели. График зависимости для плотности представлен на рис. 3, для индекса преломления - на рис. 4. График зависимости для температуры кипения выборки A (олигомеров) - на рис. 5.

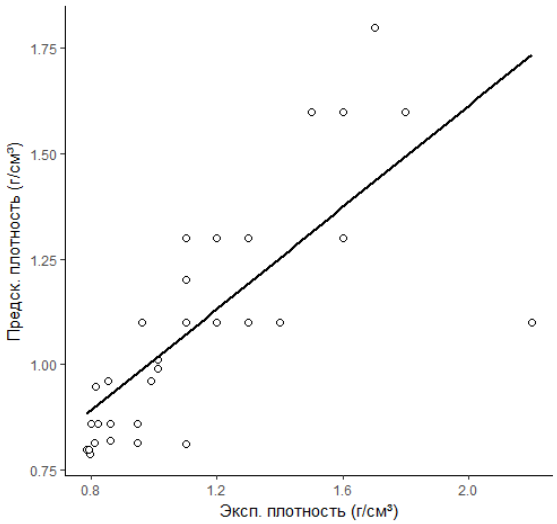

Рисунок 3. *График зависимости для плотности*

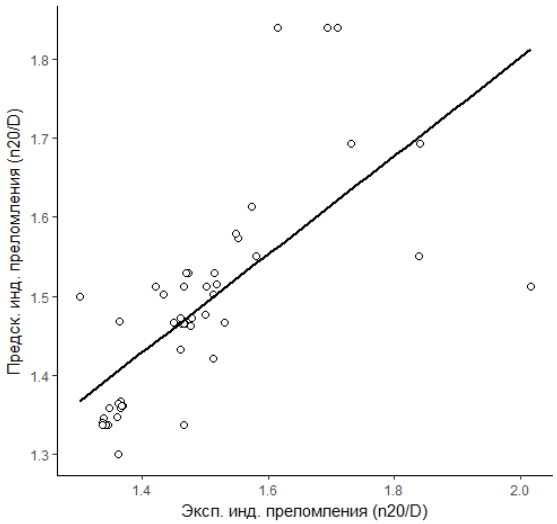

Рисунок 4. *График зависимости для индекса преломления*

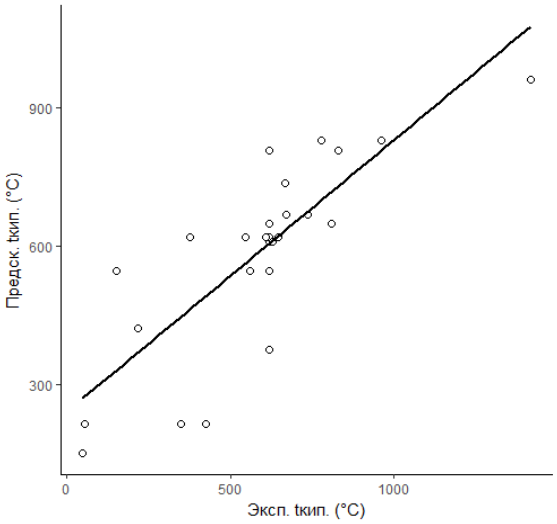

Рисунок 5. *График зависимости для температуры кипения*

Следует отметить, что прогнозирование свойств химических веществ с помощью моделей, не ограниченных группой или классом соединений, не дает удовлетворительных результатов [27]. В данном исследовании (при использовании RDKit и Rcpi дескрипторов) были получены значения коэффициентов детерминации 0,58 (для плотности) и 0,49 (для индекса преломления) с ошибками 0,19 и 0,11 соответственно; значение коэффициента детерминации 0,65 (для кипения) с ошибкой 158,59. С учетом того, что прогнозирование было проведено для смешанной выборки, содержащий вещества схожей структуры, но разных классов - результаты можно считать удовлетворительными. Стоит также учитывать то, что размеры рассматриваемых выборок были невелики - что усложняло задачу получения точного прогноза для свойств данных выборок.

ЛИТЕРАТУРА

1. Применение дендримеров [Электронный ресурс]. URL: http://thesaurus.rusnano.com/wiki/article763
2. Применение дендримеров в медицине [Электронный ресурс]. URL: http://www.iq-coaching.ru/vysokie-tehnologii/nanotehnologii/544.html
3. Ana Sousa-Herves, Ramon Novoa-Carballal, Ricardo Riguera, Eduardo Fernandez-Megia, GATS Dendrimers and PEGylated Block Copolymers: from Synthesis to Bioapplicationa. – Arlington: The AAPS Journal. – 2014. – №5

4. Олигомеры [Электронный ресурс]. URL: https://goldbook.iupac.org/html/O/O04286.html
5. QSAR/QSPR [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/QSAR
6. О. А. Раевский, Дескрипторы молекулярной структуры в компьютерном дизайне биологически активных веществ. – М.: Успехи химии. – 1999
7. Дескрипторы QSAR [Электронный ресурс]. URL: https://www.chemcomp.com/journal/descr.htm#Energy
8. Топологические дескрипторы [Электронный ресурс]. URL: http://www.scbdd.com/chemdes/list2/
9. D. W. Van Krevelen, P. J. Hoftyzer, Prediction of Polymer Densities. – Wiley: Journal of applied polymer science. – 1969. - с. 871
10. Afzal, Mohammad Atif Faiz; Cheng, Chong; Hachmann, Johannes, Accurate prediction of the refractive index of polymers using first principles and data modeling. – APS March Meeting. – 2016
11. Температура кипения [Электронный ресурс]. URL: https://www.pirika.com/ENG/TCPE/BP-Theory.html
12. База данных Sigma-Aldrich [Электронный ресурс]. URL: http://www.sigmaaldrich.com/materials-science/nanomaterials/dendrimers.html
13. База данных ChemSpider [Электронный ресурс]. URL: http://www.chemspider.com/
14. Библиотека RDKit [Электронный ресурс]. URL: http://www.rdkit.org/docs/GettingStartedInPython.html
15. Пакет Rcpi [Электронный ресурс]. URL: https://bioconductor.org/packages/release/bioc/html/Rcpi.html
16. Дескрипторы пакета Rcpi [Электронный ресурс]. URL: https://nanx.me/papers/Rcpi.pdf
17. Роберт И. Кабаков, R в действии. Анализ и визуализация данных в языке R / пер. с англ. Полины А. Волковой. – М: ДМК Пресс. – 2014. – с. 287
18. Алгоритм пошаговой регрессии в языке R [Электронный ресурс]. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html
19. Библиотека sklearn [Электронный ресурс]. URL: http://scikit-learn.org/stable/
20. Дерево решений [Электронный ресурс]. URL: http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
21. Дендримеры [Электронный ресурс]. URL: http://www.pereplet.ru/obrazovanie/stsoros/683.html

22. PAMAM-дендримеры [Электронный ресурс]. URL: http://www.dendritech.com/pamam.html
23. PPI-дендримеры [Электронный ресурс]. URL: http://www.symo-chem.nl/dendrimer.htm
24. Уровень значимости статистического теста [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/index.php?title=Уровень_значимости
25. Критерий Стьюдента [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/index.php?title=Критерий_Стьюдента
26. Роберт И. Кабаков, R в действии. Анализ и визуализация данных в языке R / пер. с англ. Полины А. Волковой. – М: ДМК Пресс. – 2014. – с. 261
27. Оценка температуры плавления низкомолекулярных органических соединений [Электронный ресурс]. URL: http://sntbul.bmstu.ru/doc/851873.html