Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Казанский национальный исследовательский технологический университет»
(ФГБОУ ВО «КНИТУ»)

На правах рукописи

ШАДРИНА ГУЗЕЛЬ РУСЛАНОВНА

АНАЛИЗ СВЯЗИ «СТРУКТУРА – ТЕМПЕРАТУРА СТЕКЛОВАНИЯ ОРГАНИЧЕСКИХ ГОМОПОЛИМЕРОВ» В РАМКАХ ТЕОРИИ ХИМИЧЕСКОГО СТРОЕНИЯ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ И ТЕОРИЙ СТЕКЛОВАНИЯ ПОЛИМЕРОВ

1.4.3. Органическая химия

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор химических наук,
профессор Н.В. Улитин

Казань-2026

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность и степень разработанности темы исследования

В настоящее время среди химических объектов повышенный интерес в аспекте моделирования «структура-свойство» методами машинного обучения[1] вызывают полимеры. Причина этому заключается в многообразии полимеров, огромной и сложно прогнозируемой вариативности их свойств и практической востребованности полимеров. Свойства полимеров зависят от большого числа факторов: химической структуры мономеров, молекулярно-массовых характеристик полимеров, разветвленности и стереорегулярности макромолекул, степени кристалличности и др. Эти зависимости являются нелинейными и часто сложно интерпретируемыми. Небольшое изменение в химической структуре мономера способно принципиально изменить свойства полимера. А даже малое улучшение свойств полимеров или снижение стоимости мономеров имеет большое коммерческое значение, поскольку полимеры являются сырьем для более сложных материалов (пластмасс, композитов, эластомеров и др.). Исследователю сложно, а зачастую невозможно, уловить такие зависимости в многомерном пространстве данных, в то время как модели «структура-свойство», построенные на основе методов машинного обучения, идеально подходят для этого. Кроме того, следует отметить, что синтез нового полимера, его выделение, очистка и определение комплекса его свойств могут требовать существенных финансовых затрат и занимать длительное время. Модели «структура-свойство», построенные на основе методов машинного обучения и обученные на существующих данных, позволяют значительно ускорить разработку новых полимеров и оптимизировать их свойства под необходимые эксплуатационные требования.

Гомополимеры – это полимеры, макромолекулы которых состоят из одинаковых повторяющихся звеньев. Исследовательский интерес к гомополимерам в аспекте моделирования «структура-свойство» методами

[1] Методы машинного обучения – это математические алгоритмы, общей идеей которых является построение статистической модели за счет ее так называемого обучения на большом наборе данных (обучающей выборке), причем модель в дальнейшем должна быть способна прогнозировать требуемые данные для новых объектов (объектов тестовой выборки).

машинного обучения обусловлен тем, что гомополимеры достаточно просты, чтобы моделировать связь «структура-свойство» с помощью ограниченного набора данных (дескрипторы описывают повторяющееся звено одного типа), но и одновременно достаточно сложны для интерпретации проявления свойств в зависимости от химического строения их повторяющихся звеньев и их многоуровневой структуры. В этой связи гомополимеры являются идеальными объектами для выявления базовых принципов, понимание которых необходимо в дальнейшем при установлении и анализе связи «структура-свойство» для более сложных объектов (сополимеров, смесей полимеров, композитов и др.). Большинство крупнотоннажных полимеров – это именно гомополимеры, причем органические: полиэтилен, полипропилен, поливинилхлорид, полистирол. Следует отметить, что органические гомополимеры представляют собой наиболее широкий класс полимеров. Оптимизация их свойств – это актуальнейшая задача органической химии, где модели «структура-свойство», построенные на основе методов машинного обучения, уже дают оптимистичные результаты. Важнейшим свойством, на величину которого опираются при подборе полимера под требуемые условия эксплуатации, является температура стеклования – температурная граница между их стеклообразным (упруготвердым) и высокоэластическим физическими состояниями. При температурах ниже температуры стеклования полимеры используются как конструкционные материалы, при температурах выше температуры стеклования, но ниже температуры текучести полимеры применяются как эластичные материалы. То есть, другими словами, температура стеклования полимеров является предельной температурой, до которой не проявляется ползучесть.

Анализ обзора работ по моделированию «структура – температура стеклования органических полимеров» методами машинного обучения показал, что все они носят чисто методический характер, то есть посвящены поиску наилучших параметров структуры, оптимального метода машинного обучения с точки зрения повышения прогностической способности модели, и совершенно не анализируют полученные результаты в рамках теории химического строения органических соединений и существующих теорий стеклования полимеров. Это

делает проведение такого исследования **актуальным**, и оно может осуществляться с привлечением инкрементального подхода к расчету свойств полимеров исходя из строения их повторяющихся звеньев – подхода, основанного на идее о том, что свойства полимера определяются суммой инкрементов атомов повторяющегося звена и инкрементов межмолекулярных взаимодействий.

В связи с чем **целью настоящей работы** стало построение модели машинного обучения, описывающей связь «структура – температура стеклования органических гомополимеров» и содержащей параметры, которые интерпретируются в рамках теории химического строения органических соединений и теорий стеклования полимеров. Для достижения поставленной цели в работе решались следующие **задачи**:

1) рассмотрение различных методов машинного обучения и выбор среди них обеспечивающего наибольшую достоверность модели, способной прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее на основе строения их повторяющихся звеньев по аналогии с инкрементальным подходом;

2) уточнение значений параметров, определяющих температуру стеклования органических гомополимеров на основе строения их повторяющихся звеньев по аналогии с инкрементальным подходом, за счет применения при моделировании на основе машинного обучения комбинированных дескрипторов[2] для обеспечения более достоверных прогнозов;

3) установление корреляций параметров, определяющих температуру стеклования органических гомополимеров на основе строения их повторяющихся звеньев по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям полимеров; анализ установленных корреляций в рамках теории химического строения органических соединений и теорий стеклования полимеров.

[2] Под дескрипторами понимаются характеристики химического объекта, которые представляют его структуру в числовой форме, понятной алгоритмам машинного обучения.

Научная новизна работы

I (п. 7 паспорта специальности 1.4.3.). На основе молекулярных отпечатков Моргана как характеристик молекулярной структуры повторяющихся звеньев полимеров и метода случайного леса как метода машинного обучения построена модель, способная прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее по аналогии с инкрементальным подходом.

II (п. 4 и 7 паспорта специальности 1.4.3.). На примере полистиролов с различными положениями (2-, 3-, 4-) заместителя (фтор-, хлор-, бром-, метил- и этил-) в ароматическом кольце показано, что учет химического строения повторяющихся звеньев органических гомополимеров (в данном случае учет положения заместителя в ароматическом кольце) является определяющим фактором повышения точности прогнозирования температуры стеклования органических гомополимеров: модели, построенные без учета и с учетом положения заместителя, имеют коэффициенты детерминации 0.12 и 0.81 соответственно.

III (п. 7 паспорта специальности 1.4.3.). По результатам моделирования на количественном уровне показано, что температура стеклования органических гомополимеров:

- прямо пропорциональна параметру, связанному с молекулярным объемом повторяющегося звена полимера и интерпретируемому как мера гибкости макромолекул (эта закономерность согласуется с термодинамическими теориями стеклования полимеров и кинетическими теориями М.В. Волькенштейна-О.Б. Птицына и Ю.Я. Готлиба-О.Б. Птицына);

- обратно пропорциональна сумме двух параметров, один из которых характеризует совокупность всех типов межмолекулярных взаимодействий (эта закономерность согласуется с теорией межмолекулярных связей С.Н. Журкова и флуктуационной теорией стеклования полимеров), а второй – долю свободного объема (эта закономерность согласуется с теорией свободного объема).

IV (п. 4 и 7 паспорта специальности 1.4.3.). Показано, что электронные свойства повторяющихся звеньев органических гомополимеров (средняя

поляризуемость, дипольный момент, потенциал ионизации, сродство к электрону, энергетический зазор между высшей занятой молекулярной орбиталью и низшей свободной молекулярной орбиталью, химический потенциал, химическая жесткость, электрофильность) обладают статистически значимой корреляцией с параметром, который в связи «структура – температура стеклования органических гомополимеров» характеризует совокупность всех типов межмолекулярных взаимодействий, включая диполь-дипольные (слабые взаимодействия), водородные связи (сильные взаимодействия) и электростатические (сильные взаимодействия).

Теоретическая и практическая значимость работы

Теоретическая значимость работы заключается в том, что построенная модель «структура – температура стеклования органических гомополимеров» позволяет получать результаты, которые могут быть проанализированы в рамках теории химического строения органических соединений и теорий стеклования полимеров.

Практическая значимость работы заключается в том, что: 1) разработано программное обеспечение для прогнозирования температуры стеклования органических гомополимеров; 2) модель может применяться в качестве прогностического модуля при технологическом моделировании промышленных процессов синтеза органических гомополимеров.

Методология и методы исследования

Исследование провели по следующей методологии: 1) формирование базы данных для модели «структура – температура стеклования органических гомополимеров», построенной на основе машинного обучения; 2) выбор дескрипторов для описания химических структурных формул повторяющихся звеньев органических гомополимеров; 3) выбор метода машинного обучения для моделирования связи «структура – температура стеклования органических гомополимеров»; 4) первый этап моделирования связи «структура – температура стеклования органических гомополимеров» на основе машинного обучения: проверка возможности прогнозирования температуры стеклования органических гомополимеров с помощью модели, построенной на основе методов машинного

обучения, не напрямую, а через параметры, определяющие температуру стеклования на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом; 5) второй этап моделирования связи «структура – температура стеклования органических гомополимеров» на основе машинного обучения: уточнение значений параметров, определяющих температуру стеклования органических гомополимеров на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом, за счет применения комбинированных дескрипторов для обеспечения более достоверных прогнозов; 6) установление корреляций уточненных параметров, определяющих температуру стеклования органических гомополимеров на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям полимеров; 7) анализ установленных корреляций в рамках теории химического строения органических соединений и теорий стеклования полимеров.

Моделирование на основе машинного обучения реализовывали на языке программирования Python 3.13.7. При моделировании рассмотрели 3 метода машинного обучения: метод случайного леса, метод k ближайших соседей и многослойный перцептрон. Для описания химических структурных формул повторяющихся звеньев органических гомополимеров выбрали и протестировали структурные ключи и молекулярные отпечатки Моргана. Показали, что модель на основе метода случайного леса, использующая молекулярные отпечатки Моргана, обладает наибольшей точностью прогнозирования температуры стеклования органических гомополимеров. Для метода случайного леса, продемонстрировавшего максимальную точность прогнозов среди испытанных, провели подбор гиперпараметров с помощью алгоритма GridSearchCV из библиотеки scikit-learn. Комбинированные дескрипторы получали, объединяя информацию о химическом строении повторяющихся звеньев органических гомополимеров и их экспериментальных и рассчитанных в рамках инкрементального подхода значениях температуры стеклования.

В качестве объектов исследования на этапе установления корреляций уточненных параметров, определяющих температуры стеклования органических гомополимеров на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям полимеров, выбрали полистиролы с различными положениями (2-, 3-, 4-) заместителя (фтор-, хлор-, бром-, метил- и этил-) в ароматическом кольце. В качестве моделей молекулярной структуры объектов исследования в квантово-химических расчетах использовали повторяющиеся звенья, в которых открытую валентность концевых групп закрывали атомами водорода. Оптимизацию структур объектов исследования выполняли в программном пакете Gaussian 16, Rev. C.01 с использованием гибридного функционала B3LYP и валентно-расщепленного базисного набора Попла 6-31G(d,p). Стабильность рассчитанных структур характеризовалась отсутствием отрицательных частот колебаний в матрице вторых производных. Рассчитывали следующие квантово-химические параметры: средняя поляризуемость, дипольный момент, потенциал ионизации, сродство к электрону, энергетический зазор между высшей занятой молекулярной орбиталью и низшей свободной молекулярной орбиталью, химический потенциал, химическая жесткость, электрофильность, молекулярный объем. Степень корреляции параметров оценивали с помощью коэффициента Пирсона.

Положения, выносимые на защиту

1. Методология построения модели машинного обучения, способной прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом.

2. Закономерности, выявленные при анализе связи «структура – температура стеклования органических гомополимеров», построенной на основе машинного обучения, в рамках теории химического строения органических соединений и теорий стеклования полимеров.

Достоверность результатов работы и обоснованность положений, выносимых на защиту, обеспечивается корректностью используемых методов

моделирования, верификацией модели по обширной базе экспериментальных данных, согласованием результатов моделирования с положениями общепринятых теорий.

Личный вклад автора заключается в сборе и анализе литературных данных, реализации решения задач исследования, анализе результатов, формулировании заключения и участии в написании и подготовке публикаций. Работа выполнена на кафедре общей химической технологии ФГБОУ ВО «КНИТУ».

Соответствие специальности

Диссертация соответствует п. 4. Развитие теории химического строения органических соединений и п. 7. Выявление закономерностей типа «структура – свойство» паспорта специальности 1.4.3. Органическая химия.

Апробация результатов работы

Результаты обсуждались на XXXIV Российской молодежной научной конференции с международным участием, посвященной 190-летию со дня рождения Д.И. Менделеева «Проблемы теоретической и экспериментальной химии» (Екатеринбург, 2024), IX Всероссийской научной конференции «Теоретические и экспериментальные исследования процессов синтеза, модификации и переработки полимеров» (Уфа, 2024), IV Всероссийской научной конференции (с международным участием) преподавателей и студентов вузов «Актуальные проблемы науки о полимерах» (Казань, 2024), Международной научной конференции «Актуальные вопросы естествознания и функциональные полимеры для фармацевтики, нефтяной промышленности, экологии, био- и нанотехнологии», посвященной 125-летию профессора К. Жубанова (Казахстан, г. Актобе, 2024).

Публикации

Результаты работы представлены в 2 статьях в рецензируемых изданиях, рекомендованных ВАК для размещения материалов диссертаций, и 5 публикациях в сборниках материалов конференций.

Структура и объем работы

Диссертация изложена на 150 страницах, содержит 18 рисунков и 14 таблиц (13 – в основной части, 1 – в приложении), состоит из введения, 3 глав, заключения, списка литературы, насчитывающего 147 наименований, 2 приложений.

ГЛАВА 1 ЛИТЕРАТУРНО-АНАЛИТИЧЕСКИЙ ОБЗОР

1.1 История развития исследований «структура-свойство» в химии

Первоначальные исследования взаимосвязей между структурами химических соединений и их свойствами основывались на систематизации экспериментальных данных [1-4]. Однако такой подход не обеспечивал строгой количественной интерпретации и не позволял точно прогнозировать свойства новых химических соединений [1-4]. С развитием данной области исследований стала очевидной необходимость создания моделей, способных объяснять и прогнозировать свойства химических соединений на основе их структуры [5-9].

Первым и наиболее ярким прообразом количественной взаимосвязи «структура-свойство» в химии можно считать Периодический закон Д.И. Менделеева (1869 г.) [10]. Д.И. Менделеев показал, что свойства химических элементов связаны со строением их атомов [10]. Это позволило не только описать свойства известных химических элементов, но и спрогнозировать существование и свойства неизвестных на тот момент химических элементов [10].

В XIX веке ученые начали накапливать данные по эмпирическим закономерностям [2-4, 11], в частности:

1868 г. – А. Крам Браун и Т.Р. Фрейзер установили связь между химическим составом солей аммониевых оснований и их физиологическим действием [2];

1869-1870 гг. – Б. Ричардсон и А. Рабуто независимо друг от друга выявили, что токсичность одноатомных спиртов и их растворимость в воде уменьшаются с ростом длины алифатической цепи [11];

1893 г. – Ш. Рише связал токсичность некоторых летучих веществ с их растворимостью в воде [3];

1899 г. – Г. Мейер показал, что анестезирующее действие веществ тем больше, чем больше их растворимость в жирах (эта закономерность нашла применение в области наркоза, поскольку мембраны нервных клеток представляют собой жироподобные вещества) [4].

Полученные закономерности не были описаны в виде математических зависимостей [2-4, 11]. Проблема заключалась в том, что исследователи не знали, как представить структуры химических соединений в виде чисел [2-4, 11]. Одним из первых примеров успешно реализованной в виде математической зависимости связи «структура-свойство» в химии является уравнение Л. Гамметта (1937 г.) [1]:

$$\lg(k/k_0) = \rho\sigma,$$

где k – константа скорости реакции с участием замещенного в бензольном кольце соединения; k_0 – константа скорости реакции с участием незамещенного в бензольном кольце соединения; ρ – константа реакционной серии, которая отражает степень восприимчивости реакции к изменению заместителя в реагенте и не зависит от свойств заместителей; σ – константа заместителя, которая отражает свойства самого заместителя и не зависит от реакционной серии (константа заместителя выражается как десятичный логарифм отношения констант кислотности замещенной и незамещенной бензойной кислоты, причем производные бензойной кислоты взяты в качестве стандартного ряда соединений, для которого $\rho = 1$).

Значительный скачок применения моделей «структура-свойство» в химии произошел в 1970-х гг. с появлением методов машинного обучения [12-14]. Методы машинного обучения – это математические алгоритмы, общей идеей которых является построение статистической модели за счет ее так называемого обучения на большом наборе данных (обучающей выборке), причем модель в дальнейшем должна быть способна прогнозировать требуемые данные для новых объектов (объектов тестовой выборки) [8]. Достоинство методов машинного обучения заключается в способности строить нелинейные и многомерные зависимости [8]. Среди важнейших достижений 1970-х гг. – привлечение перцептронов (от англ. perceptron – простейшая модель искусственной нейронной сети) для прогнозирования биологической активности химических соединений [12, 13]. Параллельно развивались методы подструктурного анализа (substructural analysis), которые позволяют выделять фрагменты молекул, критичные для

проявления молекулами определенных свойств, и использовать их для планирования синтеза новых химических соединений [14].

В настоящее время модели «структура-свойство» на основе методов машинного обучения активно применяются в фармацевтической химии, экологии, материаловедении и химической технологии [15-29]. Следует отметить, что в литературе [15-29] за областью моделирования «структура-свойство» на основе методов машинного обучения закрепились аббревиатуры SAR, QSAR и QSPR. Если исследуется принципиальное проявление химическими соединениями биологической активности (проявляет, не проявляет), то для обозначения исследования используется аббревиатура SAR (от англ. structure-activity relationships). Если биологическая активность химических соединений исследуется на количественном уровне, то – аббревиатура QSAR (от англ. quantitative structure-activity relationships). Если речь в исследовании идет о других свойствах химических соединений, то – QSPR (от англ. quantitative structure-property relationships).

1.2 Этапы построения моделей «структура-свойство» на основе методов машинного обучения

Построение моделей «структура-свойство» на основе методов машинного обучения состоит из следующих этапов [17, 30-32]:

1) формирование и подготовка базы данных;

2) выбор дескрипторов – характеристик химического объекта (химического соединения, химической реакции, смесей химических соединений, растворов, материалов и др.), которые представляют его структуру в числовой форме, понятной алгоритмам машинного обучения;

3) выбор метода машинного обучения;

4) оценка качества модели.

1.2.1 Формирование и подготовка базы данных

Основой любой модели «структура-свойство», строящейся на основе методов машинного обучения, является база данных [17, 30-32]. К данным, вошедшим в базу, предъявляются следующие требования [17, 30-32]:

1) достоверность – данные должны быть достоверными; некорректные данные могут привести к построению моделей с низкой прогностической способностью;

2) репрезентативность – данные должны отражать основные характеристики всей совокупности объектов исследования;

3) сбалансированность – данные в базе должны быть равномерно распределены как с точки зрения прогнозируемых свойств, так и с точки зрения рассматриваемых химических объектов в химическом пространстве (химическое пространство – вся совокупность химических объектов, представленных в виде векторов, или соответствующих им координат, и расположенных в пределах многомерной системы координат [8, 16]).

Подготовка базы данных для моделирования заключается в так называемых химической и математической обработках данных (подготовку базы данных для моделирования в литературе часто называют предобработкой данных) [17, 30-32]. Химическая обработка данных сосредоточена на химических объектах и включает в себя удаление дубликатов, исключение объектов, которые могут вызвать трудности с точки зрения представления в числовом виде или нерепрезентативно отражают совокупность объектов исследования, и приведение химических объектов к единой форме (нормализация специфических хемотипов, резонансных форм и таутомеров) [17, 30-32]. Математическая обработка сосредоточена на числовых данных и включает в себя удаление неполных и недостоверных данных, приведение значений к единым единицам измерения, проверку качества данных в аспекте их нормального распределения [17, 30-33].

Следует отметить, что, как правило, модель «структура-свойство», построенная на основе методов машинного обучения, хорошо прогнозирует

свойство в пределах границ значений, представленных в базе, и для типов химических объектов, представленных в базе (все это относится к области применимости модели) [8, 34, 35].

1.2.2 Выбор дескрипторов

Дескрипторы – это характеристики химического объекта, которые представляют его структуру в числовой форме, понятной алгоритмам машинного обучения [8, 16]. Абстрактно дескриптор можно представлять как вектор, описывающий координаты химического объекта в многомерном химическом пространстве [32, 36-39]. Основные критерии выбора дескрипторов следующие [37-39]:

1) возможность интерпретации (структурно-химической или физической);

2) отсутствие корреляции с другими дескрипторами;

3) наличие хорошей корреляции с интересующим свойством;

4) область применимости не должна быть слишком узкой.

Существует несколько классификаций дескрипторов [7-9, 39-41]:

1) по типу химического объекта – молекулярные дескрипторы (вычисляются из структуры молекулы, а значит, характеризуют структуру молекулы), дескрипторы ансамбля химических объектов (смесей, растворов, материалов), дескрипторы химических реакций;

2) по происхождению – вычисляемые и экспериментальные;

3) по типу числового значения – бинарные (принимают значения 0 и 1), целочисленные положительные (принимают строго целочисленные положительные значения), вещественные (принимают любые числовые значения);

4) по локальности – глобальные (относятся ко всему химическому объекту), локальные (относятся к части химического объекта), точечные (относятся к точкам физического пространства);

5) по относительности – абсолютные (относятся только к определенному химическому объекту и не зависят от других объектов) и относительные

(рассматривают отношения рассматриваемого химического объекта с другими химическими объектами);

6) по методу вычисления (для этой классификации также используется название «по функциональности») – физико-химические, фрагментные, квантово-химические, топологические и др. (табл. 1).

Таблица 1

Описание и примеры дескрипторов по методу вычисления (по функциональности) [42, 43]

Наименование	Описание	Примеры
1	2	3
Физико-химические	Характеризуют физико-химические свойства химического объекта.	Молекулярная масса, молекулярный объем, липофильность, молярная рефракция.
Топологические (иногда называются топологическими индексами)	Описывают структуру химического объекта в виде молекулярного графа (двумерное представление химического объекта, в котором вершинами графа являются атомы, а ребрами – связи между ними). Характеризуют топологические параметры, такие как кратность связей, связь атомов между собой, наличие гетероатомов и т.п.	Дескрипторы, основанные на матрице смежности, дескрипторы, основанные на матрице расстояний.
Фрагментные	Характеризуют наличие тех или иных фрагментов в рассматриваемом химическом объекте.	Структурные ключи, молекулярные отпечатки, константы заместителей.
Квантово-химические	Характеризуют квантово-химические свойства химического объекта.	Энергии молекулярных орбиталей, дипольный момент, потенциал ионизации, электростатическое поле, создаваемое молекулой.
Дескрипторы молекулярного подобия	Характеризуют степень подобия (сходства) рассматриваемых химических объектов (или степень близости объектов в многомерном химическом пространстве).	Максимальная общая подструктура, дескрипторы меры квантового молекулярного подобия.

Окончание табл.1

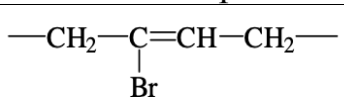
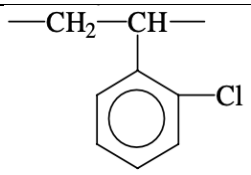
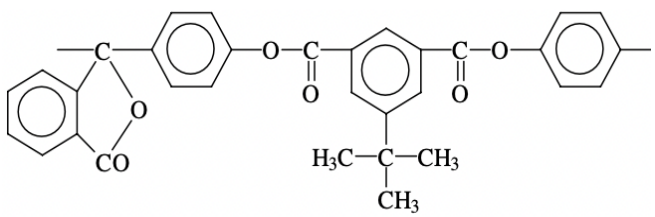
1	2	3
Дескрипторы размерности представления молекул	Характеризуют расположение атомов в пространстве.	0D – вычисляются на основе элементного состава молекул (число атомов, молекулярная масса); 1D – вычисляются из рациональной формулы или рациональной формулы в полуразвернутом виде (число функциональных групп); 2D – вычисляются, исходя из представления химических объектов в виде графов; 3D – вычисляются, исходя из представления химических объектов в виде пространственной структуры (расстояния между атомами, форма молекул); 4D – вычисляются, исходя из рассмотрения химических объектов в виде ансамбля пространственных конформеров (четвертым измерением является время).

Один и тот же дескриптор может входить в разные классификации [43]. В процессе моделирования тестируют разные типы дескрипторов с целью выбора тех, что обеспечат максимальную точность описания данных в базе [43]. Для расчета дескрипторов молекулы кодируют. Существуют различные варианты кодирования молекул: международный химический идентификатор ИЮПАК InChI (от англ. INternational CHemical Identifier); SMILES (от англ. Simplified Molecular Input Line Entry System) – система представления молекул в виде одномерной строки символов; 2D/3D координаты и др., – среди которых наибольшее распространение получила система SMILES [43, 44]. Согласно SMILES, молекула представляется в виде строки, в которую атомы записываются как символы соответствующих химических элементов [43, 44]. Одинарные связи между атомами не указываются, двойные связи указываются в виде символа =, тройные – в виде символа # [43, 44]. Боковые ответвления заключаются в круглые

скобки [43, 44]. Чтобы обозначить циклические структуры, атомы, между которыми условно разрывается цикл, нумеруются [43, 44]. Для обозначения начала и конца повторяющегося звена в полимерах используется комбинация символов [*] (табл. 2).

Таблица 2

Примеры представления повторяющегося звена полимера в виде SMILES [43, 44]

Структура повторяющегося звена полимера	SMILES
	[*]CC(Br)=CC[*]
	[*]CC(C1=CC=CC=C1Cl)[*]
	[*]C(C1=CC=CC=C1C2=O)(O2)C3=CC=C(OC(C4=CC(C(C)(C)C)=CC(C(OC5=CC=C([*])C=C5)=O)=C4)=O)C=C3

1.2.3 Методы машинного обучения, наиболее распространенные в химии для построения моделей «структура-свойство»

Построение моделей «структура-свойство» на основе методов машинного обучения осуществляется в определении такой статистической модели *F*, которая как можно более точнее свяжет значение свойства *Y* со значениями дескрипторов, описывающих структуру рассматриваемых химических объектов, и обеспечит наилучшую прогнозирующую способность в применении к другим (в идеальном случае – произвольным) химическим объектам, то есть минимальное расхождение между спрогнозированными с ее помощью значениями свойства и его экспериментально полученными значениями:

$$Y = F(X).$$

Конкретный вид функции F и способ ее определения зависит от метода машинного обучения.

Выделяют следующие основные типы задач, которые решаются в химии с помощью методов машинного обучения [8]:

1) регрессионная задача – восстановление зависимости между структурой и свойством химических объектов, принимающим непрерывный набор значений;

2) классификационная задача – разделение данных по классам в зависимости от дискретного категорирования свойств химических объектов (соединение активно или не активно; реакция протекает или не протекает и т.п.); частным случаем классификационной задачи является задача одноклассовой классификации – отнесение тестируемого объекта к целевому классу;

3) задача кластеризации – группировка данных по кластерам – группам объектов, схожих по заданному признаку;

4) задача ранжирования – сортировка объектов по рангам в зависимости от заданных признаков;

5) прогностическая задача – определение свойства для нового химического объекта на основе сведений об известных химических объектах.

Обзор методов машинного обучения, наиболее распространенных в химии для построения моделей «структура-свойство», представлен в табл. 3.

Таблица 3

Методы машинного обучения, наиболее распространенные в химии для построения моделей «структура-свойство»

Метод	Описание метода	Достоинства	Недостатки	Источники
1	2	3	4	5
Множественная линейная регрессия (от англ. multiple linear regression, или MLR)	Метод не используется для классификационных задач. Метод основан на использовании метода наименьших квадратов. Применяется, когда зависимость свойства от значений дескрипторов предполагается линейной: $y = w_0 + w_1x_1 + \ldots + w_nx_n$, где y – свойство; w_0 – свободный член; w_n – коэффициент регрессии для n-го дескриптора; x_n – n-й дескриптор; n – количество дескрипторов.	Использование метода наименьших квадратов позволяет быстро обучить модель даже при большом числе дескрипторов. Коэффициенты модели являются весовыми коэффициентами (множителями), отражающими количественный вклад каждого дескриптора в свойство, что облегчает интерпретацию и оценку значимости дескрипторов.	Метод работоспособен, только если отсутствует сильная корреляция между дескрипторами. Предположение о линейной зависимости свойства от дескрипторов ограничивает применение метода, поскольку в данных могут быть нелинейности, которые метод проигнорирует при обучении модели. В конечном итоге это приведет к пониженной прогностической способности модели.	[45]

Продолжение табл. 3

1	2	3	4	5
Метод k ближайших соседей (от англ. k nearest neighbors, или kNN)	Принцип работы метода заключается в том, что для текущего химического объекта вычисляется расстояние до всех химических объектов обучающей выборки, после чего он относится к тому классу, который наиболее часто встречается среди k ближайших соседей (для классификационной задачи), либо значение его свойства определяется как среднее по свойствам k ближайших соседей (для регрессионной задачи). Наиболее частое назначение метода – прогнозирование свойств, которыми обладают химические объекты, наиболее схожие с рассматриваемым объектом.	Концептуальная простота – единственным параметром является k, значение которого определяет прогностическую способность модели. Легкость интерпретации: например, в случае классификационной задачи – объект А относится к классу 1, так как к классу 1 относится наиболее близкий к объекту А объект В.	При большом количестве дескрипторов эффективность метода ухудшается, так как расстояния между объектами могут стать «неинформативными». Во избежание этого, для обеспечения качественного обучения, необходимо увеличить количество данных в базе. Метод чувствителен к выбросам и шумам, особенно в классификационной задаче, потому что выбросы и шумы могут быть восприняты алгоритмом в качестве ближайших соседей. Непосредственно сам метод не дает информации о значимости дескрипторов, поскольку основывается на	[45, 46]

Продолжение табл. 3

1	2	3	4	5
			поиске похожих друг на друга химических объектов без оценки вкладов дескрипторов в свойство. Для этого приходится использовать внешние методы. Метод чувствителен к несбалансированности классов, потому что он относит химический объект к тому классу, к которому относится большинство его ближайших соседей. Если в обучающей выборке один класс преобладает по количеству объектов, а другой представлен меньшим количеством объектов, то при определении класса для нового объекта его ближайшие соседи, вероятнее всего, будут из класса с преобладающим количеством объектов. В итоге объекты из	

Продолжение табл. 3

1	2	3	4	5
			класса с меньшим количеством объектов могут неправильно классифицироваться как объекты из класса с бóльшим количеством объектов. То есть метод плохо справляется с распознаванием объектов из класса с меньшим количеством объектов.	
Метод опорных векторов (от англ. support vector machine, или SVM)	Метод используется как для классификационных, так и для регрессионных задач. В классификационных задачах метод строит так называемую гиперплоскость, разделяющую объекты разных классов с максимальным зазором между ближайшими объектами разных классов, которые называются опорными векторами. Другими словами, гиперплоскость – это разделяющая разные классы граница (кривая или поверхность), а опорные векторы – это ключевые объекты, принадлежащие разным классам и определяющие позицию границы, которая разделяет эти классы (гиперплоскости). В регрессионных задачах метод строит функцию (кривую или поверхность), которая приблизительно повторяет зависимость между дескрипторами и свойством, допуская небольшую ошибку (допустимую ошибку, или эпсилон-зону). Метод строит функцию максимально гладкой,	Метод может обучаться даже в случае, когда количество дескрипторов велико, а количество данных в базе мало. Максимизация зазора обеспечивает лучшее разделение данных и повышает качество модели. В случае использования ядровых функций метод хорошо подходит для	Метод чувствителен к выбросам и шумам в базе данных (они могут стать опорными векторами и исказить гиперплоскость). При необходимости использования ядровых функций сложно выбрать правильную функцию для конкретной задачи, кроме того, параметры функции нужно тщательно настраивать.	[47]

Продолжение табл. 3

1	2	3	4	5
	проходящей максимально близко к большинству данных, при этом учитываются опорные векторы, то есть важные (ключевые) точки в базе данных. Таким образом, в регрессионных задачах метод использует тот же принцип, что и в классификационных задачах – он использует опорные векторы – ключевые точки, которые сильнее всего влияют на построение функции, затем использующейся для прогнозов. Если спрогнозированное значение свойства отличается от приведенного в базе значения не более, чем на значение эпсилон, то совпадение полагается хорошим и такая ошибка не штрафуется. Если же ошибка выходит за пределы эпсилон, модель старается уменьшить ошибку. Если данные нельзя разделить простой гиперплоскостью (например, если они «переплетены»), метод использует ядровые функции. Ядровая функция – это специальная математическая функция, которая преобразует исходные данные из низкоразмерного пространства так, словно они переносятся в пространство с бóльшим числом измерений, причем в новом пространстве они станут точно разделимыми (в идеале – линейно разделимыми).	сложных, многомерных данных и не переобучается.		

Продолжение табл. 3

1	2	3	4	5
Метод случайного леса (от англ. random forest, или RF)	Метод относится к методам деревьев принятия решений (от англ. decision trees). Деревья состоят из узлов, ветвей и листьев. Узлы представляют собой точки, в которых происходит ветвление дерева, и включают в себя: - корневой узел – отправную точку дерева, из которой выходят все последующие ветви и в которой фиксируется исходная проблема или решение, которое необходимо принять; - узлы решений – точки, из которых растут ветви решений, представляющие собой возможные варианты выбора; - вероятностные узлы – точки, из которых растут вероятностные ветви, представляющие собой события с определенными вероятностями (сумма вероятностей по всем вероятностным ветвям всегда должна быть равна 1). Ветви оканчиваются листьями, которые представляют собой конкретные окончательные результаты определенных путей решений и событий. С помощью деревьев принятия решений в химии решаются как классификационные, так и регрессионные задачи. В регрессионных задачах значение свойства прогнозируется как среднее арифметическое среди значений свойств соединений обучающей выборки, попавших в тот же лист. Обучение деревьев принятия решений называют выращиванием (от англ. growing). Прогнозирующая способность индивидуальных моделей-деревьев принятия решений может быть не очень высокой из-за так называемого	Достоинства индивидуальных моделей-деревьев принятия решений: - наглядность: методы просты для понимания и интерпретации; - вариативность: могут быть просчитаны все возможные варианты. Достоинства метода случайного леса: - высокая точность; - устойчивость к «шумам» и выбросам: «шумы» и выбросы не сильно отразятся на результате прогноза, так как повлияют только на обучение отдельных деревьев; - стойкость к переобучению, так как индивидуальные	Основной недостаток индивидуальных моделей-деревьев принятия решений – низкая точность прогнозов из-за возможного переобучения «громоздких» моделей. Недостатки метода случайного леса: - сложность интерпретации результатов в отличие от результатов, полученных индивидуальной моделью-деревом; - большие вычислительные и временные затраты: обучение и запуск совокупности моделей-деревьев требует больше вычислительной мощности и времени, чем при обучении и запуске одной модели-дерева.	[48]

Продолжение табл. 3

1	2	3	4	5
	переобучения («переподогнанности»), то есть из-за большого количества узлов и листьев в дереве по отношению к количеству данных в обучающей выборке. Во избежание этого недостатка применяются методы, объединяющие по разным принципам прогнозы индивидуальных моделей-деревьев. Именно к таким методам относится метод случайного леса. Этот метод создает множество простых индивидуальных моделей-деревьев. Каждая модель-дерево учится независимо на «своей» подвыборке из общей обучающей выборки и со «своим» набором дескрипторов, выделенных случайно из первоначального набора дескрипторов (набор дескрипторов меняется на каждом шаге выращивания дерева). Концепция метода случайного леса заключается в следующем: простая индивидуальная модель-дерево может ошибиться, но совокупность таких моделей-деревьев, ошибающихся по-разному, дает прогноз всегда точнее, чем индивидуальная модель. В методе случайного леса при решении классификационной задачи окончательный прогноз определяется тем классом, который был спрогнозирован большинством индивидуальных моделей-деревьев, при решении регрессионной задачи окончательный прогноз представляет собой среднее арифметическое прогнозов, даваемых индивидуальными моделями-деревьями.	модели-деревья учатся на «своих» данных и это не дает лесу «вызубрить» одну и ту же информацию; - способность к оценке важности дескрипторов.		

Продолжение табл. 3

1	2	3	4	5
Метод искусственных нейронных сетей (Artificial Neural Networks, или ANN)	Искусственные нейронные сети работают по аналогии с человеческим головным мозгом. Искусственная нейронная сеть состоит из искусственных нейронов (вычислительных единиц, являющихся простейшей математической моделью нейронов головного мозга) и связей между ними, имеющими определенные весовые коэффициенты, или веса связей (на эти коэффициенты умножаются сигналы, идущие от одних нейронов к другим.) Самой распространенной архитектурой нейронной сети является многослойная искусственная нейронная сеть, состоящая из слоев входных, скрытых и выходных нейронов, причем сигналы с каждого нейрона предыдущего слоя идут на каждый нейрон следующего слоя, связи между нейронами внутри слоев отсутствуют и отсутствуют прямые связи между не соседними слоями. Чаще всего для обучения многослойной нейронной сети (поиска весовых коэффициентов) используется алгоритм обратного распространения ошибки, который основан на расчете невязок нейронов (частных производных функции ошибки для каждого объекта из обучающей выборки по отношению к весу связи) в обратном направлении при движении от выходного слоя к входному. Многослойные искусственные нейронные сети обратного распространения называют также многослойные перцептроны (от англ. multilayer perceptrons). На сегодняшний день в химии получили распространение следующие модифицированные многослойные искусственные нейронные сети	Могут быть описаны сколь угодно сложные зависимости.	Возможно переобучение искусственной нейронной сети при слишком малом соотношении между количеством соединений в обучающей выборке и количеством весовых коэффициентов, значения которых находятся в результате обучения искусственной нейронной сети. Для многослойных перцептронов может иметь место неполная воспроизводимость результатов. Зачастую возникает сложность интерпретации результатов (нейронные сети часто рассматриваются как «черный ящик»).	[49]

Окончание табл. 3

1	2	3	4	5
	обратного распространения: - ассоциативные нейронные сети (от англ. ASsociative Neural Network, или ASNN); - регуляризованные по Байесу нейронные сети (от англ. Bayesian regularized neural networks, или BRNN); - автокодировщики (от англ. autoencoders), называемые также автоэнкодерами, автоассоциаторами, автоассоциативными нейронными сетями. Кроме того, в химии нашли применение следующие более сложные нейронные сети: - сети (карты) Кохонена (от англ. Kohonen neural networks) и другие сети с конкурирующими нейронами; - встречного распространения (от англ. counterpropagation neural networks) - с радиальной базисной функцией (от англ. radial basis function neuralnetworks, RBFNN, или RBF-сети); - рекуррентные (с обратными связями): сети Хопфилда (от англ. Hopfield neural networks); машина Больцмана (от англ. Boltzmann machine) – стохастический вариант сетей Хопфилда; ограниченная машина Больцмана (от англ. restricted Boltzmann machine, или RBM); - сверточные (от англ. convolutional); - для работы на графах: сеть V. Kvasnička; сети ChemNet и MolNet; рекурсивные нейронные сети (от англ. recursive neural networks); графовые машины (от англ. graph machines); - «глубокого обучения» (от англ. deep learning).			

1.2.4 Оценка качества моделей «структура-свойство», построенных на основе методов машинного обучения

Для оценки качества модели исходную базу данных сначала делят на обучающую выборку (данные, используемые для построения модели) и тестовую выборку (данные, не используемые для построения модели) [8, 43, 50]. Качественная модель должна обладать хорошей описательной способностью (от англ. fitting) и хорошей прогностической способностью (от англ. predictive performance) [50]. Описательная способность модели – способность модели воспроизводить значения свойства для объектов обучающей выборки [8, 43, 50]. Прогностическая способность модели – способность модели воспроизводить значения свойства для объектов тестовой выборки [8, 43, 50]. Для количественной оценки качества модели «структура-свойство» на основе методов машинного обучения используются специальные метрики, выбор которых зависит от типа решаемой задачи [8, 43, 50]. Формулы для расчета метрик качества моделей одни и те же как для обучающей, так и для тестовой выборок [8, 43, 50].

В классификационных задачах применяют следующие основные метрики качества моделей (рассмотрим, в частности, на примере бинарной классификации) [43, 51-53].

1. Точность (от англ. accuracy, или ACC) – отражает долю верно классифицированных объектов среди всех объектов:

$$ACC=\frac{TP+TN}{TP+TN+FP+FN},$$

где TP (от англ. true positive) – количество верно классифицированных объектов 1-го класса (традиционно называемого положительным; именно этот класс представляет интерес); TN (от англ. true negative) – количество верно классифицированных объектов 2-го класса (традиционно называемого отрицательным); FP (от англ. false positive) – количество неверно классифицированных объектов 1-го класса (то есть это количество объектов, принадлежащих ко 2-му классу, отрицательному, но классифицированных

принадлежащими к 1-му классу, положительному); FN (от англ. false negative) – количество неверно классифицированных объектов 2-го класса (то есть это количество объектов, принадлежащих к 1-му классу, положительному, но классифицированных принадлежащими ко 2-му классу, отрицательному). ACC используется, когда классы сбалансированы, а при несбалансированных классах вводит в заблуждение.

2. Прецизионность (от англ. precision) – отражает долю верно классифицированных объектов 1-го класса среди всех объектов, классифицированных принадлежащими к 1-му классу:

$$\text{Precision}=\frac{TP}{TP+FP}.$$

3. Чувствительность (от англ. sensitivity, или SEN; также обозначаемая как true positive rate, или TPR; иногда называют recall – полнота) – отражает долю верно классифицированных объектов 1-го класса среди всех объектов, реально принадлежащих к 1-му классу:

$$\text{SEN}=\text{TPR}=\text{Recall}=\frac{TP}{TP+FN}.$$

4. Специфичность (от англ. specificity, или SPC; также обозначаемая как true negative rate, или TNR) – отражает долю верно классифицированных объектов 2-го класса среди всех объектов 2-го класса:

$$\text{SPC}=\text{TNR}=\frac{TN}{TN+FP}.$$

5. Доля ложноположительных результатов (от англ. false positive rate, или FPR) – отражает долю неверно классифицированных объектов 1-го класса среди всех объектов, фактически принадлежащих ко 2-му классу:

$$\text{FPR}=1-\text{TNR}=\frac{FP}{FP+TN}.$$

6. F1-мера (от англ. F1-Score) – представляет собой гармоническое среднее между прецизионностью и чувствительностью:

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

F1-мера балансирует прецизионность и чувствительность, из-за чего она дает более сбалансированную оценку модели, особенно при несбалансированных данных F1-мера принимает значения от 0 (худший результат) до 1 (идеальная прецизионность и чувствительность). F1-мера помогает понять, насколько хорошо модель одновременно избегает ложных срабатываний и при этом не пропускает реальные положительные примеры. Это более честная метрика, чем ACC, когда распределение классов не сбалансировано.

7. ROC-кривая (от англ. receiver operating characteristic curve – кривая рабочей характеристики приемника; под приемником понимается модель классификации) – строится путем вычисления TPR и FPR при всех возможных значениях порога классификации (от 0 до 1) и отображения TPR по оси ординат и FPR по оси абсцисс. ROC-кривая показывает, насколько хорошо модель отделяет один класс от другого (по форме кривой можно делать базовые выводы: в частности, чем выше над диагональю, выпуклее и левее кривая, тем лучше модель разделяет классы; диагональная прямая – это модель, случайно определяющая класс; кривая ниже диагонали – модель ошибочно определяет класс). Площадь под ROC-кривой обозначают как AUC-ROC (от англ. area under the receiver operating characteristic curve). AUC-ROC интерпретируют как вероятность того, что модель классификации правильно различит случайно выбранный объект: при AUC-ROC = 1 модель полагают идеальной, то есть безошибочно определяющей классы, при AUC-ROC = 0.5 модель случайно определяет класс, при AUC-ROC < 0.5 модель ошибочно определяет класс, то есть чем выше AUC-ROC, тем лучше модель умеет определять классы на всех уровнях порогов классификации. ROC-кривую и AUC-ROC используют для сбалансированных данных.

8. PR-кривая (от англ. precision-recall curve) – график зависимости Precision от Recall при изменении порога классификации. PR-кривую используют для сильно несбалансированных данных в отличие от ROC-кривой. AUC-PR –

площадь под PR-кривой. Чем больше AUC-PR, тем лучше модель классифицирует.

В регрессионных задачах применяют следующие основные метрики качества моделей [43, 54].

1. Сумма квадратов остатков (от англ. residual sum of squares, или RSS):

$$\mathrm{RSS} = \sum_{i=1}^{N} (\mathrm{y}_i^{exp} - \mathrm{y}_i^{calc})^2,$$

y_i^{exp} – зафиксированное в базе данных (истинное, фактическое, реальное) значение свойства *i*-го объекта в выборке; y_i^{calc} – значение свойства *i*-го объекта в выборке, определенное регрессионной моделью; N – количество объектов в выборке.

Низкое значение RSS говорит о высокой точности модели. На основе RSS строятся метрики MSE, RMSE и R^2 (см. ниже). При минимизации RSS автоматически минимизируются MSE и RMSE. RSS также является основой метода наименьших квадратов: когда строится линейная регрессия, ищутся такие параметры регрессии, при которых RSS является минимально возможной. RSS сильно чувствительна к выбросам в данных, поскольку из-за возведения в квадрат одна огромная ошибка может сильно завысить RSS, даже если все остальные прогнозируемые значения идеально совпадают с реальными значениями.

2. Средняя абсолютная ошибка (от англ. mean absolute error, или MAE) – измеряет среднюю величину ошибки без учета ее знака:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| \mathrm{y}_i^{exp} - \mathrm{y}_i^{calc} \right|.$$

Значения MAE интерпретируются следующим образом: спрогнозированные моделью значения в среднем отклоняются от реальных значений на [значение MAE] единиц. MAE не дает представления о завышении или занижении ошибки. Низкое значение MAE говорит о высокой точности модели. MAE устойчивее к выбросам, чем RMSE (см. ниже), и хороша, если нужно оценить среднюю величину ошибки без преувеличения влияния больших отклонений.

3. Средняя квадратичная ошибка (от англ. mean squared error, или MSE) – измеряет средний квадрат ошибок:

$$\mathrm{MSE} = \frac{\mathrm{RSS}}{N} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{y}_i^{exp} - \mathrm{y}_i^{calc})^2.$$

Низкое значение MSE говорит о высокой точности модели.

4. Корень из средней квадратичной ошибки (от англ. root mean squared error, или RMSE) – измеряет стандартное отклонение ошибок прогнозирования:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{y}_i^{exp} - \mathrm{y}_i^{calc})^2}.$$

Низкое значение RMSE говорит о высокой точности модели. RMSE удобнее интерпретировать по сравнению с MSE, поскольку RMSE измеряется в тех же единицах, что и целевая переменная. RMSE интерпретируется аналогично MAE, но с акцентом на большие отклонения. RMSE чувствителен к выбросам в данных. Если RMSE значительно больше MAE, то в данных есть выбросы, которые модель плохо прогнозирует.

5. Средняя абсолютная процентная ошибка (от англ. mean absolute percentage error, или MAPE) – измеряет среднюю ошибку в процентах:

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{\mathrm{y}_i^{exp} - \mathrm{y}_i^{calc}}{\mathrm{y}_i^{exp}}\right|.$$

Значения MAPE интерпретируются следующим образом: в среднем модель ошибается на [значение MAPE]%. MAPE не определена, если в базе данных есть нулевые значения ($\mathrm{y}_i^{exp} = 0$). MAPE сильно завышена при низких фактических значениях в базе данных.

6. Коэффициент детерминации (от англ. coefficient of determination) – показатель, равный отношению «объясненной части» дисперсии свойства к его полной дисперсии (то есть показывающий, какую долю дисперсии целевой переменной «объясняет» модель):

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i^{exp} - y_i^{calc})^2}{\sum_{i=1}^{N}(y_i^{exp} - \overline{y}^{exp})^2},$$

где $TSS = \sum_{i=1}^{N}(y_i^{exp} - \overline{y}^{exp})^2$ – сумма квадратов отклонения (от англ. total sum of squares, или TSS) значения свойства *i*-го объекта в выборке из *N* объектов от среднего значения свойства $\overline{y}^{exp} = \frac{1}{N}\sum_{i=1}^{N} y_i^{exp}$ в этой выборке. Если RSS является ошибкой построенной модели, TSS является ошибкой «глупой» модели, которая всегда прогнозирует среднее значение. Чем меньше RSS относительно TSS, тем больше R^2 и тем лучше построенная модель. R^2 принимает значения от -∞ (очень плохая модель) до 1 (модель идеально описывает данные, то есть объясняет 100% дисперсии данных). Обычно удовлетворительным результатом полагают $R^2 > 0.5$, а очень хорошим – $R^2 > 0.8$. При $R^2 = 0$ модель не нашла никаких закономерностей в данных. Значения R^2 интерпретируются следующим образом: модель объясняет $[R^2 \cdot 100]$% дисперсии данных. Нужно понимать, что R^2 не показывает размер ошибки, то есть, помимо R^2, обязательно необходимо отслеживать еще и MAE или RMSE, чтобы оценить, на сколько единиц ошибается модель. Высокое значение R^2 не означает, что модель является хорошей, потому что если в данных есть выбросы или произошло переобучение модели, то значение R^2 может быть высоким, но модель будет плохо работать на новых данных.

1.3 Модели «структура-свойство», построенные на основе методов машинного обучения, для разных типов химических объектов

Большинство работ по моделированию «структура-свойство» на основе методов машинного обучения посвящены индивидуальным органическим соединениям [55, 56], есть многочисленные сообщения о смесях химических

соединений, кристаллах неорганических соединений, наноматериалах и полимерах (табл. 4), а также о химических реакциях [57].

Таблица 4

Особенности моделирования «структура-свойство» на основе методов машинного обучения для смесей химических соединений, кристаллов неорганических соединений, наноматериалов, полимеров

Объект	Особенности моделирования	Источники
Смеси химических соединений	Дескрипторы, характеризующие смесь в целом, должны характеризовать и каждый ее компонент и учитывать количественное содержание компонентов в смеси. Модель должна быть устойчива к перестановке порядка компонентов смеси при формировании дескрипторов, характеризующих смесь в целом: значение прогнозируемого свойства не должно меняться от того, какой компонент рассматривали первым, какой – вторым и т.д. Поскольку одно и то же химическое соединение может присутствовать в качестве компонента сразу в нескольких смесях, прогностическая способность модели будет сильно зависеть от того, присутствуют ли смеси с этими соединениями, для которых осуществляется прогноз, в обучающей выборке.	[58, 59]
Кристаллы неорганических соединений	В отличие от органических соединений структуру кристаллов неорганических соединений невозможно представить в виде классических молекулярных графов, то есть традиционные дескрипторы не подходят для описания их структуры. Поэтому могут быть использованы два вида дескрипторов: топологические дескрипторы, описывающие молекулярный граф гипотетической молекулы, в которой ионные связи заменяются на ковалентные, и дескрипторы атомных орбиталей для описания энергии кристаллической решетки.	[60, 61]
Наноматериалы	В целом процедура моделирования не отличается от традиционных объектов. Если тонкая настройка свойств наноматериалов осуществляется за счет модификации низкомолекулярными органическими соединениями, то для представления структуры наноматериалов применяются составные дескрипторы, включающие дескрипторы для немодифицированного наноматериала и дескрипторы для модификатора.	[62-66]
Полимеры	Процедура построения моделей «структура-свойство» для полимеров не имеет принципиальных отличий от построения моделей для низкомолекулярных веществ. Основная сложность заключается в описании многоуровневой структуры полимера. Как правило, дескрипторы соотносят с мономерами (низкомолекулярными веществами, «кирпичиками», из которых синтезируются полимеры) или повторяющимися звеньями макромолекул полимеров. Недостаток такого подхода – невозможность учета в явном виде межмолекулярных взаимодействий макромолекул.	[67-69]

В настоящее время среди химических объектов повышенный интерес в аспекте моделирования «структура-свойство» методами машинного обучения вызывают полимеры. Причина этому заключается в многообразии полимеров, огромной и сложно прогнозируемой вариативности их свойств и практической востребованности полимеров. Свойства полимеров зависят от большого числа факторов: химической структуры мономеров, молекулярно-массовых характеристик полимеров, разветвленности и стереорегулярности макромолекул, степени кристалличности и др. Эти зависимости являются нелинейными и часто сложно интерпретируемыми. Небольшое изменение в химической структуре мономера способно принципиально изменить свойства полимера. А даже малое улучшение свойств полимеров или снижение стоимости мономеров имеет большое коммерческое значение, поскольку полимеры являются сырьем для более сложных материалов (пластмасс, композитов, эластомеров и др.). Исследователю сложно, а зачастую невозможно, уловить такие зависимости в многомерном пространстве данных, в то время как модели «структура-свойство», построенные на основе методов машинного обучения, идеально подходят для этого. Кроме того, следует отметить, что синтез нового полимера, его выделение, очистка и определение комплекса его свойств могут требовать существенных финансовых затрат и занимать длительное время. Модели «структура-свойство», построенные на основе методов машинного обучения и обученные на существующих данных, позволяют значительно ускорить разработку новых полимеров и оптимизировать их свойства под необходимые эксплуатационные требования.

Гомополимеры – это полимеры, макромолекулы которых состоят из одинаковых повторяющихся звеньев. Исследовательский интерес к гомополимерам в аспекте моделирования «структура-свойство» методами машинного обучения обусловлен тем, что гомополимеры достаточно просты, чтобы моделировать связь «структура-свойство» с помощью ограниченного набора данных (дескрипторы описывают повторяющееся звено одного типа), но и одновременно достаточно сложны для интерпретации проявления свойств в

зависимости от их многоуровневой структуры. В этой связи гомополимеры являются идеальными объектами для выявления базовых принципов, понимание которых необходимо в дальнейшем при установлении и интерпретации связи «структура-свойство» для более сложных объектов (сополимеров, смесей полимеров, композитов и др.). Большинство крупнотоннажных полимеров – это именно гомополимеры, причем органические: полиэтилен, полипропилен, поливинилхлорид, полистирол. Следует отметить, что органические гомополимеры представляют собой наиболее широкий класс полимеров. Оптимизация их свойств – это актуальнейшая задача органической химии, где модели «структура-свойство, построенные на основе методов машинного обучения, уже дают оптимистичные результаты (см. п. 1.4.4). Важнейшим свойством, на величину которого опираются при подборе полимера под требуемые условия эксплуатации, является температура стеклования – температурная граница между их стеклообразным (упруготвердым) и высокоэластическим физическими состояниями. При температурах ниже температуры стеклования полимеры используются как конструкционные материалы, при температурах выше температуры стеклования, но ниже температуры текучести полимеры применяются как эластичные материалы [70]. То есть, другими словами, температура стеклования полимеров является предельной температурой, до которой не проявляется ползучесть [70].

1.4 Обзор исследований «структура – температура стеклования полимеров»

1.4.1 Основные представления о стекловании полимеров. Экспериментальные методы определения температуры стеклования полимеров

Различают структурное и механическое стеклование полимеров [71]. Под структурным стеклованием полимеров понимается фиксация положения макромолекул в пространстве при понижении температуры, под механическим стеклованием – фиксация положения макромолекул в пространстве, прежде всего, за счет действия силового поля, то есть в последнем случае при понижении

температуры можно достичь стеклования раньше, чем в отсутствие действия силового поля [71]. В настоящей диссертации под температурой стеклования полимеров понимается температура структурного стеклования, которую далее для краткости будем называть температурой стеклования, кроме случаев, где нужно различать температуры структурного и механического стеклования.

Экспериментальные методы определения температуры стеклования полимеров традиционно делятся на статические и динамические [72-86]. Статические методы служат для определения температуры структурного стеклования полимеров и основаны на измерении изменения различных свойств (механических, теплофизических, дилатометрических и др.) образца полимера при одновременном изменении температуры с определенной скоростью в отсутствии внешнего периодического воздействия [72-77]. Наиболее распространены следующие статические методы:

1) термомеханический анализ – измерение деформации (механического свойства) образца под действием постоянной нагрузки [72];

2) дифференциальный термический анализ – измерение разности температур (теплофизического свойства) исследуемого и эталонного образцов, нагреваемых одним нагревателем, причем в эталонном образце не должно происходить физических и химических превращений, сопровождающихся тепловыми эффектами, при этом в области стеклования исследуемого образца наблюдается перегиб на термограмме [73];

3) дифференциальная сканирующая калориметрия – метод, аналогичный дифференциальному термическому анализу, но измеряется тепловой поток (теплофизическое свойство), необходимый для поддержания одинаковых температур у исследуемого и эталонного образцов, которые нагреваются разными нагревателями, при этом в области стеклования исследуемого образца наблюдается изменение его теплоемкости [74];

4) дилатометрия – измерение изменения линейных размеров или объема (дилатометрических свойств) ненагруженного образца, при этом переход из

одного физического состояния в другое сопровождается изменением значения коэффициента теплового расширения образца [75];

5) люминесцентный анализ – измерение изменения интенсивности люминесцентного свечения (излучающих свойств) люминофоров, введенных в образец [76];

6) инфракрасная спектроскопия – измерение изменения оптической плотности (оптических свойств) образца в инфракрасной области спектра, связанного с ограничением подвижности сегментов макромолекул в образце при переходе полимера из одного физического состояния в другое [77].

Динамические методы служат для определения температуры механического стеклования полимеров и основаны на измерении изменения свойств образца полимера при одновременном изменении температуры с определенной скоростью и внешнем периодическом воздействии [78-86]. Наиболее распространены следующие динамические методы:

1) динамический термомеханический анализ – измерение механических свойств при их комплексном представлении: деформации, модулей, механических потерь (мнимой составляющей свойств) и тангенса угла механических потерь образца [78, 79];

2) динамический диэлектрический анализ – измерение диэлектрических свойств при их комплексном представлении: диэлектрической проницаемости, электропроводности, диэлекрических потерь (мнимой составляющей диэлектрической проницаемости) и тангенса угла диэлектрических потерь [78];

3) метод токов термостимулированной деполяризации – измерение токов деполяризации (диэлектрического свойства), возникающих при нагревании образца, который был предварительно поляризован при температурах выше температуры текучести [80, 81];

4) метод ядерного магнитного резонанса – измерение второго момента спектральной линии в спектрах ядерного магнитного резонанса (магнитного свойства) на частотах 10^4-10^8 Гц [82-84];

5) метод электронного парамагнитного резонанса – измерение ширины линии в спектре электронного парамагнитного резонанса (магнитного свойства) на частотах 10^6-10^9 Гц для радикалов или парамагнитных зондов, введенных в образец [83, 85, 86].

1.4.2 Теории стеклования полимеров

Исследования стеклования полимеров ведутся с начала XX века [87, 88], однако до сих пор природа стеклования полимеров остается предметом научных дискуссий [89-95].

Развиты следующие теории стеклования полимеров [96]:

1) термодинамические теории, согласно которым стеклование является термодинамическим переходом (в отношении порядка перехода нет общей единой точки зрения): при понижении температуры полимера до температуры стеклования конформационные переходы в нем становятся энергетически невозможными, в связи с чем тепловое движение сегментов макромолекул прекращается;

2) кинетические теории, согласно которым стеклование наступает, когда структура полимера перестает успевать перестраиваться за изменением внешних условий (температуры, давления):

- теория М.В. Волькенштейна-О.Б. Птицына – стеклование полимеров описывается как активационный процесс, в котором сегменты макромолекул переходят между потенциальными ямами, преодолевая энергетические барьеры;

- теория Ю.Я. Готлиба-О.Б. Птицына – исходит из тех же предпосылок, что и теория М.В. Волькенштейна-О.Б. Птицына, но учитывает кооперативный характер движения сегментов макромолекул;

- теория межмолекулярных связей С.Н. Журкова – развивает представления теории Ю.Я. Готлиба-О.Б. Птицына; согласно теории С.Н. Журкова, стеклование полимеров наступает при «замораживании» пространственных положений

макромолекул, причем макромолекулы связаны между собой межмолекулярными связями;

- флуктуационная теория стеклования полимеров – развивает представления С.Н. Журкова; согласно флуктуационной теории стеклования полимеров, межмолекулярные связи являются флуктуационными;

- теория свободного объема – полагает, что подвижность сегментов макромолекул связана с наличием свободного объема – пространства, необходимого для их перемещения: при повышении температуры свободный объем в полимере увеличивается, подвижность сегментов макромолекул возрастает; при понижении температуры свободный объем в полимере уменьшается до критической величины, и движение сегментов макромолекул затрудняется («замораживается»), что приводит к стеклованию полимера.

Из теорий стеклования полимеров следует, что температура стеклования полимера определяется 3 ключевыми факторами:

1) гибкостью макромолекул (термодинамические теории и кинетические теории М.В. Волькенштейна-О.Б.Птицына и Ю.Я. Готлиба-О.Б. Птицына);

2) межмолекулярными взаимодействиями (теория межмолекулярных связей С.Н. Журкова и флуктуационная теория стеклования полимеров);

3) долей свободного объема, или плотностью молекулярной упаковки при температурах выше температуры стеклования (теория свободного объема).

1.4.3 Подходы к расчету температуры стеклования полимеров исходя из строения их повторяющихся звеньев

Развиты следующие подходы к расчету свойств полимеров (в том числе, температуры стеклования) исходя из строения их повторяющихся звеньев.

1. Подход, основанный на аддитивности вкладов химических групп в свойства полимеров (подход Д.В. ван Кревелена) [97] – это один из наиболее ранних и широко известных подходов, ориентированный на линейные и слаборазветвленные полимеры. Основная идея подхода заключается в том, что

каждая химическая группа в повторяющемся звене полимера вносит аддитивный вклад в значение свойства. Эти вклады суммируются с учетом их кратности в структуре полимера. Недостатки подхода: а) невозможность рассчитать свойство полимера, если вклад какой-то химической группы неизвестен; б) в явном виде не учитываются межмолекулярные взаимодействия; в) в расчетных формулах зачастую фигурируют подгоночные коэффициенты, не имеющие физического смысла; г) не учитывается стереорегулярность в макромолекулах; д) отсутствуют формулы для расчета свойств сополимеров и сетчатых полимеров.

2. Подход индексов связанности (подход Дж. Бицерано) [98] – это подход, развитый для линейных, разветвленных и сетчатых полимеров и оперирующий регрессионными корреляциями между свойствами полимеров и индексами связанности, которые отражают связи атомов в повторяющемся звене полимера и в некоторых случаях отдельные данные об их электронной структуре. Недостатки подхода: а) в явном виде не учитываются межмолекулярные взаимодействия; б) не учитывается стереорегулярность в макромолекулах; в) отсутствуют формулы для расчета свойств сополимеров.

3. Инкрементальный подход (подход А.А. Аскадского) [99-101] – это подход, развитый для линейных, разветвленных и сетчатых полимеров и основанный на идее о том, что свойства полимера определяются суммой инкрементов атомов повторяющегося звена и инкрементов межмолекулярных взаимодействий; подход позволяет проводить анализ влияния различных учитываемых (и несущих определенный физический смысл) факторов на свойства полимеров. Недостатки подхода: а) отсутствуют формулы для расчета свойств градиентных сополимеров; б) не учитывается положение заместителей в ароматическом кольце. Согласно инкрементальному подходу, температура стеклования полимеров рассчитывается по формуле [99]:

$$T_{g\ inc} = \frac{\sum_i \Delta V_i}{\sum_i a_i \Delta V_i + \sum_j b_j}, \quad (1)$$

где ΔV_i – ван-дер-ваальсовые объемы атомов повторяющегося звена полимера; a_i – инкременты, характеризующие энергию слабого дисперсионного взаимодействия каждого атома; b_j – инкременты, характеризующие энергию сильного специфического межмолекулярного взаимодействия (например, диполь-дипольного, водородных связей).

1.4.4 Модели «структура – температура стеклования органических гомополимеров», построенные на основе методов машинного обучения

В научной литературе представлено достаточно большое количество моделей «структура – температура стеклования органических гомополимеров», построенных на основе методов машинного обучения (табл. 5).

Таблица 5

Обзор моделей «структура – температура стеклования органических гомополимеров», построенных на основе методов машинного обучения

Год	Сущность модели	Дескрипторы	Метод машинного обучения	Достоинства модели и моделирования	Источник
1	2	3	4	5	6
1996	Модель реализована в программном пакете CODESSA и позволяет прогнозировать температуру стеклования органических гомополимеров и сополимеров.	Конституциональные, геометрические, топологические, электростатические, квантово-химические и термодинамические.	Четырехпараметрическая линейная регрессия.	Модель построена на дескрипторах, имеющих ясный физический смысл. Причем выявлена значимость вклада каждого отдельного дескриптора в температуру стеклования полимеров.	[102]
1997	Модель позволяет прогнозировать температуру стеклования акрилатных и метакрилатных гомополимеров.	Энергия, объем повторяющегося звена полимера, рассчитанные с использованием классической молекулярной механики и динамики, масса повторяющегося звена полимера.	Линейная регрессия.	Простота интерпретации результатов моделирования за счет использования в качестве дескрипторов физико-химических параметров.	[103]
2002	Модели позволяют прогнозировать температуру стеклования органических гомополимеров, исходя из структур мономеров или исходя из структур повторяющихся звеньев полимеров.	Топологические, электронные, геометрические. Сначала для полимеров из базы данных рассчитывались все указанные дескрипторы. Далее проводилось сокращение количества дескрипторов путем удаления аналогичных и сильно коррелирующих.	Множественная линейная регрессия, искусственная нейронная сеть.	Возможность использования моделей для разнообразных органических гомополимеров.	[104]

Продолжение табл. 5

1	2	3	4	5	6
2004	Модель позволяет прогнозировать температуру стеклования эпоксиаминных полимеров (эпоксидные смолы на основе диглицидилового эфира бисфенола А, отвержденные аминами).	Квантово-механические дескрипторы, рассчитанные с использованием полуэмпирического квантово-механического метода в программном пакете CODESSA.	Линейная регрессия.	Простая для понимания концепция аддитивности. Использованные квантово-механические дескрипторы поддаются интуитивной интерпретации.	[105]
2008	Модели позволяют прогнозировать температуру стеклования органических гомополимеров, исходя из структур их повторяющихся звеньев.	Использованы 28 дескрипторов, разделенных на 3 группы: дескрипторы структуры повторяющегося звена полимера, дескрипторы связей основной цепи и дескрипторы водородных связей.	Множественная линейная регрессия, искусственная нейронная сеть.	Модель на основе искусственной нейронной сети позволяет получать хорошие результаты для полиакрилатов.	[106]
2009	Модели позволяют прогнозировать температуру стеклования полиакрилатов и полистиролов, исходя из структур мономеров.	Квантово-химические дескрипторы, рассчитанные на основе теории функционала плотности: молекулярная средняя поляризуемость, энергия высшей занятой молекулярной орбитали, полная тепловая энергия и полная энтропия.	Множественная линейная регрессия, искусственная нейронная сеть.	Модель на основе искусственной нейронной сети лучше воспроизводит данные, чем модель на основе множественной линейной регрессии. Использованные квантово-химические дескрипторы позволяют интерпретировать результаты моделирования с точки зрения влияния структур мономеров на температуру стеклования полимеров.	[107]

Продолжение табл. 5

1	2	3	4	5	6
2010	Модель позволяет прогнозировать температуру стеклования полиамидов и полибензимидазолов, исходя из структур их повторяющихся звеньев.	Четыре квантово-химических дескриптора, выбранные с помощью множественной линейной регрессии из 1664 сгенерированных: MAXDP, Mor17m, nArCONHR и ESpm03x.	Метод опорных векторов.	Впервые использован метод опорных векторов для прогнозирования температуры стеклования полимеров. Высокая прогностическая способность модели для выбранных объектов исследования.	[108]
2012	Модель позволяет прогнозировать температуру стеклования полиарилэфирсульфонов, исходя из структур их повторяющихся звеньев.	Квантово-химические дескрипторы: количество вращающихся связей, дипольный момент, теплота образования, энергия высшей занятой молекулярной орбитали, молярная масса, молярный объем.	Множественная линейная регрессия.	Относительно простая модель, использующая небольшое количество параметров, связанных с повторяющимися звеньями полиарилэфирсульфонов. В результате моделирования показана значимость отдельных дескрипторов в отношении их влияния на температуру стеклования полиарилэфирсульфонов.	[109]
2013	Модель позволяет прогнозировать температуру стеклования полистиролов и поли(мет)акрилатов, исходя из структур мономеров.	Физико-химические дескрипторы, полученные из структуры мономеров: дескриптор водородных связей, гибкость цепи, средняя молекулярная поляризуемость и заряд самого электроотрицательного атома.	Метод опорных векторов в сочетании с оптимизационным алгоритмом.	Модель продемонстрировала лучшие результаты по сравнению с моделью на основе искусственной нейронной сети, обученной на той же выборке.	[110]

Продолжение табл. 5

1	2	3	4	5	6
2018	Модель позволяет прогнозировать температуру стеклования органических гомополимеров, исходя из структур их повторяющихся звеньев.	2D молекулярные дескрипторы из программ Dragon и Padel.	Линейная регрессия.	Модель продемонстрировала высокую прогностическую способность.	[111]
2018	Модель позволяет прогнозировать температуру стеклования органических и элементорганических гомополимеров, исходя из структур их повторяющихся звеньев.	С использованием программы Dragon рассчитали более 4500 дескрипторов. Далее на основе генетического алгоритма сократили набор дескрипторов до 2729. Итоговое моделирование проводили на наборах из 1-10 физико-химических дескрипторов. Наилучшие результаты продемонстрировала модель на основе 7 дескрипторов.	Множественная линейная регрессия.	При моделировании отобраны дескрипторы, которые имеют наибольшую корреляцию с температурой стеклования полимеров.	[112]
2019	Модель позволяет прогнозировать температуру стеклования полигидроксиалканоатов.	Молекулярные отпечатки повторяющихся звеньев, средние молекулярные массы и коэффициент полидисперсности.	Метод случайного леса.	Возможность прогнозирования температуры стеклования для огромного множества гомо- и гетеро полигидроксиалканоатов. Учтены молекулярно-массовые характеристики полимеров.	[113]

Продолжение табл. 5

1	2	3	4	5	6
2019	Модели позволяют прогнозировать температуру стеклования гомополимеров, исходя из представления повторяющихся звеньев в виде двух молекул мономеров.	Молекулярные отпечатки, молекулярные графы и молекулярные вложения (от англ. molecular embedding).	Метод случайного леса, множественная линейная регрессия, многослойный перцептрон, метод опорных векторов.	Модели обучены и протестированы на обширной базе данных PolyInfo. Представление повторяющихся звеньев в виде двух молекул мономеров позволяет учесть стереорегулярность полимеров.	[114]
2020	Модель позволяет прогнозировать температуру стеклования полиимидов, исходя из структур повторяющихся звеньев.	1342 дескриптора разных видов для каждого полиимида в базе данных, сгенерированные с помощью программы E-dragon.	Не указан.	Продемонстрирована процедура эффективного распределения данных между обучающей и тестовой выборками с помощью алгоритмов LASSO и Bagging.	[115]
2020	Модель позволяет прогнозировать температуру стеклования полимеров (не конкретизируется, о каких полимерах идет речь), исходя из структур повторяющихся звеньев.	Молекулярные отпечатки, учитывающие типы атомов и связей, долю вращающихся связей, топологическое расстояние между ароматическими кольцами, длины боковых цепей.	Регрессия гауссовского процесса.	Высокая точность модели, связанная с обучением и тестированием последней на обширной базе данных (1321 полимер).	[116]
2020	Модель позволяет прогнозировать температуру стеклования органических гомополимеров, исходя из структур мономеров.	Кодировки SMILES, преобразованные в бинарные изображения.	Сверточные нейронные сети.	Разработана оригинальная архитектура нейронной сети под решаемую задачу.	[117]

Продолжение табл. 5

1	2	3	4	5	6
2021	Модель позволяет прогнозировать температуру стеклования теплостойких гомополимеров, исходя из структур повторяющихся звеньев.	Молекулярные отпечатки Моргана.	Глубокие нейронные сети.	Впервые рассмотрена возможность применения глубоких нейронных сетей для прогнозирования температуры стеклования полимеров. Модель обучалась на данных по 13 000 полимеров из базы PoLyInfo.	[118]
2021	Модель позволяет прогнозировать температуру стеклования полимеров (не конкретизируется, о каких полимерах идет речь), исходя из структур повторяющихся звеньев.	Кодировки SMILES, преобразованные в машиночитаемые формы.	Рекуррентные нейронные сети (разновидность Long Short-Term Memory).	Впервые рассмотрена возможность применения рекуррентных нейронных сетей (разновидность Long Short-Term Memory) для прогнозирования температуры стеклования полимеров.	[119]
2021	Модель позволяет прогнозировать температуру стеклования полиметакрилатов, исходя из структур мономеров.	Квантово-химические дескрипторы: Длина боковой цепи, абсолютная разница между длиной боковой цепи и 1.356 нм (отражает топологическое свойство), полная энергия макромолекулы, заряд атома углерода C в группе RO, средняя поляризуемость, тепловая энергия полиметакрилатов.	Регрессия гауссовского процесса.	Модель продемонстрировала высокую точность для прогнозирования температуры стеклования полиметакрилатов.	[120]

Продолжение табл. 5

1	2	3	4	5	6
2021	Модель позволяет прогнозировать температуру стеклования теплостойких органических гомополимеров, исходя из структур повторяющихся звеньев.	Кодировки SMILES, преобразованные в машиночитаемые формы.	Рекуррентная нейронная сеть.	Модель обучена и протестирована на обширной базе данных PolyInfo. Модель точно воспроизводит температуру стеклования теплостойких полимеров со сложными структурами повторяющихся звеньев.	[121]
2022	Модель позволяет прогнозировать температуру стеклования полиимидов, исходя из структур повторяющихся звеньев.	Кодировки SMILES, преобразованные в машиночитаемые формы.	Графовая сверточная нейронная сеть.	Модель обучалась на литературных данных и данных, полученных с помощью инкрементального подхода. Модель имеет MAE = 20 K, что в 1.5 раза ниже, чем MAE модели, построенной на основе инкрементального подхода, – MAE = 33 K.	[101]
2024	Модели позволяют прогнозировать температуру стеклования полимеров (не конкретизируется, о каких полимерах идет речь), исходя из структур повторяющихся звеньев.	Молекулярные отпечатки Моргана, молекулярные дескрипторы.	Дерево принятия решений, метод опорных векторов, адаптивный бустинг, метод k ближайших соседей, метод случайного леса, экстремальный градиентный бустинг, облегченный градиентный бустинг и алгоритм Extra Tree.	Наилучший результат (R^2 = 0.88) показала модель на основе алгоритма Extra Tree.	[122]
2024	Модели позволяют прогнозировать температуру стеклования гомополимеров (не конкретизируется, о каких полимерах идет речь), исходя из структур повторяющихся звеньев.	Молекулярные дескрипторы.	Множественная линейная регрессия, метод k ближайших соседей, метод опорных векторов, метод случайного леса, регрессия гауссовского процесса и многослойный перцептрон.	Наилучший результат (R^2 = 0.813) показала модель на основе метода опорных векторов.	[123]

Окончание табл. 5

1	2	3	4	5	6
2024	Модель позволяет прогнозировать температуру стеклования гомополимеров (не конкретизируется, о каких полимерах идет речь), исходя из структур повторяющихся звеньев.	3D-координаты	Эквивариантная нейронная сеть.	Модель обучена и протестирована на обширной базе данных PolyInfo. Рассмотрено представление структур полимеров в трехмерном виде.	[124]
2024	Модели позволяют прогнозировать температуру стеклования гомополимеров (не конкретизируется, о каких полимерах идет речь), исходя из структур мономеров.	Кодировки SMILES, преобразованные в машиночитаемые формы.	Метод k ближайших соседей, метод опорных векторов, экстремальный градиентный бустинг, искусственная нейронная сеть и рекуррентная нейронная сеть.	Модель на основе экстремального градиентного бустинга показала $R^2 = 0.774$, продемонстрировала наибольшую устойчивость в воспроизведении данных и более короткое время обучения.	[125]
2025	Модели позволяют прогнозировать изменение модуля упругости в зависимости от температуры во всей зоне стеклования термореактивных полимеров, исходя из структур повторяющихся звеньев.	Дескрипторы микроскопического, мезоскопического и макроскопического уровней структуры полимеров.	Метод опорных векторов, искусственная нейронная сеть и регрессия гауссовского процесса.	Модель на основе искусственной нейронной сети признана лучшей. По сравнению с предыдущими исследованиями впервые прогнозируется не конкретное значение температуры стеклования, а вся зона стеклования полимеров.	[126]

Анализ обзора работ по моделированию «структура – температура стеклования органических полимеров» методами машинного обучения (табл. 5) показывает, что все они носят чисто методический характер, то есть посвящены поиску наилучших дескрипторов, оптимального метода машинного обучения с точки зрения повышения прогностической способности модели, и совершенно не анализируют полученные результаты в аспекте теории химического строения органических соединений и существующих теорий стеклования полимеров. Это делает проведение таких исследований актуальным.

***

В настоящее время среди химических объектов повышенный интерес в аспекте моделирования «структура-свойство» методами машинного обучения вызывают полимеры. Причина этому заключается в многообразии полимеров, огромной и сложно прогнозируемой вариативности их свойств и практической востребованности полимеров. Свойства полимеров зависят от большого числа факторов: химической структуры мономеров, молекулярно-массовых характеристик полимеров, разветвленности и стереорегулярности макромолекул, степени кристалличности и др. Эти зависимости являются нелинейными и часто сложно интерпретируемыми. Небольшое изменение в химической структуре мономера способно принципиально изменить свойства полимера. А даже малое улучшение свойств полимеров или снижение стоимости мономеров имеет большое коммерческое значение, поскольку полимеры являются сырьем для более сложных материалов (пластмасс, композитов, эластомеров и др.). Исследователю сложно, а зачастую невозможно, уловить такие зависимости в многомерном пространстве данных, в то время как модели «структура-свойство», построенные на основе методов машинного обучения, идеально подходят для этого. Кроме того, следует отметить, что синтез нового полимера, его

выделение, очистка и определение комплекса его свойств могут требовать существенных финансовых затрат и занимать длительное время. Модели «структура-свойство», построенные на основе методов машинного обучения и обученные на существующих данных, позволяют значительно ускорить разработку новых полимеров и оптимизировать их свойства под необходимые эксплуатационные требования.

Гомополимеры – это полимеры, макромолекулы которых состоят из одинаковых повторяющихся звеньев. Исследовательский интерес к гомополимерам в аспекте моделирования «структура-свойство» методами машинного обучения обусловлен тем, что гомополимеры достаточно просты, чтобы моделировать связь «структура-свойство» с помощью ограниченного набора данных (дескрипторы описывают повторяющееся звено одного типа), но и одновременно достаточно сложны для интерпретации проявления свойств в зависимости от химического строения их повторяющихся звеньев и их многоуровневой структуры. В этой связи гомополимеры являются идеальными объектами для выявления базовых принципов, понимание которых необходимо в дальнейшем при установлении и анализе связи «структура-свойство» для более сложных объектов (сополимеров, смесей полимеров, композитов и др.). Большинство крупнотоннажных полимеров – это именно гомополимеры, причем органические: полиэтилен, полипропилен, поливинилхлорид, полистирол. Следует отметить, что органические гомополимеры представляют собой наиболее широкий класс полимеров. Оптимизация их свойств – это актуальнейшая задача органической химии, где модели «структура-свойство», построенные на основе методов машинного обучения, уже дают оптимистичные результаты. Важнейшим свойством, на величину которого опираются при подборе полимера под требуемые условия эксплуатации, является температура стеклования – температурная граница между их стеклообразным (упруготвердым) и высокоэластическим физическими состояниями. При температурах ниже температуры стеклования полимеры используются как

конструкционные материалы, при температурах выше температуры стеклования, но ниже температуры текучести полимеры применяются как эластичные материалы. То есть, другими словами, температура стеклования полимеров является предельной температурой, до которой не проявляется ползучесть.

Анализ обзора работ по моделированию «структура – температура стеклования органических полимеров» методами машинного обучения показал, что все они носят чисто методический характер, то есть посвящены поиску наилучших параметров структуры, оптимального метода машинного обучения с точки зрения повышения прогностической способности модели, и совершенно не анализируют полученные результаты в рамках теории химического строения органических соединений и существующих теорий стеклования полимеров. Это делает проведение такого исследования актуальным, и оно может осуществляться с привлечением инкрементального подхода к расчету свойств полимеров исходя из строения их повторяющихся звеньев – подхода, основанного на идее о том, что свойства полимера определяются суммой инкрементов атомов повторяющегося звена и инкрементов межмолекулярных взаимодействий.

В связи с чем целью настоящей работы стало построение модели машинного обучения, описывающей связь «структура – температура стеклования органических гомополимеров» и содержащей параметры, которые интерпретируются в рамках теории химического строения органических соединений и теорий стеклования полимеров. Для достижения поставленной цели в работе решались следующие задачи:

1) рассмотрение различных методов машинного обучения и выбор среди них обеспечивающего наибольшую достоверность модели, способной прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее на основе строения их повторяющихся звеньев по аналогии с инкрементальным подходом;

2) уточнение значений параметров, определяющих температуру стеклования органических гомополимеров на основе строения их повторяющихся звеньев по аналогии с инкрементальным подходом, за счет применения при моделировании на основе машинного обучения комбинированных дескрипторов для обеспечения более достоверных прогнозов;

3) установление корреляций параметров, определяющих температуру стеклования органических гомополимеров на основе строения их повторяющихся звеньев по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям полимеров; анализ установленных корреляций в рамках теории химического строения органических соединений и теорий стеклования полимеров.

ГЛАВА 2 МЕТОДОЛОГИЯ И МЕТОДЫ ИССЛЕДОВАНИЯ

Исследование провели по следующей методологии:

1) формирование базы данных для модели «структура – температура стеклования органических гомополимеров», построенной на основе машинного обучения;

2) выбор дескрипторов для описания химических структурных формул повторяющихся звеньев органических гомополимеров;

3) выбор метода машинного обучения для моделирования связи «структура – температура стеклования органических гомополимеров»;

4) первый этап моделирования связи «структура – температура стеклования органических гомополимеров» на основе машинного обучения: проверка возможности прогнозирования температуры стеклования органических гомополимеров с помощью модели, построенной на основе методов машинного обучения, не напрямую, а через параметры, определяющие температуру стеклования на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом;

5) второй этап моделирования связи «структура – температура стеклования органических гомополимеров» на основе машинного обучения: уточнение значений параметров, определяющих температуру стеклования органических гомополимеров на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом, за счет применения комбинированных дескрипторов для обеспечения более достоверных прогнозов;

6) установление корреляций уточненных параметров, определяющих температуру стеклования органических гомополимеров на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям полимеров;

7) анализ установленных корреляций в рамках теории химического строения органических соединений и теорий стеклования полимеров.

2.1 Формирование базы данных для обучения модели «структура – температура стеклования органических гомополимеров», построенной на основе машинного обучения

В качестве исходной базы данных использовали данные работы [99]. Исходная база данных [99] включает сведения о химических структурах повторяющихся звеньев и температурах стеклования 1050 полимеров (фрагмент базы приведен в Приложении А). В базе представлены органические гомополимеры различных типов, включая полимеры с разным положением заместителя в ароматическом кольце. Структуры повторяющихся звеньев органических гомополимеров в базе данных представлены в виде SMILES.

2.2 Выбор дескрипторов для описания химических структурных формул повторяющихся звеньев органических гомополимеров

В настоящей диссертации для описания химических структурных формул повторяющихся звеньев органических гомополимеров из базы данных выбраны два вида дескрипторов:

1) структурные ключи (от англ. Molecular ACCess System Keys, или MACCS Keys, или MK) [127] с их понятной и наглядной концепцией;

2) молекулярные отпечатки Моргана (от англ. Morgan fingerprints, или MF) [128], которые широко применяются в задачах прогнозирования свойств полимеров.

Структурный ключ представляет собой битовую строку фиксированной длины (обычно 166 бит), где каждому биту соответствует определенный фрагмент (то есть в молекуле может быть максимум 166 фрагментов) из специализированной библиотеки, написанной на языке программирования Python [127]. Значение «1» в строке указывает на наличие данного фрагмента в молекуле, а «0» – на его отсутствие (рис. 1). Недостатком структурных ключей является то,

что данный вид дескрипторов не содержит информацию о расположении фрагментов относительно друг друга [127].

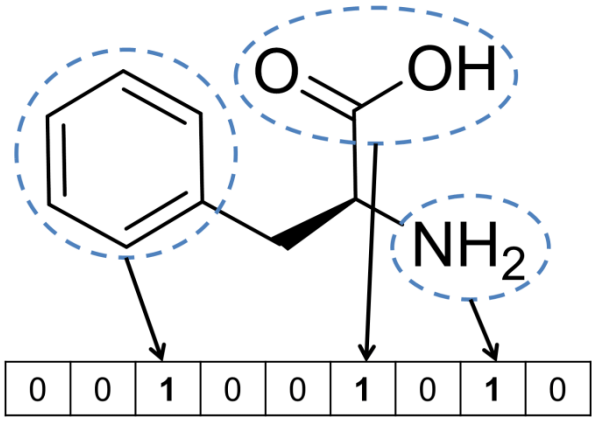


Рис. 1. Принцип формирования структурных ключей [127]

Молекулярные отпечатки Моргана используют более сложную кодировку, учитывающую не только присутствие определенных фрагментов, но и их расположение относительно друг друга [128].

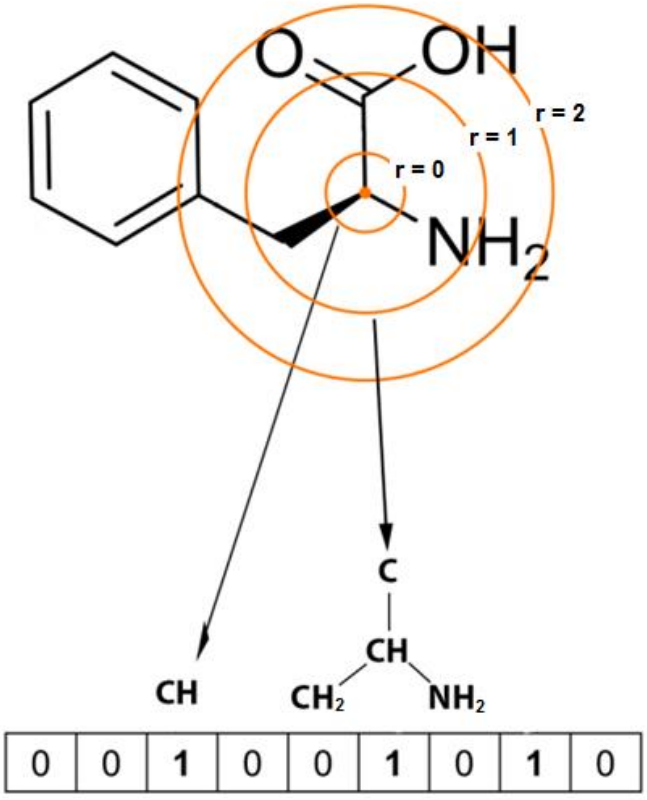


Рис. 2. Принцип формирования молекулярных отпечатков Моргана [128]

Молекулярный отпечаток Моргана представляет собой бинарный дескриптор настраиваемой длины (обычно 1024, 2048 и более бит) [128]. В отличие от генерации структурных ключей (где используется фиксированный список фрагментов), молекулярные отпечатки Моргана генерируются алгоритмически, что делает их гораздо более гибкими и способными улавливать тонкие структурные особенности [128]. Порядок генерации молекулярного отпечатка Моргана (рис. 2): каждый атом в молекуле рассматривается итеративно с постепенным увеличением радиуса ближайшего окружения, то есть на первом этапе рассматривается только сам атом, далее атом с непосредственными соседями и т.д.; все фрагменты молекулы фиксируются и хэшируются (получают определенный номер позиции) в битовой строке.

В диссертации провели тестирование обоих видов дескрипторов для определения их эффективности в прогнозировании температуры стеклования органических гомополимеров.

2.3 Выбор метода машинного обучения для моделирования связи «структура – температура стеклования органических гомополимеров»

Моделирование на основе машинного обучения реализовывали на языке программирования Python 3.13.7 (программный код приведен в Приложении Б). При моделировании рассмотрели 3 метода машинного обучения: метод случайного леса, метод k ближайших соседей и многослойный перцептрон.

Метод случайного леса в Python запускали с помощью алгоритма RandomForestRegressor из библиотеки scikit-learn [129]. Метод относится к методам деревьев принятия решений [48]. Дерево принятия решений – это способ принятия решений, который можно представить в виде схемы с развилками [48]. В частности, для классификационной задачи алгоритм на каждом шаге (узле) задает вопрос, требующий ответа «да» или «нет» [48]. Например, «Содержится ли в повторяющемся звене полимера атом галогена?». В зависимости от ответа («да» или «нет») алгоритм переходит к следующему вопросу [48]. В конце такого пути

находится лист – итоговое решение, которое алгоритм выдает после прохождения всех шагов [48] – см. рис. 3.

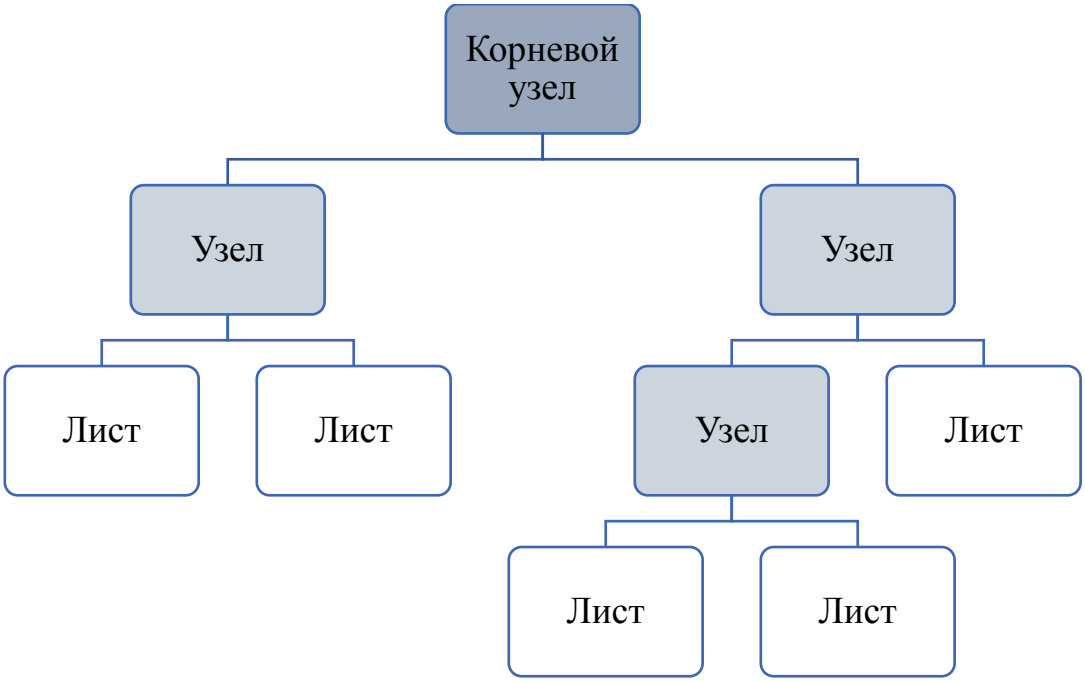


Рис. 3. Схема дерева принятия решений [48]

Отдельное дерево принятия решений дает прогнозы низкого качества [48]. В методе случайного леса создается ансамбль деревьев принятия решений, которые отличаются выбором признаков на каждом узле [48]. Объединение прогнозов всех деревьев из ансамбля позволяет получить более точные результаты [48]. Метод случайного леса из-за своей простой концепции получил широкое распространение и является одним из наиболее популярных методов машинного обучения, применяющихся в химии [48].

Метод k ближайших соседей в Python запускали с помощью алгоритма KNeighborsRegressor из библиотеки scikit-learn [129]. Метод основан на принципе, что объекты с похожими характеристиками (расположенные близко друг к другу в химическом пространстве) имеют близкие значения целевой переменной [45, 46]. Пример использования метода k ближайших соседей для решения классификационной задачи представлен на рис. 4.

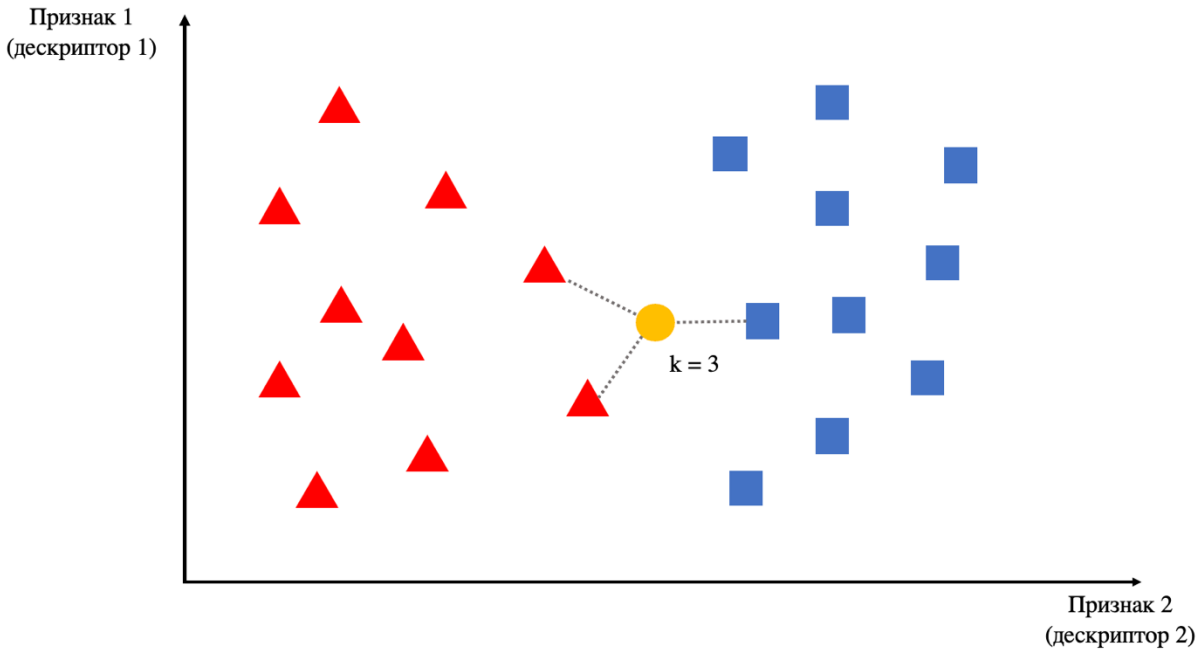


Рис. 4. Принцип работы метода k ближайших соседей в классификационной задаче [45, 46]

На рис. 4 показана диаграмма распределения объектов двух разных классов (класс «красные треугольники» и класс «синие квадраты») в зависимости от характеризующих их признаков (Признак 1 и Признак 2). Объект, для которого необходимо определить класс с помощью метода k ближайших соседей, представлен в виде желтого круга. Метод k ближайших соседей анализирует заданное число (в данном случае k = 3) ближайших к рассматриваемому объекту соседей и прогнозирует класс по наибольшему числу соседей. В данном случае класс рассматриваемого объекта – «красный треугольник» (рис. 4). При решении регрессионной задачи прогнозируемое значение рассчитывается как среднее арифметическое между значениями данного признака у k ближайших соседей [45, 46].

Многослойный перцептрон, представляющий собой искусственную нейронную сеть с одним или несколькими скрытыми слоями, в Python запускали с помощью алгоритма MLPRegressor из библиотеки scikit-learn [129].

В основе идеи работы искусственного нейрона лежит работа нейрона человеческого мозга (биологического нейрона) [49]. Биологический нейрон

состоит из ядра и отходящих от него отростков двух типов – дендритов и аксонов [49]. Входные сигналы в ядро нейрона поступают через дендриты, передача выходного сигнала происходит через аксоны [49]. Аксоны одних нейронов соединяются с дендритами других с помощью многочисленных каналов связи – синапсов [49]. То есть синапсы осуществляют передачу нервных импульсов от одних нейронов к другим [49]. Однако, не каждый нейрон принимает участие в передаче сигнала [49]. Когда совокупность электрохимических сигналов, поступающих в ядро нейрона через дендриты от соседей, превышает определенное пороговое значение, происходит активация нейрона, после которой он, в свою очередь, начинает отправлять через синапсы сигналы своим соседям, то есть нейрон способен отправлять сигналы, только если совокупность поступающих к нему сигналов будет превышать определенное пороговое значение [49]. Таким образом, возможность активации нейрона напрямую связана с сочетанием сигналов, поступающих в него в каждый конкретный момент времени [49]. Интенсивность поступающих сигналов варьируется в зависимости от активности синапса [49].

Искусственный нейрон (рис. 5) – это вычислительная единица, которая так же, как и ее биологический прототип, принимает входные сигналы $x_1, x_2, \dots, x_n$ от других нейронов [49]. В простейшей модели величины этих сигналов складываются и в случае, если их сумма превышает определенное значение b (порог активации), нейрон активируется и посылает выходной сигнал, равный 1 [49]. Если сумма сигналов меньше величины порога активации, выходной сигнал от нейрона равен 0 [49]. На практике вместо порога активации используются различные функции активации, которые осуществляют преобразование входного сигнала [49]. Для внедрения в модель идеи о возможности усиления синаптических связей между нейронами, то есть обучения нейронной сети, необходимо усложнить модель введением весовых коэффициентов w [49] В такой модели имеется возможность «усиления» или «ослабления» сигнала, передаваемого от одного нейрона к другому, как это происходит в действительности при мыслительных процессах, – каждый сигнал

характеризуется «весом» w_1, w_2, … , w_n [49]. Суммарный сигнал формируется с учетом вкладов каждого из весовых коэффициентов [49]:

$$y = \begin{cases} 1, \sum_{i=1}^{n} w_i x_i > b, \\ 0, \sum_{i=1}^{n} w_i x_i < b. \end{cases}$$

В искусственной нейронной сети нейроны собираются в слои (рис. 6), поэтому полученные от нейронов одного слоя выходные сигналы становятся входными для нейронов следующего слоя, в котором происходит аналогичная операция [49].

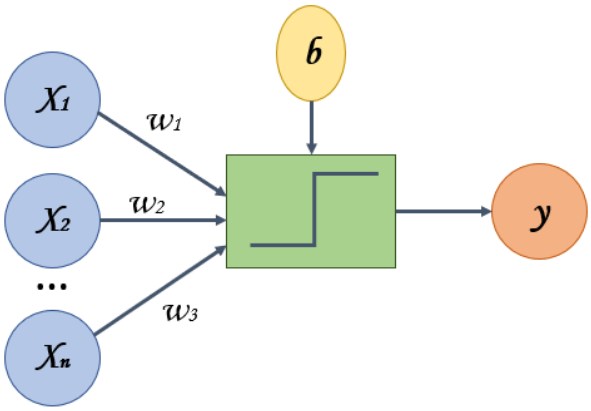


Рис. 5. Схематичное изображение искусственного нейрона: x_1, x_2, … , x_n – входные сигналы; w_1, w_2, … , w_n – весовые коэффициенты входных сигналов; b – порог активации; y – суммарный выходной сигнал из нейрона [49]

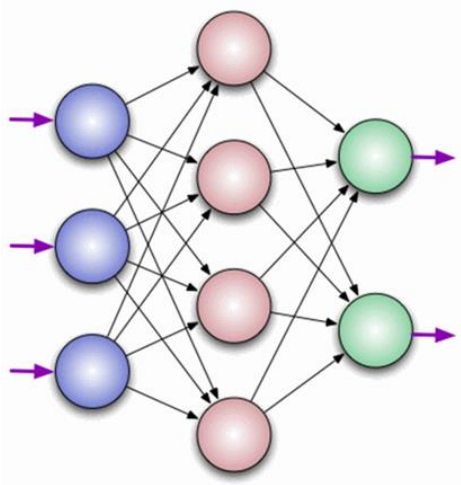


Рис. 6. Схематичное изображение строения нейронной сети: синие круги – нейроны входного слоя, зеленые круги – нейроны выходного слоя, розовые круги – нейроны скрытого слоя [49]

Каждый из этих алгоритмов имеет гиперпараметры [45, 46, 48, 49]. Гиперпараметры – это набор параметров метода машинного обучения, которые контролируют процесс обучения модели и которые можно настраивать для улучшения точности модели [45, 46, 48, 49]. В разных задачах оптимальные

значения гиперпараметров могут отличаться, поэтому их подбор является важным этапом моделирования [45, 46, 48, 49].

Для метода машинного обучения, продемонстрировавшего максимальную точность прогнозов среди испытанных (метод случайного леса – см. главу 3), провели подбор гиперпараметров с помощью алгоритма GridSearchCV из библиотеки scikit-learn [129]. GridSearchCV систематически перебирает все возможные комбинации заданных гиперпараметров, оценивая качество модели для каждого варианта [129].

2.4 Этапы моделирования связи «структура – температура стеклования органических гомополимеров» на основе машинного обучения

Моделирование проводили в два этапа. На первом этапе моделирования ставили цель проверить, можно ли прогнозировать температуру стеклования органических гомополимеров с использованием модели, построенной на основе машинного обучения, не напрямую, а через параметры, определяющие температуру стеклования органических гомополимеров по аналогии с инкрементальным подходом. То есть вместо прямого прогнозирования температуры стеклования органических гомополимеров, исследовали возможность прогнозирования параметров, от которых она зависит. Далее прогнозируемые значения параметров использовали для расчета температуры стеклования органических гомополимеров по формуле (1).

На втором этапе моделирования проводили уточнение значений параметров с использованием комбинированных дескрипторов для обеспечения более достоверных прогнозов температуры стеклования органических гомополимеров. Исходная база данных содержала изомеры с различным положением заместителя в ароматическом кольце. Инкрементальный подход не учитывает положения заместителей в ароматическом кольце (табл. 6), именно поэтому на втором этапе моделирования проводили процедуру уточнения.

Таблица 6

Значения температуры стеклования, рассчитанные в рамках инкрементального подхода по формуле (1), и их экспериментальные значения для органических полимеров изомерного ряда полихлорстирола [99]

Органический гомополимер	Температура стеклования, рассчитанная по формуле (1), K	Экспериментальная температура стеклования, K
поли-2-хлорстирол	410	392
поли-3-хлорстирол	410	363
поли-4-хлорстирол	410	388-401

Под термином «уточнение» понимается присвоение новых значений параметрам, определяющим температуру стеклования полимеров по аналогии с инкрементальным подходом и обеспечивающим более достоверный прогноз температуры стеклования.

2.5 Расчет квантово-химических параметров, относящихся к повторяющимся звеньям органических гомополимеров

В качестве объектов исследования на этапе установления корреляций уточненных параметров, определяющих температуры стеклования органических гомополимеров по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям полимеров, выбрали полистиролы с различными положениями (2-, 3-, 4-) заместителя (фтор-, хлор-, бром-, метил- и этил-) в ароматическом кольце.

В качестве моделей молекулярной структуры объектов исследования в квантово-химических расчетах использовали повторяющиеся звенья, в которых открытую валентность концевых групп закрывали атомами водорода [130].

Оптимизацию структур объектов исследования выполняли в программном пакете Gaussian 16, Rev. C.01 [131] с использованием гибридного функционала

B3LYP [132, 133] и валентно-расщепленного базисного набора Попла 6-31G(d,p) [134]. Стабильность рассчитанных структур характеризовалась отсутствием отрицательных частот колебаний в матрице вторых производных.

Параметры, с которыми ожидалась корреляция, получили из результатов квантово-химических расчетов по следующим формулам.

1. Средняя поляризуемость (α) – характеризует способность атомов или молекул деформировать свою электронную оболочку под воздействием внешнего электрического поля [107]:

$$\alpha = \frac{\alpha_{xx} + \alpha_{yy} + \alpha_{zz}}{3},$$

где α_{xx}, α_{yy}, α_{zz} – поляризуемость молекулы по x, y, z координатам, а.е.

2. Дипольный момент (d) – характеризует распределение электрического заряда в молекуле [130]:

$$d = (d^2_x + d^2_y + d^2_z)^{0.5},$$

где d_x, d_y, d_z – дипольный момент молекулы по x, y, z координатам, Д.

3. Потенциал ионизации (I) – энергия, которая необходима для удаления электрона из нейтрального атома или молекулы [135]:

$$I = -E_{ВЗМО}, \text{ эВ},$$

где $E_{ВЗМО}$ – энергия высшей занятой молекулярной орбитали (ВЗМО), эВ.

4. Сродство к электрону (*a*) характеризует изменение энергии при присоединении электрона к нейтральному атому или молекуле с образованием отрицательного иона [135]:

$$a = -E_{НСМО}, \text{ эВ},$$

где $E_{НСМО}$ – энергия низшей свободной молекулярной орбитали (НСМО), эВ.

5. Энергетический зазор между ВЗМО и НСМО (E_g) дает представление о химической стабильности структуры молекулы, ее реакционной способности и потенциале процессов переноса электронов [135]:

$$E_g = I - a, \text{ эВ}.$$

6. Химический потенциал (μ) характеризует тенденцию электронов равномерно распределяться между атомами в молекуле [135]:

$$\mu = -(I + a)/2, \text{ эВ}.$$

7. Химическая жесткость (η) количественно отражает сопротивление молекулы переносу электронов между орбиталями [135]:

$$\eta = E_g/2, \text{ эВ}.$$

8. Электрофильность (ω) характеризует способность атомов в молекуле принимать электроны [135]:

$$\omega = \mu^2/2\eta, \text{ эВ}.$$

Молекулярные объемы (V_M, $см^3$/моль) оптимизированных структур повторяющихся звеньев органических гомополимеров рассчитывали с использованием метода Монте-Карло [130] в программном пакете Gaussian 16, Rev. C.01 [131].

Степень корреляции параметров оценивали с помощью коэффициента Пирсона по формуле [136]:

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 + (y_i - \bar{y})^2}}, \quad (2)$$

где x_i – значения переменной x; y_i – значения переменной y; $\bar{x}$ – среднее арифметическое переменной x; $\bar{y}$ – среднее арифметическое переменной y.

ГЛАВА 3 РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1 Построение модели на основе методов машинного обучения, способной прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее по аналогии с инкрементальным подходом

Формулу (1) преобразовали к следующему виду:

$$T_{g\,inc} = \frac{A_{inc}}{B_{inc} + C_{inc}}, \quad (3)$$

где $A_{inc} = \sum_i \Delta V_i$, $B_{inc} = \sum_i a_i \Delta V_i$, $C_{inc} = \sum_j b_j$.

Модель на основе методов машинного обучения должна прогнозировать температуру стеклования органических гомополимеров, исходя из параметров, определяющих ее в рамках инкрементального подхода:

$$T_{g\,calc} = \frac{A}{B + C}, \quad (4)$$

где $A \approx A_{inc}$, $B \approx B_{inc}$, $C \approx C_{inc}$. В формуле (4) параметры A, B, C на первом этапе моделирования по физическому смыслу соответствуют A_{inc}, B_{inc}, C_{inc}, однако численно им не равны.

Перед моделированием базу данных разделили на обучающую выборку (80% полимеров из базы данных) и тестовую выборку (20% полимеров из базы данных). Такое разделение позволило не только обучить модель на достаточном количестве данных, но и независимо оценить ее способность к обобщению на новых, ранее не используемых данных. Обучение модели осуществляли на обучающей выборке, а тестовую выборку использовали для проверки точности прогнозирования параметров, определяющих температуру стеклования полимеров в рамках инкрементального подхода.

Для оценки точности прогнозирования температуры стеклования моделью, построенной на основе машинного обучения, проводили сравнение экспериментальных значений температуры стеклования органических

гомополимеров $T_{g\ exp}$ и соответствующих значений $T_{g\ calc}$ с использованием коэффициента детерминации R^2. Алгоритм моделирования на первом этапе схематически представлен на рис. 7.

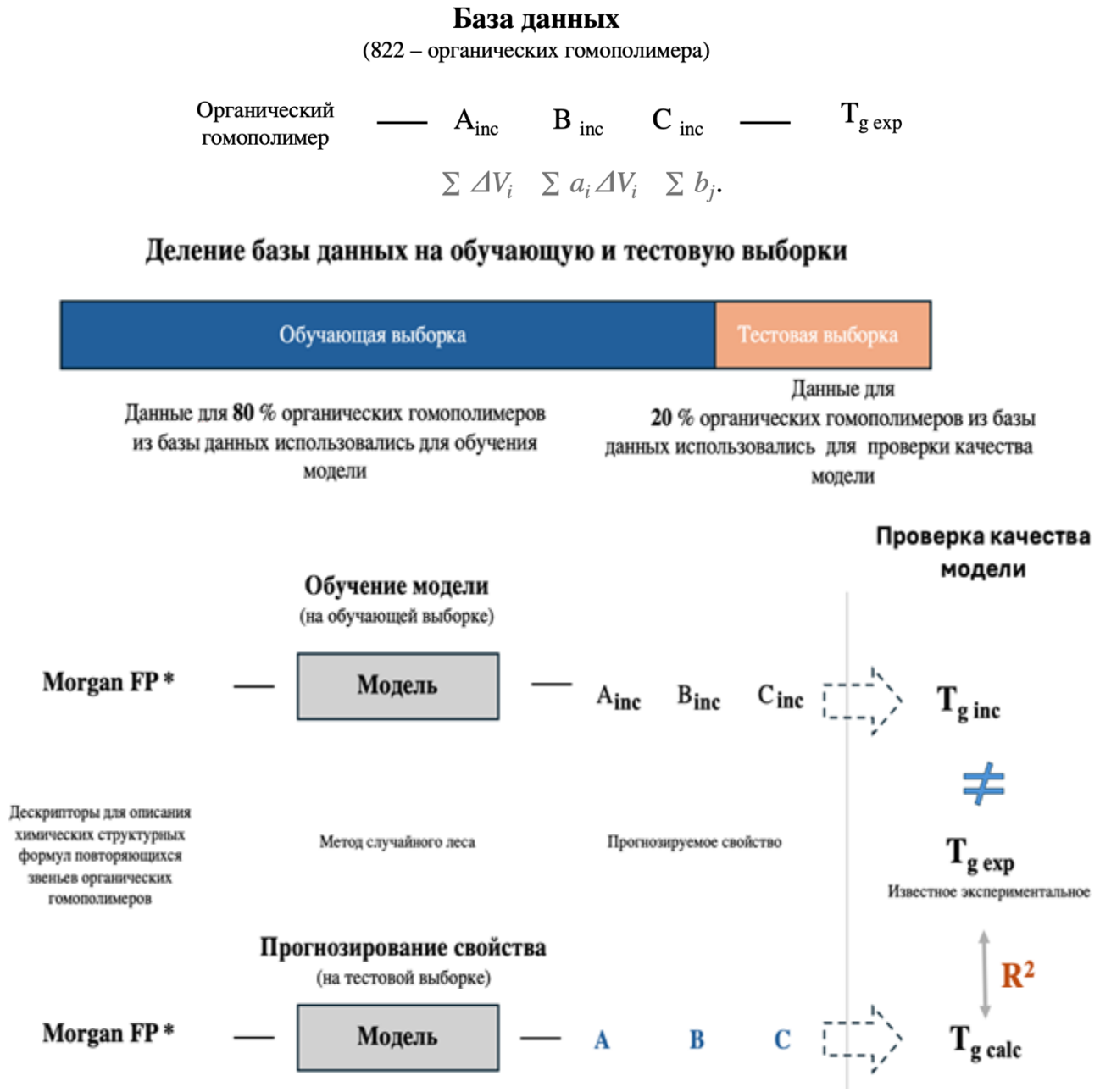


Рис. 7. Схема первого этапа моделирования связи «структура – температура стеклования органических гомополимеров» на основе машинного обучения

Значения A_{inc}, B_{inc}, C_{inc} рассчитывали по формулам, приведенным в работе [99]. Для повышения скорости расчетов реализовали алгоритм поиска фрагментов (атомов и их окружения) в повторяющемся звене органического гомополимера. В основе алгоритма лежит кодирование и распознавание структуры этих

фрагментов в формате SMARTS (от англ. SMiles ARbitrary Target Specification – основанный на SMILES язык, позволяющий специфицировать фрагменты внутри молекул при помощи шаблонов) [137], который обеспечивает гибкость и точность при описании химических структур. Следует отметить, что SMARTS [137] имеет ограничения, связанные с неоднозначной идентификацией атомов с переменной валентностью, таких как атомы серы, что может приводить к некорректной интерпретации структуры полимера. Для минимизации риска получения ложноположительных результатов и обеспечения надежности последующего анализа данных было принято решение об исключении из исходной базы данных 228 полимеров, структура которых могла быть интерпретирована неоднозначно. В результате данной процедуры основная выборка для дальнейшего исследования составила 822 гомополимера.

Экспериментальные значения температуры стеклования полимеров в базе данных получены с использованием различных экспериментальных методов, таких как термомеханический анализ, дилатометрия и др. Важно отметить, что указанные методы характеризуются различной степенью точности и специфическими особенностями определения температуры стеклования полимеров, которые могут влиять на получаемые результаты. Например, при использовании термомеханического метода температура стеклования полимера зависит от скорости изменения температуры [78]. Кроме того, даже в рамках одного метода значение температуры стеклования полимера может варьироваться в зависимости от условий проведения эксперимента [78]. В качестве примера можно привести метод радиотермолюминесценции, в котором положение переходов фиксируется с высокой точностью (до 1 K), в то время как при измерении механических свойств область стеклования полимера лежит в пределах 10-15 K в зависимости от используемого оборудования и параметров эксперимента [78].

Проблема отсутствия единообразных и стандартизированных баз данных с экспериментальными значениями температуры стеклования полимеров является недостатком большинства работ в данной области, который редко упоминается

авторами. Однако в рамках настоящей диссертации объем базы данных и разнообразие использованных экспериментальных методов формирует дисперсию, а не выбросы, что позволяет модели выстраивать корректные зависимости между структурой полимера и температурой стеклования.

Нормальное (гауссово) распределение данных полагается важным условием для высокой прогностической способности модели на основе машинного обучения [33]. Анализ распределения значений температуры стеклования полимеров в исходной базе данных [99] показал, что они соответствуют закону нормального распределения (рис. 8).

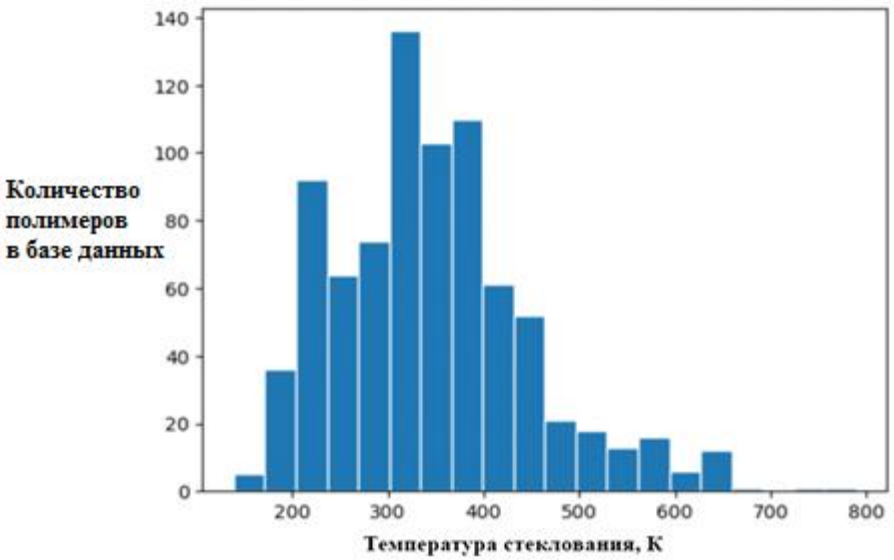


Рис. 8. Распределение значений температуры стеклования полимеров в базе данных

Сравнительная характеристика (без оптимизации гиперпараметров) методов машинного обучения, использованных в модели для прогнозирования параметров, которые определяют температуру стеклования органических гомополимеров в рамках инкрементального подхода, представлена в табл. 7. Точность модели, оцененную с помощью коэффициента детерминации R^2, определяли путем сравнения значений A, B и C, спрогнозированных моделью на основе машинного обучения, со значениями A_{inc}, B_{inc} и C_{inc}, рассчитанными в рамках инкрементального подхода.

Таблица 7

Сравнительная характеристика (без оптимизации гиперпараметров) методов машинного обучения, использованных в модели для прогнозирования параметров, которые определяют температуру стеклования органических гомополимеров в рамках инкрементального подхода

Метод машинного обучения	Дескрипторы	R^2
Метод случайного леса	молекулярные отпечатки Моргана	0.739
	структурные ключи	0.757
Метод k ближайших соседей	молекулярные отпечатки Моргана	0.660
	структурные ключи	0.690
Многослойный перцептрон	молекулярные отпечатки Моргана	0.490
	структурные ключи	0.450

Анализ результатов, представленных в табл. 7, демонстрирует, что наибольшую точность прогнозирования температуры стеклования органических гомополимеров продемонстрировала модель, основанная на методе случайного леса. Данный метод, являясь ансамблевым методом машинного обучения, эффективно справляется с задачами регрессии и классификации, обеспечивая высокую устойчивость к переобучению и способность выявлять нелинейные зависимости между дескрипторами и целевой переменной [48]. При использовании структурных ключей в качестве дескрипторов точность модели, оцененная с помощью коэффициента детерминации R^2, составила 0.757, а при использовании молекулярных отпечатков Моргана – 0.739. Полученные значения R^2 свидетельствуют о высокой прогностической способности модели на основе метода случайного леса.

Метод k ближайших соседей продемонстрировал значения коэффициента детерминации R^2, равные 0.69 и 0.66 при использовании структурных ключей и молекулярных отпечатков Моргана в качестве дескрипторов соответственно (табл. 7). Полученные результаты свидетельствуют о приемлемом уровне

точности прогнозирования температуры стеклования органических гомополимеров с использованием данного метода. Следует отметить, что точность метода k ближайших соседей существенно уступает показателям, полученным с использованием метода случайного леса, что может быть обусловлено особенностями метода k ближайших соседей, такими как чувствительность к выбору метрики расстояния и необходимость оптимизации параметра k (количество ближайших соседей).

Многослойный перцептрон, представляющий собой один из наиболее распространенных типов нейронных сетей, показал значительно более низкие значения R^2 – 0.45 и 0.49 для структурных ключей и молекулярных отпечатков Моргана соответственно (табл. 7). Более низкая точность многослойного перцептрона по сравнению с другими рассмотренными методами машинного обучения может быть связана с тем, что нейронные сети, как правило, требуют гораздо большего объема и разнообразия данных для достижения качественного обучения [49]. Кроме того, для достижения оптимальной производительности нейронных сетей необходим более тщательный подбор архитектуры сети, включая количество слоев, число нейронов в каждом слое, а также параметров обучения, таких как скорость обучения и функция активации [49].

Таким образом, метод случайного леса оказался наиболее эффективным для прогнозирования параметров, определяющих температуру стеклования органических гомополимеров в рамках инкрементального подхода. В связи с этим, на следующих этапах исследования использовали модель на основе метода случайного леса.

Далее с использованием модели на основе метода случайного леса и молекулярных отпечатков Моргана рассчитали $T_{g\ calc}$, исходя из параметров A, B и C, спрогнозированных совместно, и провели прямое прогнозирование параметров A, B и C (по отдельности) – см. табл. 8.

Таблица 8

Коэффициенты детерминации R^2 для различных параметров, прогнозируемых моделью на основе метода случайного леса и молекулярных отпечатков Моргана

Прогнозируемый параметр	R^2
A, B, C	0.739
$T_{g\ calc}$	0.770 (с $T_{g\ inc}$) 0.810 (с $T_{g\ exp}$)
A	0.770
B	0.730
C	0.720

Использование прямого прогнозирования параметров A, B и C (по отдельности) не приводит к существенному повышению точности прогнозируемых значений A, B и C по сравнению с совместным прогнозированием параметров A, B и C (табл. 8). Это может быть связано с тем, что прямое прогнозирование параметров A, B и C (по отдельности) не учитывает возможную взаимосвязанность между ними.

В связи с этим для достижения более высокой точности прогнозирования температуры стеклования органических гомополимеров приняли решение отказаться от использования подхода, основанного на прямом прогнозировании параметров A, B и C (по отдельности), и далее сосредоточиться только на совместном прогнозировании параметров A, B и C (такой подход позволяет учитывать возможную взаимосвязанность между параметрами A, B и C и, следовательно, более точно оценивать вклад каждого структурного фрагмента повторяющегося звена в температуру стеклования полимера). После чего с целью повышения коэффициента детерминации R^2 на обучающей выборке провели процедуру оптимизации гиперпараметров модели на основе метода случайного леса (табл. 9).

Таблица 9

Гиперпараметры метода случайного леса

Гиперпараметр	Описание	Оптимальное значение
'bootstrap'	Критерий обучающей выборки: True – каждое дерево обучается на случайной подвыборке данных (с повторениями); False – все деревья обучаются на полной базе данных.	False
'criterion'	Функция качества разделения узлов в деревьях.	'squared_error' (минимизирует средний квадрат ошибок)
'max_depth'	Максимальная глубина деревьев (максимальное количество уровней узлов).	None
'max_features'	Число признаков для выбора при разделении.	'sqrt'
'min_samples_leaf'	Минимальное число образцов для разделения листа.	1
'min_samples_split'	Минимальное число образцов в листе.	4
'n_estimators'	Количество деревьев принятия решений в лесу.	500

Несмотря на то, что процедура оптимизации гиперпараметров модели на основе метода случайного леса обеспечила лишь незначительное увеличение точности прогнозирования температуры стеклования для всей выборки органических гомополимеров (увеличение коэффициента детерминации R^2 на 0.01), она позволила существенно улучшить точность прогнозирования $T_{g\ calc}$ для отдельных полимеров. Данный факт указывает на наличие потенциальной возможности дальнейшего повышения точности модели за счет проведения

индивидуальной настройки гиперпараметров для различных групп полимеров в пределах одного класса (органических гомополимеров).

Для более детальной и наглядной оценки качества прогнозов разработанной модели на конкретных примерах выбрали полистиролы с различными положениями (2-, 3-, 4-) заместителя (фтор-, хлор-, бром-, метил- и этил-) в ароматическом кольце (табл. 10). Данный выбор обусловлен тем, что разное положение заместителя в ароматическом кольце обеспечивает разные значения температуры стеклования изомеров полистирола.

Таблица 10

Результаты прогнозирования температуры стеклования для полистиролов с различными положениями заместителя в ароматическом кольце

Органический гомополимер	$T_{g\,exp}$, K [99]	$T_{g\,inc}$, K [99]	$T_{g\,calc}$, K			
			молекулярные отпечатки Моргана до оптимизации гиперпараметров	молекулярные отпечатки Моргана после оптимизации гиперпараметров	структурные ключи до оптимизации гиперпараметров	структурные ключи после оптимизации гиперпараметров
поли-2-фторстирол (П-2-ФС)	-	402	400.6	402.1	400.8	402.1
поли-3-фторстирол (П-3-ФС)	-	402	397.8	397.3	400.8	402.1
поли-4-фторстирол (П-4-ФС)	368	402	384.9	383.2	395.2	396.0
поли-2-хлорстирол (П-2-ХС)	392	410	405.0	402.7	405.1	405.3
поли-3-хлорстирол (П-3-ХС)	363	410	400.0	376.0	400.0	403.0
поли-4-хлорстирол (П-4-ХС)	388-401	410	403.0	400.0	403.0	403.0
поли-2-бромстирол (П-2-БрС)	-	423	402.3	397.8	389.4	401.4
поли-3-бромстирол (П-3-БрС)	-	423	404.3	398.6	413.6	422.6
поли-4-бромстирол (П-4-БрС)	391, 414-430	423	385.3	381.6	376.7	381.1
поли-2-метилстирол (П-2-МеС)	409	400	394.1	375.5	396.0	399.9
поли-3-метилстирол (П-3-МеС)	370	400	395.5	403.2	396.4	400.0
поли-4-метилстирол (П-4-МеС)	366, 374, 382	400	386.1	399.1	388.8	400.0
поли-2-этилстирол (П-2-ЭтС)	376	354	355.8	356.3	298.9	339.5
поли-3-этилстирол (П-3-ЭтС)	> 303	354	350.7	355.8	364.6	367.7
поли-4-этилстирол (П-4-ЭтС)	300, 351	354	357.7	345.3	364.6	367.6

Спрогнозированные исходя из структурных ключей значения температуры стеклования являются практически одинаковыми для изомеров, отличающихся положением заместителя в ароматическом кольце (табл. 10). Это связано с тем, что структурные ключи в отличие от молекулярных отпечатков Моргана предоставляют информацию только о наличии или отсутствии определенных структурных фрагментов в молекуле, не учитывая их взаимное расположение, относительную ориентацию и другие важные структурные особенности [137]. То есть структурные ключи не позволяют различить рассматриваемые изомеры. Модель, использующая молекулярные отпечатки Моргана, обладает большей точностью прогнозирования температуры стеклования для рассматриваемых изомеров полистиролов (табл. 10), поскольку молекулярные отпечатки Моргана отражают связность атомов и их окружения [137], а значит, модель способна различать положение заместителя в ароматическом кольце и учитывать его влияние на температуру стеклования. Таким образом, несмотря на то, что общая точность прогнозов модели, использующей молекулярные отпечатки Моргана, несколько ниже по сравнению с моделью, использующей структурные ключи (табл. 7), использование для моделирования молекулярных отпечатков Моргана предпочтительнее из-за их возможности учитывать нюансы строения изомерных молекул.

Стоит отметить интересное наблюдение – температура стеклования $T_{g\ calc}$, рассчитанная с использованием совместно прогнозируемых моделью значений параметров A, B и C, имеет бóльшую корреляцию с экспериментальными значениями температуры стеклования $T_{g\ exp}$, чем с температурой стеклования $T_{g\ inc}$, рассчитанной в рамках инкрементального подхода (табл. 8). Данный факт дает основания полагать, что модель при совместном прогнозировании параметров A, B и C может учитывать определенные закономерности структуры повторяющегося звена, изменяя исходные, заложенные инкрементальным подходом, физические смыслы этих параметров, за счет чего увеличивается точность модели в отношении прогнозирования экспериментальных значений температуры стеклования органических гомополимеров, а не ее значений,

полученных в рамках инкрементального подхода. Эта гипотеза стала основой при реализации второго этапа моделирования с помощью модели на основе метода случайного леса при использовании молекулярных отпечатков Моргана в качестве базовых дескрипторов.

3.2 Уточнение значений параметров, определяющих температуру стеклования органических гомополимеров по аналогии с инкрементальным подходом, с использованием при моделировании на основе машинного обучения комбинированных дескрипторов

Точность модели на основе метода случайного леса при использовании молекулярных отпечатков Моргана в качестве дескрипторов оказалась несколько ниже по сравнению с традиционным инкрементальным подходом: коэффициент детерминации R^2 для $T_{g\ calc}$ (с $T_{g\ exp}$) при совместном прогнозировании параметров A, B, C составляет 0.81 (табл. 8), а для $T_{g\ inc}$ (с $T_{g\ exp}$) – 0.90. Это объясняется рядом факторов. Прежде всего, необходимо отметить, что результаты прогнозирования модели на основе метода случайного леса содержат как погрешность, связанную с допущениями и ограничениями, которые характерны для количественных моделей «структура-свойство», так и погрешность инкрементального подхода, на данные которого опирается модель при обучении. Тем не менее, несмотря на более низкую точность модели на основе метода случайного леса, полученные результаты продемонстрировали принципиальную возможность использования модели для прогнозирования температуры стеклования органических гомополимеров не напрямую, а опосредованно – через ее зависимость от параметров A, B и C. Такой подход позволяет реализовать концепцию модели с интерпретируемыми параметрами, в которой каждый параметр имеет ясный смысл.

На втором этапе моделирования провели процедуру уточнения параметров A, B и C, то есть присвоения параметрам A, B и C новых значений, которые позволят получить по формуле (4) более точные значения температур стеклования

органических гомополимеров. Алгоритм процедуры уточнения параметров A, B и C схож с алгоритмом перекрестной проверки (кросс-валидации). В процессе кросс-валидации база данных делится на k частей, и обучение проводится k раз [50]. Каждая из k частей один раз становится тестовой выборкой, в то время как остальные части используются для обучения модели [50]. Кросс-валидация позволяет избежать случайных ошибок, связанных с завышением или занижением качества модели, и дает возможность использовать все данные в базе и для обучения, и для тестирования [50].

В настоящей диссертации алгоритм кросс-валидации использовали для последовательного уточнения параметров A, B и C. Для этого базу данных разделили на выборку для уточнения (90% базы данных) и тестовую выборку (10% базы данных), которая не участвовала в процедуре уточнения. Выборка для уточнения, как и в общем случае кросс-валидации, делилась на k = 8 частей (на рис. 9 представлен принцип деления базы данных для уточнения параметров A, B и C при k = 5).

В процедуре уточнения параметров A, B и C использовали два варианта комбинированных дескрипторов. Каждый комбинированный дескриптор, построенный по первому варианту, представлял собой вектор, состоящий из битовой строки молекулярного отпечатка Моргана, экспериментального $T_{g\ exp}$ и рассчитанного по формуле (3) $T_{g\ inc}$ значений температуры стеклования органических гомополимеров (рис. 10). Разность между $T_{g\ exp}$ и $T_{g\ inc}$ давала модели информацию о «степени неточности» инкрементального подхода. Комбинированные дескрипторы, построенные по первому варианту, использовали в качестве входных значений модели в обучающей выборке.

Комбинированные дескрипторы, построенные по второму варианту, отличались от комбинированных дескрипторов, построенных по первому варианту, использованием экспериментального значения $T_{g\ exp}$ на месте значения $T_{g\ inc}$ (рис. 10). Комбинированные дескрипторы, построенные по второму варианту, использовали в качестве входных значений модели в выборке для уточнения.

Рис. 9. Принцип деления базы данных для уточнения параметров A, B и C

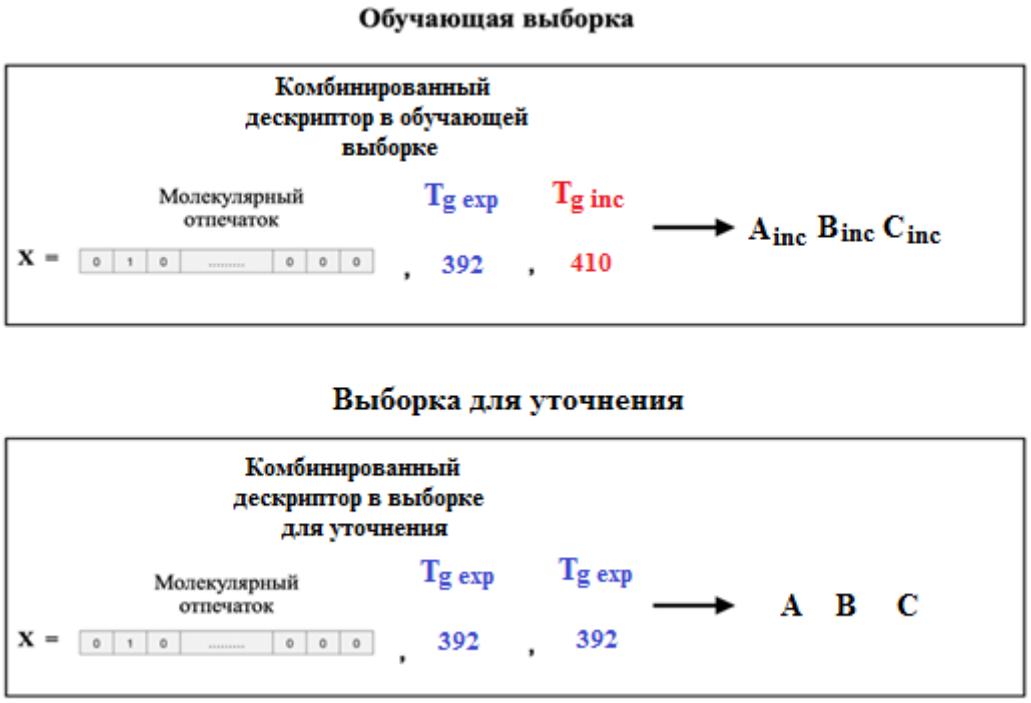


Рис. 10. Комбинированный дескриптор

Таким образом, прогнозируемыми параметрами модели на основе метода случайного леса становятся такие значения параметров A, B и C, которые

обеспечивают минимальное расхождение между экспериментальным значением $T_{g\ exp}$ и значением $T_{g\ inc}$, рассчитанным по формуле (3).

По результатам процедуры уточнения параметров A, B и C получили обновленную базу данных. Проверку точности прогнозирования модели на втором этапе моделирования проводили на заранее выделенной тестовой выборке (рис. 9). Сравнение баз данных первого и второго этапа моделирования приведено на рис. 11.

Рис. 11. Сравнение баз данных первого и второго этапов моделирования

Использование одной и той же базы данных приводит к различным результатам моделирования при перераспределении входящих в нее органических гомополимеров между обучающей и тестовой выборками. Поэтому для получения статистически достоверных результатов провели 50 независимых вычислительных экспериментов, заключающихся в повторении второго этапа моделирования: снова случайным образом выборка для уточнения (составляет 90% от исходной базы данных) делилась на k частей, далее проводилась процедура уточнения A, B и C, на основе результатов которой обучалась модель для более точного прогнозирования температуры стеклования органических гомополимеров. По результатам проведенных вычислительных экспериментов

получили 50 наборов прогнозируемых параметров A, B и C. Для дальнейшей интерпретации результатов использовали медианные значения полученных параметров. В ходе 50 независимых вычислительных экспериментов обнаружили, что в ряде случаев коэффициент детерминации R^2 на тестовой выборке достигал 0.90, что соответствует уровню точности результатов расчета с использованием инкрементального подхода. При этом среднее значение R^2 по результатам всех экспериментов составило 0.85, что свидетельствует о существенном повышении точности модели по сравнению с исходным вариантом (R^2 = 0.81, см. табл. 8). Сопоставление значений температуры стеклования полистиролов с различными положениями заместителя в ароматическом кольце:

- экспериментальных,

- рассчитанных с помощью инкрементального подхода,

- рассчитанных с помощью модели на основе метода случайного леса при использовании молекулярных отпечатков Моргана в качестве дескрипторов и совместно прогнозируемых уточненных параметров A, B и C, – приведено в табл. 11.

Таблица 11

Значения температуры стеклования полистиролов с различными положениями заместителя в ароматическом кольце: экспериментальные ($T_{g\ exp}$), рассчитанные с помощью инкрементального подхода ($T_{g\ inc}$) и рассчитанные с помощью модели на основе метода случайного леса при использовании молекулярных отпечатков Моргана в качестве дескрипторов и совместно прогнозируемых уточненных параметров A, B и C ($T_{g\ calc}$)

Органический гомополимер	$T_{g\ exp}$, K [99]	$T_{g\ inc}$, K [99]	$T_{g\ calc}$, K
П-2-ФС		402	384
П-3-ФС		402	389
П-4-ФС	368	402	380
П-2ХС	392	410	397
П-3ХС	363	410	387
П-4ХС	388-401	410	397
П-2-БрС		423	398
П-3-БрС		423	398
П-4-БрС	391, 414-430	423	422
П-2-МеС	409	400	399
П-3-МеС	370	400	380
П-4-МеС	366, 374, 382	400	375
П-2-ЭтС	376	354	365
П-3-ЭтС	> 303	354	332
П-4-ЭтС	300, 351	354	335

Модели, построенные без учета и с учетом положения заместителя, имеют коэффициенты детерминации 0.12 ($T_{g\ exp}$ с $T_{g\ inc}$) и 0.81 ($T_{g\ exp}$ с $T_{g\ calc}$) соответственно (табл. 11).

По результатам проведенной процедуры уточнения параметров A, B и C обнаружили, что соотношение между значениями параметров B и C существенно изменилось по сравнению с исходным соотношением между значениями параметров B_{inc} и C_{inc}, полученными в рамках инкрементального подхода (значения параметров A_{inc}, B_{inc} и C_{inc} показаны на рис. 12-16 красной пунктирной линией). Данный факт может свидетельствовать о том, что в результате уточнения произошло изменение факторов, оказывающих влияние на значения параметров

A, B и C, и, как следствие, изменился их смысл. Предположительно, это связано с тем, что разработанная модель на основе метода случайного леса в отличие от инкрементального подхода учитывает не только наличие или отсутствие определенных структурных фрагментов, но и их взаимное расположение, а также другие структурные особенности повторяющихся звеньев органических гомополимеров. Это приводит к тому, что параметры A, B и C начинают отражать не только вклад отдельных структурных фрагментов, но и влияние их взаимодействий на температуру стеклования органических гомополимеров.

Далее провели статистический анализ с использованием диаграммы размаха (также известной как «ящик с усами», от англ. box and whisker plot). Данный тип диаграммы выбрали для визуализации распределения спрогнозированных значений A, B и C на примере полистиролов с различными положениями заместителя в ароматическом кольце (рис. 12-16).

Диаграмма размаха наглядно представляет основные статистические характеристики распределения данных: медиану, квартили и выбросы. Непосредственно сама диаграмма изображается в виде прямоугольного «ящика», верхняя граница которого соответствует верхнему квартилю (75-му процентилю), то есть 75% значений в выборке находятся ниже данного уровня (рис. 12, поз. 1), а нижняя граница – нижнему квартилю (25-му процентилю), то есть 25% значений в выборке находятся выше данного уровня (рис. 12, поз. 2). Горизонтальная линия, расположенная внутри «ящика» (рис. 12, поз. 3), – это медиана распределения, которая является значением, разделяющим выборку на две равные части. «Усы» (рис. 12, поз. 4), выходящие из «ящика» вверх и вниз, отражают полную дисперсию значений в выборке, исключая выбросы. Выбросы визуализируются в виде точек (рис. 12-16).

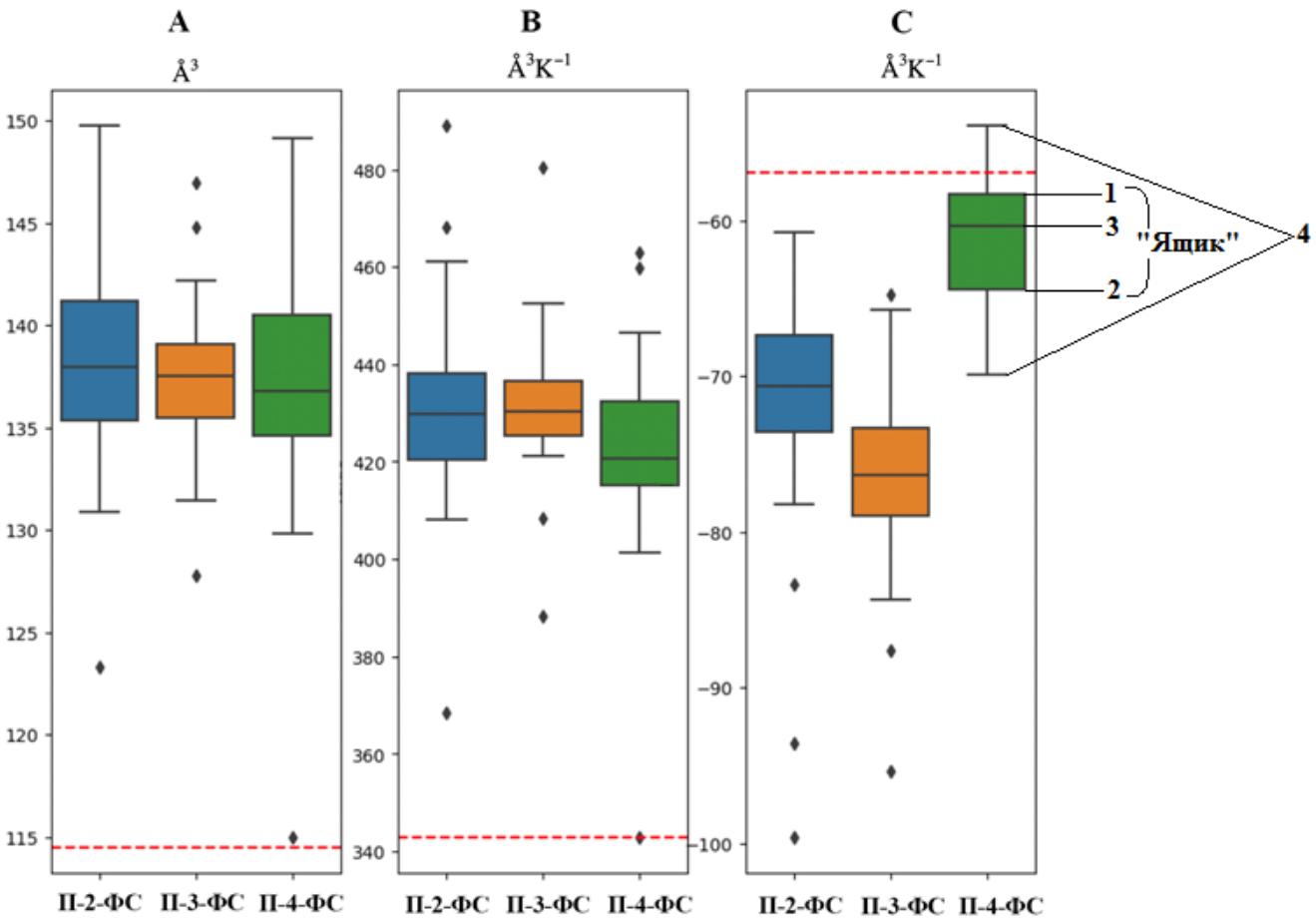


Рис. 12. Диаграмма размаха значений параметров A, B и C для фторзамещенных полистиролов (здесь и далее: 1, 2 – верхний и нижний квартили соответственно; 3 – медиана; 4 – «усы»; красная пунктирная линия – значения A_{inc}, B_{inc} и C_{inc})

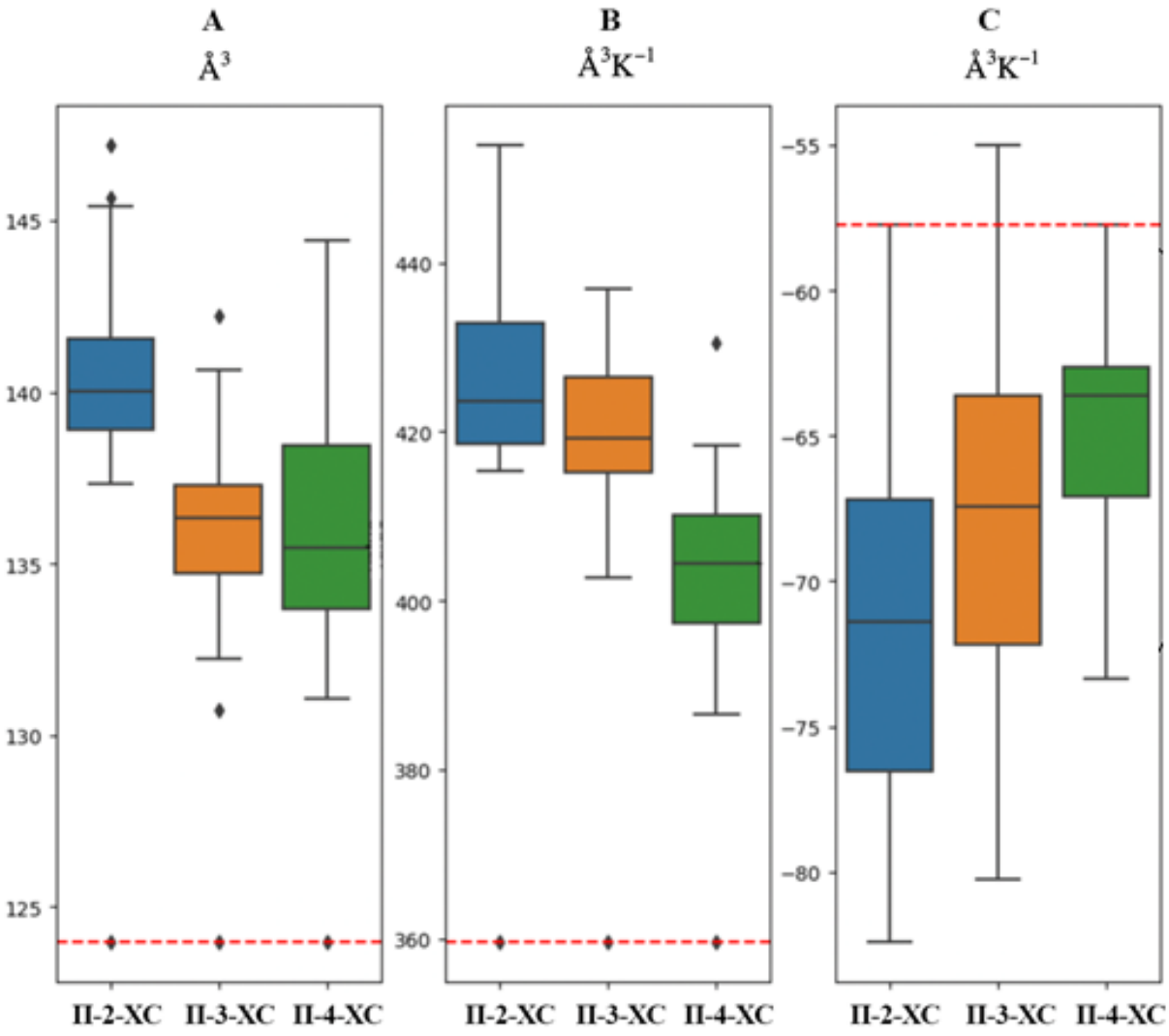


Рис. 13. Диаграмма размаха значений параметров A, B и C для хлорзамещенных полистиролов

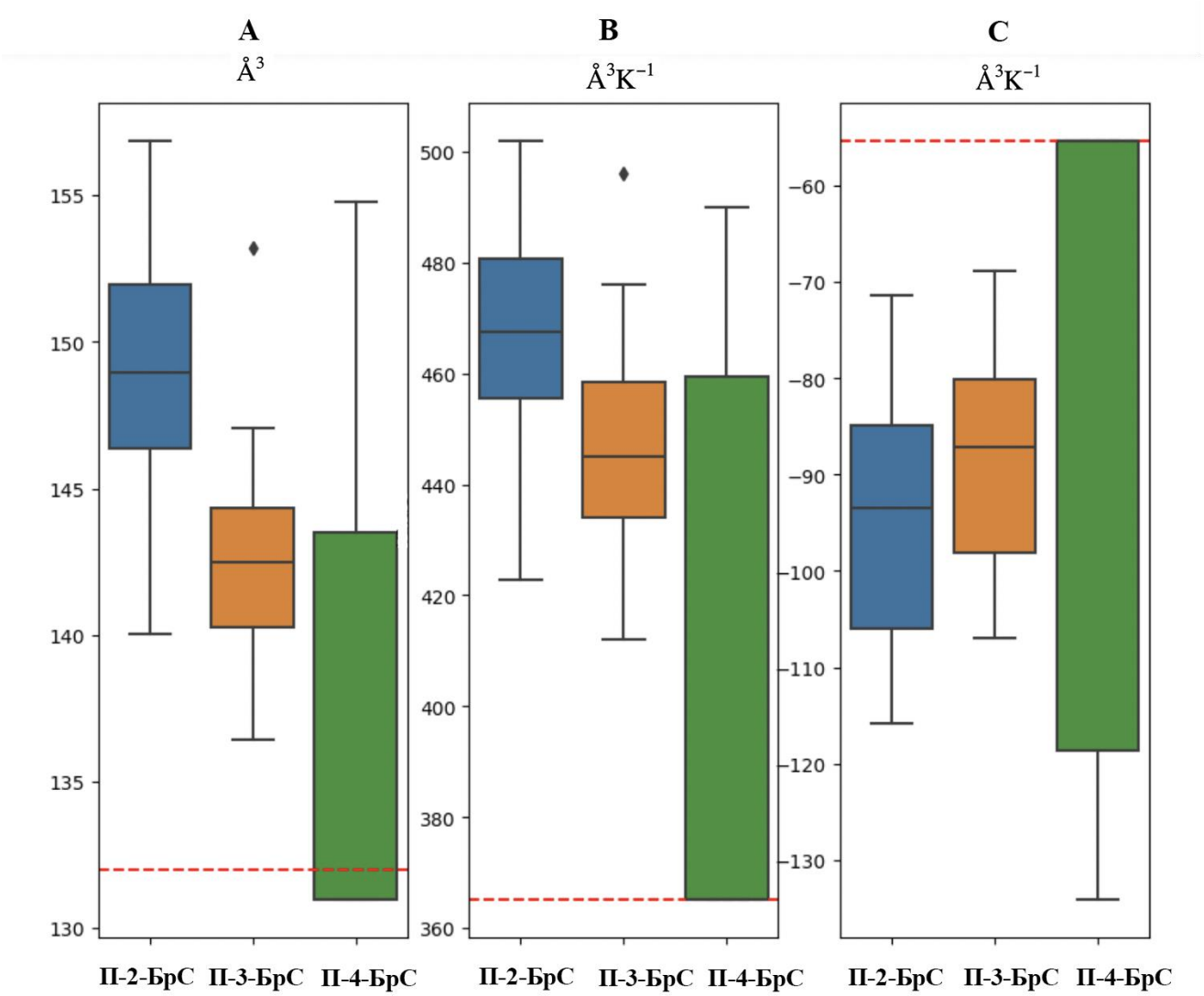


Рис. 14. Диаграмма размаха значений параметров A, B и C для бромзамещенных полистиролов

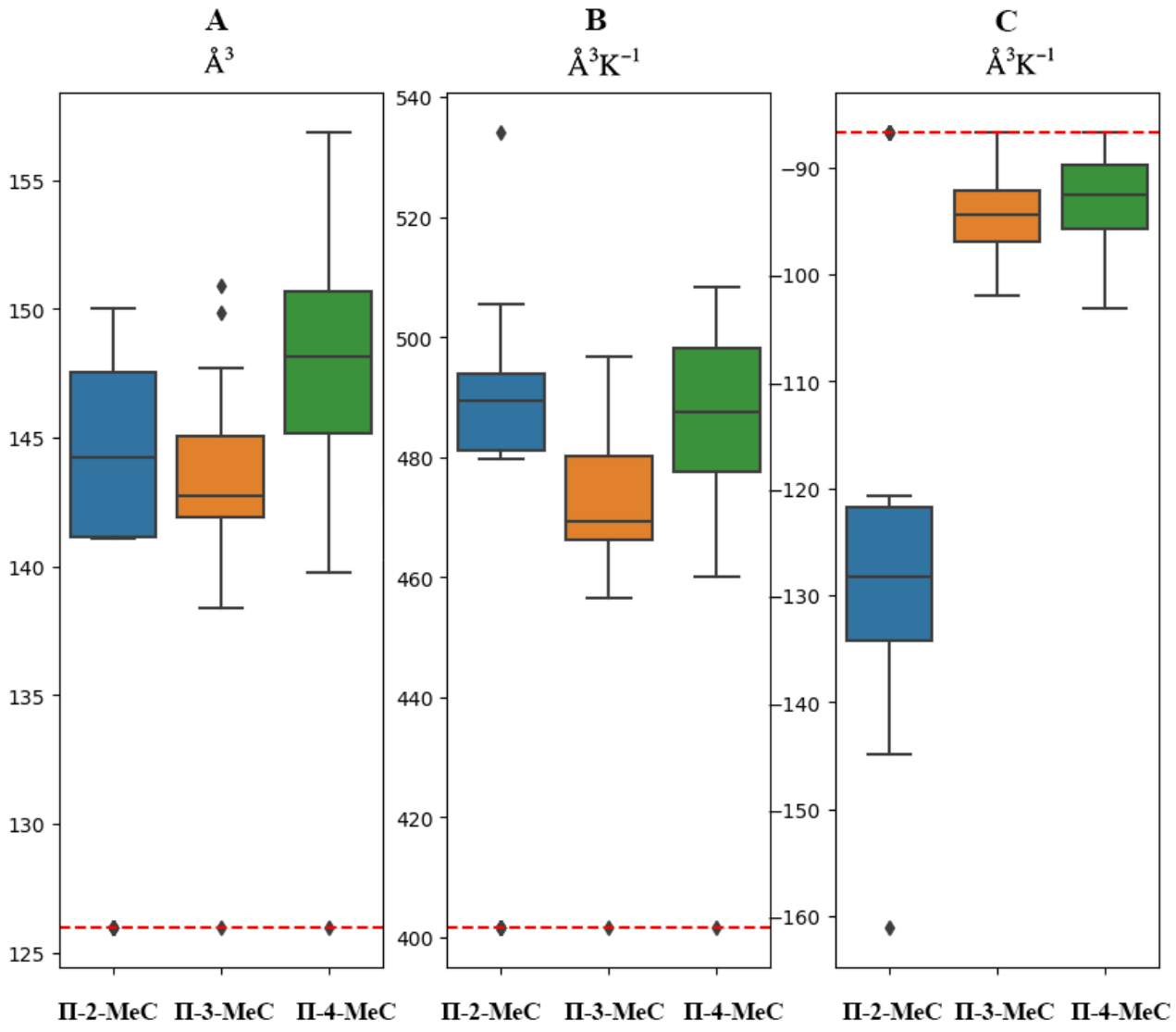


Рис. 15. Диаграмма размаха значений параметров A, B и C для метилзамещенных полистиролов

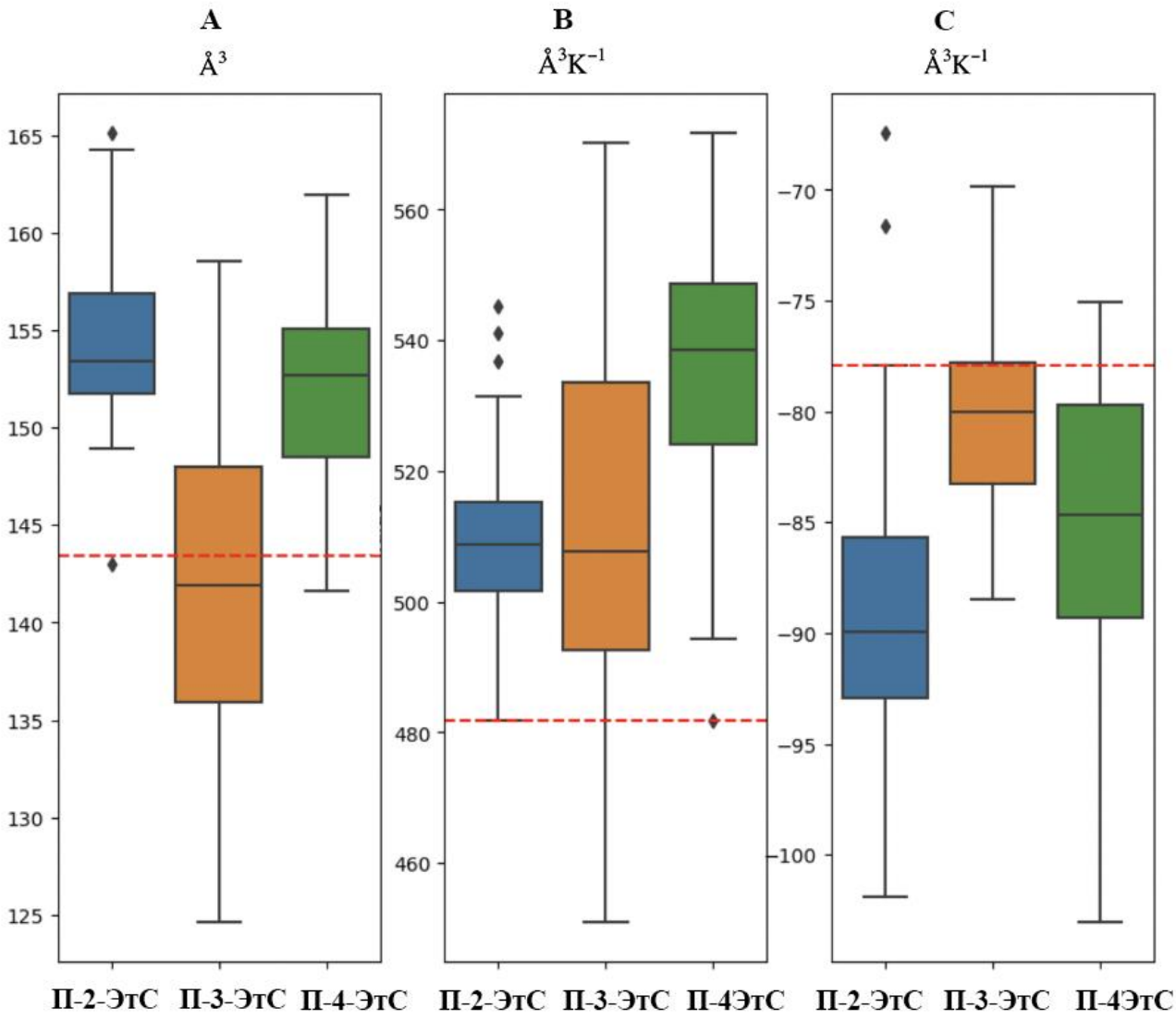


Рис. 16. Диаграмма размаха значений параметров A, B и C для этилзамещенных полистиролов

В качестве итоговых значений параметров A, B и C, полученных моделью по результатам 50 вычислительных экспериментов, принимались значения медианы, поскольку медиана является устойчивой статистической характеристикой распределения и менее чувствительна к выбросам, чем среднее арифметическое.

3.3 Установление и анализ корреляций параметров, определяющих температуру стеклования органических гомополимеров по аналогии с инкрементальным подходом, с квантово-химическими параметрами, относящимися к повторяющимся звеньям органических гомополимеров

Анализ данных, представленных в табл. 12, выявил наличие статистически значимой обратной корреляции между значениями параметра A и такими характеристиками повторяющегося звена органического гомополимера, как молекулярный объем (V_M) и дипольный момент (d).

Таблица 12

Значения молекулярных объемов и дипольных моментов (B3LYP/6-31G(d,p)) для повторяющихся звеньев полистиролов с различными положениями заместителя в ароматическом кольце и их корреляции с параметром A

Повторяющееся звено органического гомополимера	Параметр A, Å^3	V_M, см^3/моль	Коэффициент корреляции Пирсона параметра A с V_M	d, Д	Коэффициент корреляции Пирсона параметра A с d
2-фторстирол	137.770	104.230	-0.46	1.07	-0.99
3-фторстирол	136.848	101.501		1.67	
4-фторстирол	136.787	126.356		1.85	
2-хлорстирол	140.035	107.720	-0.81	1.65	-0.99
3-хлорстирол	136.325	110.330		2.28	
4-хлорстирол	135.666	116.870		2.51	
2-бромстирол	149.679	121.870	-0.99	1.57	-0.94
3-бромстирол	141.861	124.076		2.19	
4-бромстирол	131.000	126.356		2.42	
2-метилстирол	144.245	109.656	-0.75	0.67	-0.63
3-метилстирол	142.733	118.398		0.37	
4-метилстирол	148.153	109.022		0.14	
2-этилстирол	153.548	120.402	-0.99	0.75	-0.10
3-этилстирол	142.126	120.726		0.48	
4-этилстирол	153.499	120.402		0.08	

Установленный факт позволяет выдвинуть предположение о том, что параметр A, по всей видимости, характеризует ван-дер-ваальсовый объем

повторяющегося звена органического гомополимера и, вероятно, отражает влияние слабых межмолекулярных взаимодействий на температуру стеклования. Данное предположение подтверждается тем, что для метил- и этилзамещенных полистиролов, в которых метильные и этильные группы не являются полярными и оказывают слабое влияние на дипольный момент молекулы, значения коэффициента корреляции между параметром A и молекулярным объемом уменьшаются (табл. 12). Это свидетельствует о том, что вклад диполь-дипольных взаимодействий в параметр A незначителен и основной вклад вносит именно ван-дер-ваальсовый объем повторяющегося звена органического гомополимера. Таким образом, параметр A можно интерпретировать как меру гибкости макромолекул.

Наблюдаемое низкое значение коэффициента корреляции между параметром A и молекулярным объемом повторяющегося звена для поли-2-фторстирола, поли-3-фторстирола и поли-4-фторстирола (ниже значения порога статистической значимости, принятого равным 0.5) – см. табл. 12 – может быть обусловлено влиянием ошибки самовзаимодействия. Ошибка самовзаимодействия (от англ. self-interaction error, или SIE) представляет собой систематическую ошибку, возникающую при моделировании электронной структуры молекул и материалов с использованием приближенных методов теории функционала плотности [138]. Суть ошибки самовзаимодействия заключается в том, что электрон в рассматриваемой системе ненадлежащим образом взаимодействует с самим собой [138]. Это приводит к искажению результатов расчета электронной плотности, энергий и других физических свойств системы, в том числе к некорректному описанию распределения электронной плотности, завышению энергий связывания и искажению дипольных моментов молекул [138].

Ошибка самовзаимодействия особенно сильно проявляется при расчетах структур молекул, содержащих сильно электроотрицательные атомы, к которым, в частности, относится и фтор [138]. Вблизи таких атомов электронная плотность становится более локализованной, что усиливает эффект самовзаимодействия [138]. В рамках теории функционала плотности, использование GGA-методов

(аббр. от англ. generalized gradient approximations – обобщенные градиентные приближения), к которым относится и примененный в настоящей диссертации функционал B3LYP, не позволяет полностью устранить ошибку самовзаимодействия [138]. В результате, при расчете электронной структуры фторсодержащих полистиролов с использованием метода B3LYP могут возникать значительные ошибки, влияющие на точность расчета молекулярного объема и дипольного момента, что, в свою очередь, может приводить к снижению корреляции между параметром A и молекулярным объемом повторяющегося звена полимера.

Все значения электронных свойств повторяющихся звеньев исследуемых органических гомополимеров, полученные на основе квантово-химических расчетов, обладают статистически значимой корреляцией с параметром B (табл. 13).

Таблица 13

Значения электронных свойств (B3LYP/6-31G(d,p)) для повторяющихся звеньев полистиролов с различными положениями заместителя в ароматическом кольце и их корреляции с параметром B

Повторяющееся звено органического гомополимера	α, а.е.	I, эВ	*a*, эВ	E_g, эВ	ω, эВ	μ, эВ	η, эВ	d, Д
2-фторстирол	78.95	-6.40	-0.05	-0.05	-1.64	-3.23	-3.18	1.07
3-фторстирол	79.11	-6.48	-0.20	-0.20	-1.77	-3.34	-3.14	1.67
4-фторстирол	79.14	-6.31	-0.18	-0.18	-1.72	-3.25	-3.06	1.85
2-хлорстирол	89.56	-6.53	-0.24	-0.24	-1.82	-3.39	-3.15	1.65
3-хлорстирол	90.43	-6.58	-0.31	-0.31	-1.89	-3.44	-3.13	2.28
4-хлорстирол	90.88	-6.42	-0.15	-0.15	-1.72	-3.29	-3.13	2.51
2-бромстирол	96.55	-6.45	-0.27	-0.27	-1.82	-3.36	-3.09	1.57
3-бромстирол	97.81	-6.47	-0.32	-0.32	-1.87	-3.40	-3.08	2.19
4-бромстирол	98.43	-6.33	-0.32	-0.32	-1.84	-3.33	-3.01	2.42
2-метилстирол	90.45	-6.21	0.21	0.21	-1.40	-3.00	-3.21	0.67
3-метилстирол	91.33	-6.23	0.17	0.17	-1.44	-3.03	-3.20	0.37
4-метилстирол	91.81	-6.11	0.20	0.20	-1.39	-2.96	-3.16	0.14
2-этилстирол	101.72	-6.25	0.12	0.12	-1.48	-3.07	-3.19	0.75
3-этилстирол	102.64	-6.25	0.12	0.12	-1.48	-3.07	-3.19	0.48
4-этилстирол	103.18	-6.09	0.23	0.23	-1.36	-2.93	-3.16	0.08
Коэффициент корреляции Пирсона с параметром B	0.50	0.67	0.80	-0.65	0.79	0.77	-0.65	-0.86

Этот факт позволяет выдвинуть предположение о том, что параметр B характеризует совокупность всех типов межмолекулярных взаимодействий, включая как слабые взаимодействия (диполь-дипольные), так и более сильные взаимодействия (водородные связи и электростатические взаимодействия).

При температуре стеклования для всех органических гомополимеров коэффициент молекулярной упаковки один и тот же (≈ 0.667 [99]) и доля свободного объема тоже одна и та же (доля свободного объема равна разности между единицей и коэффициентом молекулярной упаковки и составляет ≈ 0.333 [99]), поэтому нижеследующие рассуждения сформулированы в привязке к температурам выше температуры стеклования органических гомополимеров. Поскольку с увеличением сил межмолекулярных взаимодействий (которые характеризуются параметром B) уменьшается плотность молекулярной упаковки, (которая характеризуется коэффициентом молекулярной упаковки) [139], можно предположить, что параметр C связан с коэффициентом молекулярной упаковки. Обратная пропорциональность между параметрами C и B наиболее сильно проявляется на рис. 13. Данная взаимосвязь может быть объяснена тем, что сильные межмолекулярные взаимодействия приводят к образованию более жестких и устойчивых связей между макромолекулами. Это, в свою очередь, может затруднять их плотную упаковку и приводить к увеличению свободного объема в органическом гомополимере. Другими словами, сильные межмолекулярные взаимодействия могут приводить к образованию более «рыхлой» структуры полимера, в которой макромолекулы не могут занять более «плотное» положение, что оказывает влияние на температуру стеклования полимера, поскольку плотность упаковки макромолекул является одним из ключевых факторов, определяющих их подвижность и, следовательно, температуру стеклования.

Несмотря на значимость коэффициента молекулярной упаковки для понимания физико-химических свойств органических гомополимеров, его количественная оценка представляет собой сложную задачу. Экспериментальные значения коэффициента молекулярной упаковки доступны лишь для

ограниченного числа органических гомополимеров [99], входящих в использованную в настоящей диссертации базу данных. Это обстоятельство существенно ограничивает возможности проведения статистического анализа и количественной оценки корреляции между параметром C и экспериментальными значениями коэффициента молекулярной упаковки.

В связи с этим для анализа влияния параметра C на коэффициент молекулярной упаковки органических гомополимеров использовали метод аналогии. В частности, для изомерного ряда поли-2-хлорстирол, поли-3-хлорстирол, поли-4-хлорстирол предположили, что плотность молекулярной упаковки уменьшается в указанном порядке по аналогии с известными данными о плотности молекулярной упаковки в соответствующих кристаллических органических гомополимерах [140]. При этом отметили, что параметр C увеличивается в том же порядке (как это видно из рис. 13), что свидетельствует об обратной пропорциональности между параметром C и коэффициентом молекулярной упаковки.

Для подтверждения данной особенности выбрали гомологический ряд полиметакрилатов, для которых известно, что уменьшение температуры стеклования с увеличением длины алифатической цепи обусловлено преимущественно уменьшением плотности молекулярной упаковки (рис. 17) [99]. В связи с этим ожидали на диаграмме размаха наблюдать увеличение значения параметра C с увеличением длины алифатического заместителя в гомологическом ряду полиметакрилатов, что соответствовало бы уменьшению плотности молекулярной упаковки. Чтобы подтвердить данную гипотезу, рассчитали значения параметра C для данного гомологического ряда с помощью построенной в настоящей диссертации модели на основе метода случайного леса. Результаты расчетов представлены на рис. 18.

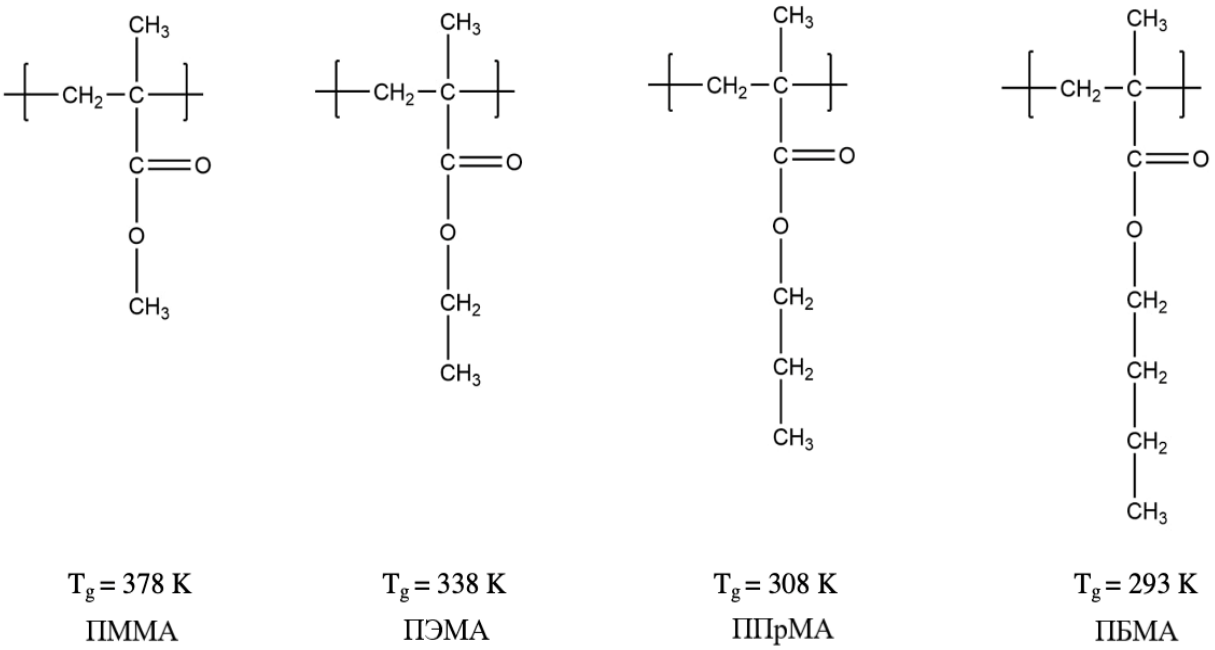


Рис. 17. Значения $T_{g\ exp}$ в гомологическом ряду полиметакрилатов [99]: ПММА – полиметилметакрилат, ПЭМА – полиэтилметакрилат, ППрМА – полипропиленметакрилат, ПБМА – полибутилметакрилат

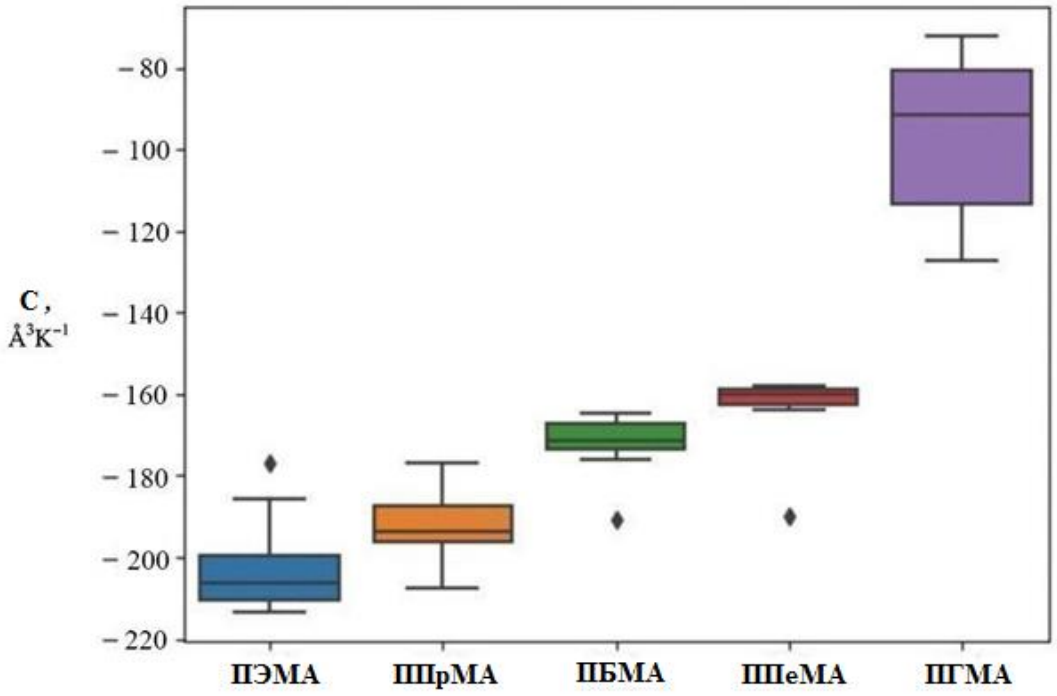


Рис. 18. Диаграмма размаха прогнозируемых значений параметра C для гомологического ряда полиметакрилатов: ПЭМА – полиэтилметакрилат, ППрМА – полипропиленметакрилат, ПБМА – полибутилметакрилат, ППеМА – полипентилметакрилат, ПГМА – полигексилметакрилат

Результаты расчетов (рис. 18) подтвердили выдвинутую гипотезу о наличии обратной пропорциональности между параметром C и коэффициентом молекулярной упаковки и, как следствие, о наличии прямой пропорциональности между параметром C и долей свободного объема. Следовательно, при идентичной скорости охлаждения образец полимера, характеризующийся более высокой долей свободного объема в высокоэластическом состоянии, переходит в стеклообразное состояние при более низкой температуре вследствие запаздывания релаксационных процессов.

***

Таким образом, в настоящей диссертации впервые представлена модель, построенная на основе молекулярных отпечатков Моргана как характеристик молекулярной структуры повторяющихся звеньев полимеров и метода случайного леса как метода машинного обучения и способная прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее по аналогии с инкрементальным подходом. На примере полистиролов с различными положениями (2-, 3-, 4-) заместителя (фтор-, хлор-, бром-, метил- и этил-) в ароматическом кольце впервые показано, что учет химического строения повторяющихся звеньев органических гомополимеров (в данном случае учет положения заместителя в ароматическом кольце) является определяющим фактором повышения точности прогнозирования температуры стеклования органических гомополимеров: модели, построенные без учета и с учетом положения заместителя, имеют коэффициенты детерминации 0.12 и 0.81 соответственно.

По результатам моделирования впервые на количественном уровне показано, что температура стеклования органических гомополимеров:

- прямо пропорциональна параметру, связанному с молекулярным объемом повторяющегося звена полимера и интерпретируемому как мера гибкости макромолекул (эта закономерность согласуется с термодинамическими теориями

стеклования полимеров и кинетическими теориями М.В. Волькенштейна-О.Б. Птицына и Ю.Я. Готлиба-О.Б. Птицына);

- обратно пропорциональна сумме двух параметров, один из которых характеризует совокупность всех типов межмолекулярных взаимодействий (эта закономерность согласуется с теорией межмолекулярных связей С.Н. Журкова и флуктуационной теорией стеклования полимеров), а второй – долю свободного объема (эта закономерность согласуется с теорией свободного объема).

Кроме того, впервые показано, что электронные свойства повторяющихся звеньев органических гомополимеров (средняя поляризуемость, дипольный момент, потенциал ионизации, сродство к электрону, энергетический зазор между высшей занятой молекулярной орбиталью и низшей свободной молекулярной орбиталью, химический потенциал, химическая жесткость, электрофильность) обладают статистически значимой корреляцией с параметром, который в связи «структура – температура стеклования органических гомополимеров» характеризует совокупность всех типов межмолекулярных взаимодействий, включая диполь-дипольные (слабые взаимодействия), водородные связи (сильные взаимодействия) и электростатические (сильные взаимодействия).

Теоретическая значимость работы заключается в том, что построенная модель «структура – температура стеклования органических гомополимеров» позволяет получать результаты, которые могут быть проанализированы с позиций теории химического строения органических соединений и теорий стеклования полимеров.

Практическая значимость работы заключается в том, что: 1) разработано программное обеспечение для прогнозирования температуры стеклования органических гомополимеров; 2) модель может применяться в качестве прогностического модуля при технологическом моделировании промышленных процессов синтеза органических гомополимеров.

Результаты диссертации полностью опубликованы в работах [141-147]. На разных этапах выполнения диссертационного исследования соавторами в публикациях являлись Н.В. Улитин, В.И. Анисимова, К.А. Терещенко,

Д.А. Шиян, А.Д. Лифанов, Е.С. Воробьев, И.А. Суворова, И.С. Родионов, А.А. Балдинов, Я.Л. Люлинская, М.Г. Казанская, А.О. Софьин, А.А. Кулыгин. Научный руководитель Н.В. Улитин поставил цель и задачи диссертации, принимал участие в обсуждении результатов и написании публикаций. К.А. Терещенко и Д.А. Шиян консультировали автора диссертации в области синтеза органических гомополимеров, их многоуровневой структуры и свойств. А.Д. Лифанов и Е.С. Воробьев консультировали автора диссертации в области машинного обучения. В.И. Анисимова, А.А. Балдинов, И.С. Родионов и И.А. Суворова консультировали автора диссертации в области квантово-химических расчетов. Я.Л. Люлинская, М.Г. Казанская, А.О. Софьин и А.А. Кулыгин занимались проверкой данных в используемой базе данных. Личный вклад автора диссертации заключается в сборе и анализе литературных данных, реализации решения задач исследования, анализе результатов, формулировании заключения и участии в написании и подготовке публикаций.

ЗАКЛЮЧЕНИЕ

По результатам диссертационного исследования можно сделать следующие выводы.

1. На основе методов машинного обучения построена модель, способная прогнозировать температуру стеклования органических гомополимеров через параметры, которые определяют ее на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом. При моделировании рассмотрено 3 метода машинного обучения: метод случайного леса, метод k ближайших соседей и многослойный перцептрон. Для описания химических структурных формул повторяющихся звеньев органических гомополимеров выбраны и протестированы структурные ключи и молекулярные отпечатки Моргана. Показано, что модель на основе метода случайного леса, использующая молекулярные отпечатки Моргана, обладает наибольшей достоверностью прогнозирования температуры стеклования органических гомополимеров.

2. Для обеспечения более достоверных прогнозов уточнены значения параметров, определяющих температуру стеклования органических гомополимеров на основе химического строения их повторяющихся звеньев по аналогии с инкрементальным подходом, за счет применения в модели на основе метода случайного леса комбинированных дескрипторов. Комбинированные дескрипторы получены путем объединения информации о химическом строении повторяющихся звеньев органических гомополимеров и их экспериментальных и рассчитанных в рамках инкрементального подхода значениях температуры стеклования. На примере полистиролов с различными положениями (2-, 3-, 4-) заместителя (фтор-, хлор-, бром-, метил- и этил-) в ароматическом кольце показано, что модель, построенная на основе комбинированных дескрипторов и метода случайного леса, в отличие от модели, построенной в рамках инкрементального подхода, достаточно точно (коэффициент детерминации в

среднем $R^2 = 0.81$) прогнозирует значения температуры стеклования органических гомополимеров с различным положением заместителя в ароматическом кольце.

3. По результатам моделирования на количественном уровне показано, что температура стеклования органических гомополимеров:

- прямо пропорциональна параметру, связанному с молекулярным объемом повторяющегося звена полимера и интерпретируемому как мера гибкости макромолекул (эта закономерность согласуется с термодинамическими теориями стеклования полимеров и кинетическими теориями М.В. Волькенштейна-О.Б. Птицына и Ю.Я. Готлиба-О.Б. Птицына);

- обратно пропорциональна сумме двух параметров, один из которых характеризует совокупность всех типов межмолекулярных взаимодействий (эта закономерность согласуется с теорией межмолекулярных связей С.Н. Журкова и флуктуационной теорией стеклования полимеров), а второй – долю свободного объема (эта закономерность согласуется с теорией свободного объема).

Показано, что электронные свойства повторяющихся звеньев органических гомополимеров (средняя поляризуемость, дипольный момент, потенциал ионизации, сродство к электрону, энергетический зазор между высшей занятой молекулярной орбиталью и низшей свободной молекулярной орбиталью, химический потенциал, химическая жесткость, электрофильность) обладают статистически значимой корреляцией с параметром, который в связи «структура – температура стеклования органических гомополимеров» характеризует совокупность всех типов межмолекулярных взаимодействий, включая диполь-дипольные (слабые взаимодействия), водородные связи (сильные взаимодействия) и электростатические (сильные взаимодействия).

Перспективы дальнейшей разработки темы диссертации: 1) учет в построенной модели стереорегулярности макромолекул, молекулярно-массовых характеристик и степени кристалличности органических гомополимеров; 2) развитие предложенной методологии построения модели на сополимеры.

СПИСОК ЛИТЕРАТУРЫ

1. Hammett, L.P. The effect of structure upon the reactions of organic compounds. Benzene derivatives / L.P. Hammett // Journal of the American Chemical Society. – 1937. – V. 59, №1. – P. 96-103. https://doi.org/10.1021/ja01280a022

2. Crum Brown, A. On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia / A. Crum Brown, T.R. Fraser // Journal of Anatomy and Physiology. – 1868. – V. 2, №2. – P. 224-242. https://pubmed.ncbi.nlm.nih.gov/17230757/

3. Richet, C. Sur le rapport entre la toxicité et les propriétés physiques des corps / C. Richet // Comptes Rendus Hebdomadaires des Séances et Mémoires de la Société de Biologie. – 1893. – 9^{e} série, T. V. – P. 775-776.

4. Meyer, H. Zur theorie der alkoholnarkose: erste mittheilung. Welche eigenschaft der anästhetica bedingt ihre narkotische wirkung? / H. Meyer // Archiv für experimentelle Pathologie und Pharmakologie. – 1899. – V. 42. – P. 109-118. https://doi.org/10.1007/BF01834479

5. Wu, Z. Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method / Z. Wu, D. Jiang, C.-Y. Hsieh, G. Chen, B. Liao, D. Cao, T. Hou // Briefings in Bioinformatics. – 2021. – V. 22, №5. – Article bbab112. https://doi.org/10.1093/bib/bbab112

6. Polishchuk, P. Interpretation of quantitative structure–activity relationship models: past, present, and future / P. Polishchuk // Journal of Chemical Information and Modeling. – 2017. – V. 57, №11. – P. 2618-2639. https://doi.org/10.1021/acs.jcim.7b00274

7. Statistical modelling of molecular descriptors in QSAR/QSPR / eds. M. Dehmer, K. Varmuza, D. Bonchev // Quantitative and Network Biology: in 2 v. V. 2. – Weinheim: Wiley-VCH Verlag & Co. KGaA, 2012. – 456 p.

8. Roy, K. A primer on QSAR/QSPR modeling: fundamental concepts / K. Roy, S. Kar, R.N. Das. – Cham: Springer, 2015. – 121 p. https://doi.org/10.1007/978-3-319-17281-1

9. QSPR/QSAR analysis using SMILES and quasi-SMILES / eds. A.P. Toropova, A.A. Toropov. – Cham: Springer, 2023. – 467 p. https://doi.org/10.1007/978-3-031-28401-4

10. Менделеев, Д.И. Периодический закон / Д.И. Менделеев; под ред. Б.М. Кедрова. – М.: Изд-во АН СССР, 1958. – 839 с.

11. Dearden, J.C. The history and development of quantitative structure-activity relationships (QSARs) / J.C. Dearden // International Journal of Quantitative Structure-Property Relationships. – 2016. – V. 1, №1. – P. 1-44. https://doi.org/10.4018/IJQSPR.2016010101

12. Гиллер, С.А. Распознавание физиологической активности химических соединений на перептроне со случайной адаптацией структуры / С.А. Гиллер, А.Б. Глаз, Л.А. Растригин, А.Б. Розенблит // Доклады Академии наук СССР. – 1971. – Т. 199, №4. – С. 851-853.

13. Hiller, S.A. Cybernetic methods of drug design. I. Statement of the problem – the perceptron approach / S.A. Hiller, V.E. Golender, A.B. Rosenblit, L.A. Rastrigin, A.B. Glaz // Computers and Biomedical Research. – 1973. – V. 6, №5. – P. 411-421. https://doi.org/10.1016/0010-4809(73)90074-8

14. Cramer III, R.D. Substructural analysis. A novel approach to the problem of drug design / R.D. Cramer III, G. Redl, C.E. Berkoff // Journal of Medicinal Chemistry. – 1974. – V. 17, №5. – P. 533-535. https://pubs.acs.org/doi/abs/10.1021/jm00251a014

15. Toropov, A.A. QSPR/QSAR: state-of-art, weirdness, the future / A.A. Toropov, A.P. Toropova // Molecules. – 2020. – V. 25, №6. – Article 1292. https://doi.org/10.3390/molecules25061292

16. Li, J. A review of quantitative structure-activity relationship: the development and current status of data sets, molecular descriptors and mathematical models / J. Li, T. Zhao, Q. Yang, S. Du, L. Xu // Chemometrics and Intelligent Laboratory Systems. – 2025. – V. 256. – Article 105278. https://doi.org/10.1016/j.chemolab.2024.105278

17. van den Maagdenberg, H.W. QSPRpred: a flexible open-source quantitative structure-property relationship modelling tool / H.W. van den Maagdenberg, M. Šícho, D.A. Araripe, S. Luukkonen, L. Schoenmaker, M. Jespers, O.J. M. Béquignon, M.G. González, R.L. van den Broek, A. Bernatavicius, J.G.C. van Hasselt, P.H. van der Graaf, G.J.P. van Westen // Journal of Cheminformatics. – 2024. – V. 16. – Article 128. https://doi.org/10.1186/s13321-024-00908-y

18. Bongers, B.J. Proteochemometrics – recent developments in bioactivity and selectivity modeling / B.J. Bongers, A.P. IJzerman, G.J.P. Van Westen // Drug Discovery Today: Technologies. – 2019. – V. 32-33. – P. 89-98. https://doi.org/10.1016/j.ddtec.2020.08.003

19. Fluetsch, A. Adapting deep learning QSPR models to specific drug discovery projects / A. Fluetsch, E. Di Lascio, G. Gerebtzoff, R. Rodríguez-Pérez // Molecular Pharmaceutics. – 2024. – V. 21, №4. – P. 1817-1826. https://doi.org/10.1021/acs.molpharmaceut.3c01124

20. Samadi, A. Development of remediation technologies for organic contaminants informed by QSAR/QSPR models / A. Samadi, A.K. Pour, R. Jamieson // Environmental Advances. – 2021. – V. 5. – Article 100112. https://doi.org/10.1016/j.envadv.2021.100112

21. Morrill, J.A. Development of quantitative structure–property relationships for predictive modeling and design of energetic materials / J.A. Morrill, E.F.C. Byrd // Journal of Molecular Graphics and Modelling. – 2008. – V. 27, №3. – P. 349-355. https://doi.org/10.1016/j.jmgm.2008.06.003

22. Vasilev, B. A (comprehensive) review of the application of quantitative structure–activity relationship (QSAR) in the prediction of new compounds with anti-breast cancer activity / B. Vasilev, M. Atanasova // Applied Sciences. – 2025. – V. 15, №3. – Article 1206. https://doi.org/10.3390/app15031206

23. Piir, G. Best practices for QSAR model reporting: physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints / G. Piir, I. Kahn, A.T. García-Sosa, S. Sild, P. Ahte, U. Maran // Environmental Health

Perspectives. – 2018. – V. 126, №12. – Article 126001. https://doi.org/10.1289/EHP3264

24. Chen, C.-H. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications / C.-H. Chen, K. Tanaka, M. Kotera, K. Funatsu // Journal of Cheminformatics. – 2020. – V. 12. – Article 19. https://doi.org/10.1186/s13321-020-0417-9

25. Pham, T.H. A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques / T.H. Pham, P.K. Le, D.N. Son // RSC Advances. – 2024. – V. 14, №16. – P. 11157-11168. https://doi.org/10.1039/d4ra02159b

26. Soares, T.A. The (re)-evolution of quantitative structure–activity relationship (QSAR) studies propelled by the surge of machine learning methods / T.A. Soares, A. Nunes-Alves, A. Mazzolari, F. Ruggiu, G.-W. Wei, K. Merz // Journal of Chemical Information and Modeling. – 2022. – V. 62, №22. – P. 5317-5320. https://doi.org/10.1021/acs.jcim.2c01422

27. Banchero, M. Comparison between multi-linear- and radial-basis-function-neural-network-based QSPR models for the prediction of the critical temperature, critical pressure and acentric factor of organic compounds / M. Banchero, L. Manna // Molecules. – 2018. – V. 23, №6. – Article 1379. https://doi.org/10.3390/molecules23061379

28. Winkler, D.A. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR / D.A. Winkler, T.C. Le // Molecular Informatics. – 2017. – V. 36, №1-2. – Article 1600118. https://doi.org/10.1002/minf.201600118

29. Wu, J. Deep-learning architecture in QSPR modeling for the prediction of energy conversion efficiency of solar cells / J. Wu, S. Wang, L. Zhou, X. Ji, Y. Dai, Y. Dang, M. Kraft // Industrial & Engineering Chemistry Research. – 2020. – V. 59, №42. – P. 18991-19000. https://doi.org/10.1021/acs.iecr.0c03880

30. Clark, R.D. Building a quantitative structure-property relationship (QSPR) model / R.D. Clark, P.R. Daga // Bioinformatics and Drug Discovery: Methods in

Molecular Biology / eds. R.S. Larson, T.I. Oprea. – New York: Humana Press, 2019. – P. 139-159. https://doi.org/10.1007/978-1-4939-9089-4_8

31. Hemmateenejad, B. Quantitative structure–retention relationship for the Kovats retention indices of a large set of terpenes: a combined data splitting-feature selection strategy / B. Hemmateenejad, K. Javadnia, M. Elyasi // Analytica Chimica Acta. – 2007. – V. 592, №1. – P. 72-81. https://doi.org/10.1016/j.aca.2007.04.009

32. Rybińska-Fryca, A. Representation of the structure – a key point of building QSAR/QSPR models for ionic liquids / A. Rybińska-Fryca, A. Sosnowska, T. Puzyn // Materials. – 2020. – V. 13, №11. – Article 2500. https://doi.org/10.3390/ma13112500

33. Limpert, E. Problems with using the normal distribution – and ways to improve quality and efficiency of data analysis / E. Limpert, W.A. Stahel // PLOS ONE. – 2011. – V. 6, №7. – Article e21403. https://doi.org/10.1371/journal.pone.0021403

34. Sheridan, R.P. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR / R.P. Sheridan, B.P. Feuston, V.N. Maiorov, S.K. Kearsley // Journal of Chemical Information and Computer Sciences. – 2004. – V. 44, №6. – P. 1912-1928. https://doi.org/10.1021/ci049782w

35. Netzeva, T.I. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52 / T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang // Alternatives to Laboratory Animals. – 2005. – V. 33, №2. – P. 155-173. https://doi.org/10.1177/026119290503300209

36. Teixeira, A.L. Random forests for feature selection in QSPR models – an application for predicting standard enthalpy of formation of hydrocarbons / A.L. Teixeira, J.P. Leal, A.O. Falcao // Journal of Cheminformatics. – 2013. – V. 5, №1. – Article 9. https://doi.org/10.1186/1758-2946-5-9

37. Varmuza, K. Multivariate linear QSPR/QSAR models: rigorous evaluation of variable selection for PLS / K. Varmuza, P. Filzmoser, M. Dehmer // Computational

and Structural Biotechnology Journal. – 2013. – V. 5, №6. – Article e201302007. http://doi.org/10.5936/csbj.201302007

38. Rácz, A. Intercorrelation limits in molecular descriptor preselection for QSAR/QSPR / A. Rácz, D. Bajusz, K. Héberger // Molecular Informatics. – 2019. – V. 38, №8-9. – Article 1800154. https://doi.org/10.1002/minf.201800154

39. Martínez, M.J. Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods / M.J. Martínez, I. Ponzoni, M.F. Díaz G.E. Vazquez, A.J. Soto // Journal of Cheminformatics. – 2015. – V. 7, №1. – Article 39. https://doi.org/10.1186/s13321-015-0092-4

40. Baskin, I. Ch 1. Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening / I. Baskin, A. Varnek // Chemoinformatics Approaches to Virtual Screening / ed. by A. Varnek, A. Tropsha. – Cambridge: The Royal Society of Chemistry, 2008. – P. 1-43.

41. Orosz, Á. Comparison of descriptor- and fingerprint sets in machine learning models for ADME-tox targets / Á. Orosz, K. Héberger, A. Rácz // Frontiers in Chemistry. – 2022. – V. 10. – Article 852893. https://doi.org/10.3389/fchem.2022.852893

42. Todeschini, R. Molecular descriptors for chemoinformatics / R. Todeschini, V. Consonni. – 2nd ed. – Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA, 2009. – 1220 p.

43. Баскин, И.И. Введение в хемоинформатику: 3. Моделирование «структура-свойство» / И.И. Баскин, Т.И. Маджидов, А.А. Варнек. – Казань: Изд-во Казан. ун-та, 2015. – 304 с.

44. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules / D. Weininger // Journal of Chemical Information and Computer Sciences. – 1988. – V. 28, №1. – P. 31-36. https://doi.org/10.1021/ci00057a005

45. Баскин, И.И. Введение в хемоинформатику: 4. Методы машинного обучения / И.И. Баскин, Т.И. Маджидов, А.А. Варнек. – Казань: Изд-во Казан. ун-та, 2016. – 330 с.

46. Zhang, Z. Introduction to machine learning: k-nearest neighbors / Z. Zhang // Annals of Translational Medicine. – 2016. – V. 4, №11. – Article 218. http://doi.org/10.21037/atm.2016.03.37

47. Yao, X.J. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression / X.J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan // Journal of Chemical Information and Computer Sciences. – 2004. – V. 44, №4. – P. 1257-1266. https://doi.org/10.1021/ci049965i

48. Salman, H.A. Random forest algorithm overview / H.A. Salman, A. Kalakech, A. Steiti // Babylonian Journal of Machine Learning. – 2024. – V. 2024. – P. 69-79. https://doi.org/10.58496/BJML/2024/007

49. Aggarwal, C.C. Neural networks and deep learning / C.C. Aggarwal. – Cham : Springer International Publishing AG, 2018. – 497 p. https://doi.org/10.1007/978-3-319-94463-0

50. Stone, M. Cross-validatory choice and assessment of statistical predictions / M. Stone // Journal of the Royal Statistical Society. Series B (Methodological). – 1974. – V. 36, №2. – P. 111-133. https://doi.org/10.1111/j.2517-6161.1976.tb01573.x

51. Powers, D.M.W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation / D.M.W. Powers // International Journal of Machine Learning Technology. – 2011. – V. 2, №1. – P. 37-63. https://doi.org/10.48550/arXiv.2010.16061

52. Davis, J. The relationship between Precision-Recall and ROC curves / J. Davis, M. Goadrich // ICML '06: The 23rd International Conference on Machine learning. – New York: Association for Computing Machinery, 2006. – P. 233-240. https://doi.org/10.1145/1143844.1143874

53. Espíndola, R.P. On extending F-measure and G-mean metrics to multi-class problems / R.P. Espíndola, N.F.F. Ebecken // WIT Transactions on Information and Communication Technologies. – 2005. – V. 35. Data Mining VI. – P. 25-34. https://doi.org/10.2495/DATA050031

54. Plevris, V. Investigation of performance metrics in regression analysis and machine learning-based prediction models / V. Plevris, G. Solorzano, N.P. Bakas, M.E.A.B. Seghier // 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022, 5 – 9 June 2022). – Oslo, 2022. – https://www.scipedia.com/public/Plevris_et_al_2022a, https://doi.org/10.23967/eccomas.2022.155

55. Niu, H. Property estimation of organic compounds based on QSPR models with norm indices / H. Niu, Y. Zhang, Q. Jia, Q. Wang, F. Yan // Chemical Engineering Science. – 2024. – V. 288. – Article 119835. https://doi.org/10.1016/j.ces.2024.119835

56. Шулаева, Н.А. Модели связи «структура-свойство» органических соединений на основе молекулярных графов с элементами пространственного строения молекул / Н.А. Шулаева, М.И. Скворцова, Н.А. Михайлова // Тонкие химические технологии. – 2020. – Т. 15, №6. – С. 84-103. https://doi.org/10.32362/2410-6593-2020-15-6-84-103

57. Baskin, I.I. Artificial intelligence in synthetic chemistry: achievements and prospects / I.I. Baskin, T.I. Madzhidov, I.S. Antipin, A.A. Varnek // Russian Chemical Reviews. – 2017. – V. 86, №11. – P. 1127-1156. https://doi.org/10.1070/RCR4746

58. Ajmani, S. Application of QSPR to mixtures / S. Ajmani, S.C. Rogers, M.H. Barley, D.J. Livingstone // Journal of Chemical Information and Modeling. – 2006. – V. 46, №5. – P. 2043-2055. https://doi.org/10.1021/ci050559o

59. Muratov, E.N. Existing and developing approaches for QSAR analysis of mixtures / E.N. Muratov, E.V. Varlamova, A.G. Artemenko, P.G. Polishchuk, V.E. Kuz'min // Molecular Informatics. – 2012. – V. 31, №3-4. – P. 202-221. https://doi.org/10.1002/minf.201100129

60. Toropova, A.P. QSPR modeling mineral crystal lattice energy by optimal descriptors of the graph of atomic orbitals / A.P. Toropova, A.A. Toropov, S.Kh. Maksudov // Chemical Physics Letters. – 2006. – V. 428, №1-3. – P. 183-186. https://doi.org/10.1016/j.cplett.2006.06.084

61. Kong, C.S. Information-theoretic approach for the discovery of design rules for crystal chemistry / C.S. Kong, W. Luo, S. Arapan, P. Villars, S. Iwata, R. Ahuja,

K. Rajan // Journal of Chemical Information and Modeling. – 2012. – V. 52, №7. – P. 1812-1820. https://doi.org/10.1021/ci200628z

62. Fourches, D. Quantitative nanostructure-activity relationship (QNAR) modeling / D. Fourches, D. Pu, C. Tassa, R. Weissleder, S.Y. Shaw, R.J. Mumper, A. Tropsha// ACS Nano. – 2010. – V. 4, №10. – P. 5703-5712. https://doi.org/10.1021/nn1013484

63. González-Nilo, F. Nanoinformatics: an emerging area of information technology at the intersection of bioinformatics, computational chemistry and nanobiotechnology / F. González-Nilo, T. Pérez-Acle, S. Guínez-Molinos, D.A. Geraldo, C. Sandoval, A. Yévenes, L.S. Santos, V.F. Laurie, H. Mendoza, R.E. Cachau // Biological Research. – 2011. – V. 44, №1. – P. 43-51. https://doi.org/10.4067/S0716-97602011000100006

64. Puzyn, T. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles / T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T.P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska, J. Leszczynski // Nature Nanotechnology. – 2011. – V. 6, №3. – P. 175-178. https://doi.org/10.1038/nnano.2011.10

65. Fourches, D. Exploring quantitative nanostructure-activity relationships (QNAR) modeling as a tool for predicting biological effects of manufactured nanoparticles / D. Fourches, D. Pu, A. Tropsha // Combinatorial Chemistry & High Throughput Screening. – 2011. – V. 14, №3. – P. 217-225. https://doi.org/10.2174/138620711794728743

66. Maojo, V. Nanoinformatics: a new area of research in nanomedicine / V. Maojo, M. Fritts, D. de La Iglesia, R.E.Cachau, M. Garcia-Remesal, J.A. Mitchell, C. Kulikowski // International Journal of Nanomedicine. – 2012. – V. 7. – P. 3867-3890. http://doi.org/10.2147/IJN.S24582

67. Xu, J. Prediction of dielectric dissipation factors of polymers from cyclic dimer structure using multiple linear regression and support vector machine / J. Xu, L. Zhu, D. Fang, L. Liu, L. Wang, W. Xu // Colloid and Polymer Science. – 2013. – V. 291. – P. 551-561. https://doi.org/10.1007/s00396-012-2743-6

68. Chen, L. Polymer informatics: Current status and critical next steps / L. Chen, G. Pilania, R. Batra, T.D. Huan, C. Kim, C. Kuenneth, R. Ramprasad // Materials Science & Engineering R. – 2021. – V. 144. – Article 100595. https://doi.org/10.1016/j.mser.2020.100595

69. Chew, A.K. Designing the next generation of polymers with machine learning and physics-based models / A.K. Chew, M.A.F. Afzal, A. Chandrasekaran, J.H. Kamps, V. Ramakrishnan // Machine Learning: Science and Technology. – 2024. – V. 5, №4. – Article 045031. https://doi.org/10.1088/2632-2153/ad88d7

70. Бортников, В.Г. Теоретические основы и технология переработки пластических масс: учебник. – 3-е изд. – М.: ИНФРА-М, 2015. – 480 с.

71. Rieger, J. The glass transition temperature T_g of polymers – Comparison of the values from differential thermal analysis (DTA, DSC) and dynamic mechanical measurements (torsion pendulum) / J. Rieger // Polymer Testing. – 2001. – V. 20, №2. – P. 199-204. https://doi.org/10.1016/S0142-9418(00)00023-4

72. Belukhichev, E.V. Films based on a blend of PVC with copolymer of 3-hydroxybutyrate with 3-hydroxyhexanoate / E.V. Belukhichev, V.E. Sitnikova, E.O. Samuylova, M.V. Uspenskaya, D.M. Martynova // Polymers. – 2020. – V. 12, №2. – Article 270. https://doi.org/10.3390/polym12020270

73. Zainal, N.F.A. Thermal analysis: basic concept of differential scanning calorimetry and thermogravimetry for beginners / N.F.A. Zainal, J.M. Saiter, S.I.A. Halim, R. Lucas, C.H. Chan// Chemistry Teacher International. – 2021. – V. 3, №2. – P. 59-75. https://doi.org/10.1515/cti-2020-0010

74. Qian, Z. Challenge and solution of characterizing glass transition temperature for conjugated polymers by differential scanning calorimetry / Z. Qian, L. Galuska, W.W. McNutt, M.U. Ocheje, Y. He, Z. Cao, S. Zhang, J. Xu, K. Hong, R.B. Goodman, S. Rondeau-Gagné, J. Mei, X. Gu // Journal of Polymer Science Part B: Polymer Physics. – 2019. – V. 57, №23. – P. 1635-1644. https://doi.org/10.1002/polb.24889

75. Kamasa, P. Experimental aspects of temperature-modulated dilatometry of polymers / P. Kamasa, P. Myśliński, M. Pyda // Thermochimica Acta. – 2006. – V. 442, №1-2. – P. 48-51. https://doi.org/10.1016/j.tca.2005.11.017

76. Ma, Y. Luminescent molecularly-imprinted polymer nanocomposites for sensitive detection / Y. Ma, S. Xu, S. Wang, L. Wang // Trends in Analytical Chemistry. – 2015. – V. 67. – P. 209-216. https://doi.org/10.1016/j.trac.2015.01.012

77. Zhang, Z. Application of infrared spectroscopy in research on aging of silicone rubber in harsh environment / Z. Zhang, T. Liang, Z. Jiang, X. Jiang, J. Hu, G. Pang // Polymers. – 2022. – V. 14, №21. – Article 4728. https://doi.org/10.3390/polym14214728

78. Аверко-Антонович, И.Ю. Методы исследования структуры и свойств полимеров / И.Ю. Аверко-Антонович, Р.Т. Бикмуллин. – Казань: Изд-во Казанского государственного технологического университета, 2002. – 604 с.

79. Herzog, B. Glass-transition temperature based on dynamic mechanical thermal analysis techniques as an indicator of the adhesive performance of vinyl ester resin / B. Herzog, D.J. Gardner, R. Lopez-Anido, B. Goodell // Journal of Applied Polymer Science. – 2005. – V. 97, №6. – P. 2221-2229. . https://doi.org/10.1002/app.21868

80. Xu, H. Thermally stimulated discharge current analysis of polymeric solid-state ionic conductors / H. Xu, Q. Gu, M. Fan, C. Yang // Physica Status Solidi (A). – 1997. – V. 161, №2. – P. 343-348. https://doi.org/10.1002/1521-396X(199706)161:2%3C343::AID-PSSA343%3E3.0.CO;2-0

81. Gu, Q. Thermally stimulated discharge current analysis of solid polystyrenesulfonic acid films / Q. Gu, W. Ye, Q. Dai, Z. Wang, K. Sun, W.M. Risen, Jr. // Solid State Communications. – 1987. – V. 63, №10. – P. 881-883. https://doi.org/10.1016/0038-1098(87)90331-0

82. Monti, G.A. Solid state nuclear magnetic resonance of polymers / G.A. Monti, R.H. Acosta, A.K. Chattah, Y.G. Linck // Journal of Magnetic Resonance Open. – 2023. – V. 16-17. – Article 100119. https://doi.org/10.1016/j.jmro.2023.100119

83. Shankarayya Wadi, V.K. NMR and EPR structural analysis and stability study of inverse vulcanized sulfur copolymers / V.K. Shankarayya Wadi, K.K. Jena, S.Z. Khawaja, K. Yannakopoulou, M. Fardis, G. Mitrikas, M. Karagianni, G. Papavassiliou, S.M. Alhassan // ACS Omega. – 2018. – V. 3, №3. – P. 3330-3339. https://doi.org/10.1021/acsomega.8b00031

84. Zachmann, H.G. Investigation of the glass transition and melting of polymers by nuclear magnetic resonance / H.G. Zachmann // Journal of Polymer Science: Polymer Symposia. – 1973. – V. 43, №1. – P. 111-123. https://doi.org/10.1002/polc.5070430112

85. Naveed, K.-ur-R. Recent progress in the electron paramagnetic resonance study of polymers / K.-R. Naveed, L. Wang, H. Yu, R.S. Ullah, M. Haroon, S. Fahad, J. Li, T. Elshaarani,R.U. Khan, A. Nazir // Polymer Chemistry. – 2018. – V. 9, №24. – P. 3306-3335. https://doi.org/10.1039/C8PY00689J

86. Uddin, M.A. Recent progress in EPR study of spin labeled polymers and spin probed polymer systems / M.A. Uddin, H. Yu, L. Wang, K.-ur-R. Naveed, F. Haq, B.U. Amin, S. Mehmood, A. Nazir, Y. Xing, D. Shen // Journal of Polymer Science. – 2020. – V. 58, №14. – P. 1924-1948. https://doi.org/10.1002/pol.20200039

87. Jenckel, E. Zur temperaturabhängigkeit der viscosität von schmelzen / E. Jenckel // Zeitschrift für Physikalische Chemie. – 1939. – V. 184A, №1. – P. 309-319. https://doi.org/10.1515/zpch-1939-18425

88. Tammann, G. Die abhängigkeit der viscosität von der temperatur bei unterkühlten flüssigkeiten / G. Tammann, W. Hesse // Zeitschrift für anorganische und allgemeine Chemie. – 1926. – V. 156, №1. – P. 245-257. https://doi.org/10.1002/zaac.19261560121

89. Kozlov, G.V. A cluster model for the polymer amorphous state / G.V. Kozlov, V.U. Novikov // Physics – Uspekhi. – 2001. – V. 44, №7. – P. 681-724. https://doi.org/10.1070/PU2001v044n07ABEH000832

90. Aksenov, V.L. Kinetic equations for describing the liquid-glass transition in polymers / V.L. Aksenov, T.V. Tropin, J.V.P. Schmelzer // Theoretical and Mathematical Physics. – 2018. – V. 194, №1. – P. 142-147. https://doi.org/10.1134/S0040577918010105

91. Sanditov, D.S. Relaxation aspects of the liquid–glass transition / D.S. Sanditov, M.I. Ojovan // Physics – Uspekhi. – 2019. – V. 62, №2. – P. 111-130. https://doi.org/10.3367/UFNe.2018.04.038319

92. Baldanza, A. Modelling changes in glass transition temperature in polymer matrices exposed to low molecular weight penetrants / A. Baldanza, V. Loianno, G. Mensitieri, G. Scherillo // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. – 2023. – V. 381, №2240. – Article 20210216. https://doi.org/10.1098/rsta.2021.0216

93. Banerjee, A. Data-driven identification and analysis of the glass transition in polymer melts / A. Banerjee, H.-P. Hsu, K. Kremer, O. Kukharenko // ACS Macro Letters. – 2023. – V. 12, №6. – P. 679-684. https://doi.org/10.1021/acsmacrolett.2c00749

94. Xu, W.-S. A thermodynamic perspective on polymer glass formation / W.-S. Xu, Z.-Y. Sun // Chinese Journal of Polymer Science. – 2023. – V. 41, №9. – P. 1329-1341. https://doi.org/10.1007/s10118-023-2951-1

95. Liang, J. Glass transition in monolayers of rough colloidal ellipsoids / J. Liang, X. Feng, N. Zheng, H. Wang, R. Ni, Z. Zhang // Physical Review Letters. – 2025. – V. 134, №3. – Article 038202. https://doi.org/10.1103/PhysRevLett.134.038202

96. Ростиашвили, В.Г. Стеклование полимеров / В.Г. Ростиашвили, В.И. Иржак, Б.А. Розенберг. – Л.: Химия, 1987. – 192 с.

97. van Krevelen, D.W. Properties of polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions / D.W. van Krevelen, revised by K. te Nijenhuis. – 4th ed. – Amsterdam: Elsevier, 2009. – 1004 p.

98. Bicerano, J. Prediction of polymer properties / J. Bicerano. – 3rd ed. – New York: Marcel Dekker, Inc., 2002. – 746 p.

99. Askadskii, A.A. Computational materials science of polymers / A.A. Askadskii. – Cambridge: Cambridge International Science Publishing, 2003. – 696 p.

100. Askadskii, A.A. Further research on the improvement of models and computer programs for the prediction and analysis of the physical properties of polymers / A.A. Askadskii, T.A. Matseevich // Physics-Uspekhi. – 2023. – V. 66, №6. – P. 586-627. https://doi.org/10.3367/UFNe.2021.12.039124

101. Volgin, I.V. Machine learning with enormous «synthetic» data sets: predicting glass transition temperature of polyimides using graph convolutional neural networks / I.V. Volgin, P.A. Batyr, A.V. Matseevich, A.Yu. Dobrovskiy, M.V. Andreeva, V.M. Nazarychev, S.V. Larin, M.Ya. Goikhman, Y.V. Vizilter, A.A. Askadskii, S.V. Lyulin// ACS Omega. – 2022. – V. 7, №48. – P. 43678-43691. https://doi.org/10.1021/acsomega.2c04649

102. Katritzky, A.R. Prediction of polymer glass transition temperatures using a general quantitative structure−property relationship treatment / A.R. Katritzky, P. Rachwal, K.W. Law, M. Karelson, V.S. Lobanov // Journal of Chemical Information and Computer Sciences. – 1996. – V. 36, №4. – P. 879-884. https://doi.org/10.1021/ci950156w

103. Camelio, P. A novel approach toward the prediction of the glass transition temperature: application of the EVM model, a designer QSPR equation for the prediction of acrylate and methacrylate polymers / P. Camelio, C.C. Cypcar, V. Lazzeri, B. Waegell // Journal of Polymer Science Part A: Polymer Chemistry. – 1997. – V. 35, №13. – P. 2579-2590. https://doi.org/10.1002/(SICI)1099-0518(19970930)35:13%3C2579::AID-POLA5%3E3.0.CO;2-M

104. Mattioni, B.E. Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks / B.E. Mattioni, P.C. Jurs // Journal of Chemical Information and Computer Sciences. – 2002. – V. 42, №2. – P. 232-240. https://doi.org/10.1021/ci010062o

105. Morrill, J.A. Prediction of the formulation dependence of the glass transition temperatures of amine-epoxy copolymers using a QSPR based on the AM1 method / J.A. Morrill, R.E. Jensen, P.H. Madison, C.F. Chabalowski // Journal of Chemical Information and Computer Sciences. – 2004. – V. 44, №3. – P. 912-920. https://doi.org/10.1021/ci030290d

106. Chen, X. A neural network approach to prediction of glass transition temperature of polymers / X. Chen, L. Sztandera, H.M. Cartwright // International Journal of Intelligent Systems. – 2008. – V. 23. – P. 22-32. https://doi.org/10.1002/int.20256

107. Liu, W. Artificial neural network prediction of glass transition temperature of polymers / W. Liu, C. Cao // Colloid and Polymer Science. – 2009. – V. 287. – P. 811-818. https://doi.org/10.1007/s00396-009-2035-y

108. Yu, X. Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers / X. Yu // Fibers and Polymers. – 2010. – V. 11, №5. – P. 757-766. https://doi.org/10.1007/s12221-010-0757-6

109. Hamerton, I. Predicting glass transition temperatures of polyarylethersulphones using QSPR methods / I. Hamerton, B.J. Howlin, G. Kamyszek // PLoS ONE. – 2012. – V. 7, №6. – Article e38424. https://doi.org/10.1371/journal.pone.0038424

110. Pei, J.F. Modeling and predicting the glass transition temperature of vinyl polymers by using hybrid PSO-SVR method / J.F. Pei, C.Z. Cai, Y.M. Zhu // Journal of Theoretical and Computational Chemistry. – 2013. – V. 12, №3. – Article 1350002. https://doi.org/10.1142/S0219633613500028

111. Khan, P.M. QSPR modelling for prediction of glass transition temperature of diverse polymers / P.M. Khan, K. Roy // SAR and QSAR in Environmental Research. – 2018. – V. 29, №12. – P. 935-956. https://doi.org/10.1080/1062936X.2018.1536078

112. Chen, M. A computational structure–property relationship study of glass transition temperatures for a diverse set of polymers / M. Chen, F. Jabeen, B. Rasulev, M. Ossowski, P. Boudjouk // Journal of Polymer Science, Part B: Polymer Physics. – 2018. – V. 56, №11. – P. 877-885. https://doi.org/10.1002/polb.24602

113. Pilania, G. Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers / G. Pilania, C.N. Iverson, T. Lookman, B.L. Marrone // Journal of Chemical Information and Modeling. – 2019. – V. 59, №12. – P. 5013-5025. https://doi.org/10.1021/acs.jcim.9b00807

114. Ma, R. Evaluating polymer representations via quantifying structure–property relationships / R. Ma, Z. Liu, Q. Zhang, Z. Liu, T. Luo // Journal of Chemical Information and Modeling. – 2019. – V. 59, №7. – P. 3110-3119. https://doi.org/10.1021/acs.jcim.9b00358

115. Wen, C. Determination of glass transition temperature of polyimides from atomistic molecular dynamics simulations and machine-learning algorithms / C. Wen, B. Liu, J. Wolfgang, T.E. Long, R. Odle, S. Cheng // Journal of Polymer Science. – 2020. – V. 58, №11. – P. 1521-1534. https://doi.org/10.1002/pol.20200050

116. Ramprasad, M. Assessing and improving machine learning model predictions of polymer glass transition temperatures / M. Ramprasad, C. Kim // Journal of Emerging Investigators. – 2020. – V. 3. – P. 1-5. https://doi.org/10.59720/19-097

117. Miccio, L.A. From chemical structure to quantitative polymer properties prediction through convolutional neural networks / L.A. Miccio, G.A. Schwartz // Polymer. – 2020. – V. 193. – Article 122341. https://doi.org/10.1016/j.polymer.2020.122341

118. Tao, L. Machine learning discovery of high-temperature polymers / L. Tao, G. Chen, Y. Li // Patterns. – 2021. – V. 2, №4. – Article 100225. https://doi.org/10.1016/j.patter.2021.100225

119. Goswami, S. Deep learning based approach for prediction of glass transition temperature in polymers / S. Goswami, R. Ghosh, A. Neog, B. Das // Materials Today: Proceedings. – 2021. – V. 46, №12. – P. 5838-5843. https://doi.org/10.1016/j.matpr.2021.02.730

120. Zhang, Y. Machine learning glass transition temperature of polymethacrylates / Y. Zhang, X. Xu // Molecular Crystals and Liquid Crystals. – 2021. – V. 730, №1. – P. 9-22. https://doi.org/10.1080/15421406.2021.1946348

121. Chen, G. Predicting polymers' glass transition temperature by a chemical language processing model / G. Chen, L. Tao, Y. Li // Polymers. – 2021. – V. 13, №11. – Article 1898. https://doi.org/10.3390/polym13111898

122. Uddin, M.J. Interpretable machine learning framework to predict the glass transition temperature of polymers / M.J. Uddin, J. Fan // Polymers. – 2024. – V. 16, №8. – Article 1049. https://doi.org/10.3390/polym16081049

123. Casanola-Martin, G.M. Machine learning analysis of a large set of homopolymers to predict glass transition temperatures / G.M. Casanola-Martin, A. Karuth, H. Pham-The, H. González-Díaz, D.C. Webster, B. Rasulev //

Communications Chemistry. – 2024. – V. 7. – Article 226. https://doi.org/10.1038/s42004-024-01305-0

124. Long, Z. Large-scale glass-transition temperature prediction with an equivariant neural network for screening polymers / Z. Long, H. Lu, Z. Zhang // ACS Omega. – 2024. – V. 9, №5. – P. 5452-5462. https://doi.org/10.1021/acsomega.3c06843

125. Fatriansyah, J.F. Prediction of glass transition temperature of polymers using simple machine learning / J.F. Fatriansyah, B.D.P. Linuwih, Y. Andreano, I.S. Sari, A. Federico, M. Anis, S.N. Surip, M. Jaafar// Polymers. – 2024. – V. 16, №17. – Article 2464. https://doi.org/10.3390/polym16172464

126. Yan, C. Forecast of glass transition zone of thermoset polymers using a multiscale machine learning approach / C. Yan, X. Feng, P. Mensah, G. Li // The Journal of Physical Chemistry B. – 2025. – V. 129, №9. – P. 2621-2636. https://doi.org/10.1021/acs.jpcb.4c07666

127. Durant, J.L. Reoptimization of MDL keys for use in drug discovery / J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse // Journal of Chemical Information and Computer Sciences. – 2002. – V. 42, №6. – P. 1273-1280. https://doi.org/10.1021/ci010132r

128. Rogers, D. Extended-connectivity fingerprints / D. Rogers, M. Hahn // Journal of Chemical Information and Modeling. – 2010. – V. 50, №5. – P. 742-754. https://doi.org/10.1021/ci100050t

129. Pedregosa, F. Scikit-learn: machine learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay // Journal of Machine Learning Research. – 2011. – V. 12. – P. 2825-2830.

130. Navarro, A. A quantum mechanical study on polymer flexibility: extended model from monomer to tetramer of 2- and 4-bromostyrenes / A. Navarro, M.P. Fernández-Liencres, T. Peña-Ruiz, J.M. Granadino-Roldán, M. Fernández-Gómez, G. Domínguez-Espinosa, M.J. Sanchís // Polymer. – 2009. – V. 50, №1. – P. 317-327. https://doi.org/10.1016/j.polymer.2008.10.041

131. Gaussian 16, Revision C.01, Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Petersson, G.A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.V.; Bloino, J.; Janesko, B.G.; Gomperts, R.; Mennucci, B.; Hratchian, H.P.; Ortiz, J. V.; Izmaylov, A.F.; Sonnenberg, J.L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V.G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J.A., Jr.; Peralta, J.E.; Ogliaro, F.; Bearpark, M.J.; Heyd, J.J.; Brothers, E.N.; Kudin, K.N.; Staroverov, V.N.; Keith, T.A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.P.; Burant, J.C.; Iyengar, S.S.; Tomasi, J.; Cossi, M.; Millam, J.M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J.W.; Martin, R.L.; Morokuma, K.; Farkas, O.; Foresman, J.B.; Fox, D.J. Gaussian, Inc., Wallingford CT, 2016.

132. Becke, A.D. Density-functional thermochemistry. III. The role of exact exchange / A.D. Becke // The Journal of Chemical Physics. – 1993. – V. 98, №7. – P. 5648-5652. https://doi.org/10.1063/1.464913

133. Lee, C. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density / C. Lee, W. Yang, R.G. Parr // Physical Review B. – 1988. – V. 37, №2. – P. 785-789. https://doi.org/10.1103/PhysRevB.37.785

134. Frisch, M.J. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets / M.J. Frisch, J.A. Pople, J.S. Binkley // The Journal of Chemical Physics. – 1984. – V. 80, №7. – P. 3265-3269. https://doi.org/10.1063/1.447079

135. Bhatia, M. An overview of conceptual-DFT based insights into global chemical reactivity of volatile sulfur compounds (VSCs) / M. Bhatia // Computational Toxicology. – 2024. – V. 29. – Article 100295. https://doi.org/10.1016/j.comtox.2023.100295

136. Šverko, Z. Complex Pearson correlation coefficient for EEG connectivity analysis / Z. Šverko, M. Vrankić, S. Vlahinić, P. Rogelj // Sensors. – 2022. – V. 22, №4.

– Article 1477. https://doi.org/10.3390/s22041477

137. Маджидов, Т.И. Введение в хемоинформатику: 1. Компьютерное представление химических структур / Т.И. Маджидов, И.И. Баскин, И.С. Антипин, А.А. Варнек. – Казань: Изд-во Казан. ун-та, 2013. – 174 с.

138. Kim, H.-J. Fluorine-induced local magnetic moment in graphene: a hybrid DFT study / H.-J. Kim, J.-H. Cho // Physical Review B. – 2013. – V. 87, №17. – Article 174435. https://doi.org/10.1103/PhysRevB.87.174435

139. Privalko, V.P. On the molecular packing density in crystalline polymers / V.P. Privalko // Polymer Journal. – 1975. – V. 7, №2. – P. 202-206. https://doi.org/10.1295/polymj.7.202

140. Askadskii, A.A. Methods for calculating the physical properties of polymers / A.A. Askadskii // Review Journal of Chemistry. – 2015. – V. 5, №2. – P. 83-142. https://doi.org/10.1134/S2079978015020016

141. Шадрина, Г.Р. Гибридный подход при прогнозировании температур стеклования органических гомополимеров: сочетание модели QSPR и метода инкрементов / Г.Р. Шадрина, В.И. Анисимова, И.С. Родионов, А.А. Балдинов, Н.В. Улитин, Я.Л. Люлинская, К.А. Терещенко, Д.А. Шиян // Вестник технологического университета. – 2025. – Т. 28, №3. – С. 68-74.

142. Улитин, Н.В. Интерпретация закономерности «структура-температура стеклования» для органических гомополимеров с использованием методов инкрементов и «случайного леса», а также теории функционала плотности / Н.В. Улитин, Г.Р. Шадрина, В.И. Анисимова, И.С. Родионов, А.А. Балдинов, Я.Л. Люлинская, К.А. Терещенко, Д.А. Шиян // Журнал структурной химии. – 2025. – Т. 66, №5. – Статья 147275. (англ. версия: Ulitin, N.V. Interpretation of the structure–glass transition temperature relationship for organic homopolymers with the use of increment, random forest, and density functional theory methods / N.V. Ulitin, G.R. Shadrina, V.I. Anisimova, I.S. Rodionov, A.A. Baldinov, Ya.L. Lyulinskaya, K.A. Tereshchenko, D.A. Shiyan // Journal of Structural Chemistry. – 2025. – V. 66, №5. – P. 1095-1109.)

143. Шадрина, Г.Р. Прогнозирование температуры стеклования органических гомополимеров с использованием молекулярных отпечатков и алгоритма случайного леса/ Г.Р. Шадрина, Н.В. Улитин, А.Д. Лифанов, К.А. Терещенко, А.А. Балдинов, И.С. Родионов, М.Г. Казанская, А.О. Софьин, Д.А. Шиян // XXXIV Рос. молодеж. науч. конф. с междунар. уч-м, посвящ. 190-летию со дня рожд. Д.И. Менделеева «Проблемы теоретической и экспериментальной химии» (Екатеринбург, Уральский федеральный университет, 23-26 апреля 2024 г.): сб. тез. докл. – Екатеринбург: Изд-во Урал. ун-та, 2024. – С. 61.

144. Шадрина, Г.Р. О влиянии межмолекулярных взаимодействий на температуру стеклования органических гомополимеров / Г.Р. Шадрина, А.А. Балдинов, Н.В. Улитин, В.И. Анисимова, А.Д. Лифанов, Д.А. Шиян, К.А. Терещенко, И.С. Родионов, М.Г. Казанская, Е.С. Воробьев, И.А. Суворова // IX Всерос. науч. конф. «Теоретические и экспериментальные исследования процессов синтеза, модификации и переработки полимеров» (Уфа, Уфимский университет науки и технологий, 3-4 июня 2024 г.): сб. тез. докл. – Уфа: РИЦ УУНиТ, 2024. – С. 38-40.

145. Шадрина, Г.Р. Анализ факторов, влияющих на температуру стеклования органических гомополимеров, с привлечением машинного обучения / Г.Р. Шадрина, Н.В. Улитин, А.Д. Лифанов, К.А. Терещенко, Д.А. Шиян, М.Г. Казанская, А.А. Балдинов, И.С. Родионов, Е.С. Воробьев // IX Всерос. науч. конф. «Теоретические и экспериментальные исследования процессов синтеза, модификации и переработки полимеров» (Уфа, Уфимский университет науки и технологий, 3-4 июня 2024 г.): сб. тез. докл. – Уфа: РИЦ УУНиТ, 2024. – С. 75-77.

146. Шадрина, Г.Р. Модель для прогнозирования температуры стеклования органических гомополимеров на основе метода инкрементов с использованием алгоритма случайного леса / Г.Р. Шадрина, Н.В. Улитин, В.И. Анисимова, А.А. Балдинов, И.С. Родионов, Д.А. Шиян, К.А. Терещенко, Я.Л. Люлинская // IV Всерос. науч. конф. (с междунар. уч-м) преп-й и студ-в вузов «Актуальные проблемы науки о полимерах» (Казань, Казанский национальный

исследовательский технологический университет, 23-26 сентября 2024 г.): сб. тез. докл. – Казань: Изд-во КНИТУ, 2024. – С. 202-205.

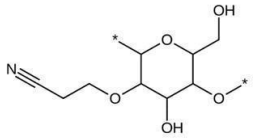
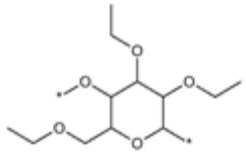
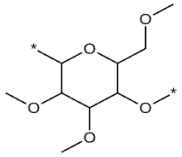
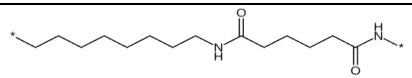
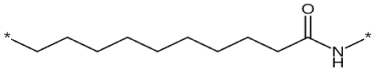
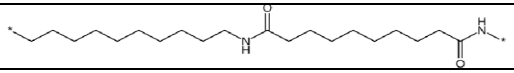
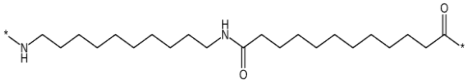
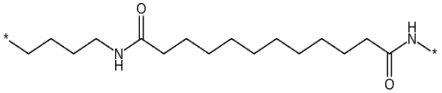
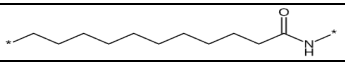
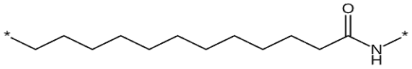
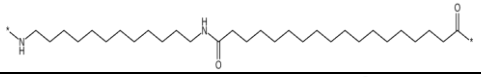
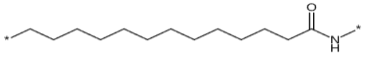
147. Шадрина, Г.Р. Использование дескрипторов и алгоритма случайного леса для прогнозирования температуры стеклования органических гомополимеров / Г.Р. Шадрина, А.Д. Лифанов, Н.В. Улитин, А.А. Балдинов, И.С. Родионов, Д.А. Шиян, М.Г. Казанская, А.А. Кулыгин, А.О. Софьин, К.А. Терещенко // Междунар. науч. конф. «Актуальные вопросы естествознания и функциональные полимеры для фармацевтики, нефтяной промышленности, экологии, био- и нанотехнологии», посвященная 125-летию проф. К. Жубанова (Казахстан, г. Актобе, Актюбинский региональный университет им. К. Жубанова, 25-27 сентября 2024 г.): сб. мат-в. – Актобе: Изд-во Актюб. рег. ун-та им. К. Жубанова, 2024. – С. 160-161.

Приложение А

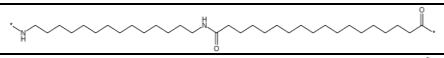
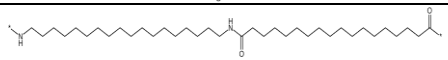
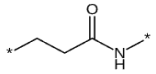
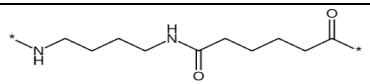
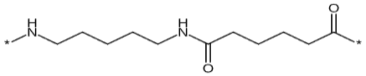
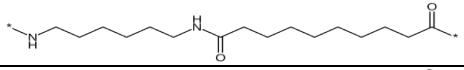
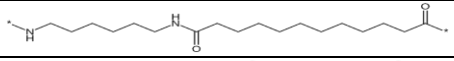
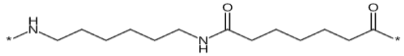
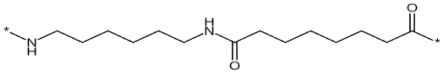
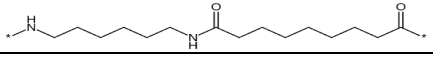
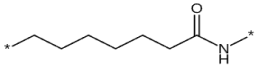
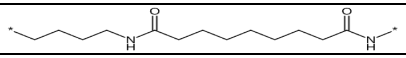
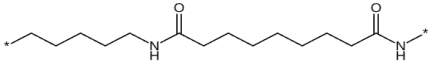
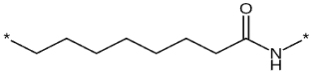
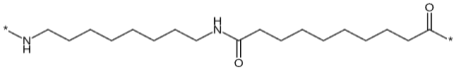
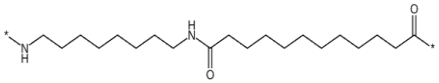
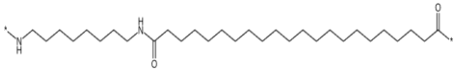
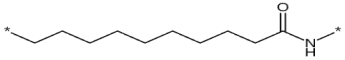
(обязательное)

Фрагмент итоговой базы данных

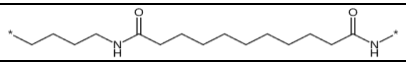
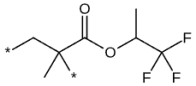
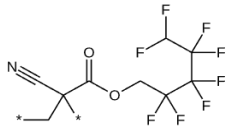
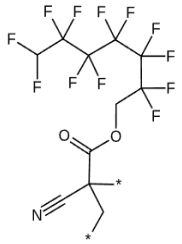
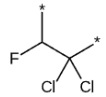
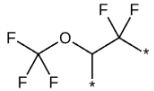
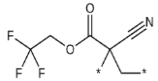
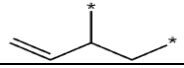
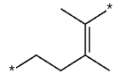
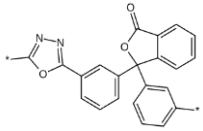
Таблица А.1 – Фрагмент итоговой базы данных

№[а]	SMILES[б]	Химическое строение повторяющегося звена полимера	$A_{inc} = \sum_i \Delta V_i$, Å^3	$B_{inc} = \sum_i a_i \Delta V_i$, Å^3/K	$C_{inc} = \sum_j b_j$, Å^3/K	$T_{g\,exp}$, K
1	2	3	4	5	6	7
1	O(C1C(CO)OC(C(OCCC#N)C1O)[*])[*]		188	1184.506	-701.506	453
2	CCOCC1OC(C(OCC)C(OCC)C1O[*])[*]		168	1272.874	-618.874	316
3	COCC1OC(C(OC)C(OC)C1O[*])[*]		151	1032.328	-536.328	423
4	N(C(CCCCC(NCCCCCCCCC[*])=O)=O)[*]		266	1279.956	-468.956	318
5	C(CCCCCCCCC(=O)N[*])[*]		184	880.524	-300.524	315
6	C(CCCCCCCCCNC(=O)CCCCCCCCC(=O)N[*])[*]		368	1761.048	-600.048	333
7	N(CCCCCCCCCCNC(=O)CCCCCCCCCCC(=O)[*])[*]		402	1921.412	-641.412	322
8	C(CCCNC(=O)CCCCCCCCCCC(=O)N[*])[*]		300	1440.32	-511.32	313
9	C(CCCCCCCCCC(=O)N[*])[*]		201	960.706	-320.706	315
10	C(CCCCCCCCCCC(=O)N[*])[*]		218	1040.888	-341.888	313
11	N(CCCCCCCCCCCCNC(=O)CCCCCCCCCCCCCCCCC(=O)[*])[*]		539	2562.868	-812.868	323
12	C(CCCCCCCCCCCC(=O)N[*])[*]		235	1121.07	-365.07	314

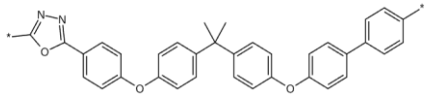
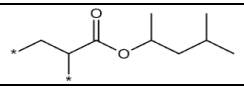
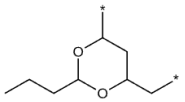
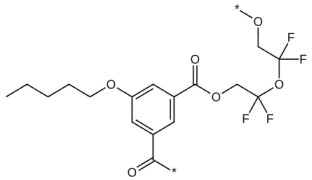
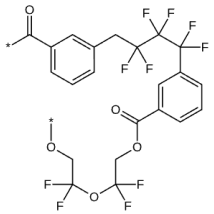
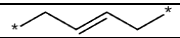
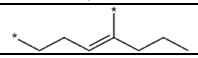
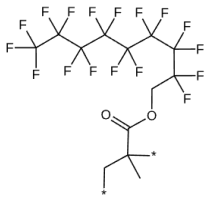
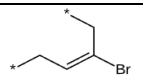
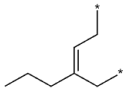
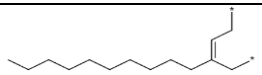
Продолжение табл. A1

1	2	3	4	5	6	7
13	N(CCCCCCCCCCCCCCCNC(=O)CCCCCCCCCCCCCCCCCC(=O)[*])[*]		573	2723.232	-857.232	321
14	N(CCCCCCCCCCCCCCCCCCCCNC(=O)CCCCCCCCCCCCCCCCCC(=O)[*])[*]		641	3043.96	-941.96	323
15	C(CC(=O)N[*])[*]		64.7	319.25	-149.25	384
16	N(CCCCNC(=O)CCCCC(=O)[*])[*]		198	959.228	-382.228	316
17	N(CCCCCNC(=O)CCCCC(=O)[*])[*]		215	1039.41	-405.41	318
18	N(CCCCCCNC(=O)CCCCCCCCC(=O)[*])[*]		300	1440.32	-511.32	323
19	N(CCCCCCNC(=O)CCCCCCCCCCC(=O)[*])[*]		334	1600.684	-556.684	319
20	N(CCCCCCNC(=O)CCCCCC(=O)[*])[*]		249	1199.774	-447.774	331
21	N(CCCCCCNC(=O)CCCCCCC(=O)[*])[*]		266	1279.956	-468.956	330
22	N(CCCCCCNC(=O)CCCCCCCC(=O)[*])[*]		283	1360.138	-492.138	330
23	C(CCCCCC(=O)N[*])[*]		133	639.978	-234.978	331
24	C(CCCNC(=O)CCCCCCCC(=O)N[*])[*]		249	1199.774	-447.774	333
25	C(CCCCNC(=O)CCCCCCCC(=O)N[*])[*]		266	1279.956	-468.956	328
26	C(CCCCCCC(=O)N[*])[*]		150	720.16	-256.16	323
27	N(CCCCCCCCNC(=O)CCCCCCCCC(=O)[*])[*]		334	1600.684	-556.684	333
28	N(CCCCCCCCNC(=O)CCCCCCCCCCC(=O)[*])[*]		368	1761.048	-600.048	323
29	N(CCCCCCCCNC(=O)CCCCCCCCCCCCCCCCCCCCC(=O)[*])[*]		539	2562.868	-812.868	321
30	C(CCCCCCCCC(=O)N[*])[*]		167	880.524	-358.524	319

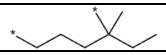
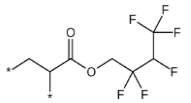
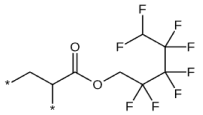
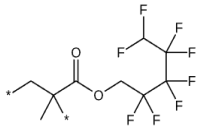
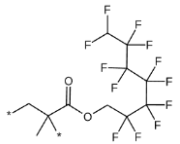
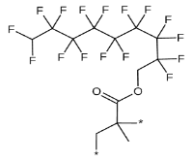
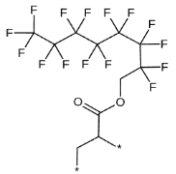
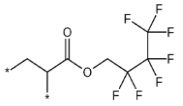
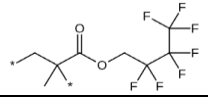
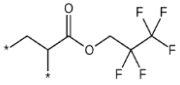
Продолжение табл. A1

1	2	3	4	5	6	7
31	C(CCCNC(=O)CCCCCCCCCC(=O)N[*])[*]		283	1360.138	-492.138	318
32	[*]CC(C)(C(=O)OC(C)C(F)(F)F)[*]		146	648.576	-254.576	354
33	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)F)=O)(C#N)[*]		209	863.06	-222.06	363
34	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)F)=O)(C#N)[*]		264	1111.824	-219.824	330
35	[*]C(F)C(Cl)(Cl)[*]		68.9	261.222	-56.222	320
36	[*]C(F)(F)C(OC(F)(F)F)[*]		86.6	405.572	-87.572	273
37	[*]CC(C(OCC(F)(F)F)=O)(C#N)[*]		132	511.986	-197.986	373
38[в)]	[*][Si](C)(C)N[*]		76.1	534.84	-125.84	191
39	[*]C(C)(CCC[*])C		102	481.094	-115.094	253
41	[*]CC(C=C)[*]		64.2	240.806	-0.806	269
42	[*]\C(C)=C(\CC[*])C		98.3	401.172	24.828	262
43	[*]C5=NN=C(C1=CC(=CC=C1)C3(C2=CC=CC=C2C(=O)O3)C4=CC(=CC=C4)[*])O5		306	633.111	-139.111	653

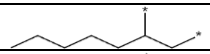
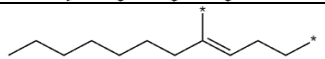
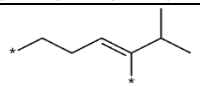
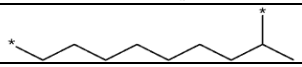
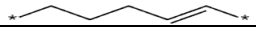
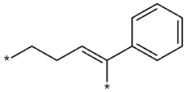
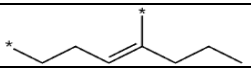
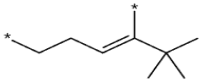
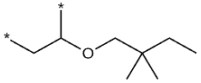
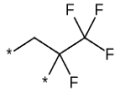
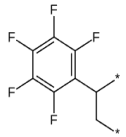
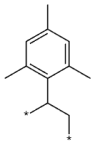
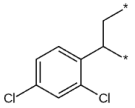
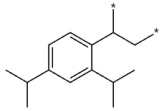
Продолжение табл. A1

1	2	3	4	5	6	7
45	[*]C1=NN=C(C2=CC=C(OC3=CC=C(C(C)(C4=CC=C(OC5=CC=C(C6=CC=C([*])C=C6)C=C5)C=C4)C)C=C3)C=C2)O1		418	1191.315	-260.315	453
46	[*]CC(C(OC(CC(C)C)C)=O)[*]		165	742.638	-163.638	258
47	[*]CC1CC([*])OC(CCC)O1		144	670.42	-172.42	322
48	[*]OCC(F)(F)OC(F)(F)COC(C1=CC(=CC(=C1)OCCCCC)C([*])=O)=O		324	1342.318	-110.318	287
49	[*]OCC(F)(F)OC(F)(F)COC(C1=CC(=CC=C1)C(F)(F)C(F)(F)C(F)(F)CC2=CC(=CC=C2)C([*])=O)=O		387	1512.596	-149.596	303
51	[*]CC=CC[*]		64.2	240.808	134.192	171
52	[*]C(CCC)=CCC[*]		98.3	481.354	7.646	197
53	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)=O)(C)[*]		321	1439.066	-200.066	310
54	[*]CC(Br)=CC[*]		86	288.953	79.047	241
55	[*]CC(CCC)=CC[*]		132	481.354	185.646	192
56	[*]CC(CCCCCCCCCC)=CC[*]		235	1042.628	162.372	220

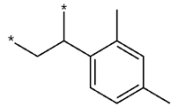
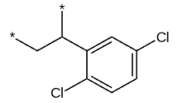
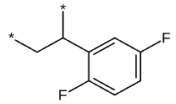
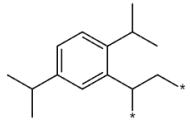
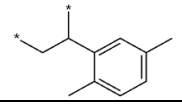
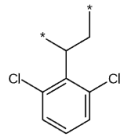
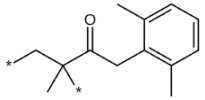
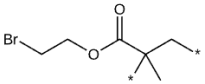
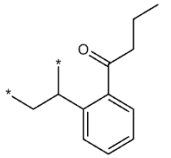
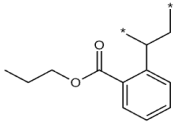
Продолжение табл. A1

1	2	3	4	5	6	7
57	[*]C(C)(CC)CCC[*]		119	561.276	-107.276	250
58	[*]CC(C(OCC(F)(F)C(F)C(F)(F)F)=O)[*]		162	714.874	-91.874	251
59	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)F)=O)[*]		189	839.282	-115.282	238
60	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)F)=O)(C)[*]		206	919.466	-223.466	309
61	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)F)=O)(C)[*]		261	1168.23	-222.23	286
62	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)F)=O)(C)[*]		316	1416.994	-219.994	258
63	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)=O)[*]		277	1234.5	-89.5	256
64	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)F)=O)[*]		167	736.972	-91.972	243
65	[*]CC(C(OCC(F)(F)C(F)(F)C(F)(F)F)=O)(C)[*]		184	817.156	-294.156	330
66	[*]CC(C(OCC(F)(F)C(F)(F)F)=O)[*]		139	612.59	-93.59	247

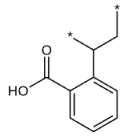
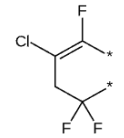
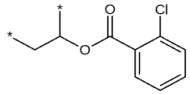
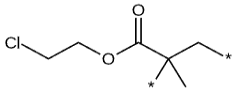
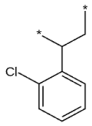
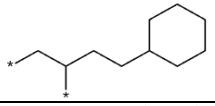
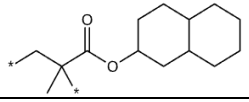
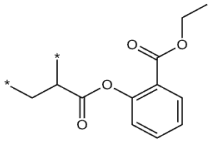
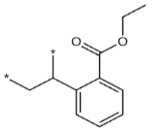
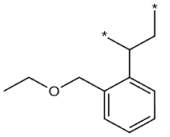
Продолжение табл. А1

1	2	3	4	5	6	7
67	[*]CC(CCCCC)[*]		119	561.274	-23.274	226
68	[*]C(CCCCCCC)=CCC[*]		184	802.082	136.918	190
69	[*]C(C(C)C)=CCC[*]		115	481.354	20.646	221
70	[*]C(C)CCCCCCC[*]		154	721.638	-57.638	215
71	[*]C=CCCC[*]		81.3	320.99	136.01	183
72	[*]C(C1=CC=CC=C1)=CCC[*]		140	402	107	283
73	[*]C(CCC)=CCC[*]		115	481.354	96.646	196
74	[*]C(C(C)(C)C)=CCC[*]		132	561.538	-33.538	293
75	[*]CC(OCC(C)(C)CC)[*]		146	695.984	-158.984	282
76	[*]C(F)(C(F)(F)F)C[*]		71.9	328.948	-88.948	315
77	[*]C(C1=C(F)C(F)=C(F)C(F)=C1F)C[*]		135	428.594	-57.594	378
78	[*]CC(C1=C(C)C=C(C)C=C1C)[*]		160	562.07	-196.07	435
79	[*]CC(C1=CC=C(Cl)C=C1Cl)[*]		137	397.884	-59.884	406
80	[*]C(C1=CC=C(C(C)C)C=C1C(C)C)C[*]		211	802.626	-254.626	435

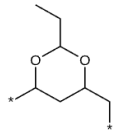
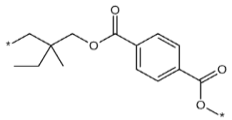
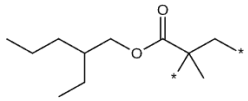
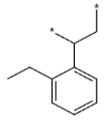
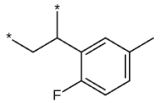
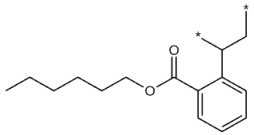
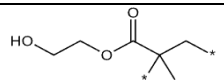
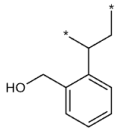
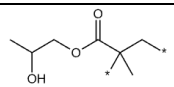
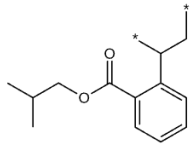
Продолжение табл. А1

1	2	3	4	5	6	7
81	[*]C(C1=CC=C(C)C=C1C)C[*]		143	481.898	-141.898	385
82	[*]C(C1=CC(=CC=C1Cl)Cl)C[*]		137	397.884	-59.884	379
83	[*]C(C1=CC(=CC=C1F)F)C[*]		120	364.37	-57.37	374
84	[*]C(C1=CC(=CC=C1C(C)C)C(C)C)C[*]		211	802.626	-254.626	441
85	[*]C(C1=CC(=CC=C1C)C)C[*]		143	481.898	-141.898	416
86	[*]C(C1=C(Cl)C=CC=C1Cl)C[*]		137	397.884	-59.884	440
87	[*][C](C)(C(CC1=C(C)C=CC=C1C)=O)C[*]		188	688.92	-257.92	440
88	[*][C](C)(C(OCCBr)=O)C[*]		135	550.239	-170.239	325
89	[*]C(C1=CC=CC=C1C(CCC)=O)C[*]		206	608.746	15.854	339
90	[*]C(C1=CC=CC=C1C(OCCC)=O)C[*]		204	663.272	47.728	340

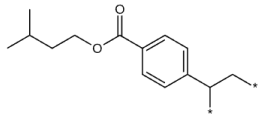
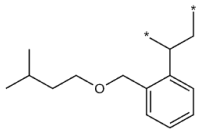
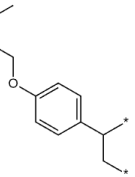
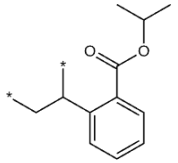
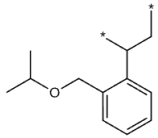
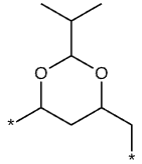
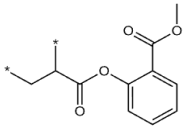
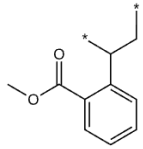
Продолжение табл. А1

1	2	3	4	5	6	7
91	[*]C(C1=CC=CC=C1C(O)=O)C[*]		135	511.728	-162.728	450
92	[*]\C(F)=C(/CC(F)(F)[*])Cl		94.5	346.484	25.516	256
93	[*]CC(OC(C1=CC=CC=C1Cl)=O)[*]		152	460.903	-35.903	335
94	[*]C[C](C)(C(OCCCl)=O)[*]		128	541.468	-167.468	365
95	[*]CC(C1=CC=CC=C1Cl)[*]		124	359.719	-57.719	392
96	[*]CC(CCC1CCCCC1)[*]		159	721.736	-159.736	313
97	[*]C[C](C)(C(OC2CC1CCCCC1CC2)=O)[*]		226	983.382	-465.382	418
98	[*]CC(C(OC1=C(C(OCC)=O)C=CC=C1)=O)[*]		200	673.078	-56.078	303
99	[*]CC(C1=CC=CC=C1C(OCC)=O)[*]		172	583.09	-138.49	391
100	[*]CC(C1=CC=CC=C1COCC)[*]		170	616.612	-85.612	347

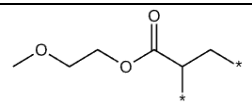
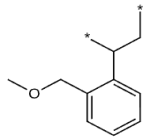
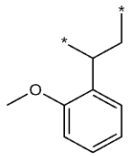
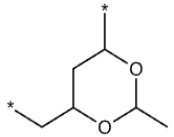
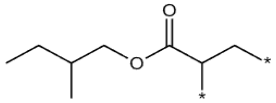
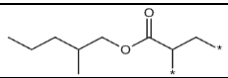
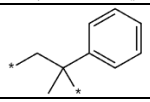
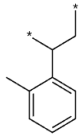
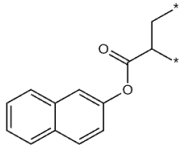
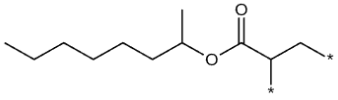
Продолжение табл. А1

1	2	3	4	5	6	7
101	[*]CC1CC([*])OC(CC)O1		126	590.238	-184.238	345
102	CC(COC(C1=CC=C(C(O[*])=O)C=C1)=O)(CC)C[*]		231	891.898	-232.898	340
103	[*]C[C](C)(C(OCC(CC)CCC)=O)[*]		216	903.002	-106.002	263
104	[*]CC(C1=CC=CC=C1CC)[*]		143	481.908	-77.908	376
105	[*]CC(C1=CC(=CC=C1F)C)[*]		131	423.134	-86.134	384
106	[*]CC(C1=CC=CC=C1C(OCCCCCC)=O)[*]		240	903.818	-102.218	318
107	[*]C[C](C)(C(OCCO)=O)[*]		121	645.618	-299.618	358
108	[*]CC(C1=CC=CC=C1CO)[*]		134	545.25	-218.25	433
109	[*]C[C](C)(C(OCC(O)C)=O)[*]		138	725.802	-292.802	328
110	[*]CC(C1=CC=CC=C1C(OCC(C)C)=O)[*]		206	743.454	-194.854	400

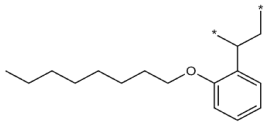
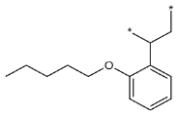
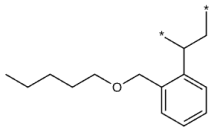
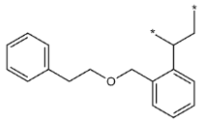
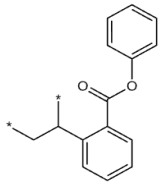
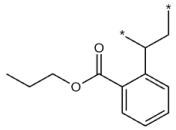
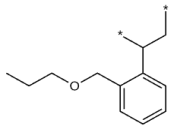
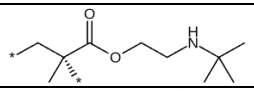
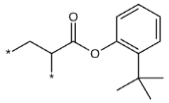
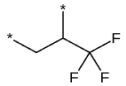
Продолжение табл. А1

1	2	3	4	5	6	7
111	[*]CC(C1=CC=C(C(OCCC(C)C)=O)C=C1)[*]		223	823.636	-186.036	341
112	[*]CC(C1=CC=CC=C1COCCC(C)C)[*]		221	857.158	-132.158	351
113	[*]CC(C1=CC=C(OCCC(C)C)C=C1)[*]		204	765.786	-129.786	330
114	[*]CC(C1=CC=CC=C1C(OC(C)C)=O)[*]		189	663.274	-193.674	419
115	[*]CC(C1=CC=CC=C1COC(C)C)[*]		187	696.796	-141.796	361
116	[*]CC1CC([*])OC(C(C)C)O1		144	670.42	-246.42	329
117	[*]CC(C(OC1=CC=CC=C1C(OC)=O)=O)[*]		182	592.896	-58.896	319
118	[*]CC(C1=CC=CC=C1C(OC)=O)[*]		155	502.908	-81.908	403

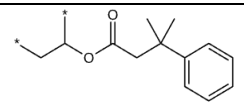
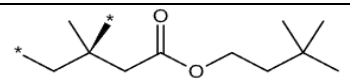
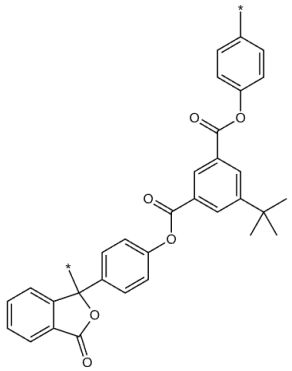
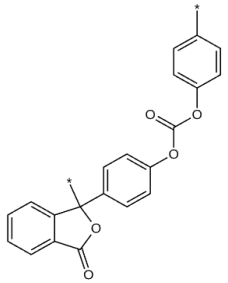
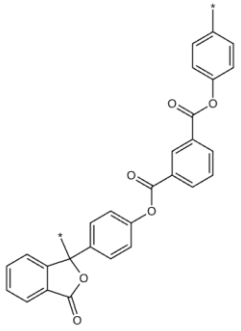
Продолжение табл. А1

1	2	3	4	5	6	7
119	[*]CC(C(OCCOC)=O)[*]		125	556.614	-60.614	223
120	[*]CC(C1=CC=CC=C1COC)[*]		153	536.43	-85.43	362
121	[*]CC(C1=CC=CC=C1OC)[*]		136	445.058	-82.058	348
122	[*]CC1CC([*])OC(C)O1		109	510.056	-192.056	355
123	[*]CC(C(OCC(CC)C)=O)[*]		148	662.454	-106.454	241
124	[*]CC(C(OCC(CCC)C)=O)[*]		165	742.636	-97.636	235
125	[*]CC(C1=CC=CC=C1)(C)[*]		126	401.738	-87.738	375
126	[*]CC(C1=CC=CC=C1C)[*]		126	401.726	-86.726	409
127	[*]CC(C(OC2=CC1=CC=CC=C1C=C2)=O)[*]		183	492.294	-5.294	358
128	[*]CC(C(OC(CCCCCC)C)=O)[*]		199	903.002	-70.002	228

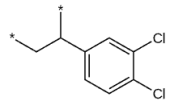
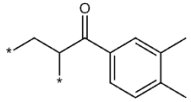
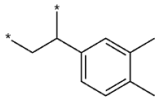
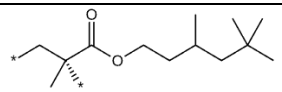
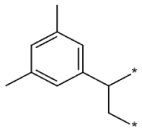
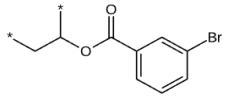
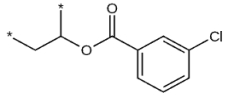
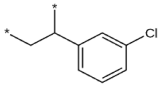
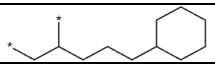
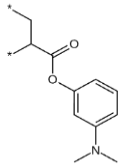
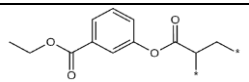
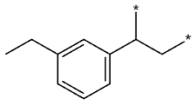
Продолжение табл. А1

1	2	3	4	5	6	7
129	[*]CC(C1=CC=CC=C1OCCCCCCCC)[*]		255	1006.332	-29.332	286
130	[*]CC(C1=CC=CC=C1OCCCCC)[*]		223	765.786	-53.786	365
131	[*]CC(C1=CC=CC=C1COCCCCC)[*]		221	857.158	-59.158	320
132	[*]CC(C1=CC=CC=C1COCCC2=CC=CC=C2)[*]		246	777.802	-47.802	336
133	[*]CC(C1=CC=CC=C1C(OC2=CC=CC=C2)=O)[*]		213	572.724	-110.724	397
134	[*]CC(C1=CC=CC=C1C(OCCC)=O)[*]		189	663.272	-129.672	381
135	[*]CC(C1=CC=CC=C1COCCC)[*]		187	696.794	-75.794	370
136	[*]C[C@](C)(C(OCCNC(C)(C)C)=O)[*]		194	923.799	-276.799	306
137	[*]CC(C(OC1=C(C(C)(C)C)C=CC=C1)=O)[*]		206	732.262	-169.262	345
138	[*]CC(C(F)(F)F)[*]		66.7	306.846	-89.846	300

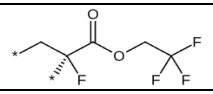
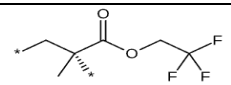
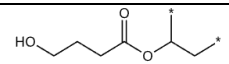
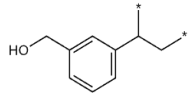
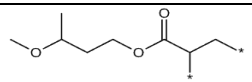
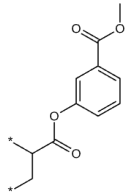
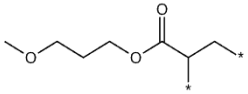
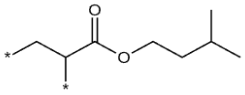
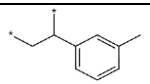
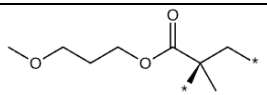
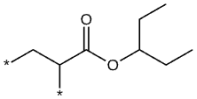
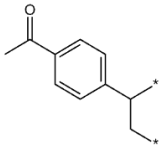
Продолжение табл. А1

1	2	3	4	5	6	7
139	[*]CC(OC(CC(C)(C)C1=CC=CC=C1)=O)[*]		206	743.466	-134.466	293
140	[*]C[C@@](CC(OCCC(C)(C)C)=O)(C)[*]		182	903.004	-350.004	318
141	[*]C(C1=CC=CC=C1C2=O)(O2)C3=CC=C(OC(C4=CC(C(C)(C)C)=CC(C(OC5=CC=C([*])C=C5)=O)=C4)=O)C=C3		457	1284.326	-414.326	552
142	[*]C(C1=CC=CC=C1C2=O)(O2)C3=CC=C(OC(OC4=CC=C([*])C=C4)=O)C=C3		296	755.766	-205.766	538
143	[*]C(C1=CC=CC=C1C2=O)(O2)C3=CC=C(OC(C4=CC=CC(C(OC5=CC=C([*])C=C5)=O)=C4)=O)C=C3		390	963.606	-243.606	543

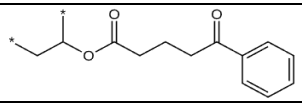
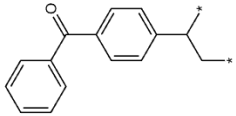
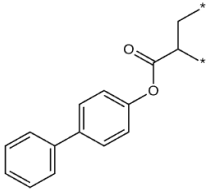
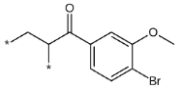
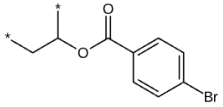
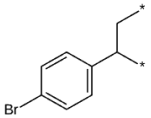
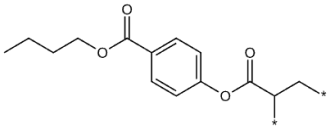
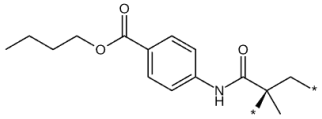
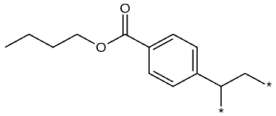
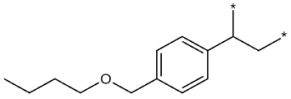
Продолжение табл. А1

1	2	3	4	5	6	7
144	[*]CC(C1=CC=C(Cl)C(Cl)=C1)[*]		137	397.884	-59.884	401
145	[*]CC(C(C1=CC=C(C)C(C)=C1)=O)[*]		162	528.554	-117.554	315
146	[*]CC(C1=CC=C(C)C(C)=C1)[*]		143	481.898	-141.898	384
147	[*]C[C@](C)(C(OCCC(CC(C)(C)C)C)=O)[*]		233	1063.368	-316.368	274
148	[*]CC(C1=CC(C)=CC(C)=C1)[*]		143	481.898	-141.898	377
149	[*]CC(OC(C1=CC=CC(Br)=C1)=O)[*]		159	466.461	-30.461	331
150	[*]CC(OC(C1=CC=CC(Cl)=C1)=O)[*]		151	460.903	-32.903	338
151	[*]CC(C1=CC=CC(Cl)=C1)[*]		124	359.719	-57.719	363
152	[*]CC(CCCC1CCCCC1)[*]		176	801.918	-149.918	248
153	[*]CC(C(OC1=CC(N(C)C)=CC=C1)=O)[*]		184	620.299	-115.299	320
154	[*]CC(C(OC1=CC(C(OCC)=O)=CC=C1)=O)[*]		200	673.078	-56.078	297
155	[*]CC(C1=CC=CC(CC)=C1)[*]		143	481.908	-77.908	303

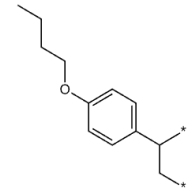
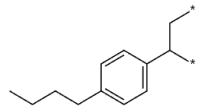
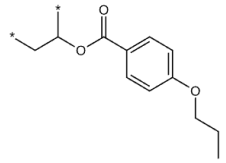
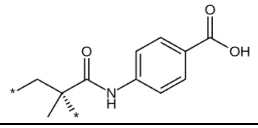
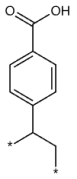
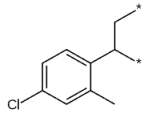
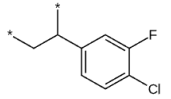
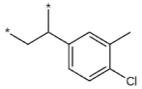
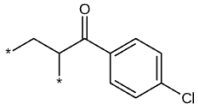
Продолжение табл. А1

1	2	3	4	5	6	7
156	[*]C[C@](F)(C(OCC(F)(F)F)=O)[*]		117	510.31	-199.31	394
157	[*]C[C@](C)(C(OCC(F)(F)F)=O)[*]		129	568.392	-199.392	355
158	[*]CC(OC(CCCO)=O)[*]		121	645.618	-194.618	278
159	[*]CC(C1=CC=CC(CO)=C1)[*]		134	545.25	-218.25	398
160	[*]CC(C(OCCC(OC)C)=O)[*]		157	716.98	-105.98	217
161	[*]CC(C(OC1=CC(C(OC)=O)=CC=C1)=O)[*]		182	592.896	-58.896	331
162	[*]CC(C(OCCCOC)=O)[*]		140	636.796	-50.796	198
163	[*]CC(C(OCCC(C)C)=O)[*]		148	662.454	-106.454	228
164	[*]CC(C1=CC=CC(C)=C1)[*]		126	401.726	-86.726	370
165	[*]C[C@@](C(OCCCOC)=O)(C)[*]		157	716.98	-155.98	289
166	[*]CC(C(OC(CC)CC)=O)[*]		148	662.456	-97.456	267
167	[*]CC(C1=CC=C(C(C)=O)C=C1)[*]		145	448.382	-84.982	389

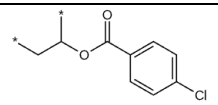
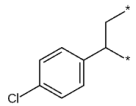
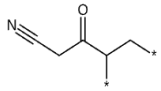
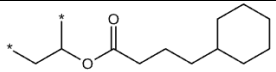
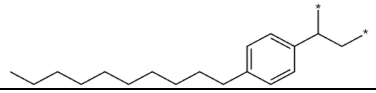
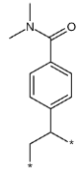
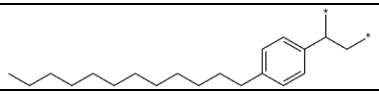
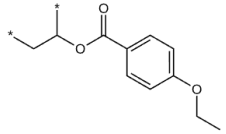
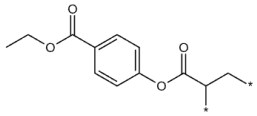
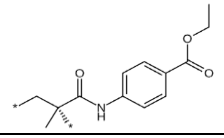
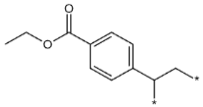
Продолжение табл. А1

1	2	3	4	5	6	7
168	[*]CC(OC(CCCC(C1=CC=CC=C1)=O)=O)[*]		208	709.938	-77.938	318
169	[*]CC(C1=CC=C(C(C2=CC=CC=C2)=O)C=C1)[*]		204	529.392	-1.392	371
170	[*]CC(C(OC1=CC=C(C2=CC=CC=C2)C=C1)=O)[*]		213	572.722	-33.722	383
171	[*]CC(C(C1=CC=C(Br)C(OC)=C1)=O)[*]		175	535.437	-59.437	317
172	[*]CC(OC(C1=CC=C(Br)C=C1)=O)[*]		159	466.461	-30.461	365
173	[*]CC(C1=CC=C(Br)C=C1)[*]		131	365.277	-55.277	422
174	[*]CC(C(OC1=CC=C(C(OCCCC)=O)C=C1)=O)[*]		234	833.442	75.558	286
175	[*]C[C@@](C(NC1=CC=C(C(OCCCC)=O)C=C1)=O)(C)[*]		254	972.046	-303.046	401
176	[*]CC(C1=CC=C(C(OCCCC)=O)C=C1)[*]		206	743.454	-118.454	349
177	[*]CC(C1=CC=C(COCCCC)C=C1)[*]		204	776.976	-65.976	283

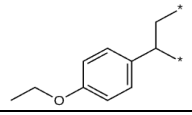
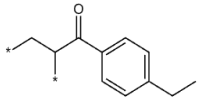
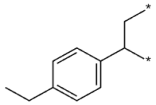
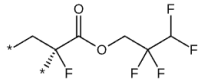
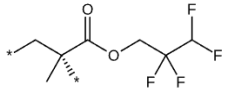
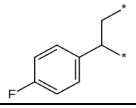
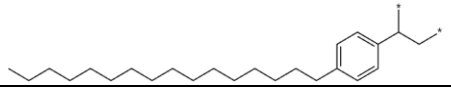
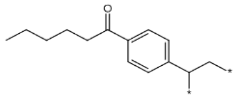
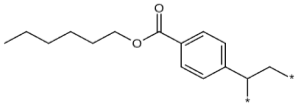
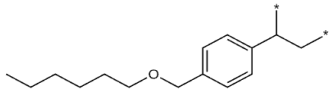
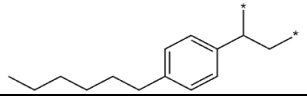
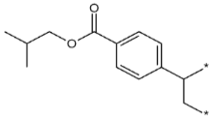
Продолжение табл. А1

1	2	3	4	5	6	7
178	[*]CC(C1=CC=C(OCCCC)C=C1)[*]		187	685.604	-64.604	320
179	[*]CC(C1=CC=C(CCCC)C=C1)[*]		178	642.272	-56.272	279
180	[*]CC(OC(C1=CC=C(OCCC)C=C1)=O)[*]		217	706.606	8.394	324
181	[*]C[C@](C)(C(NC1=CC=C(C(O)=O)C=C1)=O)[*]		183	740.32	-456.32	473
182	[*]CC(C1=CC=C(C(O)=O)C=C1)[*]		135	511.728	-218.728	386
183	[*]CC(C1=CC=C(Cl)C=C1C)[*]		140	439.891	-85.891	418
184	[*]CC(C1=CC=C(Cl)C(F)=C1)[*]		129	381.127	-57.127	395
186	[*]CC(C1=CC=C(Cl)C(C)=C1)[*]		140	439.891	-85.891	387
187	[*]CC(C(C1=CC=C(Cl)C=C1)=O)[*]		147	406.375	-34.375	336

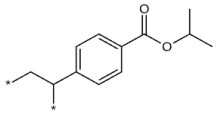
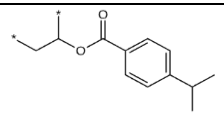
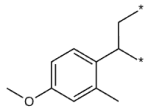
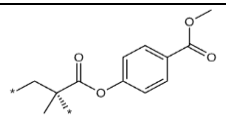
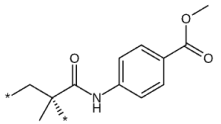
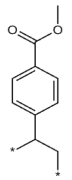
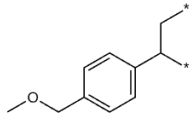
Продолжение табл. А1

1	2	3	4	5	6	7
188	[*]CC(OC(C1=CC=C(Cl)C=C1)=O)[*]		151	460.903	-32.903	357
189	[*]CC(C1=CC=C(Cl)C=C1)[*]		124	359.719	-57.719	395
190	[*]CC(C(CC#N)=O)[*]		150	310.978	313.022	236
191	[*]CC(OC(CCCC1CCCCC1)=O)[*]		204	903.1	-172.1	263
192	[*]CC(C1=CC=C(CCCCCCCCCC)C=C1)[*]		280	1123.364	-3.364	208
193	[*]CC(C1=CC=C(C(N(C)C)=O)C=C1)[*]		174	583.942	-146.942	398
194	[*]CC(C1=CC=C(CCCCCCCCCCCC)C=C1)[*]		314	1283.728	-145.728	221
195	[*]CC(OC(C1=CC=C(OCC)C=C1)=O)[*]		181	626.424	-58.424	343
196	[*]CC(C(OC1=CC=C(C(OCC)=O)C=C1)=O)[*]		200	673.078	-56.078	310
197	[*]C[C@@](C(NC1=CC=C(C(OCC)=O)C=C1)=O)(C)[*]		220	811.682	-320.682	441
198	[*]CC(C1=CC=C(C(OCC)=O)C=C1)[*]		172	583.09	-138.09	367

Продолжение табл. А1

1	2	3	4	5	6	7
199	[*]CC(C1=CC=C(OCC)C=C1)[*]		153	525.24	-83.24	359
200	[*]CC(C(C1=CC=C(CC)C=C1)=O)[*]		162	528.564	-52.564	325
201	[*]CC(C1=CC=C(CC)C=C1)[*]		143	481.908	-77.908	326
202	[*]C[C@@](C(OCC(F)(F)C(F)F)=O)(F)[*]		139	612.62	-222.62	368
203	[*]C[C@@](C(OCC(F)(F)C(F)F)=O)(C)[*]		151	670.702	-222.702	353
204	[*]CC(C1=CC=C(F)C=C1)[*]		115	342.962	-56.962	368
205	[*]CC(C1=CC=C(CCCCCCCCCCCCCCCCC)C=C1)[*]		382	1604.456	-429.456	278
206	[*]CC(C1=CC=C(C(CCCCC)=O)C=C1)[*]		213	769.11	-47.71	339
207	[*]CC(C1=CC=C(C(OCCCCCC)=O)C=C1)[*]		240	903.818	-102.218	339
208	[*]CC(C1=CC=C(COCCCCCC)C=C1)[*]		238	937.34	-49.34	253
209	[*]CC(C1=CC=C(CCCCCC)C=C1)[*]		212	802.636	-37.636	246
210	[*]CC(C1=CC=C(C(OCC(C)C)=O)C=C1)[*]		206	743.454	-194.454	363

Окончание табл. А1

1	2	3	4	5	6	7
211	[*]CC(C1=CC=C(C(OC(C)C)=O)C=C1)[*]		189	663.274	-193.274	368
212	[*]CC(OC(C1=CC=C(C(C)C)C=C1)=O)[*]		189	663.274	-143.274	342
213	[*]CC(C1=CC=C(OC)C=C1C)[*]		152	525.23	-139.23	361
214	[*]C[C@](C)(C(OC1=CC=C(C(OC)=O)C=C1)=O)[*]		200	673.08	-191.08	379
215	[*]C[C@](C)(C([NH]C1=CC=C(C(OC)=O)C=C1)=O)[*]		202	731.5	-321.5	453
216	[*]CC(C1=CC=C(C(OC)=O)C=C1)[*]		155	502.908	-81.908	386
217	[*]CC(C1=CC=C(COC)C=C1)[*]		153	536.43	-85.43	350

а) Указанная нумерация повторяющихся звеньев полимеров соответствует таковой в табл. 19', стр. 117 [Askadskii, A.A. Computational materials science of polymers / A.A. Askadskii. – Cambridge: Cambridge International Science Publishing, 2003. – 696 p.].

б) SMILES (от англ. Simplified Molecular Input Line Entry System) – система представления молекул в виде одномерной строки символов (подробнее см. главу 1). $A_{inc} = \sum_i \Delta V_i$, $B_{inc} = \sum_i a_i \Delta V_i$, $C_{inc} = \sum_j b_j$ – компоненты формулы (3), см. главу 3; в программном коде (см. Приложение Б): $V = A_{inc}$, $aV = B_{inc}$, $b = C_{inc}$. $T_{g\,exp}$ – экспериментальные значения температуры стеклования полимеров.

в) База данных по органическим гомополимерам дополнена кремнийорганическими гомополимерами с целью проверки возможности потенциального расширения применения модели для прогнозирования температуры стеклования полимеров других классов.

Приложение Б

(обязательное)

Программный код

Характеристики вычислительных ресурсов: Apple M1 (16 ядер), ОЗУ 8 ГБ. Используемое программное обеспечение: среда разработки – Jupyter Notebook, язык программирования – Python 3.13.7.

Импорт библиотек для использования кодом функций, классов и переменных, определенных в этих библиотеках

```
from sklearn.metrics import r2_score, mean_squared_error, balanced_accuracy_score, roc_curve # Импорт функций для оценки метрик моделей (регрессия, классификация).

from sklearn.model_selection import RepeatedKFold, cross_val_predict, cross_val_score, GridSearchCV, train_test_split # Импорт функций для разделения данных, кросс-валидации и оптимизации гиперпараметров.

from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor # Импорт ансамблевых моделей регрессии (Random Forest, Gradient Boosting).

from sklearn.neighbors import KNeighborsClassifier, KNeighborsRegressor # Импорт моделей ближайших соседей (KNeighbors) для классификации и регрессии.

from sklearn.neural_network import MLPRegressor # Импорт модели многослойного перцептрона (MLPRegressor) для регрессии.

from sklearn.datasets import make_regression # Импорт функции для генерации синтетических регрессионных наборов данных.

from sklearn.preprocessing import StandardScaler, FunctionTransformer # Импорт функций для стандартизации данных и применения пользовательских преобразований.

# from sklearn import tree # Закомментированный импорт модуля для моделей дерева решений.

import matplotlib.pyplot as plt # Импорт библиотеки для визуализации данных.

# from matplotlib import rcParams # Закомментированный импорт модуля для настройки параметров графиков.

import pandas as pd # Импорт библиотеки для работы с табличными данными (DataFrame).

from rdkit.Chem import Descriptors, AllChem # Импорт модулей RDKit для вычисления молекулярных дескрипторов и работы с химическими структурами.

from rdkit import DataStructs # Импорт модуля RDKit для работы со структурами данных (например, отпечатки молекул).
```

 # from sklearn.compose import ColumnTransformer # Закомментированный импорт инструмента для применения различных преобразований к столбцам.

import numpy as np # *Импорт библиотеки для работы с численными массивами.*

from numpy import zeros, array, unique, arange, trapz # *Импорт специфических функций из NumPy для создания массивов, получения уникальных значений, генерации последовательностей и численного интегрирования.*

from rdkit import Chem # *Импорт основного модуля RDKit для работы с химическими структурами.*

Генерация молекулярных отпечатков Моргана из структур повторяющихся звеньев полимеров, представленных в базе данных в виде SMILES, для использования данного вида дескрипторов в целях обучения моделей

data = pd.read_excel(open('db_check/28_12_2023_res/29_12_23_data_color.xlsx','rb')) # *Загрузка данных из файла Excel в объект DataFrame.*]

smiles = data.SMILES # *Извлечение SMILES-строк из столбца 'SMILES' DataFrame.*

mols = [Chem.MolFromSmiles(smi) for smi in smiles] # *Конвертация каждой SMILES-строки в соответствующий объект молекулы RDKit.*

"**MorganFingerprint**"

X = [] # *Инициализация пустого списка для хранения генерируемых молекулярных отпечатков.*

for m in mols: # *Итерация по каждому объекту молекулы в коллекции.*

arr = zeros((1,), dtype=int) # *Инициализация одномерного массива NumPy для временного хранения битового вектора отпечатка.*

DataStructs.ConvertToNumpyArray(AllChem.GetMorganFingerprintAsBitVect(m, 2, 1024), arr) # *Генерация битового вектора Morgan (радиус 2, 1024 бита) для молекулы и его запись в массив NumPy.*

X.append(arr) # *Добавление полученного битового отпечатка в список признаков X.*

Генерация структурных ключей из структур повторяющихся звеньев полимеров, представленных в базе данных в виде SMILES, для использования данного вида дескрипторов в целях обучения моделей

"**Structural keys**"

X = [] # *Инициализация пустого списка для хранения генерируемых структурных ключей.*

for m in mols: # *Итерация по каждому объекту молекулы в коллекции.*

fp_MACCS = AllChem.GetMACCSKeysFingerprint(m) # *Генерация структурного ключа MACCS для текущей молекулы.*

vector = (fp_MACCS) # *Присвоение сгенерированного структурного ключа переменной 'vector'.*

X.append(vector) # *Добавление полученного структурного ключа в список признаков X.*

V, aV, b, Tg_exp = data.V, data.aV, data.b, data.Tg # *Извлечение данных из соответствующих столбцов DataFrame и присвоение их переменным.*

data = list(zip(V, aV, b, Tg_exp)) # *Объединение извлеченных данных в список кортежей.*

Y = list(zip(V, aV, b)) # *Создание списка кортежей Y, содержащего подмножество извлеченных данных.*

plt.hist(V, bins = 30) # *Построение гистограммы распределения переменной V с 30 интервалами для проверки соответствия распределения Гаусса*.

plt.hist(aV, bins = 30) # *Построение гистограммы распределения переменной aV с 30 интервалами для проверки соответствия распределения Гаусса*.

plt.hist(b, bins = 30) # *Построение гистограммы распределения переменной b с 30 интервалами для проверки соответствия распределения Гаусса*.

Деление базы данных на обучающую и тестовую выборки

X_train, X_test, data_train, data_test = train_test_split(X, data, test_size=0.2) # *Разделение признаков (X) и полных данных (data) на обучающую и тестовую выборки с долей 20% для тестовой выборки.*

V_train, aV_train, b_train, Tg_exp_train = zip(*data_train) # *Распаковка обучающей выборки данных на отдельные переменные V, aV, b и Tg_exp.*

y_train = list(zip(V_train, aV_train, b_train)) # *Формирование целевой переменной y_train из V, aV и b для обучающей выборки.*

V_test, aV_test, b_test, Tg_exp_test = zip(*data_test) # *Распаковка тестовой выборки данных на отдельные переменные V, aV, b и Tg_exp.*

y_test = list(zip(V_test, aV_test, b_test)) # *Формирование целевой переменной y_test из V, aV и b для тестовой выборки.*

Прогнозирование с использованием MLPRegressor (проверка прогностической способности модели на основе многослойного перцептрона)

"**MLPRegressor**"

MLP = MLPRegressor(random_state=1, max_iter=2000) # *Инициализация модели многослойного перцептрона (MLPRegressor) с фиксированным случайным состоянием и максимальным числом итераций.*

MLP.fit(X_train, y_train) # *Обучение модели MLPRegressor на обучающих данных (признаки X_train, целевые значения y_train).*

y_pred = MLP.predict(X_test) # *Выполнение прогнозирования целевых значений для тестовых данных (X_test) с использованием обученной модели.*

q2, rmse = r2_score(y_test, y_pred), mean_squared_error(y_test, y_pred) # *Вычисление коэффициента детерминации (R²) и среднеквадратичной ошибки (RMSE) между фактическими (y_test) и предсказанными (y_pred) значениями.*

print(f'Q\N{SUPERSCRIPT TWO}: {q2:.2f}\nRMSE: {rmse:.2f}') # *Вывод вычисленных метрик Q² и RMSE с форматированием до двух знаков после запятой.*

Прогнозирование с использованием KNeighborsRegressor (проверка прогностической способности модели на основе метода ближайших соседей)

"**KNeighborsRegressor**"

kNN = KNeighborsRegressor() # *Инициализация модели регрессии K-ближайших соседей (KNeighborsRegressor) с параметрами по умолчанию.*

kNN.fit(X_train, y_train) # *Обучение модели KNeighborsRegressor на обучающих данных (признаки X_train, целевые значения y_train).*

y_pred = kNN.predict(X_test) # *Выполнение прогнозирования целевых значений для тестовых данных (X_test) с использованием обученной модели.*

q2, rmse = r2_score(y_test, y_pred), mean_squared_error(y_test, y_pred) # *Вычисление коэффициента детерминации (R²) и среднеквадратичной ошибки (RMSE) между фактическими (y_test) и предсказанными (y_pred) значениями.*

print(f'Q\N{SUPERSCRIPT TWO}: {q2:.2f}\nRMSE: {rmse:.2f}') # *Вывод вычисленных метрик Q² и RMSE с форматированием до двух знаков после запятой.*

Прогнозирование с использованием RandomForestRegressor (проверка прогностической способности модели на основе метода случайного леса)

"**Random Forest**"

rfr = RandomForestRegressor(n_estimators=500) # *Инициализация модели регрессии случайного леса (RandomForestRegressor) с 500 деревьями.*

rfr.fit(X_train, y_train) # *Обучение модели RandomForestRegressor на обучающих данных (признаки X_train, целевые значения y_train).*

y_pred = rfr.predict(X_test) # *Выполнение прогнозирования целевых значений для тестовых данных (X_test) с использованием обученной модели.*

q2, rmse = r2_score(y_test, y_pred), mean_squared_error(y_test, y_pred) # *Вычисление коэффициента детерминации (R²) и среднеквадратичной ошибки (RMSE) между фактическими (y_test) и предсказанными (y_pred) значениями.*

print(f'Q\N{SUPERSCRIPT TWO}: {q2:.2f}\nRMSE: {rmse:.2f}') # *Вывод вычисленных метрик Q² и RMSE с форматированием до двух знаков после запятой.*

Tg_calc = [(y[0]/(y[1]+y[2])*1000) for y in y_pred] # *Расчет значений Tg_calc на основе предсказанных значений y_pred по определенной формуле.*

r2_score(Tg_exp_test, Tg_calc) *# Вычисление коэффициента детерминации (R²) между фактическими значениями Tg_exp_test и расчетными Tg_calc.*

fig, ax = plt.subplots(figsize=(10, 6)) *# Создание объекта фигуры и осей для графика с заданным размером.*

ax.scatter(x = Tg_exp_test, y = Tg_calc) *# Построение диаграммы рассеяния, сопоставляющей фактические (Tg_exp_test) и расчетные (Tg_calc) значения.*

plt.show() *# Отображение построенного графика.*

Проверка качества прогнозирования отдельных параметров (V, aV, b) с использованием модели на основе RandomForestRegressor

"**Checking individual parameters**"

rfr = RandomForestRegressor(n_estimators=500) *# Инициализация RandomForestRegressor с 500 деревьями.*

rfr.fit(X_train, V_train) *# Обучение модели RandomForestRegressor для прогнозирования V_train на основе признаков X_train.*

V_pred = rfr.predict(X_test) *# Выполнение прогнозирования значений V для тестовых данных (X_test) с использованием обученной модели.*

q2, rmse = r2_score(V_test, V_pred), mean_squared_error(V_test, V_pred) *# Вычисление коэффициента детерминации (R²) и среднеквадратичной ошибки (RMSE) для прогноза V.*

print(f'Q\N{SUPERSCRIPT TWO}: {q2:.2f}\nRMSE: {rmse:.2f}') *# Вывод вычисленных метрик Q² и RMSE для прогноза V.*

rfr = RandomForestRegressor(n_estimators=500) *# Инициализация новой RandomForestRegressor с 500 деревьями.*

rfr.fit(X_train, aV_train) *# Обучение модели RandomForestRegressor для прогнозирования aV_train на основе признаков X_train.*

aV_pred = rfr.predict(X_test) *# Выполнение прогнозирования значений aV для тестовых данных (X_test) с использованием обученной модели.*

q2, rmse = r2_score(aV_test, aV_pred), mean_squared_error(aV_test, aV_pred) *# Вычисление коэффициента детерминации (R²) и среднеквадратичной ошибки (RMSE) для прогноза aV.*

print(f'Q\N{SUPERSCRIPT TWO}: {q2:.2f}\nRMSE: {rmse:.2f}') *# Вывод вычисленных метрик Q² и RMSE для прогноза aV.*

rfr = RandomForestRegressor(n_estimators=500) *# Инициализация новой модели RandomForestRegressor с 500 деревьями.*

rfr.fit(X_train, b_train) *# Обучение модели RandomForestRegressor для прогнозирования b_train на основе признаков X_train.*

b_pred = rfr.predict(X_test) *# Выполнение прогнозирования значений b для тестовых данных (X_test) с использованием обученной модели.*

q2, rmse = r2_score(b_test, b_pred), mean_squared_error(b_test, b_pred) # *Вычисление коэффициента детерминации (R^2) и среднеквадратичной ошибки (RMSE) для прогноза b.*

print(f'Q\N{SUPERSCRIPT TWO}: {q2:.2f}\nRMSE: {rmse:.2f}') # *Вывод вычисленных метрик Q^2 и RMSE для прогноза b.*

Подбор гиперпараметров для модели на основе RandomForestRegressor

"**Hyperparameter selection**"

from sklearn.model_selection import GridSearchCV # *Импорт класса GridSearchCV для подбора оптимальных гиперпараметров модели.*

param_grid = { # *Определение словаря 'param_grid' с диапазонами значений гиперпараметров для оптимизации.*

'bootstrap': [True, False], # *Гиперпараметр 'bootstrap': выбор метода формирования выборок для деревьев.*

'max_depth': [15, 40, 50], # *Гиперпараметр 'max_depth': максимальная глубина каждого дерева.*

'max_features': [2, 10, 15, 20], # *Гиперпараметр 'max_features': количество признаков для рассмотрения при каждом разбиении.*

'min_samples_leaf': [1, 2, 5, 10], # *Гиперпараметр 'min_samples_leaf': минимальное количество выборок, необходимых в листовом узле.*

'min_samples_split': [2, 3, 4], # *Гиперпараметр 'min_samples_split': минимальное количество выборок, необходимых для разделения внутреннего узла.*

'n_estimators': [50, 100, 150, 200] # *Гиперпараметр 'n_estimators': количество деревьев в лесу.*

}

Создадим базовую модель

rf = RandomForestRegressor(n_estimators=500, random_state=123) # *Инициализация базовой модели RandomForestRegressor с фиксированным числом деревьев и случайным состоянием.*

Создадим экземпляр модели с помощью GridSearch

grid = GridSearchCV(estimator = rf, param_grid = param_grid, # *Инициализация GridSearchCV для поиска лучших гиперпараметров.*

cv = 5, n_jobs = -1, verbose = 2, refit = True) # *Конфигурация GridSearchCV: 5-кратная кросс-валидация, использование всех ядер ЦП, подробный вывод, переобучение лучшей моделью.*

grid.fit(X_train, y_train) # *Обучение модели с подбором гиперпараметров на обучающих данных.*

grid_predictions = grid.predict(X_test) # *Выполнение прогнозирования на тестовых данных с использованием лучшей модели, найденной GridSearchCV.*

q2, rmse = r2_score +(y_test, grid_predictions), mean_squared_error(y_test, grid_predictions, squared=False) # *Вычисление коэффициента детерминации (R²) и среднеквадратичной ошибки (RMSE) для прогнозов.*

print(q2) # *Вывод значения коэффициента детерминации R².*

print(rmse) # *Вывод значения среднеквадратичной ошибки RMSE.*

grid.best_params_ # *Вывод оптимальных гиперпараметров, найденных GridSearchCV.*

Прогнозирование с использованием RandomForestRegressor с учетом кросс-валидации

"**Cross validation**"

rfr = RandomForestRegressor(n_estimators=500) # *Инициализация модели RandomForestRegressor с 500 деревьями.*

y_pred = cross_val_predict(rfr, X, Y, cv=10) # *Выполнение прогнозирования с использованием 10-кратной кросс-валидации для модели RandomForestRegressor на полных данных.*

Tg_calc2 = [(y[0]/(y[1]+y[2])*1000) for y in y_pred] # *Расчет значений Tg_calc2 на основе предсказанных значений y_pred по заданной формуле.*

r2_score(Tg_exp, Tg_calc2) # *Вычисление коэффициента детерминации (R²) между фактическими значениями Tg_exp и расчетными Tg_calc2.*

fig, ax = plt.subplots(figsize=(10, 6)) # *Создание объекта фигуры и осей для графика с заданным размером.*

ax.scatter(x = Tg_exp, y = Tg_calc2) # *Построение диаграммы рассеяния, сопоставляющей фактические (Tg_exp) и расчетные (Tg_calc2) значения.*

plt.show() # *Отображение построенного графика.*