

Аннотация

Объектом настоящей работы являются насыщенные углеводороды, а целью – разработка метода прогнозирования вязкости чистых веществ на основе параметров уравнения состояния CP-PC-SAFT с точностью, превышающей модель MYS.

В ходе исследования была проведена очистка и агрегация экспериментальных данных, полученных из базы ThermoML, и реализован расчет производных термодинамических параметров, включая мольный объем, избыточную энтропию и вязкость идеального газа. Всего было собрано более 2000 точек по вязкости для алканов от бутана до додекана (за исключением ундекана) в широких диапазонах температур (от 273.15 до 693.7К) и давлений (от 57кПа до 245МПа). На основе полученных данных была реализована генерация новых признаков, дающих больше информации о системе, а также разработан итеративный алгоритм их отбора.

Основное внимание в работе уделено обучению моделей машинного обучения – таких как линейная регрессия, случайный лес и метод ближайших соседей – с целью оценки их точности и устойчивости на разных диапазонах условий. Проведено сравнение с моделью MYS, а также протестированы способности моделей к обобщению на новых веществах и в условиях, слабо представленных в обучающей выборке.

Полученные результаты показали, что случайный лес стабильно демонстрирует наименьшие значения MRE на всем диапазоне тестовых данных. Линейная регрессия, в свою очередь, достигает наилучших значений RMSE и MAE при использовании расширенных признаков. Этим было доказано, что машинное обучение, в сочетании с параметрами уравнения состояния, может быть использовано для создания точных и интерпретируемых моделей прогнозирования вязкости.

Работа изложена на 54 страницах, содержит 19 рисунков, 8 таблиц и использует 92 источника.

Abstract

In this work saturated hydrocarbons are studied, and the objective is to develop a method for predicting the viscosity of pure substances based on the parameters of the CP-PC-SAFT equation of state, with accuracy surpassing that of the MYS model.

During the research, experimental data obtained from the ThermoML database were cleaned and aggregated, and the calculation of derivative thermodynamic parameters was implemented, including molar volume, excess entropy, and ideal gas viscosity. A total of over 2000 viscosity data points were collected for alkanes from butane to dodecane (excluding undecane), covering a wide range of temperatures (from 273.15 to 693.7 K) and pressures (from 57 kPa to 245 MPa). Based on the obtained data, new features were generated that provide more information about the system, and an iterative algorithm for their selection was developed.

The main focus of the study is on training machine learning models – such as linear regression, random forest, and k-nearest neighbors – with the aim of evaluating their accuracy and robustness across different ranges of conditions. A comparison with the MYS model was conducted, and the models' generalization capabilities were tested on new substances and under conditions sparsely represented in the training dataset.

The results demonstrated that the random forest consistently showed the lowest MRE values across the entire range of test data. Linear regression, in turn, achieved the best RMSE and MAE values when extended features were used. This confirmed that machine learning, when combined

with equation-of-state parameters, can be employed to create accurate and interpretable viscosity prediction models.

The thesis is presented on 54 pages, contains 19 figures, 8 tables, and references 92 sources.

Содержание

| | |
|--|-----------|
| Введение | 8 |
| 1 Обзор литературы и постановка задачи | 9 |
| 1.1 Свойства вязкости и ее роль в инженерных задачах | 9 |
| 1.1.1 Вязкость в промышленных приложениях | 9 |
| 1.1.2 Значение точного прогнозирования вязкости | 9 |
| 1.2 Обзор подходов к предсказанию вязкости | 10 |
| 1.2.1 Ограничения эмпирических формул | 10 |
| 1.2.2 Модель MYS | 10 |
| 1.2.3 Уравнение состояния CP-PC-SAFT | 11 |
| 1.2.4 Параметры веществ в CP-PC-SAFT | 13 |
| 1.2.5 Трудности машинного обучения | 14 |
| 1.2.6 Символьная регрессия | 15 |
| 1.3 Цель и задачи исследования | 16 |
| 1.3.1 Цель исследования | 16 |
| 1.3.2 Задачи исследования | 16 |
| 1.4 Выводы по главе | 17 |
| 2 Сбор и обработка данных | 19 |
| 2.1 Работа с базой данных | 19 |
| 2.1.1 Источник данных: база ThermoML | 19 |
| 2.1.2 ThermoPyL для чтения данных | 19 |
| 2.2 Выбор веществ и критерии включения | 19 |
| 2.3 Распределение экспериментальных данных | 20 |
| 2.3.1 Фазовое состояние | 20 |
| 2.3.2 Температура | 20 |
| 2.3.3 Давление | 21 |
| 2.3.4 Вязкость | 21 |
| 2.4 Добавление параметров уравнения CP-PC-SAFT | 22 |
| 2.5 Вычисление производных величин | 23 |
| 2.5.1 Вязкость идеального газа | 23 |
| 2.5.2 Мольный объем | 25 |
| 2.5.3 Избыточная энтропия | 27 |
| 2.6 Выводы по главе | 28 |
| 3 Интерпретируемые модели и генерация признаков | 29 |
| 3.1 Модель MYS | 29 |
| 3.2 Символьная регрессия | 31 |
| 3.3 Алгоритм автоматической генерации признаков | 33 |
| 3.4 Выводы по главе | 35 |
| 4 Обучение и сравнение ML моделей | 36 |
| 4.1 Метод ближайших соседей | 36 |
| 4.1.1 Оценка точности на большом обучающем наборе | 37 |
| 4.1.2 Оценка на ограниченных данных | 37 |
| 4.1.3 Влияние расширенных признаков | 38 |

| | | |
|-------|---|-----------|
| 4.1.4 | Вывод по методу KNN | 39 |
| 4.2 | Случайный лес | 39 |
| 4.2.1 | Оценка точности на большом обучающем наборе | 39 |
| 4.2.2 | Оценка на ограниченных данных | 39 |
| 4.2.3 | Влияние расширенных признаков | 40 |
| 4.2.4 | Вывод по методу случайного леса | 41 |
| 4.3 | Линейная регрессия | 41 |
| 4.3.1 | Оценка точности на большом обучающем наборе | 42 |
| 4.3.2 | Оценка на ограниченных данных | 42 |
| 4.3.3 | Влияние расширенных признаков | 42 |
| 4.3.4 | Вывод по линейной регрессии | 44 |
| 4.4 | Сравнение подходов | 44 |
| 4.5 | Применимость к новым условиям | 45 |
| 4.5.1 | Новые вещества | 45 |
| 4.5.2 | Новые температуры | 46 |
| 4.5.3 | Новые давления | 46 |
| 4.6 | Выводы по главе | 47 |
| | Заключение | 48 |
| | Список литературы | 50 |

Введение

Точные методы оценки вязкости играют ключевую роль в инженерных расчетах, связанных с транспортировкой, обработкой и хранением жидкостей. Вязкость влияет на сопротивление потоку, теплообмен, гидродинамику и даже на фазовое поведение веществ, поэтому ее предсказание имеет прикладное значение для химической промышленности, энергетики и смежных областей.

На сегодняшний день большинство широко применяемых подходов к прогнозированию вязкости основаны на эмпирических зависимостях или полуэмпирических формулах, таких как модель Yarranton–Satyro (MYS). Несмотря на свою проверенность и простоту, подобные модели имеют ограничения, особенно при выходе за пределы условий, для которых они калиброваны. Это делает задачу расширения диапазона применимости и повышения точности моделей особенно актуальной.

Появление доступных термодинамических баз данных, таких как ThermoML, а также рост интереса к применению машинного обучения (ML) в физико-химическом моделировании открывают возможности для новых подходов. Модели машинного обучения могут выявлять сложные зависимости между параметрами вещества и реологическими свойствами, не требуя ручного выбора аналитических формул. При этом существует и ряд ограничений: риск переобучения, трудности с интерпретацией и зависимость от качества данных.

Цель данной работы – разработать подход к предсказанию вязкости чистых веществ на основе параметров уравнения состояния CP-PC-SAFT, обеспечивающий точность, превосходящую модель MYS. Основное внимание уделяется применению моделей машинного обучения и их оценке.

Для достижения поставленной цели в работе были последовательно решены следующие задачи:

- сбор, фильтрация и стандартизация экспериментальных данных о вязкости;
- расчет производных термодинамических величин (мольного объема, избыточной энтропии, вязкости идеального газа);
- генерация и отбор производных признаков, потенциально влияющих на вязкость;
- построение и сравнение моделей машинного обучения (линейной регрессии, случайного леса, метода ближайших соседей);
- оценка точности моделей на тестовых данных, в том числе за пределами обучающего диапазона.

Объектом исследования являются насыщенные углеводороды, а предметом – зависимость вязкости от температуры, давления и параметров вещества, описывающих его в рамках уравнения состояния CP-PC-SAFT. В качестве методов использовались вычислительная термодинамика, методы статистического анализа, алгоритмы машинного обучения и символьной регрессии.

Новизна работы заключается в объединении физико-химических параметров и алгоритмов машинного обучения для предсказания вязкости. В частности, был предложен и реализован алгоритм автоматической генерации признаков, позволивший повысить точность моделей и расширить их интерпретируемость. Также получено компактное аналитическое выражение на основе символьной регрессии дающее неплохое приближение.

Работа состоит из четырех глав. В первой главе дан обзор существующих методов предсказания вязкости и сформулированы цели и задачи исследования. Вторая глава посвящена

сбору, фильтрации и обработке данных, а также расчету необходимых производных величин. В третьей главе рассматриваются интерпретируемые подходы – в том числе модель MYS и символьная регрессия – и описан алгоритм генерации признаков. Наконец, в четвертой главе проводится обучение и тестирование моделей машинного обучения, их сравнение между собой и с моделью MYS, а также оценка способности моделей работать вне диапазона использованных при обучении параметров.

Основные результаты работы включают:

- создание и оценка полезности алгоритма генерации признаков;
- сравнение ML моделей и MYS по метрикам RMSE, MAE и MRE;
- демонстрацию практической применимости моделей ML в инженерных задачах;
- формирование рекомендаций по использованию интерпретируемых моделей для оценки вязкости.

Результаты работы, выносимые на защиту, включают оценку пригодности методов машинного обучения для предсказания вязкости, описание созданного алгоритма генерации признаков и сравнительный анализ моделей.

1. Обзор литературы и постановка задачи

1.1. Свойства вязкости и ее роль в инженерных задачах

1.1.1. Вязкость в промышленных приложениях

Вязкость – ключевой реологический параметр, определяющий сопротивление жидкости деформации при сдвиге. Ее значение напрямую влияет на гидродинамические характеристики систем, включая распределение давления, скорости потока и теплопередачу. В ряде инженерных процессов, таких как транспортировка нефти и газа, проектирование теплообменников, химическая переработка и производство смазочных материалов, точное знание вязкости определяет эффективность и безопасность эксплуатации оборудования.

В многофазных или сложных многокомпонентных системах, например, в синтетических нефтяных смесях, вязкость влияет на фазовое поведение, устойчивость эмульсий и общую производительность процессов. В микрофлюидике и фармацевтике контроль вязкости необходим для предсказуемого управления потоками и дозированием веществ.

1.1.2. Значение точного прогнозирования вязкости

Точное моделирование вязкости представляет собой одну из ключевых задач в термодинамическом моделировании жидкостей. На практике невозможно измерить вязкость для всех возможных условий и составов, особенно в системах, содержащих десятки или сотни компонентов, как, например, в нефтяной промышленности. В таких случаях используются предсказательные модели, основанные на молекулярных или термодинамических представлениях, которые позволяют получить значения вязкости при отсутствии экспериментальных данных.

Среди современных подходов особое внимание уделяется моделям, связывающим вязкость с остаточной энтропией и другими термодинамическими величинами, так как они

обеспечивают высокую переносимость и физическую обоснованность. Связь между вязкостью и избыточной энтропией подмечали во многих работах, например [1]. Особенno актуальны такие модели для расчетов в широком диапазоне температур и давлений, а также при переходе к сверхкритическим и конденсированным фазам.

Таким образом, разработка универсальных моделей вязкости, не требующих индивидуальной подгонки параметров, является приоритетной задачей в области физико-химического моделирования сложных жидкостей.

1.2. Обзор подходов к предсказанию вязкости

1.2.1. Ограничения эмпирических формул

Большинство классических подходов к описанию вязкости основаны на эмпирических уравнениях, содержащих подобранные коэффициенты для конкретных веществ или их классов. Такие модели, как правило, работают в узком диапазоне температур и давлений и требуют предварительных измерений вязкости. Это делает их трудноприменимыми для новых или редких веществ, особенно в условиях, когда экспериментальные данные недоступны или ограничены.

1.2.2. Модель MYS

Одной из наиболее универсальных и известных современных моделей для описания динамической вязкости жидкостей в широком диапазоне условий является модель, предложенная Ильей Полищуком в 2015 году – так называемая *Modified Yarranton–Satyro (MYS)* модель [2]. Ее основное назначение – обеспечить точное предсказание вязкости для широкого спектра органических жидкостей, включая углеводороды, в диапазонах температур от тройной точки до высоких значений, а также при изменяющемся давлении.

Предпосылки модели. Изначально модель Yarranton–Satyro (YS) создавалась для описания вязкости как функции плотности, температуры и параметров молекулярного строения. Polishuk модифицировал ее, чтобы избежать необходимости экспериментальной подгонки и позволить использовать только параметры, выводимые из уравнений состояния (в частности, SAFT). Основной целью стало создание аналитической корреляции, не требующей знания вязкости как входного параметра, но использующей только предсказуемые или известные характеристики вещества.

Структура модели. Модель MYS представляет собой сложную аналитическую формулу, включающую молекулярные параметры, полученные из теории SAFT:

$$\eta = 0.1 \left(\exp \left(\frac{c_1 \sqrt{m} + \ln \left(1 + \frac{M_w^4}{c_2 v^4 m^3} \right) \ln \left(\frac{T}{T_p + 120} \right)}{\exp \left(\frac{1.04v}{N_{av} m \sigma^3} \exp \left(\frac{c_3 P}{\varepsilon_k \sqrt{-\frac{dP}{dv}} \cdot \sqrt{mv}} \right) - 1 \right) - 1} \right) - 1 \right) + \eta_0 \quad (1)$$

где:

- η – динамическая вязкость (Па·с)
- η_0 – вязкость идеального газа (обычно берется как постоянная или ноль)
- m – молярная масса
- M_w – молекулярная масса

- v – мольный объем
- T – температура
- T_{tp} – температура тройной точки
- P – давление
- ε_k – параметр межмолекулярного взаимодействия
- dP/dv – производная давления по объему (рассчитанная из уравнения состояния)
- σ – эффективный диаметр молекулы
- N_{av} – число Авогадро
- $c_1 = 0.27, c_2 = 2.5 \cdot 10^{11}, c_3 = 2.1$ – эмпирические коэффициенты (подобраны автором)

Особенности модели. - Формула MYS не требует априорного знания вязкости, но использует величины, производные из термодинамических функций, так как она связана с параметрами уравнения состояния CP-PC-SAFT. - Автор демонстрирует среднюю ошибку модели на уровне менее 10% для широкой выборки жидкостей. - Подход особенно эффективен для органических и углеводородных жидкостей.

Применение в данной работе. Модель MYS была выбрана в качестве основного ориентира для сравнения, поскольку:

- Она является одной из наиболее точных универсальных формул, не требующих подбора индивидуальных коэффициентов для каждого вещества.
- Она опирается на термодинамически интерпретируемые параметры, аналогично нашему подходу.
- Ее точность считается приемлемой для инженерных и прикладных задач.
- Она сохраняет свою точность на достаточно большом диапазоне давлений и температур

В настоящем исследовании модели машинного обучения и символьной регрессии сравниваются по точности предсказания вязкости с моделью MYS. Наши модели демонстрируют в ряде случаев ошибку до 2% – значительно ниже, чем у MYS – что позволяет рассматривать их как более точную альтернативу в условиях, когда имеются данные для обучения модели.

1.2.3. Уравнение состояния CP-PC-SAFT

В данной работе уравнение состояния CP-PC-SAFT используется с целью увеличения набора признаков для экспериментальных точек. Оно служит для вычисления как параметров компонентов, так и производных термодинамических величин. Описанные ниже формулы реализованы на языке Julia в библиотеке CP-PC-SAFT [3], в рамках которой и использовались.

Уравнение состояния PC-SAFT выражается через вклад свободной энергии Гельмгольца a :

$$A(v, T) = A^{\text{id}} + A^{\text{hs}} + A^{\text{chain}} + A^{\text{disp}} + A^{\text{assoc}}$$

Где:

- A^{id} – вклад идеального газа;
- A^{hs} – вклад жестких сфер:

$$A^{\text{hs}} = RT \frac{m}{\zeta_0} \left(\frac{3\zeta_1\zeta_2}{1-\zeta_3} + \frac{\zeta_2^3}{\zeta_3(1-\zeta_3)^2} \right) + \left(\frac{\zeta_2^3}{\zeta_3^2} - \zeta_0 \right) \ln[1-\zeta_3] \sqrt{\frac{d^3(\zeta_3-1)}{\zeta_3\sigma^3-d^3}}$$

- A^{chain} – вклад цепных взаимодействий:

$$A^{\text{chain}} = RT \sum_i x_i x_j (1-m_{ij}) \ln[g_{ij}(d_{ij})]^{hs}$$

- A^{disp} – вклад дисперсионных сил:

$$\begin{aligned} A^{\text{disp}} = & -RN_{\text{Av}} \left(\frac{2\pi(\varepsilon/k)m^2\sigma^3}{\nu} I_1 \right. \\ & \left. + \frac{\pi(\varepsilon/k)^2 m^3 \sigma^3}{\nu T \left(1 + \frac{m(8\zeta_3-2\zeta_3^2)}{(1-\zeta_3)^4} + \frac{(1-m)(20\zeta_3-27\zeta_3^2+12\zeta_3^3-2\zeta_3^4)}{((1-\zeta_3)(2-\zeta_3))^2} \right)} I_2 \right) \end{aligned}$$

- A^{assoc} – вклад ассоциативных взаимодействий;
- d – эффективный диаметр молекулы, связанный с σ через θ :

$$d = \theta * \sigma$$

- σ – характерный размер молекулы (диаметр сферического сегмента). В случае смеси нескольких веществ:

$$\sigma = \sqrt[3]{\frac{\sum_i x_i x_j m_i m_j \sigma_{ij}^3}{(\sum_i x_i m_{ii})^2}}$$

- θ – коэффициент масштабирования длины связи:

$$\theta = \frac{1 + 0.2977(k/\varepsilon)T}{1 + 0.33163(k/\varepsilon)T + 0.0010477(k/\varepsilon)^2 T^2}$$

- ζ_k – параметр упаковки молекул:

$$\zeta_k = \frac{\pi N_{av}}{6\nu} \sum_i x_i m_i \sigma_i^{dk}$$

- ε/k_B – приведенная энергия взаимодействия между частицами. В случае смеси нескольких веществ:

$$\varepsilon/k = \frac{\sum_i x_i x_j m_i m_j \sigma_{ij}^3 (\varepsilon/k)_{ij}}{\sigma^3 (\sum_i x_i m_{ii})^2}$$

- k_B – постоянная Больцмана;
- N_{Av} – число Авогадро;

- m – количество сегментов в молекуле. В случае смеси нескольких веществ:

$$m = \sum_i x_i m_i$$

- x_i – мольная доля компонента i ;
- P – давление;
- T – температура;
- v – мольный объем;
- d_{ij} – средний эффективный диаметр молекул i и j ;
- g_{ij} – функция радиальной корреляции молекул i и j :

$$g_{ij}(d_{ij})^{hs} = \frac{1}{1 - \zeta_3} + \frac{3d_i d_j \zeta_2}{(d_{ii} + d_{jj})(1 - \zeta_3)^2} + 2 \left(\frac{d_i d_j}{d_{ii} + d_{jj}} \right)^2 \frac{\zeta_2^2}{(1 - \zeta_3)^3}$$

- ν – молярный объем системы;
- I_1, I_2 – интегралы в дисперсионном вкладе;
- $(\varepsilon/k)_{ij}$ – приведенная энергия взаимодействия между разными компонентами i и j :

$$(\varepsilon/k)_{ij} = (1 - k_{ij}) \sqrt{(\varepsilon/k)_{ii} (\varepsilon/k)_{jj}}$$

- m_{ij} – эффективное среднее число сегментов при взаимодействии компонентов i и j :

$$m_{ij} = (1 - l_{ij}) \frac{m_{ii} + m_{jj}}{2}$$

- σ_{ij} – характерный средний диаметр сегментов при взаимодействии компонентов i и j , рассчитанный как простое среднее арифметическое:

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2}$$

- k_{ij}, l_{ij} – коэффициент парного взаимодействия и параметр согласования размеров сегментов. Считываются равными нулю, либо подбираются на основании экспериментальных данных.

1.2.4. Параметры веществ в CP-PC-SAFT

В рамках CP-PC-SAFT каждое вещество описывается тремя основными параметрами:

- m – число сегментов в молекуле;
- σ – эффективный диаметр сегмента молекулы;
- ε – энергия взаимодействия между сегментами.

Эти параметры определяются через критическую точку и температуру кипения вещества, а не через эмпирическую подгонку, что делает модель более универсальной.

В модели CP-PC-SAFT параметры вещества корректируются так, чтобы соответствовать следующим условиям [4]:

$$\left(\frac{\partial P}{\partial v}\right)_{T_c} = 0, \quad \left(\frac{\partial^2 P}{\partial v^2}\right)_{T_c} = 0$$

$$P_c = P_{c, \text{exp}}$$

$$\rho_{\text{liq, triple}} = \rho_{\text{liq, triple, exp}}$$

Здесь:

- P_c – давление в критической точке;
- T_c – критическая температура;
- $\rho_{\text{liq, triple}}$ – жидкостная плотность в тройной точке.

Условиями для поиска параметров являются критическая температура, критическое давление и плотность жидкости при температуре плавления, соответствующие экспериментальным значениям. Необходимость знания значений всего трех экспериментальных величин для вычисления всех параметров вещества делает модель CP-PC-SAFT крайне привлекательной в вопросах предсказания свойств новых веществ и смесей на большом диапазоне условий.

1.2.5. Трудности машинного обучения

Применение методов машинного обучения (ML) к задачам физико-химического моделирования, включая предсказание вязкости, становится все более распространенным. Такие методы действительно позволяют выявлять сложные зависимости в данных и могут достигать высокой точности без необходимости ручного выбора функциональной формы модели. Однако, несмотря на растущую популярность подхода, существуют важные причины, по которым машинное обучение не следует воспринимать как универсальное решение для задач, связанных с предсказанием реологических свойств веществ.

- a) **Ограниченнная интерпретируемость.** Большинство моделей машинного обучения, особенно ансамблевые методы и нейросети, по своей природе слабо интерпретируемые. Это может быть критично в задачах молекулярной термодинамики, где важна не только точность, но и понимание, какие параметры и в каком виде влияют на результат. На этом фоне выигрывают подходы, сохраняющие аналитическую форму – например, модель MYS или символьная регрессия, использованная в данной работе.
- b) **Ограниченнная обобщающая способность.** Как показали эксперименты, даже сравнительно простые модели (линейная регрессия, случайный лес) могут терять точность при работе с веществами или условиями, не представленными в обучающей выборке. Особенно заметно это на примере бутана, где поведение в газовой фазе значительно отличается. Это говорит о том, что модели машинного обучения склонны к переобучению и не всегда способны интерполировать или экстраполировать вне привычных областей данных.

- в) **Зависимость от структуры и полноты данных.** Методы ML чувствительны к объему и качеству данных. В случае вязкости, как и других термодинамических величин, данные часто собраны в разных условиях, с разной точностью и неравномерным покрытием пространства параметров. Это усложняет обучение и требует либо значительной фильтрации, либо введения новых признаков, зачастую основанных на физике (например, через параметры уравнения состояния), чтобы повысить устойчивость модели.
- г) **Нарушение физических ограничений.** В отличие от уравнений, созданных с учетом физической природы процесса, модели машинного обучения не всегда гарантируют выполнение очевидных условий. Например, они могут предсказывать отрицательные значения вязкости или вести себя нестабильно при экстремальных температурах. Такие случаи сложно контролировать, если не вводить специальных ограничений в архитектуру модели.
- д) **Ограничения при практическом применении.** Для использования в инженерной практике важны не только точность и обобщающая способность, но и простота реализации, воспроизводимость и вычислительная эффективность. Многие ML-модели требуют больше ресурсов и сложнее в поддержке, чем простые аналитические формулы. В этом смысле, компактные и интерпретируемые выражения, такие как те, что можно получить с помощью символьной регрессии, представляют собой более удобный формат для прикладных расчетов.

В целом, подходы машинного обучения открывают серьезные перспективы, особенно при работе с обширными и хорошо структуризованными данными. Однако в задачах, где физическая интерпретация важна не меньше, чем численная точность, такие методы требуют осознанного применения. Перспективным направлением представляется интеграция моделей машинного обучения с физически обоснованными величинами и признаками – такой подход способен объединить гибкость современных алгоритмов с устойчивостью и интерпретируемостью классических методов.

1.2.6. Символьная регрессия

Символьная регрессия (symbolic regression, SR) – это подход к построению моделей, при котором алгоритм не просто находит численные параметры в заданной структуре формулы как в линейной регрессии, а ищет саму структуру уравнения, наиболее точно описывающее зависимость между входными и выходными переменными. В отличие от традиционных методов машинного обучения, символьная регрессия формирует аналитическое выражение в виде комбинаций математических операций и входных признаков.

Мотивация применения. Основным преимуществом SR является высокая интерпретируемость получаемых моделей. Вместо подбора большого количества слабоинтерпретируемых коэффициентов, как в случае с нейросетями, модель находит короткие выражения, состоящие из качественно различающихся блоков (операторов). Каждое уравнение, сформированное таким способом, может быть проанализировано, сопоставлено с физическими законами, а также непосредственно использовано в инженерных расчетах. Это особенно важно при работе с термодинамическими и молекулярными свойствами, где значение имеет не только точность, но и физический смысл модели.

В рамках настоящего исследования SR использовалась для поиска наиболее компактных и точных выражений, связывающих вязкость с параметрами уравнения состояния CP-PC-SAFT и производными термодинамическими величинами, включая мольный объем, избыточную энтропию и производную давления по объему.

Используемый инструмент: PySR. Для реализации символьной регрессии применялась современная библиотека PySR (Python Symbolic Regression) [5]. Эта библиотека сочетает возможности языка Python для настройки задач и быстродействие языка Julia, где реализована основная эволюционная оптимизация. PySR использует алгоритмы на основе генетического программирования, направленные на поиск выражений с оптимальным компромиссом между точностью и сложностью.

Основные особенности PySR:

- Многофункциональная система операторов: поддерживаются стандартные математические функции (логарифм, экспонента, степень и др.).
- Поддержка многоцелевой оптимизации: минимизация ошибки при контроле за длиной формулы.
- Устойчивость к переобучению: регуляризация сложности встроена в процесс поиска.
- Высокая производительность благодаря параллельному выполнению на CPU и GPU.

Результаты применения. В ходе экспериментов символьная регрессия позволила получить компактные формулы с ошибкой, сравнимой с ошибкой модели MYS. Таким образом, символьная регрессия показала себя как перспективный инструмент, позволяющий извлекать простые закономерности. Ее применение может дополнить использование других моделей и усилить акцент на физическую обоснованность результатов.

1.3. Цель и задачи исследования

1.3.1. Цель исследования

Разработка метода прогнозирования вязкости жидкостей на основе параметров уравнения состояния CP-PC-SAFT с точностью, превосходящей существующую модель MYS.

1.3.2. Задачи исследования

Для реализации поставленной цели были намечены следующие задачи:

a) Сбор и подготовка экспериментальных данных

- Получение доступа к данным термодинамических параметров веществ.
- Формирование базы данных свойств жидкостей и газов на основе ThermoML.
- Фильтрация выборки для получения только интересующих веществ с замеренными температурой, давлением и вязкостью.
- Предварительная обработка данных для приведения интересующих параметров к единому формату.

б) Расчет производных параметров

- Вычисление мольного объема по известным температуре и давлению.
- Определение избыточной энтропии s^{ex} по известным мольному объему и температуре через производную избыточной энергии Гельмгольца по температуре.
- Вычисление вязкости идеального газа.

в) Генерация и отбор новых признаков, проверка их значимости

- Создание производных признаков, отражающих физические закономерности вязкости.
- Оценка их статистической значимости для модели, и отбор наиболее полезных.
- Валидация роли различных признаков в прогнозировании вязкости, в том числе избыточной энтропии.

г) Обучение, сравнение и оптимизация моделей машинного обучения

- Построение и проверка эффективности различных моделей ML, таких как линейная регрессия, градиентный бустинг, случайный лес и нейросетевые методы.
- Оптимизация гиперпараметров для повышения точности наиболее перспективных моделей.
- Использование символьной регрессии (PySR) для поиска интерпретируемых аналитических зависимостей.

д) Сравнение разработанной модели с существующей моделью MYS

- Оценка точности предложенного метода на тестовых данных.
- Анализ преимуществ и недостатков разработанных моделей по сравнению с MYS.
- Формирование рекомендаций по дальнейшему развитию метода прогнозирования вязкости жидкостей.

1.4. Выводы по главе

В первой главе был проведен обзор современных представлений о вязкости как важном термодинамическом параметре и существующих подходах к ее моделированию. Показано, что вязкость оказывает существенное влияние на поведение жидких и многофазных систем в различных инженерных приложениях – от нефтепереработки до микрофлюидики.

Рассмотрены эмпирические формулы, применяемые для расчета вязкости, а также универсальные аналитические модели, такие как MYS (Modified Yarranton–Satyro). Несмотря на широкую применимость первых, они имеют ограничения в точности при отсутствии экспериментальных данных для подгонки или при переходе к новым веществам. В свою очередь, MYS не всегда дает достаточные для прикладных задач оценки ввиду своей универсальности.

Особое внимание было уделено уравнению состояния CP-PC-SAFT, которое позволяет получить необходимые параметры вещества без тщательной подгонки, опираясь лишь на критические и тройные точки. Это делает его особенно подходящим для предсказательных моделей, работающих на широком спектре условий и веществ, таких как MYS или разрабатываемая в рамках данной работы.

Кроме того, рассмотрены возможности машинного обучения и его ограничения в задачах физико-химического моделирования, включая проблемы интерпретируемости, переобучения и нарушения физических ограничений. На этом фоне была обоснована целесообразность использования символьной регрессии и системы генерации новых признаков как методов, сочетающих гибкость ML с интерпретируемостью и физической обоснованностью аналитических формул.

Таким образом, глава обосновывает научную и практическую значимость задачи предсказания вязкости на основе параметров уравнения состояния и подтверждает пользу разработки нового подхода, сочетающего машинное обучение и следование физическим принципам. Далее в работе рассматривается реализация этого подхода на практике.

2. Сбор и обработка данных

2.1. Работа с базой данных

2.1.1. Источник данных: база ThermoML

Для получения экспериментальных данных о вязкости в данной работе использовалась база данных **ThermoML**, разработанная Национальным институтом стандартов и технологий США (NIST). ThermoML представляет собой машиночитаемый формат на основе XML, предназначенный для хранения и распространения термодинамических и физико-химических данных, включая фазовые равновесия, плотности, давления, температуры и вязкости чистых веществ и смесей.

Одной из ключевых особенностей ThermoML является строго стандартизированная структура описания данных, что позволяет автоматизировать их извлечение и минимизировать необходимость ручной предобработки. Это особенно важно при формировании выборок для задач машинного обучения, где требуется согласованность единиц измерения и форматов.

2.1.2. ThermoPyL для чтения данных

Для взаимодействия с базой ThermoML была использована специализированная Python-библиотека **ThermoPyL**. Данная библиотека предоставляет инструменты для:

- автоматического чтения и парсинга XML-файлов ThermoML;
- фильтрации данных по фазе и составу;
- преобразования данных в формат pandas.DataFrame для последующей обработки в Python;
- агрегации данных по условиям экспериментов;
- экспорта результатов в формат CSV.

Использование библиотеки ThermoPyL позволило автоматизировать процесс подготовки данных, не углубляясь в анализ и обработку XML файлов из ThermoML.

2.2. Выбор веществ и критерии включения

Для анализа были выбраны чистые вещества, относящиеся к классу насыщенных углеводородов. Они обладают простой молекулярной структурой и широко используются в промышленности, что делает их удобным объектом для моделирования. Основным критерием отбора было наличие значительного объема экспериментальных данных по вязкости в базе ThermoML. В финальную выборку вошли восемь веществ:

- Бутан: получено 311 экспериментальных точек из 2 работ [6, 7]
- Пентан: получено 50 экспериментальных точек из 3 работ [8—10]
- Гексан: получено 127 экспериментальных точек из 25 работ [7, 9, 11—33]
- Гептан: получено 443 экспериментальных точек из 28 работ [9, 11, 15, 16, 20, 22, 26, 31, 34—53]

- Октан: получено 362 экспериментальных точек из 32 работ [7—9, 15, 16, 18—20, 22, 27, 31, 34, 36, 38, 39, 41, 44, 45, 47, 53—65]
- Нонан: получено 115 экспериментальных точек из 13 работ [8, 9, 15, 16, 34, 36, 39, 45, 48, 61, 66—68]
- Декан: получено 224 экспериментальных точек из 25 работ [8, 9, 15, 16, 18, 19, 27, 32, 34, 36, 39, 45, 55, 58, 61, 63—65, 67, 69—74]
- Додекан: получено 412 экспериментальных точек из 25 работ [8, 15, 16, 28, 40, 45, 54, 55, 58, 66, 70, 75—88]

Для включения записи в выборку требовалось наличие всех трех параметров:

- температуры (T , К),
- давления (P , Па),
- вязкости (η , Па·с).

Записи, в которых отсутствовала хотя бы одна из этих величин, исключались. Дополнительная фильтрация по фазе не производилась, поскольку, например, для бутана полезны данные как в жидкой, так и в газообразной фазе. Каждая выборка сохранялась в виде отдельного CSV-файла для дальнейшей обработки. В результате было собрано более 2000 уникальных записей.

2.3. Распределение экспериментальных данных

2.3.1. Фазовое состояние

Большая часть данных принадлежала к жидкой фазе. Только 295 точек для бутана находятся в состоянии газа. Так как модель MYS рассчитана на данные как в жидкой, так и в газовой фазе, было решено оставить точки, принадлежащие газовой фазе, несмотря на то, что они представлены только бутаном. Хотя в ходе дальнейших экспериментов бутан зачастую исключался, так как сильно мешал анализу относительной точности.

2.3.2. Температура

Значения температуры итоговых данных распределены следующим образом Рис. 1:

- **Минимальная температура:** 273.15 K
- **Максимальная температура:** 693.7 K
- **Среднее значение температуры:** 406 K
- **Медианное значение температуры:** 403 K
- **Явный пик в районе:** 300 K

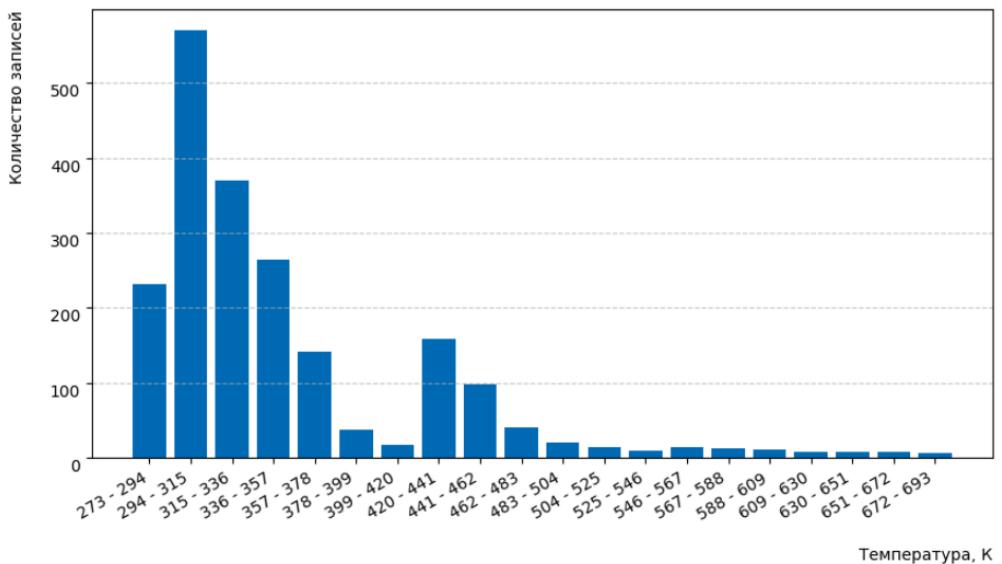


Рис. 1: Распределение температуры в экспериментальных данных

2.3.3. Давление

Значения давления итоговых данных распределены следующим образом Рис. 2:

- **Минимальное давление:** 57.183 кПа
- **Максимальное давление:** 245.16 МПа
- **Среднее значение давления:** 23 МПа
- **Медианное значение давления:** 6 МПа
- **Резкий пик наблюдается в районе:** 1 атмосферы

Данные удобнее отображать в логарифмической шкале, так как в ней они относительно равномерно распределены.

2.3.4. Вязкость

Значения вязкости итоговых данных распределены следующим образом Рис. 3:

- **Минимальная вязкость:** 7.3 мкПа·с
- **Максимальная вязкость:** 5.8 мПа·с
- **Среднее значение вязкости:** 500 мкПа·с
- **Медианное значение вязкости:** 385 мкПа·с
- **Небольшой пик наблюдается в районе:** 10 мкПа·с (эксперименты с бутаном)

Распределение вязкости, аналогично распределению давления, удобнее отображать в логарифмическом масштабе.

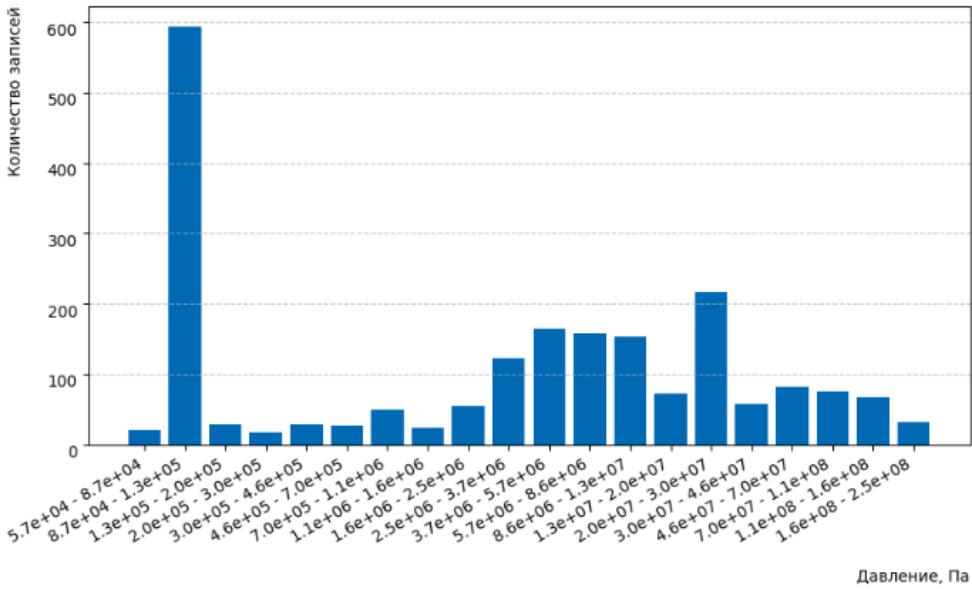


Рис. 2: Распределение давления в экспериментальных данных (логарифмическая шкала по горизонтали)

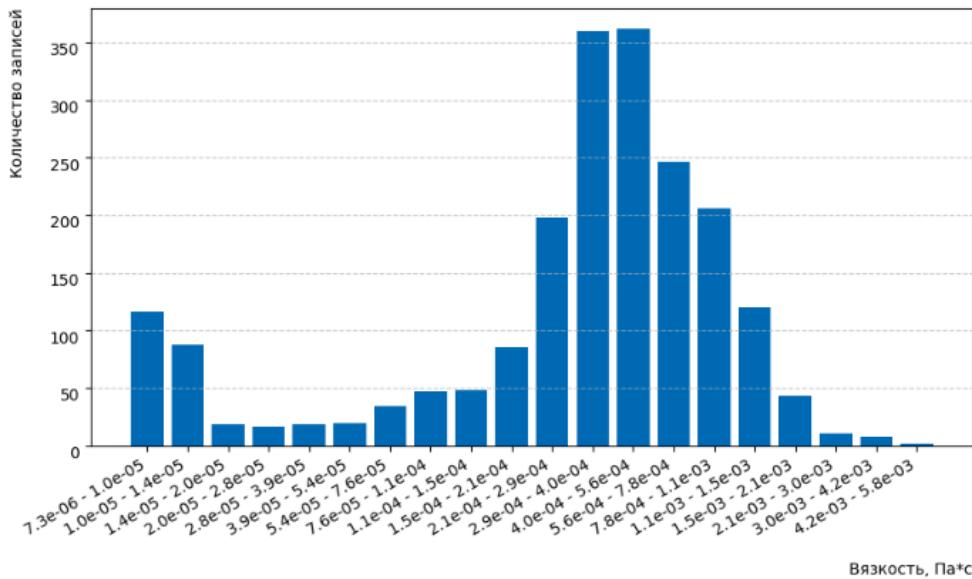


Рис. 3: Распределение вязкости в экспериментальных данных (логарифмическая шкала по горизонтали)

2.4. Добавление параметров уравнения CP-PC-SAFT

Как уже было упомянуто выше, у каждого вещества есть единственный набор параметров для уравнения CP-PC-SAFT: σ , δ , ε/k_b и m . Для всех исследуемых веществ данные признаки были добавлены в датасет. Эти параметры представляют собой ключевые свойства каждого соединения, поскольку их можно точно определить, что позволяет назвать их основными характеристиками. Значения данных параметров для рассматриваемых веществ можно увидеть в Табл. 1.

Таблица 1

Основные характеристики компонентов

| Вещество | Молярная масса, (г/моль) | m | $\varepsilon/k_b, \text{К}$ | $\sigma, \text{\AA}$ | δ |
|----------|--------------------------|----------|-----------------------------|----------------------|----------|
| бутан | 58.120 | 2.482 62 | 209.446 | 3.650 40 | 1.159 76 |
| пентан | 72.150 | 3.064 24 | 212.528 | 3.624 21 | 1.163 85 |
| гексан | 86.180 | 3.510 81 | 218.238 | 3.655 75 | 1.160 91 |
| гептан | 100.210 | 4.070 32 | 220.494 | 3.635 15 | 1.166 31 |
| октан | 114.230 | 4.454 75 | 225.287 | 3.678 68 | 1.179 34 |
| нонан | 128.257 | 4.851 00 | 229.271 | 3.704 69 | 1.187 22 |
| декан | 142.284 | 5.270 13 | 232.262 | 3.721 88 | 1.202 60 |
| додекан | 170.338 | 6.012 09 | 238.240 | 3.769 83 | 1.225 22 |

2.5. Вычисление производных величин

В ходе работы возникла потребность в дополнении данных новыми величинами, которые можно вычислить используя уже известные величины. Это нужно как для увеличения количества признаков, которыми могли бы пользоваться модели машинного обучения, так и для возможности пользоваться некоторыми формулами.

2.5.1. Вязкость идеального газа

Вязкость идеального газа является одним из параметров уравнения MYS (1). Можно было бы опустить этот коэффициент и использовать вместо него ноль, и формально формула осталась бы вполне точной для жидкой фазы. Однако для газовой области именно этот коэффициент дает наибольший вклад. Пренебречь им – значит закрыть глаза на половину картины. Поэтому было решено вычислить вязкость идеального газа для всех экспериментальных точек, чтобы обеспечить максимальную точность модели в целом. Поскольку уравнение MYS служит основным критерием точности, задача вычисления η_0 приобрела особую значимость.

Для расчета вязкости идеального газа использовался классический подход на основе параметров Леннарда–Джонса – $(\varepsilon/k)_{LJ}$ и σ_{LJ} , которые были определены по температуре плавления и мольному объему вещества. Этот подход, несмотря на свою простоту, оказывается весьма работоспособным при корректной выборке входных данных. Температура плавления (T_m) и плотность в жидком состоянии (ρ) использовались для расчета мольного объема v по формуле:

$$v = \frac{M}{\rho},$$

где M – молярная масса.

Далее параметры ε/k и σ определялись следующим образом:

$$(\varepsilon/k)_{LJ} = 1.92 \cdot T_m, \quad \sigma_{LJ} = \left(\frac{2.3 \cdot v}{\pi N_A \cdot 2/3} \right)^{1/3},$$

где N_A – число Авогадро. Стоит отметить, что здесь v переводился в кубические метры, а итоговое значение σ – в ангстремы.

Параметры Леннард–Джонса для бутана, пентана, гексана, гептана, октана и нонана были взяты из справочника по вязкости газовых смесей [89]. Параметры для декана и додекана

Таблица 2

Параметры Леннарда–Джонса для выбранных алканов

| Вещество | $(\varepsilon/k_b)_{LJ}, \text{К}$ | $\sigma_{LJ}, \text{\AA}$ |
|----------|------------------------------------|---------------------------|
| бутан | 410.00 | 4.9970 |
| пентан | 345.00 | 5.7690 |
| гексан | 413.00 | 5.9090 |
| гептан | 350.59 | 6.4406 |
| октан | 345.00 | 5.7690 |
| нонан | 413.00 | 5.9090 |
| декан | 467.52 | 7.0836 |
| додекан | 506.11 | 7.4540 |

были вычисленны на основании температуры тройной точки и плотности при нормальных условиях с помощью формулы из той же книги. Использованные в работе коэффициенты указаны в Табл. 2

Коэффициенты для одного и того же вещества могут различаться в зависимости от интервала температур, в котором они были вычислены. При этом отмечается, что при расчетах ошибка, допущенная в определении величины $(\varepsilon/k_b)_{LJ}$, дает значительно меньшую ошибку в определении величины вязкости.

Указанные величины использовались для расчета вязкости идеального газа. Формула основана на кинетической теории газов и включает поправку на столкновения. Ее можно найти в продвинутой литературе по молекулярной динамике [89]. Итоговое выражение имеет вид:

$$\eta_0 = \frac{5}{16} \cdot \frac{\sqrt{MRT}}{\pi^{1/2} \cdot \sigma_{LJ}^2 \cdot \Omega(T^*) \cdot N_A},$$

где T^* – приведенная температура, а Ω – коэффициент столкновений, зависящий от T^* :

$$T^* = \frac{RT}{(\varepsilon/k_b)_{LJ}}, \quad \Omega(T^*) = \frac{A}{(T^*)^B} + Ce^{-DT^*} + Ee^{-FT^*} + RT^{*B} \sin(ST^{*W} - P).$$

Ниже в Табл. 3 приведены значения эмпирических коэффициентов, использованных в выражении для функции столкновений $\Omega(T^*)$. Эти значения были взяты из работы [90] и использованы без модификации.

Расчеты выполнялись для каждой строки из экспериментальной выборки с помощью функции ‘apply()’ библиотеки ‘pandas’. Итоговая вязкость η_0 вносилась в результирующий CSV-файл, на основе которого строилось дальнейшее сравнение с экспериментальными и модельными данными.

Распределение вычисленной вязкости имеет следующие свойства:

- **Минимальная вязкость:** 4.913×10^{-6} Па·с
- **Максимальная вязкость:** 1.653×10^{-5} Па·с
- **Среднее значение:** 8.711×10^{-6} Па·с
- **Медианное значение:** 8.602×10^{-6} Па·с

Распределение вязкости имеет максимальную частоту значений вокруг 7.1×10^{-6} Па·с. Отмечаются пики около 6.1×10^{-6} Па·с и 1.2×10^{-5} Па·с, что указывает на неравномерное распределение вязкости.

Таблица 3

Коэффициенты для расчета функции столкновений $\Omega(T^*)$

| Коэффициент | Значение |
|-------------|--------------|
| A | 1.161 45 |
| B | 0.148 74 |
| C | 0.524 87 |
| D | 0.773 20 |
| E | 2.161 78 |
| F | 2.437 87 |
| R | -0.000 643 5 |
| S | 18.0323 |
| W | -0.768 30 |
| P | 7.273 71 |

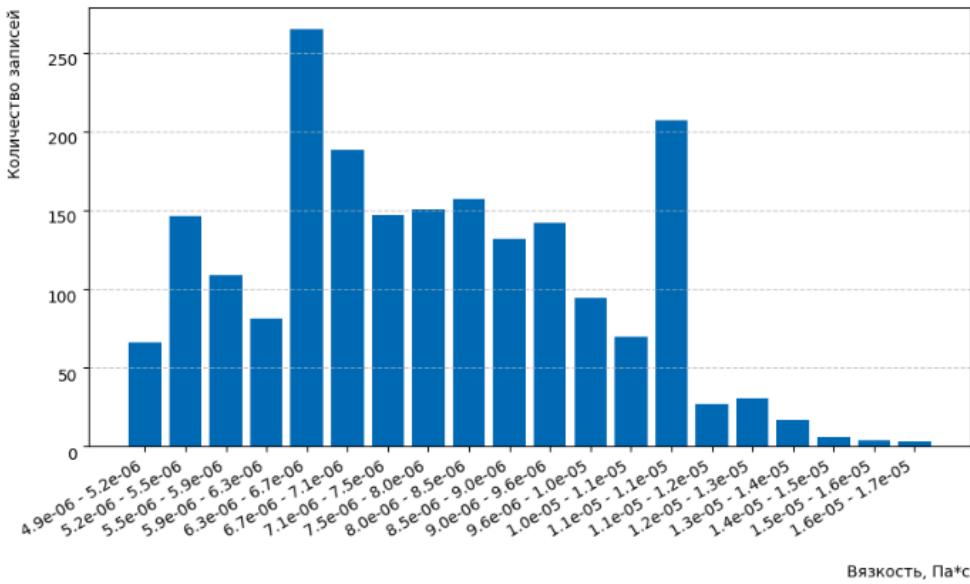


Рис. 4: Распределение вязкости идеального газа в экспериментальных данных (логарифмическая шкала по горизонтали)

2.5.2. Мольный объем

Мольный объем v является одной из трех ключевых переменных уравнения состояния SAFT наряду с температурой T и давлением P . Однако, чтобы получить энергию системы в явном виде, можно обойтись без давления. Тем не менее, в большинстве экспериментальных данных давление указано, а плотность – нет, поэтому мольный объем приходится восстанавливать численно.

Пусть задана точка с известными температурой T и давлением P . Зная аналитическое выражение для энергии системы $A(T, v)$, можно вычислить давление как производную:

$$P = - \left(\frac{\partial A}{\partial v} \right)_T$$

Тогда задача нахождения мольного объема v сводится к решению уравнения:

$$P(v, T) = P_{\text{эксп}}$$

Решение этой задачи проводится численно, например, методом половинного деления или методом Ньютона. При этом важно дополнительно проверить, что производная давления по объему отрицательна:

$$\left(\frac{\partial P}{\partial v} \right)_T < 0$$

Это необходимо для физической осмысленности и устойчивости найденного состояния.

На практике был реализован алгоритм, основанный на бинарном поиске в заданном диапазоне объемов. Для каждой экспериментальной точки из базы данных была рассчитана пара: мольный объем v и производная давления по объему при данной температуре. Алгоритм корректно отработал на всех точках: он всегда находил первое значение объема, при котором давление совпадает с заданным и производная отрицательна. Для всех вычислений использовался язык программирования Julia и реализация уравнения состояния в библиотеке *CP_PCA_SAFT* [91]. Распределение вычисленного мольного объема имеет следующие свойства:

- **Минимальный мольный объем:** 8.940×10^{-5} м³/моль
- **Максимальный мольный объем:** 5.703×10^{-2} м³/моль
- **Среднее значение мольного объема:** 4.700×10^{-4} м³/моль
- **Медианное значение мольного объема:** 1.721×10^{-4} м³/моль
- **Главный пик наблюдается в районе медианного значения**

На гистограмме распределение мольного объема представлено в логарифмическом масштабе Рис. 5.

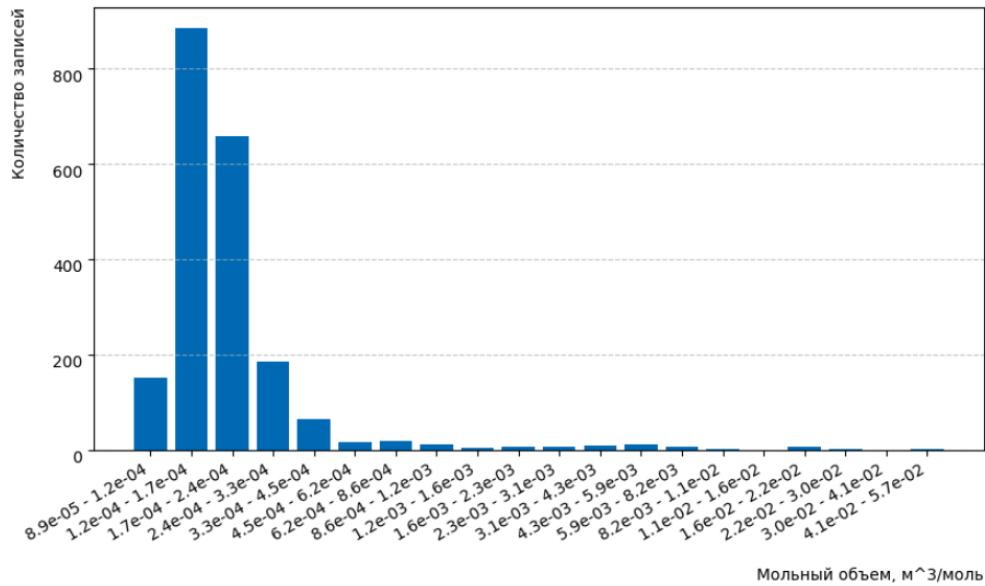


Рис. 5: Распределение мольного объема в экспериментальных данных (логарифмическая шкала)

2.5.3. Избыточная энтропия

Избыточная энтропия – важная термодинамическая характеристика, определяемая как производная избыточной свободной энергии Гельмгольца по температуре со знаком минус.

Избыточная энергия представляется в виде полной энергии без вклада идеального газа A^{id} :

$$A^{\text{res}}(v, T) = A^{\text{hs}} + A^{\text{chain}} + A^{\text{disp}},$$

Тогда избыточная мольная энтропия принимает вид:

$$s^{\text{res}}(v, T) = - \left(\frac{\partial A^{\text{res}}}{\partial T} \right)_v.$$

В работе [92] было высказано предположение, согласно которому логарифм безразмерной вязкости жидкости можно аппроксимировать полиномом от нормированной избыточной энтропии:

$$\ln(\eta^*) = A + Bs^* + C(s^*)^2 + D(s^*)^3,$$

где нормированная избыточная энтропия определяется как:

$$s^* = \frac{s^{\text{res}}(v, T)}{mk_B}.$$

Здесь:

- A^{res} – избыточная свободная энергия Гельмгольца (участвует в первых двух уравнениях);
- η_0 – характеристическая (референсная) вязкость;
- $\eta^* = \eta/\eta_0$ – безразмерная вязкость;
- A, B, C, D – эмпирические параметры, определяемые регрессией.

Основываясь на этом, можно сказать, что избыточная энтропия является важным признаком для предсказания вязкости. Разумно ожидать, что модели машинного обучения будут активно применять данные значения для улучшения результата.

Для расчета избыточной энтропии использовалась библиотека CP_PC_SAFT на языке Julia для вычисления уравнения CP-PC-SAFT, а также другие библиотеки.

Распределение вычисленной величины имеет следующие свойства:

- **Минимальное значение:** -1.295×10^2
- **Максимальное значение:** -4.671×10^{-2}
- **Среднее значение:** -6.291×10^1
- **Медианное значение:** -6.740×10^1

Распределение имеет максимальную частоту значений вокруг -5.8×10^1 . Отмечаются пики около -7.1×10^1 и -4.5×10^1 , что может указывать на неравномерное распределение величины.

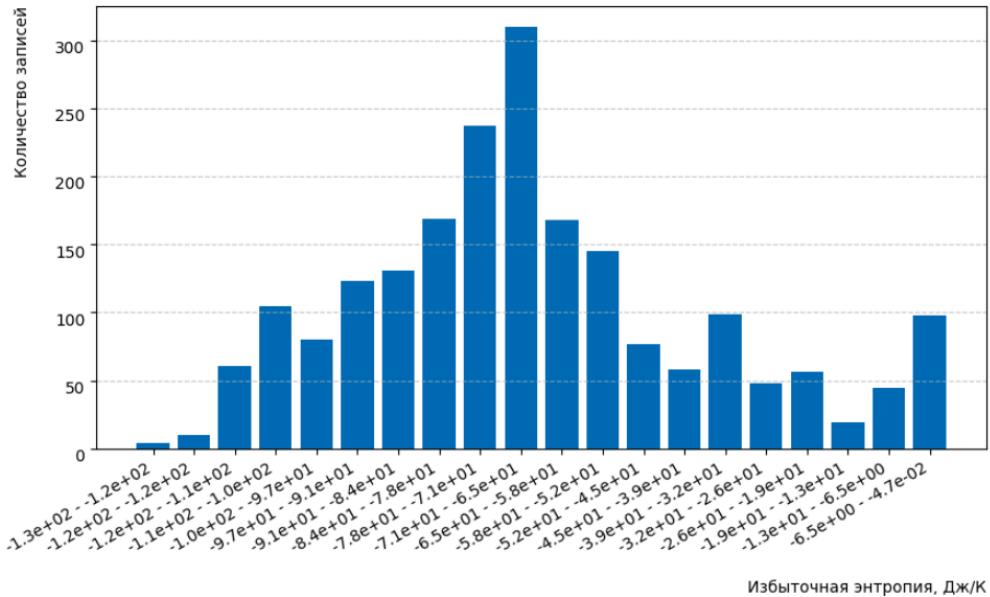


Рис. 6: Распределение вычисленной избыточной энтропии

2.6. Выводы по главе

В данной главе был описан процесс сбора и предварительной обработки экспериментальных данных о вязкости для насыщенных углеводородов с использованием базы ThermoML и библиотеки ThermoPyL. Отбор и агрегация данных производились с учетом полноты необходимых параметров (температура, давление, вязкость). Основной акцент был сделан на чистые вещества, так как анализ данных систем проще, а количество экспериментов больше, чем для смесей.

Анализ распределений показал, что явных выбросов, артефактов или дубликатов в собранных данных не наблюдается. Однако гораздо более существенную проблему представляет собой смещение распределения в сторону типичных условий, таких как область около 300 Кельвинов или 1 атмосферы. Это, вероятно, связано с предпочтениями экспериментаторов и ограничениями лабораторной аппаратуры. Подобные смещения могут повлиять на обобщающую способность моделей, особенно в предсказании вне хорошо охваченных областей параметров.

Таким образом, несмотря на общее высокое качество данных и их стандартизованный формат, уже на этапе сбора можно отметить наличие неравномерности, которую важно учитывать при последующем анализе и обучении моделей.

3. Интерпретируемые модели и генерация признаков

Одной из ключевых задач настоящей работы является не только получение точных предсказаний вязкости, но и выявление физически осмысленных закономерностей. Для этого был проведен анализ интерпретируемых моделей – от теоретически обоснованной модели MYS до автоматически полученных выражений с помощью символьной регрессии. Наряду с этим была разработана система генерации новых признаков, направленная на улучшение точности моделей и выявление скрытых зависимостей в данных.

3.1. Модель MYS

Первой была проанализирована основная в данной работе модель для сравнения с новыми моделями – MYS (Modified Yarranton–Satyro).

Для оценки точности ее предсказаний были использованы следующие метрики:

- **RMSE (корень из среднеквадратичной ошибки):**

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\eta_i^{\text{model}} - \eta_i^{\text{exp}})^2}$$

где η_i^{model} – значение вязкости, предсказанное моделью, η_i^{exp} – экспериментальное значение, n – общее число точек.

- **Среднее относительное отклонение:**

$$MRE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\eta_i^{\text{model}} - \eta_i^{\text{exp}}}{\eta_i^{\text{exp}}} \right|$$

- **Среднее абсолютное отклонение:**

$$MAE = \frac{1}{n} \sum_{i=1}^n |\eta_i^{\text{model}} - \eta_i^{\text{exp}}|$$

Для оценки эффективности модели использованы две метрики выше: среднее относительное отклонение и среднее абсолютное отклонение. Бутан представлен в основном точками газовой фазы, у которых значения вязкости значительно ниже, чем у остальных данных и маленькая абсолютная ошибка порождает очень большую относительную. Поэтому бутан был исключен из рассмотрения, так как график относительной ошибки становится абсолютно неинформативным. На Рис. 7 и Рис. 8 представлены графики этих двух метрик, иллюстрирующих поведение модели MYS на экспериментальных данных:

Для обобщения результатов были рассчитаны средние и медианные значения ошибок по каждому веществу. Данные представлены в таблице 4.

Собранные данные позволяют наглядно оценить, в каких условиях модель MYS работает особенно хорошо или наоборот – начинает терять точность, что важно при дальнейшем сравнении с моделями машинного обучения. Можно сделать предположение, что ошибка моделей может оказаться выше на данных для бутана, додекана, октана и пентана по сравнению с остальными данными, так как для этих данных значения ошибок наиболее велики.

Для данных без бутана значения основных трех метрик для модели MYS следующие:

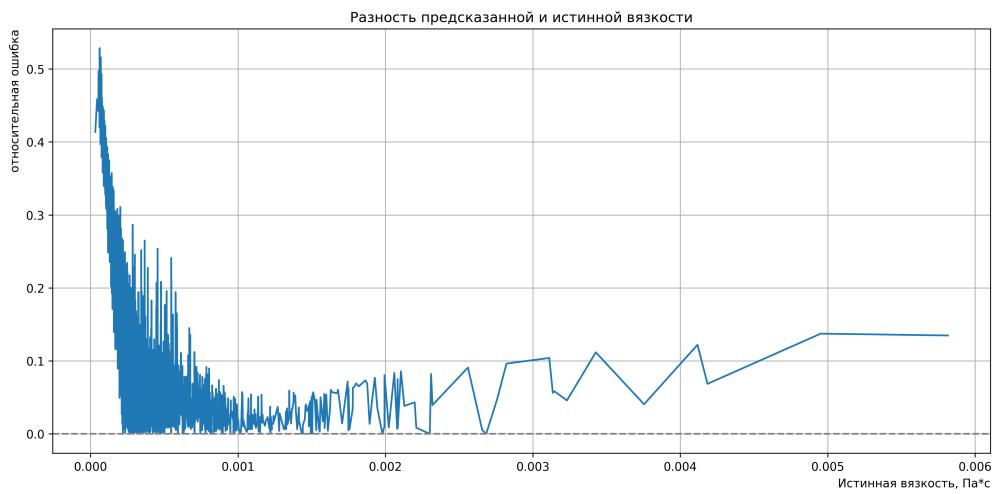


Рис. 7: Зависимость относительной ошибки от истинного значения вязкости для модели MYS

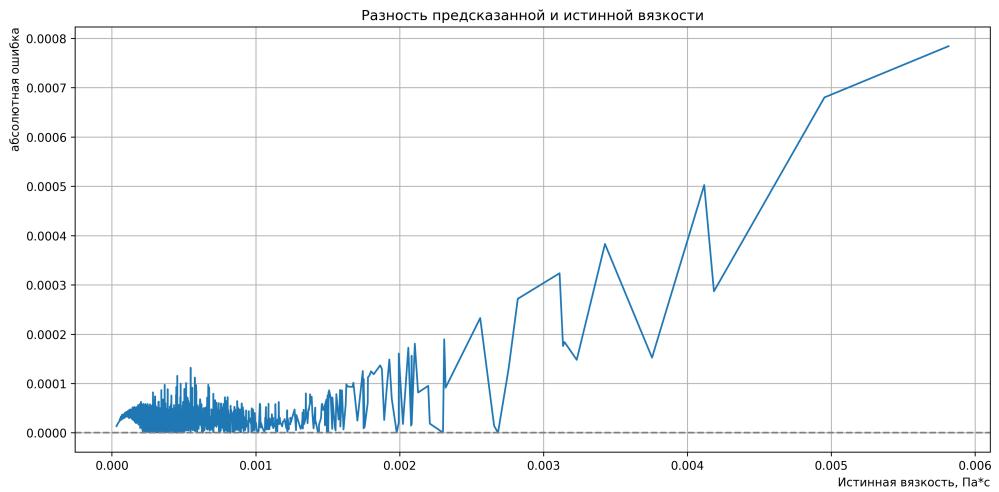


Рис. 8: Зависимость абсолютной ошибки от истинного значения вязкости для модели MYS

Таблица 4

Средние и медианные значения относительной и абсолютной ошибки модели MYS

| Вещество | Отн. ошибка, ср. | Отн. ошибка, мед. | Абс. ошибка, ср. | Абс. ошибка, мед. |
|----------|------------------|-------------------|-----------------------|-----------------------|
| бутан | 3.796 | 0.482 | 3.63×10^{-5} | 2.45×10^{-5} |
| пентан | 0.098 | 0.077 | 1.93×10^{-5} | 1.98×10^{-5} |
| гексан | 0.048 | 0.043 | 1.40×10^{-5} | 1.22×10^{-5} |
| гептан | 0.069 | 0.055 | 2.63×10^{-5} | 2.25×10^{-5} |
| октан | 0.086 | 0.069 | 3.05×10^{-5} | 2.94×10^{-5} |
| нонан | 0.040 | 0.044 | 2.65×10^{-5} | 2.80×10^{-5} |
| декан | 0.069 | 0.031 | 2.57×10^{-5} | 2.26×10^{-5} |
| додекан | 0.106 | 0.032 | 4.25×10^{-5} | 3.04×10^{-5} |

- **RMSE:** 4.85×10^{-5} Па·с
- **Средняя относительная ошибка:** 7.88×10^{-2}
- **Средняя абсолютная ошибка:** 2.99×10^{-5} Па·с

Дополнительно можно оценить, насколько важным было вычисление коэффициента вязкости идеального газа. Если его занулить, то метрики модели становятся следующими:

- **RMSE**: 5.03×10^{-5} Па·с
- **Средняя относительная ошибка**: 9.11×10^{-2}
- **Средняя абсолютная ошибка**: 3.14×10^{-5} Па·с

Все метрики стали хуже, что и ожидалось. Особенно сильно увеличилась относительная ошибка. Это подтверждает то, что вязкость идеального газа дает наибольший вклад в случаях, когда итоговое значение вязкости мало.

3.2. Символьная регрессия

В качестве дополнительного подхода была протестирована символьная регрессия на базе библиотеки PySR. Метод направлен на построение компактных аналитических выражений, аппроксимирующих целевую функцию, строя математические выражения на основе признаков. В результате был получен аналитический вид зависимости вязкости от нескольких параметров вещества:

Символьная регрессия была выполнена с использованием библиотеки PySR, основанной на генетическом поиске выражений с минимальной сложностью и максимальной точностью. В качестве признаков использовались параметры уравнения состояния CP-PC-SAFT, а в качестве целевой переменной – вязкость в Паскалях·секундах.

Для запуска модели использовались следующие параметры:

- число итераций: niterations = 1000;
- бинарные операторы: $\{+, -, \times, \div\}$;
- унарные операторы: $\{\exp, \log\}$;
- функция потерь: $(x - y)^2$;
- максимальная сложность выражения: maxsize = 40;
- максимальная глубина дерева выражения: maxdepth = 5;
- штраф за сложность (парсимония): $\lambda = 0.001$;
- оценка сложности операторов: \exp и \log – 3, остальные – 1.

Обучение модели проводилось на процессоре 11th Gen Intel® Core™ i7-1165G7 с использованием 8 потоков на 4 физических ядрах. Общая продолжительность вычислений составила около 30 минут. По завершении поиска модель автоматически выбрала наилучшее выражение по балансу между точностью и сложностью, которое приведено далее:

$$\hat{\eta} = \left(\frac{(v - 0.0003949) \cdot \sigma}{\frac{\epsilon/k_B}{s^{\text{res}}} + 1.7387} \right) - 0.0001168,$$

где:

- $\hat{\eta}$ – предсказанное значение вязкости [Па·с];
- v – мольный объем [$\text{м}^3/\text{моль}$];
- σ – диаметр сегмента для уравнения CP-PC-SAFT [\AA];
- ϵ/k_B – средняя энергия взаимодействия для уравнения CP-PC-SAFT [К];
- s^{res} – избыточная энтропия [Дж/(моль·К)].

Полученное выражение является достаточно компактным и легко интерпретируемым с точки зрения размерностей и физического смысла. Оно отражает функциональную зависимость вязкости от четырех ключевых параметров: мольного объема, размера сегмента, избыточной энтропии и глубины потенциальной ямы. При этом сложно дать однозначное объяснение данной формуле, что указывает на потенциальное наличие перекрестных эффектов между ними. Например, числитель увеличивается с ростом мольного объема и диаметра сегмента, что может быть связано с ростом внутреннего трения. Знаменатель, содержащий отношение энергии взаимодействия к энтропии, отражает сложную связь между термодинамической упорядоченностью и глубиной межмолекулярных взаимодействий. В целом, модель подчеркивает, что вязкость определяется не отдельным параметром, а их соотношением в конкретных условиях.

На Рис. 10 и Рис. 9 представлены графики абсолютной и относительной ошибки модели на основе символьной регрессии:

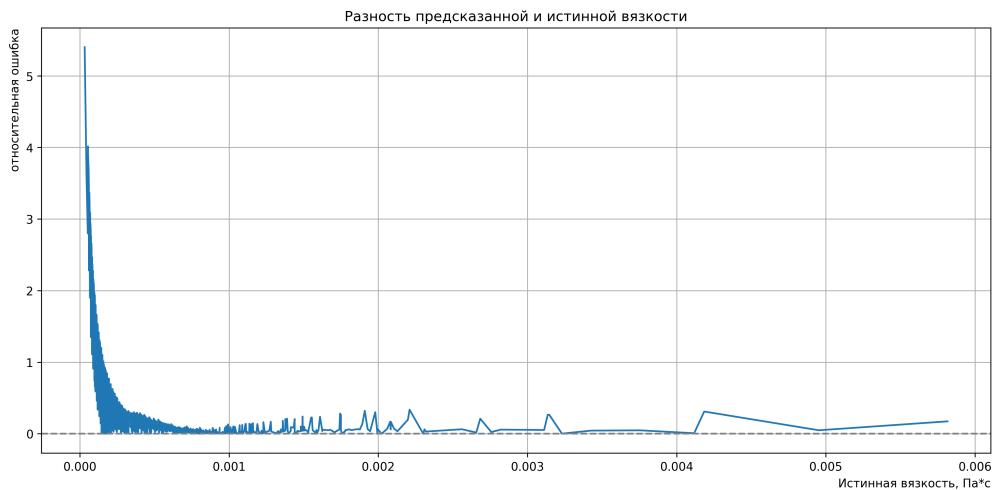


Рис. 9: Относительная ошибка символьной модели в зависимости от истинной вязкости

Несмотря на простоту выражения, модель показала относительно приемлемую точность:

- RMSE: 9.87×10^{-5} Па·с;
- Средняя абсолютная ошибка (MAE): 6.67×10^{-5} Па·с;
- Средняя относительная ошибка (MRE): 0.231.

Также стоит отметить, что символьная регрессия тренировалась на полном наборе данных без разбиения на тестовую и тренировочную выборку. Вряд ли это сильно уменьшило

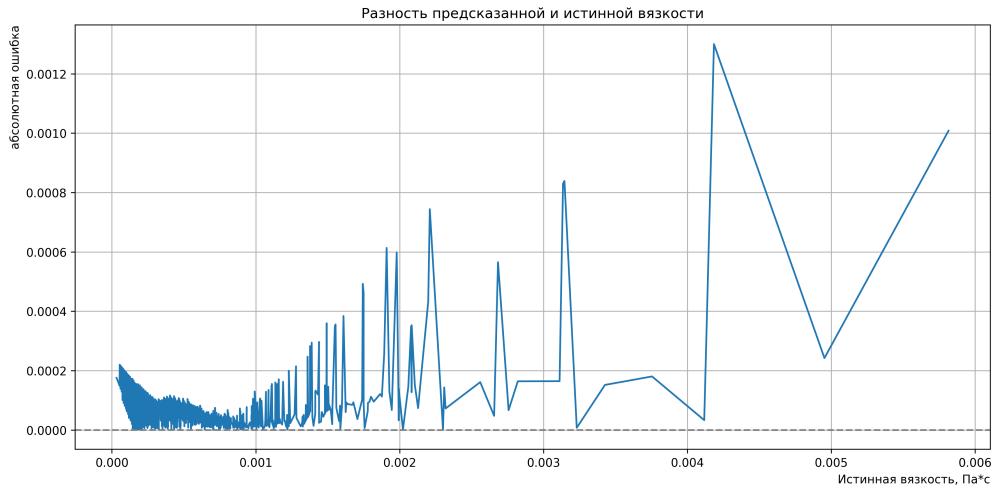


Рис. 10: Абсолютная ошибка предсказания символьной модели

значения ошибок ввиду сильной устойчивости к переобучению у данной модели, но результат мог бы быть потенциально хуже.

Таким образом, символьная регрессия может быть полезным инструментом для построения простых аналитических аппроксимаций, особенно в условиях ограниченного количества данных или требований к интерпретируемости. Метод оказался полезным как дополнительный инструмент анализа, но не подошел в качестве основной модели. Отдельный интерес представляет оптимизация параметров запуска модели символьной регрессии или запуск алгоритма на более долгое время, так как генетические алгоритмы могут внезапно находить целые классы решений и улучшать свою точность на протяжении долгого времени без угрозы переобучения.

3.3. Алгоритм автоматической генерации признаков

Начальные эксперименты с обучением простых моделей показали, что использование только исходных признаков приводит к значительно более низкой точности по сравнению с расширенным набором, включающим попарные произведения признаков. Краткий анализ продемонстрировал, что наибольший вклад в прирост точности дают лишь немногие из добавленных признаков. Это наблюдение мотивировало разработку алгоритма автоматической генерации признаков, направленного на отбор и итеративное улучшение наиболее информативных преобразований признаков.

Алгоритм генерации новых признаков реализован в виде последовательности этапов:

- Генерация новых признаков с использованием элементарных операторов.** Унарные операторы (например, экспонента, логарифм, квадратный корень) применяются к каждому признаку индивидуально. Бинарные операторы (+, -, *, /) применяются ко всем возможным парам признаков.
- Удаление некорректных или численно нестабильных признаков.** Исключаются признаки, нарушающие элементарные арифметические правила, например, ведущие к делению на ноль или взятию логарифма от отрицательных чисел. Также отбрасываются признаки с чрезмерно большими значениями, способные вызвать переполнение при хранении в формате float32.

в) **Удаление избыточных признаков.** Исключаются признаки, обладающие высокой взаимной корреляцией или практически идентичные. В частности:

- Если векторная норма разности нормализованных признаков меньше порогового значения (например, 10^{-8}).
- Если коэффициент корреляции Пирсона между признаками превышает заданный уровень (например, 0.998).

В случае дублирования предпочтение отдается более короткому признаку (с меньшим числом операций в формуле).

г) **Снижение общего числа признаков.** Если число новых признаков превышает установленный лимит (в нашем случае — не более $1/5$ от общего количества объектов в данных), выбирается подмножество наиболее компактных признаков. Отбор осуществляется с использованием функции `random.choices` из стандартной библиотеки Python, где каждому признаку присваивается вес, пропорциональный $1/(1 + \text{len})^3$, где `len` — количество операций, необходимых для вычисления признака.

д) **Обучение моделей на расширенном наборе признаков.** Данные делятся на обучающую и тестовую выборки (примерно 80% и 20% соответственно) с использованием `sklearn.model_selection.train_test_split`, при этом значение `seed` фиксируется для воспроизводимости результатов. Тестовая выборка не участвует в обучении и используется исключительно для оценки качества моделей.

е) **Оценка важности признаков.** Для каждой модели вычисляется список признаков, отсортированных по убыванию значимости. Требуется, чтобы каждая модель поддерживала метод оценки важности признаков. Методы расчета будут рассмотрены далее в описании моделей.

ж) **Добавление информативных признаков.** В итоговый набор добавляются признаки, попавшие в число 10 наиболее важных хотя бы для одной из моделей (если они еще не были добавлены ранее). Выбор размера этого списка балансирует между скоростью заполнения признакового пространства и устойчивостью моделей к переизбыточности.

з) **Удаление неинформативных признаков.** Из признакового пространства исключаются признаки, оказавшиеся в числе 20 наименее значимых одновременно для всех моделей. Это достигается через пересечение списков слабейших признаков, сформированных для каждой модели. Выбор размера списка влияет на баланс между уменьшением шума и сохранением разнообразия признаков.

и) **Сохранение текущего состояния.** Через фиксированные интервалы итераций полный список активных признаков сохраняется в CSV-файл, что позволяет продолжать генерацию признаков в дальнейшем с текущего состояния.

к) **Итеративное повторение.** Описанный алгоритм повторяется заданное число раз, после чего проводится итоговая оценка качества моделей.

л) **Строгий отсев.** По окончании итеративного улучшения производится строгий отсев коэффициентов. На этом этапе промежуточные параметры, которые могут быть важны, как составные части сложных признаков становятся не важны, поэтому их можно

отбросить. Оставляется примерно 50 коэффициентов, которые будут использованы для итоговой валидации методов.

Полученным формулам трудно дать глубокий смысл. Их смысл не в том, чтобы точно или универсально описывать систему, а чтобы лучше всего исправлять ошибку, которую допускают модели используя все остальные коэффициенты. Примеры сгенерированных параметров можно увидеть ниже.

$$\frac{T_{tp} \cdot \sigma \cdot m}{\exp^2 \left(\frac{s^{res}}{8.314462618 m} \right) \cdot (\varepsilon/k + s^{res})}$$

$$m \cdot P'^2 \cdot (M - (\varepsilon/k + s^{res}))$$

$$T_{tp} \cdot \frac{s^{res}(M_w - T_{tp})P'}{T - s^{res}}$$

Где P' это частная производная давления по объему при постоянной температуре, найденная в процессе вычисления мольного объема:

$$P' = \left(\frac{\partial P}{\partial v} \right)_T$$

Данные формулы применялись к базовым признакам, а их результаты записывались как новые признаки. Таким образом, набор входных данных был увеличен с 12 признаков до 50. Большинство тестов проходило с 53 входными параметрами. Стоит уточнить, что данные параметры не подбирались каждый раз при обучении моделей а были сгенерированы заранее. Признаки выбирались в зависимости от их эффективности на тренировочном наборе данных по описанному выше алгоритму. Валидация моделей, обученных на найденных параметрах, показала, что переобучение не наблюдается. Следует отметить, что все численные параметры, использованные в процессе – включая количество добавляемых и удаляемых признаков, пороговые значения для корреляции и различия признаков, ограничения на общее количество признаков и веса в функции выбора – были подобраны эмпирически на основе анализа поведения моделей в ходе десятков независимых запусков. Эти значения показали хорошую эффективность в контексте текущего исследования, однако не претендуют на универсальность и не являются строгими рекомендациями. Их настройка должна осуществляться с учетом специфики конкретной задачи и свойств исходных данных.

3.4. Выводы по главе

В данной главе были рассмотрены интерпретируемые подходы к моделированию вязкости и представлен алгоритм генерации новых признаков. Анализ модели MYS показал, что, несмотря на физическую обоснованность, ее точность ограничена, особенно на участках с низкой вязкостью, где вклад идеального газа становится критичен. Полученное с помощью символьной регрессии аналитическое выражение продемонстрировало разумный баланс между точностью и компактностью, подтверждая потенциал метода для первичного анализа данных.

Разработанный алгоритм автоматической генерации признаков позволил обогатить исходный набор параметров новыми комбинациями, значимость которых оценивалась с точки зрения влияния на точность моделей. Итеративная природа алгоритма, а также строгий отбор и удаление признаков, должны позволить сохранить интерпретируемость и устойчивость моделей. Полученные признаки и формулы станут основой для дальнейшего обучения моделей и сравнения с MYS друг другом.

4. Обучение и сравнение ML моделей

На этапе построения моделей машинного обучения рассматривались различные алгоритмы: линейная регрессия, решающие деревья, метод случайных соседей, случайный лес, а также нейросети. Однако от нейросетей было решено отказаться – они показали избыточную сложность в тренировке и интерпретации результатов, не обеспечивая при этом существенного выигрыша в точности.

Для повышения качества обучения и анализа обобщающей способности моделей были проведены несколько экспериментов с разным составом данных. В частности, в большинстве экспериментов бутан был исключен из обучающей выборки. Это объясняется тем, что он представлен в базе в основном в газовой фазе, и на таких точках относительная ошибка теряет смысл: даже небольшое абсолютное отклонение приводит к очень высокой относительной погрешности. Хоть удаление бутана приводило к увеличению RMSE и MAE, но делало результаты более интерпретируемыми и устойчивыми.

Признаки:

- Базовый набор – температура, давление, параметры CP-PC-SAFT;
- Расширенный набор – признаки, сгенерированные автоматически и отобранные по важности.

Модели:

- Метод ближайших соседей;
- Линейная регрессия;
- Случайный лес.

Варианты разбиения данных:

- Классическое случайное разбиение 80% / 20% – для оценки максимальной точности на тестовой выборке;
- Обратное разбиение 20% / 80% – для оценки устойчивости и обобщающей способности.

Метрики качества:

- RMSE (среднеквадратичная ошибка);
- MAE (средняя абсолютная ошибка);
- MRE (средняя относительная ошибка).

Для получения наиболее достоверных значений метрик качества было использовано усреднение по множеству итераций. Как правило проводилось по 20 независимых обучений/тестов. Ошибки были подсчитаны на совместных предсказаниях со всех итераций.

4.1. Метод ближайших соседей

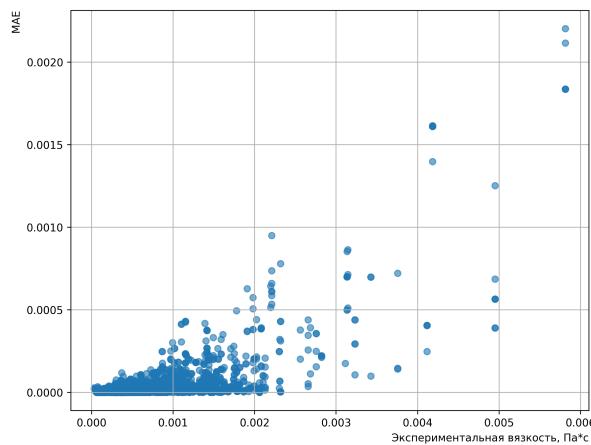
Метод k ближайших соседей (KNN) является одним из самых простых и интерпретируемых методов машинного обучения. Он не требует явной фазы обучения: предсказание для новой точки основывается на значениях целевой переменной у k ближайших по признаковому пространству соседей. В данной работе этот метод использовался как базовая отправная точка, позволяющая оценить нижнюю границу точности, которой можно достичь даже

без сложной внутренней структуры модели. Предполагалось, что обобщающая способность KNN будет крайне ограниченной.

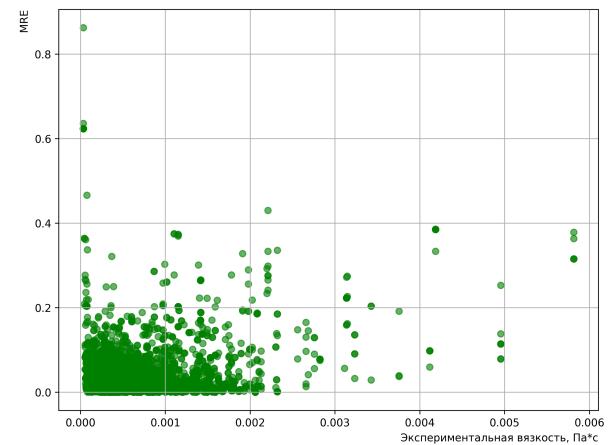
Для реализации метода была использована реализация KNeighborsRegressor из sklearn.neighbors на языке Python. Параметр количества ближайших соседей был выбран равным 3. Такой параметр давал лучшее значение RMSE во всех тестах далее. Данные были отнормированы перед использованием этого метода.

4.1.1. Оценка точности на большом обучающем наборе

На первом этапе метод был протестирован на базовом наборе признаков (температура, давление и параметры CP-PC-SAFT) при обучении на 80% данных. В этом случае модель показала хорошие значения средней относительной ошибки, однако RMSE оказался сравнительно высоким. Это указывает на наличие небольшого числа точек с особенно высокой ошибкой предсказания.



(а) Абсолютная ошибка



(б) Относительная ошибка

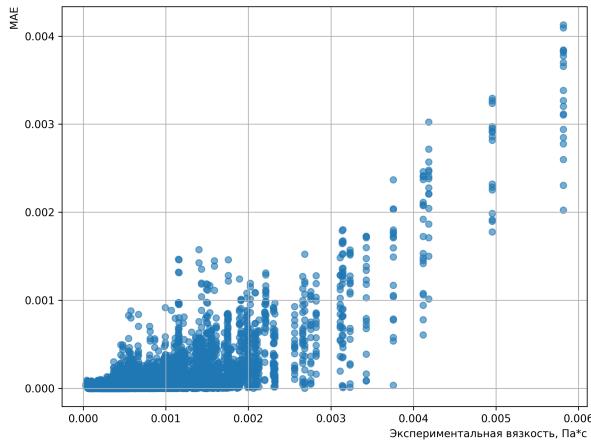
Рис. 11: Зависимость ошибки модели KNN от истинных значений вязкости. Базовый набор данных, большая обучающая выборка.

Метрики качества (базовые признаки, обучающая выборка 80%):

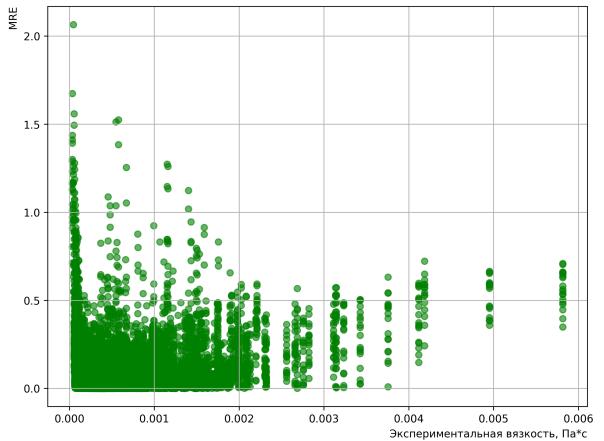
- RMSE: 9.09×10^{-5} Па·с;
- MAE: 2.48×10^{-5} Па·с;
- MRE: 0.0325.

4.1.2. Оценка на ограниченных данных

Чтобы оценить способность модели к обобщению, объемы тренировочной и тестовой выборок были поменяны местами (20% обучения, 80% теста). Как и ожидалось, точность предсказаний существенно снизилась: модель стала хуже справляться с ранее не встречавшимися данными, уступая модели MYS по всем метрикам.



(а) Абсолютная ошибка



(б) Относительная ошибка

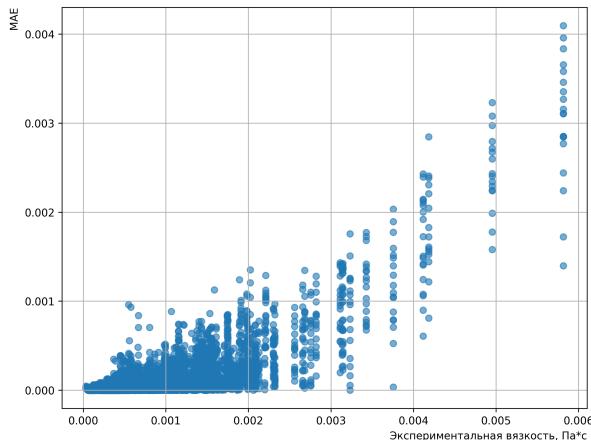
Рис. 12: Зависимость ошибки модели KNN от истинных значений вязкости. Базовый набор данных, малая обучающая выборка.

Метрики качества (базовые признаки, обучающая выборка 20%):

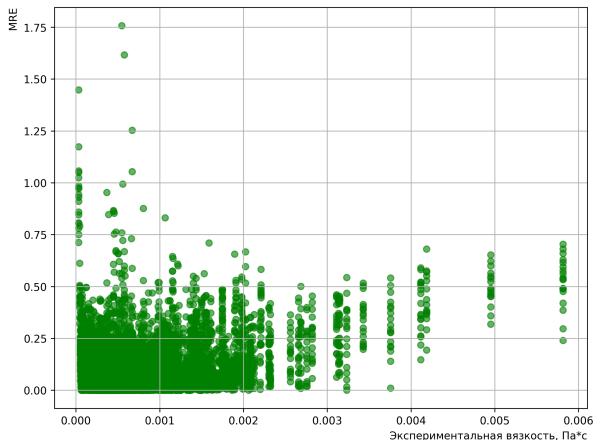
- RMSE: 1.78×10^{-4} Па·с;
- MAE: 4.95×10^{-5} Па·с;
- MRE: 0.0658.

4.1.3. Влияние расширенных признаков

Для проверки, может ли расширенный набор признаков (сгенерированные комбинации и трансформации) улучшить точность в условиях малого числа обучающих данных, модель была протестирована при тех же условиях, но с новым набором признаков. Качество предсказаний немного выросло, но все еще не достигло уровня, наблюдавшегося при большом тренировочном наборе.



(а) Абсолютная ошибка



(б) Относительная ошибка

Рис. 13: Зависимость ошибки модели KNN от истинных значений вязкости. Расширенный набор данных, малая обучающая выборка.

Метрики качества (расширенные признаки, обучающая выборка 20%):

- RMSE: 1.62×10^{-4} Па·с;
- MAE: 4.65×10^{-5} Па·с;
- MRE: 0.0594.

4.1.4. Вывод по методу KNN

Модель KNN ожидаемо показала хорошую точность на обучающих данных и слабую обобщающую способность. Это подтверждает предположение, что метод плохо переносится на новые области пространства признаков. Использование расширенных признаков позволило немного улучшить результаты, но не устранило фундаментальные ограничения метода. KNN может использоваться как базовая оценка и для валидации корректности данных, но не подходит в качестве основной модели для задач обобщения и предсказания вязкости по параметрам вещества.

4.2. Случайный лес

Модель случайного леса относится к ансамблевым методам машинного обучения и строится на объединении большого количества решающих деревьев. В отличие от метода ближайших соседей, который делает предсказания, опираясь исключительно на локальные участки пространства признаков, случайный лес формирует обобщенные предсказания, агрегируя решения деревьев, обученных на разных подмножествах данных. Благодаря этому, модель хорошо справляется с задачами регрессии в условиях сложных нелинейных зависимостей между признаками.

В данной работе использовалась реализация RandomForestRegressor из библиотеки sklearn.ensemble. Модель тестировалась с базовыми параметрами по умолчанию ($n_estimators = 100$ и др.), так как изменение настроек не приводило к устойчивому улучшению качества, а в отдельных случаях даже снижало точность.

4.2.1. Оценка точности на большом обучающем наборе

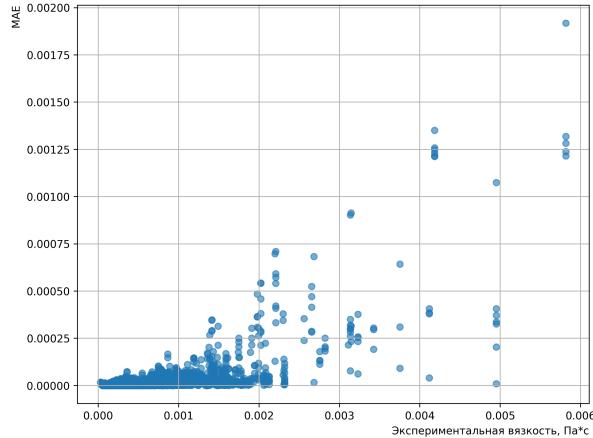
При обучении модели на 80% данных и использовании только базовых признаков случайный лес показал хорошие результаты. По сравнению с методом ближайших соседей, была достигнута существенно более высокая точность, особенно по RMSE и MAE, хотя значения относительной ошибки оставались сравнимыми. Результаты представлены на рисунке Рис. 14.

Метрики качества (базовые признаки, обучающая выборка 80%):

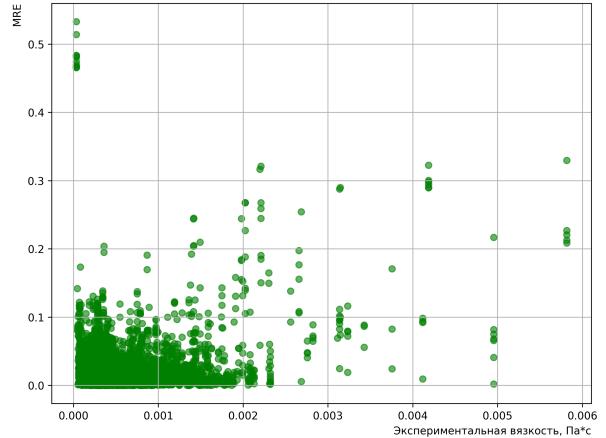
- RMSE: 6.40×10^{-5} Па·с;
- MAE: 1.63×10^{-5} Па·с;
- MRE: 0.0212.

4.2.2. Оценка на ограниченных данных

Для оценки способности модели обобщать данные была проведена обратная процедура: обучение на 20% выборки и тестирование на оставшихся 80%. При этом значения ошибок



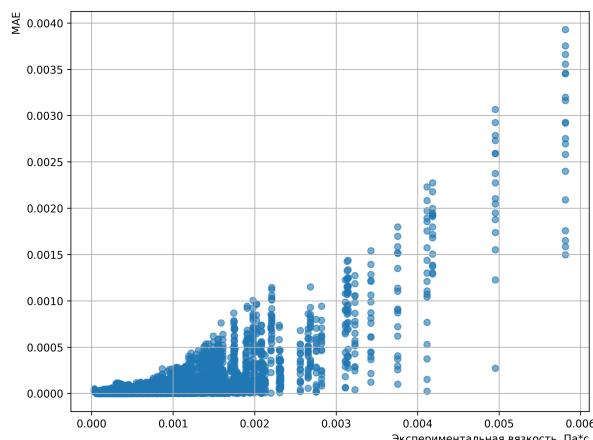
(а) Абсолютная ошибка



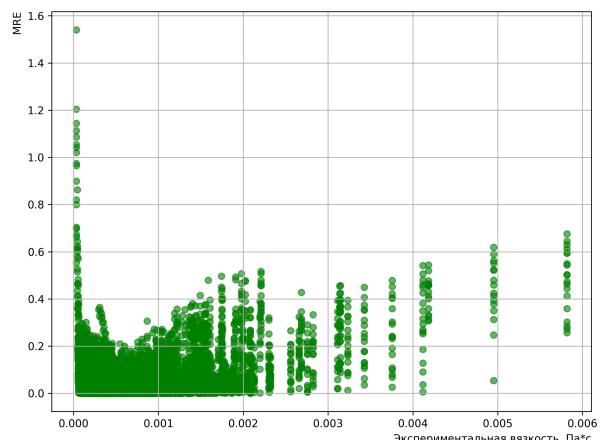
(б) Относительная ошибка

Рис. 14: Зависимость ошибки модели случайного леса от истинных значений вязкости. Базовый набор признаков, большая обучающая выборка.

возросли, как и ожидалось, однако модель сохраняла уверенную точность. Относительная ошибка по-прежнему оставалась ниже 5%, что свидетельствует о приемлемой обобщающей способности Рис. 15.



(а) Абсолютная ошибка



(б) Относительная ошибка

Рис. 15: Зависимость ошибки модели случайного леса от истинных значений вязкости. Базовый набор признаков, малая обучающая выборка.

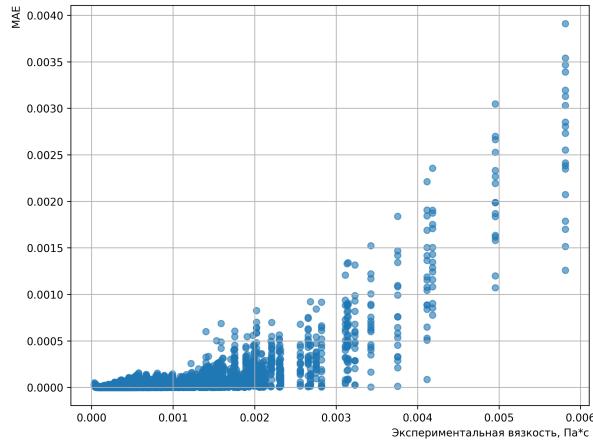
Метрики качества (базовые признаки, обучающая выборка 20%):

- RMSE: 1.36×10^{-4} Па·с;
- MAE: 3.22×10^{-5} Па·с;
- MRE: 0.0390.

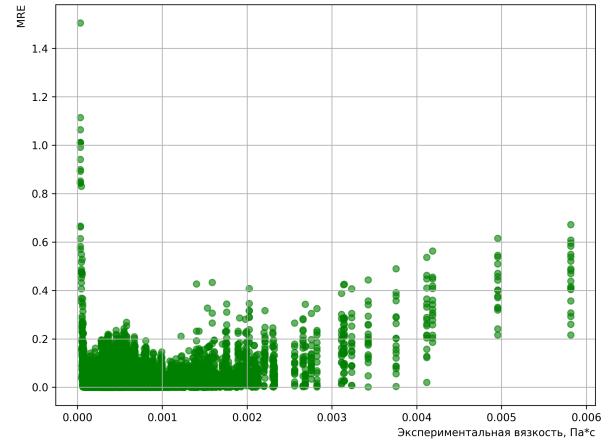
4.2.3. Влияние расширенных признаков

Далее была проверена эффективность модели на расширенном наборе признаков, включая автоматически сгенерированные комбинации и преобразования. Обучение вновь

проводилось на 20% выборки. Результаты показали, что такие признаки действительно улучшают качество предсказания: значения всех ошибок снизились по сравнению с базовым вариантом Рис. 16.



(а) Абсолютная ошибка



(б) Относительная ошибка

Рис. 16: Зависимость ошибки модели случайного леса от истинных значений вязкости. Расширенный набор признаков, малая обучающая выборка.

Метрики качества (расширенные признаки, обучающая выборка 20%):

- RMSE: 1.10×10^{-4} Па·с;
- MAE: 2.26×10^{-5} Па·с;
- MRE: 0.0282.

4.2.4. Вывод по методу случайного леса

Модель случайного леса показала высокую точность и хорошую устойчивость к переобучению. Даже при сокращении обучающей выборки в четыре раза, она сохранила приемлемый уровень ошибок, что делает ее надежным инструментом для регрессии в условиях ограниченных данных. Использование расширенных признаков дополнительно повысило точность, особенно по метрике MAE. Как и в случае с методом ближайших соседей, RMSE оказался сравнительно выше, что свидетельствует о наличии отдельных точек с крупной ошибкой предсказания. Однако общее поведение модели позволяет считать случайный лес одной из самых надежных и интерпретируемых моделей в рамках данного исследования.

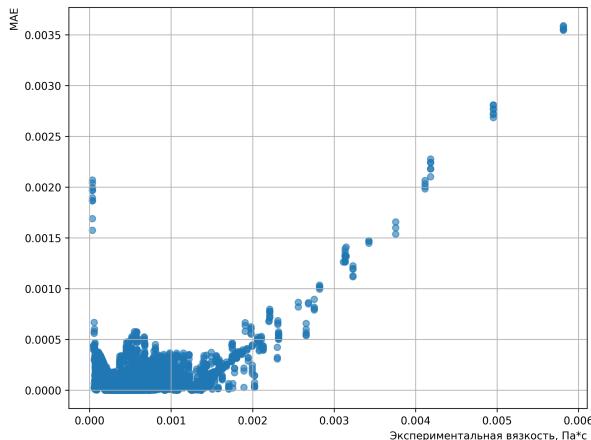
4.3. Линейная регрессия

Линейная регрессия является одной из наиболее простых и интерпретируемых моделей машинного обучения. Ее основное преимущество заключается в прозрачной структуре: каждое значение признака вносит линейный вклад в итоговое предсказание, а коэффициенты модели легко интерпретируются как меры влияния соответствующих признаков.

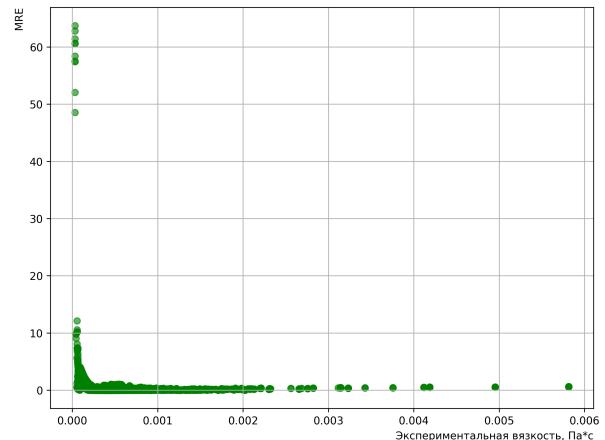
В данной работе использовалась реализация Ridge из библиотеки `sklearn.linear_model`. Модель с параметром регуляризации $\alpha = 0.003$ показала наибольшую устойчивость. Остальные параметры остались без изменений. Перед использованием модели входные данные проходили нормализацию.

4.3.1. Оценка точности на большом обучающем наборе

При обучении модели на 80% данных и использовании только базовых признаков линейная модель показала очень слабый результат. Оба предыдущих метода значительно превзошли линейную регрессию в данном случае. Результаты представлены на рисунке Рис. 17.



(а) Абсолютная ошибка



(б) Относительная ошибка

Рис. 17: Зависимость ошибки линейной регрессии от истинных значений вязкости. Базовый набор признаков, большая обучающая выборка.

Метрики качества (базовые признаки, обучающая выборка 80%):

- RMSE: 2.35×10^{-4} Па·с;
- MAE: 1.11×10^{-4} Па·с;
- MRE: 0.3885.

4.3.2. Оценка на ограниченных данных

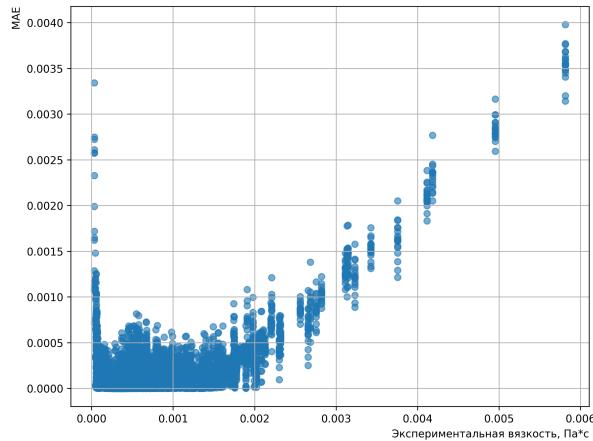
При обучении на 20% данных модель не стала хуже, сохранив значения всех ошибок примерно на том же уровне. Это может говорить о том, что модель ограничена количеством признаков, а не количеством точек. Небольшую разницу в поведении модели можно увидеть на Рис. 18.

Метрики качества (базовые признаки, обучающая выборка 20%):

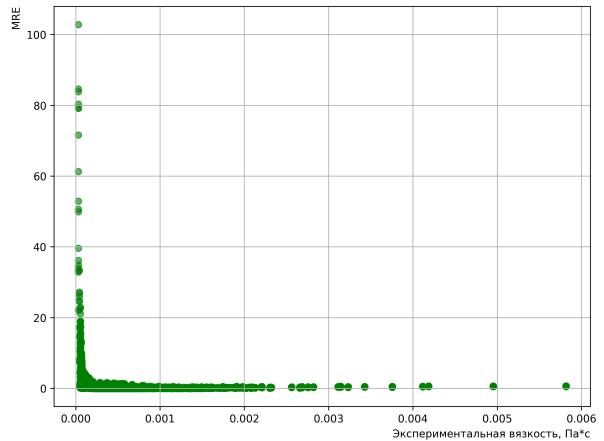
- RMSE: 2.29×10^{-4} Па·с;
- MAE: 1.08×10^{-4} Па·с;
- MRE: 0.3628.

4.3.3. Влияние расширенных признаков

При использовании расширенных признаков и обучении на 20% выборки результаты стали гораздо лучше. Модель значительно превзошла корреляцию MYS по всем трем метрикам.



(а) Абсолютная ошибка

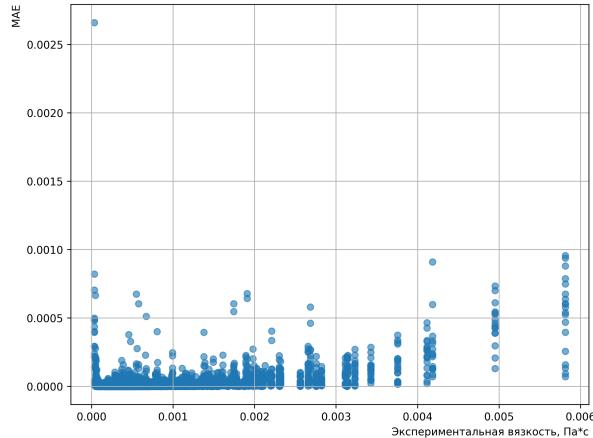


(б) Относительная ошибка

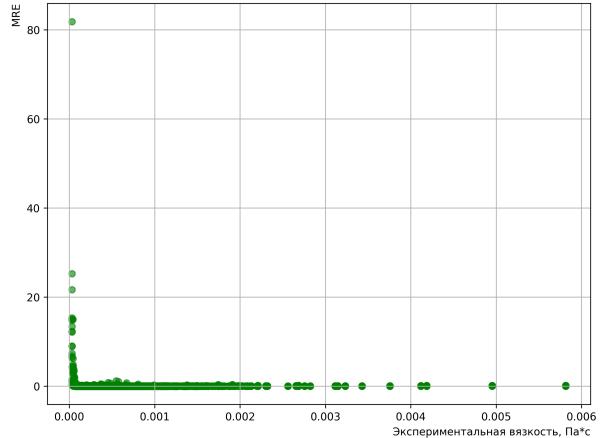
Рис. 18: Зависимость ошибки линейной регрессии от истинных значений вязкости. Базовый набор признаков, малая обучающая выборка.

При этом значение ошибки MRE выше, чем в методе случайного леса. Однако и здесь относительная ошибка менее 5%, что позволяет заявить, что данный метод пригоден для прикладных задач.

На графиках, демонстрирующих метрики Рис. 19 можно заметить много точек, дающих большую абсолютную ошибку, несмотря на маленькие значения истинной вязкости. Недолгий анализ выявил, что эти точки находятся вблизи критической точки. Большая часть из них имеют давление около 3 МПа и температуру от 400 градусов. На основании этого можно сделать вывод, что линейная регрессия может испытывать трудности при определении свойств смесей вблизи критической точки.



(а) Абсолютная ошибка



(б) Относительная ошибка

Рис. 19: Зависимость ошибки линейной регрессии от истинных значений вязкости. Расширенный набор признаков, малая обучающая выборка.

Метрики качества (расширенные признаки, обучающая выборка 20%):

- RMSE: 3.39×10^{-5} Па·с;
- MAE: 1.05×10^{-5} Па·с;
- MRE: 0.0320.

4.3.4. Вывод по линейной регрессии

Линейная регрессия, как и ожидалось, показала ограниченные возможности на базовом наборе признаков. Наиболее ярко это проявилось в эксперименте с большой обучающей выборкой, где модель не смогла захватить сложные зависимости между параметрами, несмотря на большое количество данных. Это указывает скорее на фундаментальную слабость модели в предсказании признаков на основании малого числа входных параметров, чем на недостаточный размер обучающих данных. Интересным наблюдением стало и то, что уменьшение обучающей выборки практически не повлияло на качество линейной модели.

На фоне этого особенно показательной стала роль расширенных признаков. Добавление сгенерированных признаков позволило резко улучшить точность предсказаний, особенно по метрикам RMSE и MAE. Это подтверждает предположение о том, что основной ограничивающий фактор линейной модели, это ее ограниченная гибкость – точность в большей степени определялась признаками, а не объемом данных.

Отдельного внимания заслуживают зоны, в которых линейная модель особенно плохо справлялась с предсказанием – вблизи критических точек. В частности, ошибки резко возрастили при давлениях около 3 МПа и температурах порядка 400 К. Скорее всего, это можно исправить, если генерировать признаки с упором на описание данных областей или если увеличить количество похожих данных.

В результате, несмотря на простоту, линейная модель, обученная на расширенном признаковом пространстве, продемонстрировала наилучшие метрики среди всех протестированных подходов. Особенно важно, что средняя относительная ошибка не превышала 5%, что делает подход пригодным для прикладного использования, в том числе в инженерных расчетах и быстрых приближенных оценках.

4.4. Сравнение подходов

На данном этапе обобщим результаты, полученные при тестировании всех рассмотренных моделей. Для максимально справедливого сравнения использовались следующие условия:

- обучение проводилось на ограниченном объеме данных (20% всей выборки), чтобы выявить обобщающую способность моделей;
- в качестве признаков использовался расширенный набор, включающий автоматически сгенерированные признаки, поскольку на нем модели демонстрировали наилучшую точность;
- модели сравнивались по RMSE, MAE и MRE, полученным при проверке на тестовой части выборки;
- каждая модель обучалась и тестировалась на протяжении 20 итераций;

Таблица 5**Характеристики ошибок для различных моделей**

| Модель | RMSE ($\times 10^{-4}$ Па·с) | | MAE ($\times 10^{-5}$ Па·с) | | MRE | |
|-------------------------------|-------------------------------|------|------------------------------|------|---------|--------|
| | Среднее | Std | Среднее | Std | Среднее | Std |
| Метод ближайших соседей (KNN) | 1.62 | 0.33 | 4.65 | 0.53 | 0.0594 | 0.0044 |
| Случайный лес | 1.10 | 0.35 | 2.24 | 0.46 | 0.0280 | 0.0027 |
| Линейная регрессия | 0.34 | 0.14 | 1.05 | 0.14 | 0.0320 | 0.0173 |
| MYS | 0.49 | 0.02 | 2.99 | 0.04 | 0.0783 | 0.0013 |

Также, стоит уточнить, что модель MYS не обучалась, а отклонения ошибок обусловлены случайнм разбиением данных на каждой итерации.

Как видно из таблицы Табл. 5, наиболее высокую точность по всем трем метрикам показала линейная модель, обученная на расширенных признаках. Несмотря на свою простоту, она смогла обойти даже сложные ансамблевые методы, благодаря качественно сконструированным входным данным. Особенno важно, что ее средняя относительная ошибка не превышала 5%, что делает ее пригодной для практического применения.

Случайный лес продемонстрировал стабильность и высокую точность на малых выборках, что делает его универсальным инструментом в условиях, где сложно обеспечить большой обучающий набор. Метод ближайших соседей, как и ожидалось, показал слабую способность к обобщению, несмотря на хорошую точность на тренировочных данных. Модель MYS, несмотря на простоту, показала устойчивые и интерпретируемые результаты, служа надежной точкой отсчета для оценки новых моделей.

В совокупности, результаты подтверждают идею, что правильная генерация и отбор признаков могут радикально изменить эффективность даже самых простых моделей.

4.5. Применимость к новым условиям

Чтобы окончательно протестировать обобщающую способность моделей был придуман еще один набор тестов. Их смысл в исключении данных не случайнм образом, а согласно некоторому критерию. Далее будет рассмотрено, как модели предсказывают вязкость на примерах, подобных которым не было в обучающей выборке.

В экспериментах далее был использован бутстрэппинг, так как разделение данных по какому-то признаку часто приводило к сильным диспропорциям выборок. Это было сделано, чтобы минимизировать влияние изменения размеров обучающего и тренировочного набора данных и сохранить их размеры одинаковыми. Это также позволило дать оценки стандартного отклонения в процессе обучения. Для обозначения стандартного отклонения в таблице используется общепринятое обозначение Std.

4.5.1. Новые вещества

В этот раз целое вещество было удалено из выборки, а затем валидация проводилась только по нему. При этом модели обучались на всем остальном наборе данных. Результаты представлены в Табл. 6.

В ходе тестирования пришлось значительно увеличить параметр регуляризации линейной регрессии до значения $alpha = 0.6$. Иначе модель сильно отклонялась и уступала всем остальным. Нужно держать в уме необходимость регуляризации при использовании подобных моделей на новых данных. В свою очередь, случайный лес показал чрезвычайно хоро-

Таблица 6

Характеристики ошибок при оценке на новых веществах для различных моделей

| Алкан | MYS | | KNN | | Случайный лес | | Линейная регрессия | |
|---------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | RMSE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| пентан | 2.32×10^{-5} | 6.02×10^{-7} | 3.40×10^{-5} | 2.97×10^{-6} | 1.08×10^{-5} | 1.10×10^{-6} | 2.04×10^{-5} | 1.02×10^{-5} |
| гексан | 1.80×10^{-5} | 8.29×10^{-7} | 4.94×10^{-5} | 1.34×10^{-6} | 1.36×10^{-5} | 1.65×10^{-6} | 1.99×10^{-5} | 6.43×10^{-6} |
| гептан | 3.35×10^{-5} | 1.33×10^{-6} | 8.40×10^{-5} | 7.92×10^{-6} | 2.17×10^{-5} | 2.32×10^{-6} | 6.58×10^{-5} | 2.13×10^{-5} |
| октан | 3.70×10^{-5} | 9.87×10^{-7} | 1.05×10^{-4} | 8.86×10^{-6} | 3.25×10^{-5} | 4.80×10^{-6} | 4.07×10^{-5} | 1.17×10^{-5} |
| нонан | 3.15×10^{-5} | 8.54×10^{-7} | 8.88×10^{-5} | 3.66×10^{-6} | 1.65×10^{-5} | 1.66×10^{-6} | 1.78×10^{-5} | 4.11×10^{-6} |
| декан | 3.65×10^{-5} | 4.17×10^{-6} | 1.42×10^{-4} | 4.45×10^{-5} | 4.69×10^{-5} | 1.82×10^{-5} | 4.18×10^{-5} | 1.63×10^{-5} |
| додекан | 7.46×10^{-5} | 1.11×10^{-5} | 3.25×10^{-4} | 4.27×10^{-5} | 2.38×10^{-4} | 3.92×10^{-5} | 8.70×10^{-5} | 1.98×10^{-5} |
| | MAE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| пентан | 1.93×10^{-5} | 6.56×10^{-7} | 3.05×10^{-5} | 2.40×10^{-6} | 9.13×10^{-6} | 9.46×10^{-7} | 1.81×10^{-5} | 1.04×10^{-5} |
| гексан | 1.41×10^{-5} | 6.55×10^{-7} | 4.62×10^{-5} | 1.48×10^{-6} | 1.00×10^{-5} | 1.48×10^{-6} | 1.73×10^{-5} | 6.90×10^{-6} |
| гептан | 2.65×10^{-5} | 1.02×10^{-6} | 6.94×10^{-5} | 4.04×10^{-6} | 1.24×10^{-5} | 1.31×10^{-6} | 3.61×10^{-5} | 1.29×10^{-5} |
| октан | 3.05×10^{-5} | 9.58×10^{-7} | 6.68×10^{-5} | 3.29×10^{-6} | 1.96×10^{-5} | 1.72×10^{-6} | 2.99×10^{-5} | 7.60×10^{-6} |
| нонан | 2.64×10^{-5} | 8.55×10^{-7} | 7.86×10^{-5} | 3.26×10^{-6} | 1.29×10^{-5} | 1.34×10^{-6} | 1.47×10^{-5} | 4.54×10^{-6} |
| декан | 2.58×10^{-5} | 1.28×10^{-6} | 5.37×10^{-5} | 1.08×10^{-5} | 1.82×10^{-5} | 3.30×10^{-6} | 1.98×10^{-5} | 5.71×10^{-6} |
| додекан | 4.20×10^{-5} | 3.04×10^{-6} | 1.63×10^{-4} | 1.79×10^{-5} | 1.25×10^{-4} | 3.13×10^{-5} | 3.60×10^{-5} | 4.09×10^{-6} |
| | MRE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| пентан | 0.0978 | 0.0032 | 0.1370 | 0.0101 | 0.0457 | 0.0045 | 0.0893 | 0.0512 |
| гексан | 0.0478 | 0.0021 | 0.1553 | 0.0057 | 0.0334 | 0.0051 | 0.0612 | 0.0248 |
| гептан | 0.0690 | 0.0034 | 0.1608 | 0.0047 | 0.0264 | 0.0024 | 0.0908 | 0.0347 |
| октан | 0.0858 | 0.0049 | 0.1373 | 0.0059 | 0.0392 | 0.0026 | 0.1092 | 0.0463 |
| нонан | 0.0401 | 0.0011 | 0.1236 | 0.0047 | 0.0208 | 0.0021 | 0.0249 | 0.0085 |
| декан | 0.0692 | 0.0054 | 0.0613 | 0.0074 | 0.0233 | 0.0025 | 0.0290 | 0.0059 |
| додекан | 0.1035 | 0.0084 | 0.1263 | 0.0116 | 0.0938 | 0.0258 | 0.0599 | 0.0131 |

шие результаты, превзойдя все остальные модели на всех веществах кроме додекана. Только на самом сложном для этой модели случае ее относительная ошибка была больше 5%.

4.5.2. Новые температуры

Аналогичным образом мы проверили устойчивость моделей на новых температурных диапазонах. Указанный диапазон полностью убирался из обучающей выборки и использовался для валидации. Правые и левые границы диапазонов не совпадают, так как разбиение данных происходило по непрерывной шкале, а в таблице указаны реальные минимальное и максимальное значения температуры в тестовой выборке.

Данный эксперимент дал неоднозначные результаты. Как случайный лес, так и линейная регрессия показали слабую точность в некоторых диапазонах, особенно при низких температурах. Возможно, причина этого заключается в том, что самые низкие температуры соответствуют наибольшему числу записей в базе данных. Таким образом, помимо трудностей с обобщением на новый температурный диапазон, модели могли столкнуться с проблемой малого обучающего подмножества. Тем не менее, в целом Табл. 7 показывает, что модели справились удовлетворительно и дают точность, сравнимую с моделью MYS.

4.5.3. Новые давления

В последнем эксперименте мы проверили устойчивость моделей на новых диапазонах давлений. Указанный диапазон полностью убирался из обучающей выборки и использовался для валидации. Правые и левые границы диапазонов не совпадают по той же причине, что и при разбиении по температурам.

Таблица 7

Характеристики ошибок при оценке на новых температурах для различных моделей

| Диапазон Т | MYS | | KNN | | Случайный лес | | Линейная регрессия | |
|-----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | | | | | RMSE | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| 273.15 – 353.6 | 5.03×10^{-5} | 1.30×10^{-5} | 3.74×10^{-4} | 5.48×10^{-5} | 2.61×10^{-4} | 6.86×10^{-5} | 1.87×10^{-4} | 4.61×10^{-5} |
| 358.15 – 436.86 | 3.78×10^{-5} | 1.08×10^{-6} | 1.22×10^{-4} | 2.26×10^{-5} | 2.78×10^{-5} | 1.89×10^{-6} | 3.83×10^{-5} | 6.76×10^{-6} |
| 443.2 – 523.7 | 5.56×10^{-5} | 1.51×10^{-6} | 6.13×10^{-5} | 7.76×10^{-6} | 2.12×10^{-5} | 3.18×10^{-6} | 5.67×10^{-5} | 1.56×10^{-5} |
| 533.1 – 603.2 | 4.19×10^{-5} | 4.06×10^{-7} | 1.89×10^{-5} | 1.84×10^{-6} | 6.70×10^{-6} | 9.62×10^{-7} | 9.36×10^{-5} | 3.60×10^{-5} |
| 612.9 – 683.3 | 3.59×10^{-5} | 2.49×10^{-7} | 8.26×10^{-6} | 1.34×10^{-6} | 4.27×10^{-6} | 9.60×10^{-7} | 2.08×10^{-5} | 4.88×10^{-6} |
| | MAE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| 273.15 – 353.6 | 2.71×10^{-5} | 2.24×10^{-6} | 2.14×10^{-4} | 1.93×10^{-5} | 7.88×10^{-5} | 1.32×10^{-5} | 7.06×10^{-5} | 9.53×10^{-6} |
| 358.15 – 436.86 | 3.33×10^{-5} | 9.23×10^{-7} | 7.90×10^{-5} | 8.03×10^{-6} | 1.71×10^{-5} | 1.36×10^{-6} | 2.28×10^{-5} | 2.75×10^{-6} |
| 443.2 – 523.7 | 5.18×10^{-5} | 1.22×10^{-6} | 4.60×10^{-5} | 4.56×10^{-6} | 1.41×10^{-5} | 2.50×10^{-6} | 2.93×10^{-5} | 5.67×10^{-6} |
| 533.1 – 603.2 | 4.10×10^{-5} | 4.58×10^{-7} | 1.48×10^{-5} | 1.48×10^{-6} | 4.41×10^{-6} | 5.73×10^{-7} | 2.35×10^{-5} | 8.95×10^{-6} |
| 612.9 – 683.3 | 3.56×10^{-5} | 2.70×10^{-7} | 6.59×10^{-6} | 1.22×10^{-6} | 3.60×10^{-6} | 8.69×10^{-7} | 1.75×10^{-5} | 3.50×10^{-6} |
| | MRE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| 273.15 – 353.6 | 0.0435 | 0.0018 | 0.2439 | 0.0072 | 0.0767 | 0.0078 | 0.0618 | 0.0042 |
| 358.15 – 436.86 | 0.0930 | 0.0021 | 0.1640 | 0.0097 | 0.0354 | 0.0025 | 0.0438 | 0.0035 |
| 443.2 – 523.7 | 0.2492 | 0.0022 | 0.1999 | 0.0165 | 0.0606 | 0.0104 | 0.1113 | 0.0154 |
| 533.1 – 603.2 | 0.3692 | 0.0034 | 0.1331 | 0.0121 | 0.0609 | 0.0102 | 0.5507 | 0.2595 |
| 612.9 – 683.3 | 0.4368 | 0.0013 | 0.0813 | 0.0154 | 0.0486 | 0.0123 | 0.2327 | 0.0507 |

Таблица 8

Характеристики ошибок при оценке на новых давлениях для различных моделей

| Диапазон Р | MYS | | KNN | | Случайный лес | | Линейная регрессия | |
|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | | | | | RMSE | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| 100 кПа – 180 кПа | 3.02×10^{-5} | 9.64×10^{-7} | 4.70×10^{-5} | 5.62×10^{-6} | 1.69×10^{-5} | 1.31×10^{-6} | 1.36×10^{-5} | 1.05×10^{-6} |
| 670 кПа – 2.2 МПа | 3.34×10^{-5} | 1.52×10^{-6} | 2.12×10^{-5} | 6.39×10^{-6} | 9.70×10^{-6} | 1.21×10^{-6} | 1.03×10^{-5} | 1.10×10^{-6} |
| 2.3 МПа – 10.4 МПа | 3.47×10^{-5} | 1.64×10^{-6} | 3.29×10^{-5} | 3.32×10^{-6} | 1.36×10^{-5} | 1.94×10^{-6} | 5.32×10^{-5} | 4.11×10^{-5} |
| 11.2 МПа – 51.1 МПа | 3.74×10^{-5} | 3.32×10^{-6} | 8.74×10^{-5} | 1.06×10^{-5} | 1.61×10^{-5} | 2.26×10^{-6} | 1.54×10^{-5} | 2.36×10^{-6} |
| 51.5 МПа – 245.2 МПа | 1.01×10^{-4} | 1.69×10^{-5} | 5.75×10^{-4} | 4.74×10^{-5} | 3.84×10^{-4} | 6.55×10^{-5} | 9.73×10^{-4} | 1.39×10^{-4} |
| | MAE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| 100 кПа – 180 кПа | 2.48×10^{-5} | 1.04×10^{-6} | 3.06×10^{-5} | 3.06×10^{-6} | 1.22×10^{-5} | 9.15×10^{-7} | 9.72×10^{-6} | 6.58×10^{-7} |
| 670 кПа – 2.2 МПа | 2.55×10^{-5} | 1.20×10^{-6} | 1.24×10^{-5} | 1.82×10^{-6} | 7.04×10^{-6} | 7.60×10^{-7} | 7.24×10^{-6} | 4.81×10^{-7} |
| 2.3 МПа – 10.4 МПа | 2.91×10^{-5} | 1.37×10^{-6} | 2.29×10^{-5} | 1.64×10^{-6} | 1.03×10^{-5} | 1.17×10^{-6} | 1.52×10^{-5} | 5.98×10^{-6} |
| 11.2 МПа – 51.1 МПа | 2.55×10^{-5} | 1.32×10^{-6} | 5.18×10^{-5} | 4.23×10^{-6} | 1.11×10^{-5} | 8.77×10^{-7} | 1.01×10^{-5} | 9.71×10^{-7} |
| 51.5 МПа – 245.2 МПа | 5.25×10^{-5} | 5.91×10^{-6} | 3.22×10^{-4} | 2.24×10^{-5} | 1.44×10^{-4} | 2.43×10^{-5} | 4.98×10^{-4} | 6.72×10^{-5} |
| | MRE | | | | | | | |
| | Среднее | Std | Среднее | Std | Среднее | Std | Среднее | Std |
| 100 кПа – 180 кПа | 0.0463 | 0.0021 | 0.0461 | 0.0027 | 0.0215 | 0.0016 | 0.0184 | 0.0011 |
| 670 кПа – 2.2 МПа | 0.0919 | 0.0045 | 0.0350 | 0.0054 | 0.0197 | 0.0017 | 0.0261 | 0.0045 |
| 2.3 МПа – 10.4 МПа | 0.1466 | 0.0059 | 0.0936 | 0.0136 | 0.0585 | 0.0130 | 0.1909 | 0.1480 |
| 11.2 МПа – 51.1 МПа | 0.0551 | 0.0025 | 0.0823 | 0.0035 | 0.0236 | 0.0017 | 0.0196 | 0.0024 |
| 51.5 МПа – 245.2 МПа | 0.0607 | 0.0034 | 0.2441 | 0.0120 | 0.0979 | 0.0065 | 0.5009 | 0.0684 |

Табл. 8 показывает картину, похожую на предыдущие случаи. В некоторых точках модели показывают себя хуже, в некоторых лучше. В общем точность на уровне MYS. Можно отметить самый большой диапазон давлений. На нем все модели показывают результаты значительно хуже, однако модель MYS сохраняет точность в отличие от моделей машинного обучения, компенсируя недостаток точности в других экспериментах.

4.6. Выводы по главе

В данной главе были рассмотрены как аналитические, так и машинные методы моделирования вязкости. Начальный анализ модели MYS подтвердил ее статус надежного базового ориентира: несмотря на простоту, она показывает стабильную точность в широком диапазоне температур и давлений.

зоне условий, особенно вблизи нормальных параметров.

Применение символьной регрессии продемонстрировало, что даже при ограниченных признаках и небольшом размере модели возможно получить интерпретируемое выражение с приемлемой точностью. Однако его универсальность остается ограниченной.

Наибольшее влияние на качество моделей оказала генерация признаков. Разработка и реализация итеративного алгоритма генерации позволили выявить наиболее информативные комбинации параметров, значительно улучшившие точность всех рассмотренных моделей.

Сравнительный анализ методов машинного обучения показал:

- Метод ближайших соседей подтвердил свои ограничения в условиях обобщения и оказался полезен скорее как вспомогательный инструмент.
- Случайный лес показал устойчивость и высокую точность при большем объеме данных, а также оказался лучшим по метрике MRE.
- Линейная регрессия, несмотря на слабые результаты на базовых признаках, в комбинации с расширенными данными продемонстрировала наилучшие метрики. Достигнутая точность показывает потенциал данного метода для применения в прикладных задачах.

Таким образом, результаты экспериментов позволяют с уверенностью говорить о практической применимости методов машинного обучения для предсказания вязкости. Модели продемонстрировали устойчивость даже в условиях, слабо представленных в обучающей выборке.

Наибольшие трудности возникают на точках, полностью выходящих за границы диапазонов обучающих данных – особенно, если значения существенно превышают или, наоборот, находятся ниже известных. Однако в пределах умеренного отклонения от обучающего диапазона модели показывают точность, сопоставимую или превышающую ту, что достигается с помощью модели MYS.

Это создает основу для практического использования таких подходов в инженерных задачах, особенно в тех случаях, когда имеются данные, охватывающие интересующий диапазон условий. Машинное обучение, основанное на корректно отобранных признаках, может стать универсальным инструментом анализа, способным эффективно дополнять или даже заменять традиционные эмпирические корреляции.

Заключение

В данной работе была поставлена цель разработать метод прогнозирования вязкости чистых веществ на основе параметров уравнения состояния CP-PC-SAFT с точностью, превосходящей классическую модель MYS. Для ее достижения были решены задачи, включающие сбор и очистку экспериментальных данных, расчет производных термодинамических величин, генерацию новых признаков, обучение моделей машинного обучения, их сравнение и оценка устойчивости моделей.

На этапе подготовки данных был собран объемный и качественный датасет по насыщенным углеводородам, сформированный на основе базы ThermoML. Хотя в нем не было выявлено явных артефактов или дубликатов, анализ показал выраженное смещение распределений в сторону стандартных лабораторных условий, что было учтено при оценке обобщающих свойств моделей.

Основная часть работы была посвящена обучению и анализу моделей машинного обучения. Проведенное сравнение показало, что:

- a) **случайный лес** продемонстрировал устойчивость и высокую точность даже в случаях большой разницы между обучающим и тренировочным наборами данных. В большинстве тестов именно случайный лес давал наименьшую относительную ошибку, стабильно обходя модель MYS по этой метрике.
- б) **линейная регрессия**, несмотря на ограниченность на базовом наборе данных, достигла наилучших метрик при использовании расширенных признаков. Линейная регрессия стабильно достигает наименьших значений RMSE и MAE, если обучающий набор данных достаточно хорошо описывает тестовую выборку.

Результаты экспериментов позволяют говорить о практической применимости моделей машинного обучения для прогнозирования вязкости на основе параметров уравнения состояния. Особенно важно, что модели демонстрируют устойчивость даже в условиях, слабо представленных в обучающей выборке, а при умеренном отклонении от известного диапазона показывают точность, сравнимую или лучшую, чем MYS.

Таким образом, можно заключить, что машинное обучение, при условии корректного отбора признаков и тщательной подготовки данных, может стать надежным инструментом для оценки вязкости чистых веществ. В целом, проведенный анализ демонстрирует, что интеграция физико-химических параметров с подходами машинного обучения открывает перспективный путь к созданию точных и интерпретируемых моделей для оценки вязкости.

Список литературы

- [1] Malyanah Binti Mohd Taib and J. P. Martin Trusler. «Residual entropy model for predicting the viscosities of dense fluid mixtures.» In: *The Journal of Chemical Physics* 152.16 (2020), p. 164104. DOI: 10.1063/5.0002242. URL: <https://pubs.aip.org/aip/jcp/article/152/16/164104/198785/Residual-entropy-model-for-predicting-the>.
- [2] Ilya Polishuk. «A Modeling Framework for Predicting and Correlating Viscosities of Liquids in Wide Range of Conditions.» In: *Industrial & Engineering Chemistry Research* 54.27 (2015), pp. 6999–7003. DOI: 10.1021/acs.iecr.5b01468. URL: <https://doi.org/10.1021/acs.iecr.5b01468>.
- [3] Vasily Pisarev. *CP_PC_SAFTjl - Julia package for CP-PC-SAFT equation of state*. 2025. URL: https://github.com/vvpisarev/CP_PC_SAFT.jl.
- [4] Ilya Polishuk. «Standardized critical point-based numerical solution of statistical association fluid theory parameters: the perturbed chain-statistical association fluid theory equation of state revisited.» In: *Industrial & Engineering Chemistry Research* 53.36 (2014), pp. 14127–14141.
- [5] Miles Cranmer. *PySR: Fast & Interpretable Symbolic Regression in Python and Julia*. 2023. arXiv: 2305.01582 [cs.LG].
- [6] Sebastian Herrmann and Eckhard Vogel. In: *J. Chem. Eng. Data* 60 (2015), pp. 3703–3720. DOI: 10.1021/acs.jced.5b00654.
- [7] A. Kumagai, D. Tomida, and C. Yokoyama. «Measurements of the Liquid Viscosities of Mixtures of n-Butane, n-Hexane, and n-Octane with Squalane to 30 MPa.» In: *International Journal of Thermophysics* 27 (2006), pp. 376–393. DOI: 10.1007/s10765-006-0053-2.
- [8] Alfredo Pimentel-Rodas, Luis A. Galicia-Luna, and José J. Castro-Arellano. In: *J. Chem. Eng. Data* 62 (2017), pp. 3946–3957. DOI: 10.1021/acs.jced.7b00650.
- [9] Hossein Iloukhani, Mahdi Rezaei-Sameti, and Jalal Basiri-Parsa. In: *The Journal of Chemical Thermodynamics* 38 (2006), pp. 975–982. DOI: <https://doi.org/10.1016/j.jct.2005.10.011>.
- [10] H. Iloukhani and M. Rezaei-Sameti. In: *J. Chem. Eng. Data* 50 (2005), pp. 1928–1931. DOI: 10.1021/je0501944.
- [11] Alfredo Pimentel-Rodas, Luis A. Galicia-Luna, and José J. Castro-Arellano. In: *J. Chem. Eng. Data* 61 (2016), pp. 45–55. DOI: 10.1021/acs.jced.5b00152.
- [12] Waqar Ahmad et al. In: *J. Chem. Eng. Data* 64 (2019), pp. 459–470. DOI: 10.1021/acs.jced.8b00589.
- [13] Pedro Morgado et al. In: *Fluid Phase Equilibria* 358 (2013), pp. 161–165. DOI: <https://doi.org/10.1016/j.fluid.2013.07.060>.
- [14] S.G. Mohammed Hussain et al. In: *Fluid Phase Equilibria* 476 (2018), pp. 139–146. DOI: <https://doi.org/10.1016/j.fluid.2018.08.001>.
- [15] Adel S. Al-Jimaz et al. In: *The Journal of Chemical Thermodynamics* 37 (2005), pp. 631–642. DOI: <https://doi.org/10.1016/j.jct.2004.09.021>.
- [16] J.G. Baragi et al. In: *The Journal of Chemical Thermodynamics* 38 (2006), pp. 75–83. DOI: <https://doi.org/10.1016/j.jct.2005.03.024>.

- [17] A. Rodríguez et al. In: *The Journal of Chemical Thermodynamics* 38 (2006), pp. 505–519. DOI: <https://doi.org/10.1016/j.jct.2005.07.008>.
- [18] Gyan Prakash Dubey, Monika Sharma, and Neelima Dubey. In: *The Journal of Chemical Thermodynamics* 40 (2008), pp. 309–320. DOI: <https://doi.org/10.1016/j.jct.2007.05.016>.
- [19] Gyan P. Dubey and Monika Sharma. In: *The Journal of Chemical Thermodynamics* 40 (2008), pp. 991–1000. DOI: <https://doi.org/10.1016/j.jct.2008.02.005>.
- [20] K.V.N. Suresh Reddy, G. Sankara Reddy, and A. Krishnaiah. In: *Thermochimica Acta* 440 (2006), pp. 43–50. DOI: <https://doi.org/10.1016/j.tca.2005.10.008>.
- [21] M. Domínguez et al. In: *J. Chem. Eng. Data* 48 (2003), pp. 302–307. DOI: [10.1021/je0201141](https://doi.org/10.1021/je0201141).
- [22] A. Rodríguez et al. In: *J. Chem. Eng. Data* 48 (2003), pp. 146–151. DOI: [10.1021/je020131a](https://doi.org/10.1021/je020131a).
- [23] Jyoti N. Nayak, Mrityunjaya I. Aralaguppi, and Tejraj M. Aminabhavi. In: *J. Chem. Eng. Data* 48 (2003), pp. 1152–1156. DOI: [10.1021/je030107c](https://doi.org/10.1021/je030107c).
- [24] Mikhail F. Bolotnikov and Yurij A. Neruchev. In: *J. Chem. Eng. Data* 48 (2003), pp. 739–741. DOI: [10.1021/je0340021](https://doi.org/10.1021/je0340021).
- [25] Ignacio Gascón et al. In: *J. Chem. Eng. Data* 50 (2005), pp. 722–726. DOI: [10.1021/je049576k](https://doi.org/10.1021/je049576k).
- [26] Changsheng Yang, Wei Xu, and Peisheng Ma. In: *J. Chem. Eng. Data* 49 (2004), pp. 1802–1808. DOI: [10.1021/je049777o](https://doi.org/10.1021/je049777o).
- [27] Gyan P. Dubey and Monika Sharma. In: *J. Chem. Eng. Data* 52 (2007), pp. 449–453. DOI: [10.1021/je060389r](https://doi.org/10.1021/je060389r).
- [28] Qinglai Tian and Huizhou Liu. In: *J. Chem. Eng. Data* 52 (2007), pp. 892–897. DOI: [10.1021/je060491o](https://doi.org/10.1021/je060491o).
- [29] Ana Carrasco et al. In: *J. Chem. Eng. Data* 53 (2008), pp. 1223–1227. DOI: [10.1021/je800048f](https://doi.org/10.1021/je800048f).
- [30] Sheng Fang et al. In: *J. Chem. Eng. Data* 53 (2008), pp. 2718–2720. DOI: [10.1021/je8006138](https://doi.org/10.1021/je8006138).
- [31] Manapragada V. Rathnam, Sharad Mankumare, and M. S. S. Kumar. In: *J. Chem. Eng. Data* 55 (2010), pp. 1354–1358. DOI: [10.1021/je9006597](https://doi.org/10.1021/je9006597).
- [32] N. Tripathi. «Densities, Viscosities, and Refractive Indices of Mixtures of Hexane with Cyclohexane, Decane, Hexadecane, and Squalane at 298.15 K.» In: *International Journal of Thermophysics* 26 (2005), pp. 693–703. DOI: [10.1007/s10765-005-5572-8](https://doi.org/10.1007/s10765-005-5572-8).
- [33] Montserrat Domínguez-Pérez et al. «Experimental Study of the Dynamic Viscosity Deviations in the Binary Systems: Hexane + Ethylbenzene, + o-Xylene, + m-Xylene, + p-Xylene at 298.15 K.» In: *International Journal of Thermophysics* 30 (2009), pp. 1197–1201. DOI: [10.1007/s10765-009-0622-2](https://doi.org/10.1007/s10765-009-0622-2).
- [34] Fátima Reyes-García and Gustavo A. Iglesias-Silva. In: *J. Chem. Eng. Data* 62 (2017), pp. 2726–2739. DOI: [10.1021/acs.jced.7b00121](https://doi.org/10.1021/acs.jced.7b00121).
- [35] Claus K. Zéberg-Mikkelsen et al. In: *Fluid Phase Equilibria* 245 (2006), pp. 6–19. DOI: <https://doi.org/10.1016/j.fluid.2006.01.030>.

- [36] Emila M. Živković et al. In: *Fluid Phase Equilibria* 299 (2010), pp. 191–197. DOI: <https://doi.org/10.1016/j.fluid.2010.10.009>.
- [37] Wanchalerm Srirachat et al. In: *Fluid Phase Equilibria* 434 (2017), pp. 117–129. DOI: <https://doi.org/10.1016/j.fluid.2016.11.029>.
- [38] I.M. Abdulagatov and N.D. Azizov. In: *The Journal of Chemical Thermodynamics* 38 (2006), pp. 1402–1415. DOI: <https://doi.org/10.1016/j.jct.2006.01.012>.
- [39] Sheng Fang et al. In: *The Journal of Chemical Thermodynamics* 68 (2014), pp. 281–287. DOI: <https://doi.org/10.1016/j.jct.2013.09.017>.
- [40] Johnny R. Zambrano et al. In: *The Journal of Chemical Thermodynamics* 96 (2016), pp. 104–116. DOI: <https://doi.org/10.1016/j.jct.2015.12.021>.
- [41] Zhuqiang Yang et al. In: *Thermochimica Acta* 617 (2015), pp. 1–7. DOI: <https://doi.org/10.1016/j.tca.2015.08.005>.
- [42] Chao Su et al. In: *Thermochimica Acta* 670 (2018), pp. 211–218. DOI: <https://doi.org/10.1016/j.tca.2018.10.018>.
- [43] Alfonso S. Pensado et al. In: *J. Chem. Eng. Data* 50 (2005), pp. 849–855. DOI: [10.1021/je049662k](https://doi.org/10.1021/je049662k).
- [44] Changsheng Yang, Wei Xu, and Peisheng Ma. In: *J. Chem. Eng. Data* 49 (2004), pp. 1794–1801. DOI: [10.1021/je049776w](https://doi.org/10.1021/je049776w).
- [45] Hong Liu and Lin Zhu. In: *J. Chem. Eng. Data* 59 (2014), pp. 369–375. DOI: [10.1021/je400835u](https://doi.org/10.1021/je400835u).
- [46] Dianne J. Luning Prak, Jim S. Cowart, and Paul C. Trulove. In: *J. Chem. Eng. Data* 59 (2014), pp. 3842–3851. DOI: [10.1021/je5007532](https://doi.org/10.1021/je5007532).
- [47] Diana C. Landaverde-Cortes et al. In: *J. Chem. Eng. Data* 52 (2007), pp. 1226–1232. DOI: [10.1021/je600554h](https://doi.org/10.1021/je600554h).
- [48] Zhi-Fei Wang, Li-Sheng Wang, and Tian-Bo Fan. In: *J. Chem. Eng. Data* 52 (2007), pp. 1866–1871. DOI: [10.1021/je700202h](https://doi.org/10.1021/je700202h).
- [49] Sheng Fang, Chun-Xia Zhao, and Chao-Hong He. In: *J. Chem. Eng. Data* 53 (2008), pp. 2244–2246. DOI: [10.1021/je8003707](https://doi.org/10.1021/je8003707).
- [50] Xi Yang et al. In: *J. Chem. Eng. Data* 55 (2010), pp. 2914–2916. DOI: [10.1021/je900969u](https://doi.org/10.1021/je900969u).
- [51] Hanan E. M. El-Sayed and Abdul-Fattah A. Asfour. «Viscometric and Volumetric Properties of 10 Regular Binary Systems at 308.15 K and 313.15 K.» In: *International Journal of Thermophysics* 30 (2009), pp. 1773–1790. DOI: [10.1007/s10765-009-0667-2](https://doi.org/10.1007/s10765-009-0667-2).
- [52] D. I. Sagdeev et al. «Experimental Study of the Density and Viscosity of

n

- Heptane at Temperatures from 298 K to 470 K and Pressure upto 245 MPa.» In: *International Journal of Thermophysics* 34 (2013), pp. 1–33. DOI: [10.1007/s10765-012-1373-z](https://doi.org/10.1007/s10765-012-1373-z).
- [53] Omar El-Hadad, Ruo Cai, and Abdul-Fattah A. Asfour. «Densities and Kinematic Viscosities of a Quinary Regular Liquid System and Its Five Quaternary Subsystems at 293.15 K and 298.15 K.» In: *International Journal of Thermophysics* 36 (2015), pp. 69–80. DOI: [10.1007/s10765-014-1759-1](https://doi.org/10.1007/s10765-014-1759-1).

- [54] Zhuqiang Yang, Qincheng Bi, and Song Feng. In: *J. Chem. Eng. Data* 61 (2016), pp. 3472–3480. DOI: 10.1021/acs.jced.6b00391.
- [55] B. González, A. Domínguez, and J. Tojo. In: *The Journal of Chemical Thermodynamics* 36 (2004), pp. 267–275. DOI: <https://doi.org/10.1016/j.jct.2003.12.005>.
- [56] Lei Yue et al. In: *The Journal of Chemical Thermodynamics* 81 (2015), pp. 26–33. DOI: <https://doi.org/10.1016/j.jct.2014.09.015>.
- [57] José L. Trenzado et al. In: *J. Chem. Eng. Data* 48 (2003), pp. 1004–1014. DOI: 10.1021/je034017j.
- [58] Begoña González et al. In: *J. Chem. Eng. Data* 49 (2004), pp. 1225–1230. DOI: 10.1021/je034208m.
- [59] Changsheng Yang, Wali Yu, and Peisheng Ma. In: *J. Chem. Eng. Data* 50 (2005), pp. 1197–1203. DOI: 10.1021/je049572f.
- [60] Anthony R. H. Goodwin et al. In: *J. Chem. Eng. Data* 51 (2006), pp. 190–208. DOI: 10.1021/je0503296.
- [61] Xianjie Gong et al. In: *J. Chem. Eng. Data* 57 (2012), pp. 3278–3282. DOI: 10.1021/je300899n.
- [62] A. Estrada-Baltazar, G. A. Iglesias-Silva, and C. Caballero-Cerón. In: *J. Chem. Eng. Data* 58 (2013), pp. 3351–3363. DOI: 10.1021/je4004806.
- [63] Fengjun Yang et al. In: *J. Chem. Eng. Data* 53 (2008), pp. 2237–2240. DOI: 10.1021/je800348s.
- [64] Derek R. Caudwell et al. In: *J. Chem. Eng. Data* 54 (2009), pp. 359–366. DOI: 10.1021/je800417q.
- [65] Yoshiyuki Sato et al. «A Digital Variable-Angle Rolling-Ball Viscometer for Measurement of Viscosity, Density, and Bubble-Point Pressure of CO₂ and Organic Liquid Mixtures.» In: *International Journal of Thermophysics* 31 (2010), pp. 1896–1903. DOI: 10.1007/s10765-008-0542-6.
- [66] Lingling Zhang et al. In: *J. Chem. Eng. Data* 56 (2011), pp. 4268–4273. DOI: 10.1021/je200757a.
- [67] Hai Chi et al. In: *J. Chem. Eng. Data* 58 (2013), pp. 2224–2232. DOI: 10.1021/je400250u.
- [68] Diana C. Landaverde-Cortes et al. In: *J. Chem. Eng. Data* 53 (2008), pp. 288–292. DOI: 10.1021/je700428f.
- [69] Zhaojun Liu, J. P. Martin Trusler, and Qincheng Bi. In: *J. Chem. Eng. Data* 60 (2015), pp. 2363–2370. DOI: 10.1021/acs.jced.5b00270.
- [70] Dianne J. Luning Prak et al. In: *J. Chem. Eng. Data* 62 (2017), pp. 169–187. DOI: 10.1021/acs.jced.6b00542.
- [71] Aravind R. Mahajan and Sunil R. Mirgane. In: *The Journal of Chemical Thermodynamics* 64 (2013), pp. 159–166. DOI: <https://doi.org/10.1016/j.jct.2013.05.014>.
- [72] Seema Kapoor and Virender K. Rattan. In: *J. Chem. Eng. Data* 50 (2005), pp. 1891–1896. DOI: 10.1021/je0501585.
- [73] Xiaomei Qin et al. In: *J. Chem. Eng. Data* 59 (2014), pp. 775–783. DOI: 10.1021/je4008926.

- [74] Zhiping Fang et al. In: *J. Chem. Eng. Data* 53 (2008), pp. 2787–2792. DOI: 10.1021/je800635g.
- [75] Jing Zhao et al. In: *J. Chem. Eng. Data* 62 (2017), pp. 643–652. DOI: 10.1021/acs.jced.6b00688.
- [76] Xinxin Cheng et al. In: *J. Chem. Eng. Data* 62 (2017), pp. 2330–2339. DOI: 10.1021/acs.jced.7b00201.
- [77] Luning et al. In: *J. Chem. Eng. Data* 62 (2017), pp. 3452–3472. DOI: 10.1021/acs.jced.7b00466.
- [78] Song Feng et al. In: *J. Chem. Eng. Data* 63 (2018), pp. 671–678. DOI: 10.1021/acs.jced.7b00866.
- [79] Dianne J. Luning Prak et al. In: *J. Chem. Eng. Data* 63 (2018), pp. 1642–1656. DOI: 10.1021/acs.jced.8b00008.
- [80] Yitong Dai et al. In: *J. Chem. Eng. Data* 63 (2018), pp. 4052–4060. DOI: 10.1021/acs.jced.8b00438.
- [81] Dianne J. Luning Prak et al. In: *J. Chem. Eng. Data* 64 (2019), pp. 1550–1568. DOI: 10.1021/acs.jced.8b01135.
- [82] Dianne J. Luning Prak et al. In: *J. Chem. Eng. Data* 64 (2019), pp. 1725–1745. DOI: 10.1021/acs.jced.8b01233.
- [83] Fernando Czubinski et al. In: *J. Chem. Eng. Data* 64 (2019), pp. 3375–3384. DOI: 10.1021/acs.jced.9b00187.
- [84] Iria Rodríguez-Escontrela et al. In: *Fluid Phase Equilibria* 405 (2015), pp. 124–131. DOI: <https://doi.org/10.1016/j.fluid.2015.07.022>.
- [85] Marek Blahušiak and Štefan Schlosser. In: *The Journal of Chemical Thermodynamics* 72 (2014), pp. 54–64. DOI: <https://doi.org/10.1016/j.jct.2013.12.022>.
- [86] Sara Lago et al. In: *The Journal of Chemical Thermodynamics* 75 (2014), pp. 63–68. DOI: <https://doi.org/10.1016/j.jct.2014.02.012>.
- [87] Mani Lal Singh et al. In: *J. Chem. Eng. Data* 59 (2014), pp. 1130–1139. DOI: 10.1021/je400493x.
- [88] Dianne J. Luning Prak et al. In: *J. Chem. Eng. Data* 59 (2014), pp. 1334–1346. DOI: 10.1021/je5000132.
- [89] Илья Федорович Голубев. *Вязкость газовых смесей*. Рипол Классик, 2013.
- [90] Philip D. Neufeld, A. R. Janzen, and R. A. Aziz. «Empirical Equations to Calculate 16 of the Transport Collision Integrals $\Omega(l, s)^*$ for the Lennard-Jones (12–6) Potential.» In: *The Journal of Chemical Physics* 57.3 (Aug. 1972), pp. 1100–1102. ISSN: 0021-9606. DOI: 10.1063/1.1678363. eprint: https://pubs.aip.org/aip/jcp/article-pdf/57/3/1100/18880310/1100_1_online.pdf. URL: <https://doi.org/10.1063/1.1678363>.
- [91] Stepan Zakharov. *cp_pc_saft: Julia Library for SAFT Equations of State*. 2024. URL: <https://github.com/zmeri/PC-SAFT>.
- [92] Houman B. Rokni et al. «Entropy scaling based viscosity predictions for hydrocarbon mixtures and diesel fuels up to extreme conditions.» In: *Fuel* 241 (2019), pp. 1203–1213. ISSN: 0016-2361. DOI: <https://doi.org/10.1016/j.fuel.2018.12.043>. URL: <https://www.sciencedirect.com/science/article/pii/S0016236118321069>.