

Предсказание температуры плавления органических веществ методами машинного обучения

А. Л. Варламов² и М. А. Орехов¹

¹ Объединенный институт высоких температур РАН, Ижорская ул., 13, стр.2, Москва 125412, Россия

² Московский физико-технический институт (государственный университет), Институтский пер., 9, Долгопрудный 141701, Россия

E-mail: mo2411@mail.ru

Статья поступила в редакцию 12 ноября 2024 г.

Аннотация. Для многих органических веществ известно, что при комнатной температуре (25 °C) они находятся в жидком состоянии, однако точное значение их температуры плавления остается неизвестным. Рассматривается задача прогнозирования температуры плавления для таких веществ, а также более общая задача — предсказание температуры плавления для всех органических соединений. В построении модели используются методы машинного обучения, основанные на анализе молекулярной структуры веществ. Алгоритм, предназначенный для веществ с температурой плавления до 25 °C, демонстрирует среднеквадратичную ошибку, которая вдвое меньше, чем у универсального алгоритма, применимого ко всем органическим соединениям. Кроме того, разработан алгоритм, способный определять агрегатное состояние вещества (жидкое или кристаллическое) при 25 °C, основываясь на структуре молекулы. <https://doi.org/10.33849/2024403>

1. ВВЕДЕНИЕ

Температура плавления органических веществ играет важную роль во многих процессах [1]. В частности, температура плавления позволяет проводить идентификацию веществ и анализ чистоты получаемого в синтезе вещества [2]. Важное значение температура плавления имеет для определения условий безопасного хранения и транспортировки веществ. В рамках задач фармакологии особое значение имеет корреляция между температурой плавления веществ и их растворимостью [3, 4].

Температура плавления T_{melt} органических веществ также важна в задачах разработки электрохимических источников энергии [5]. Для работы таких систем важное значение имеет используемый растворитель. Он должен быть жидким во всем диапазоне рабочих температур системы. Для систем, работающих при комнатной температуре, это означает минимальное требование $T_{melt} < 25$ °C. Обычно требуется температура плавления в существенно более широком диапазоне отрицательных температур [5]. Таким образом, в рамках разработки электрохимических систем возникает отдельная задача предсказания температуры плавления.

Для расчета температуры плавления активно используются различные теоретические методы, включая метод молекулярной динамики [6–8]. Молекулярно-динамическое моделирование затруднено сложностью предсказания кристаллической структуры веществ. Знание кристаллической структуры необходимо для расчета энергии кристалла и, соответственно, равновесной температуры плавления. Подобное моделирование выполняется для небольших наборов различных молекул в силу высокой вычислительной сложности.

Широкое распространение в предсказании температуры плавления получили методы машинного обучения [9–11]. Они позволяют быстро и эффективно проводить предсказания температур плавления для сотен тысяч веществ. Данные методы относятся к классу методов QSPR (quantitative structure–property

relationship). Средняя точность таких методов составляет 30–50 K [9, 10]. Методы машинного обучения применяются и для температуры плавления смеси молекул [12].

Применимость методов машинного обучения обусловлена наличием достаточно больших баз экспериментальных данных по температуре плавления. Размер доступных баз варьируется в диапазоне от 2 000 [13] до 230 000 молекул [11].

Несмотря на большое число работ по предсказанию температуры плавления органических молекул методами машинного обучения, большинство исследований ориентировано на построение моделей для фармакологических молекул с высокой температурой плавления. При этом модели для молекул с низкой температурой плавления отсутствуют.

В данной работе мы проводим построение и анализ модели, разработанной специально для органических веществ с низкой температурой плавления (жидкостей). Показано, что такой подход позволяет существенно повысить точность модели по сравнению с общей задачей рассмотрения всего диапазона температур плавления. Кроме того, построен классификатор, позволяющий с высокой точностью определить — является ли вещество жидкостью при комнатной температуре или нет, что может быть полезно в задачах электрохимии.

При описании фазового перехода могут использоваться термины: температура плавления и температура кристаллизации. В данной работе будет применяться термин температура плавления, исходя из того, что рассматривается конгруэнтный фазовый переход, при котором температуры плавления и кристаллизации совпадают.

В разделе 2 проведено описание методики машинного обучения. В разделе 3 описывается построение стандартного метода предсказания температуры плавления и анализ имеющихся массивов данных по температурам плавления. В разделе 4 описывается построенный метод классификации веществ на жидкости ($T_{melt} < 25$ °C) и твердые тела ($T_{melt} > 25$ °C).

В разделе 5 построена модель предсказания температуры плавления, специализированная для диапазона $T_{melt} < 25\text{ }^{\circ}\text{C}$.

2. МЕТОД РАСЧЕТА

Использованы как существующие базы экспериментальных данных, так и разработанная самостоятельно с использованием базы PubChem [14]. В PubChem указаны не только температуры плавления веществ, но и оценки из литературы, говорящие — является ли данное вещество жидкостью или твердым телом. Предполагается, что эти данные указаны в справочниках для температуры, при которой вещество исследовалось. Обычно предполагается, что это комнатная температура. Выбраны 1500 записей базы PubChem, в которых указано, как в физико-химической литературе классифицируется данное вещество (жидкость/твердое тело) и указана температура плавления. На их основе в разделе 4 показано, что, говоря о том — является ли вещество жидкостью или твердым телом, в большинстве случаев авторы справочников имеют в виду состояние вещества при температуре примерно $T = 25\text{ }^{\circ}\text{C}$. Обучение моделей проведено на выверенной по точности выборке Bradley et al. [15].

Для построения модели использовался метод градиентного бустинга. Градиентный бустинг [16, 17] — это метод ансамблирования [18] моделей машинного обучения, который объединяет множество простых моделей для создания более мощного предсказателя. Ключевая идея — последовательно строить модели, где каждая следующая модель корректирует ошибку имеющейся на данный момент суммарной модели. Самая распространенная версия бустинга использует решающие деревья [19] в качестве базовых моделей. Одна из самых распространенных реализаций алгоритма бустинга, которая получила свою популярность благодаря возможности успешно работать с категориальными признаками, сделана во фреймворке Catboost [20].

В нашей работе мы использовали модель классификации CatBoostClassifier и регрессии CatBoostRegressor. Для задачи классификации фазового состояния вещества в качестве целевой функции выбирается бинарная кросс-энтропия, а для задачи регрессии — среднеквадратичная ошибка.

Модель градиентного бустинга на решающих деревьях [19] обладает большим плюсом: ее результаты хорошо интерпретируются. Интерпретировать результаты можно при помощи нахождения важности признаков (feature importance — fi) [16, 20], которые подаются на вход модели. Один из основных методов нахождения важности признаков основан на том, чтобы определять вклад каждого признака в итоговое предсказание (prediction value change). В библиотеке Catboost итоговое предсказание нормируют от 0 до 1.

Формула для вычисления важности признака P — fi_P , выглядит так:

$$fi_P = \sum_{\text{деревья, листья}} ((v_1 - \text{avr})^2 \cdot c_1 + (v_2 - \text{avr})^2 \cdot c_2),$$

где суммирование проводится по всем деревьям и всем терминальным вершинам дерева, в разделении которых участвует признак P . Параметры c_1 и c_2 представляют

число объектов в левом и правом листьях дерева соответственно, v_1 и v_2 соответствуют предсказаниям цели (метка или значение функции) для объектов, попавших в эти листья, avr — это взвешенное среднее предсказанных значений из левого и правого листьев, вычисляемое по формуле:

$$\text{avr} = \frac{v_1 \cdot c_1 + v_2 \cdot c_2}{c_1 + c_2}.$$

Также в библиотеке Catboost важность признаков нормируется на единицу.

Другая, использованная нами модель, основанная на ансамблировании моделей машинного обучения — случайный лес (Random Forest) [21]. Случайный лес — это бэггинг [17, 18] моделей решающих деревьев [19] с небольшой глубиной. Бэггинг основан на усреднении предсказания нескольких моделей, обученных на бутстрепных выборках, полученных из исходной. В нашей работе мы использовали имплементацию решающего леса из библиотеки Scikit-learn [22] для Python.

В построении моделей использованы 93 молекулярных признака из библиотеки RDKit [23]. Наибольшую значимость имеют признаки: NumHBD — число доноров водородной связи, NumHeavyAtoms — число тяжелых атомов, MolWt — вес молекулы, LogP — расчетное значение липофильности, TPSA — суммарная площадь поверхности полярных атомов, NumRotatableBonds — число химических связей с торсионными углами, FractionCSP3 — доля sp^3 углеродов, NumAromaticRings — число ароматических колец, FractionRotatableBonds — доля химических связей с торсионными углами, NumRings — число колец в молекуле, NumHeteroatoms — число гетероатомов, Chi(0/1/2/3/4)v — индексы чи [24]. Индексы чи ранга i ($i = 0, 1, 2, 3, 4$) [24] соответствуют сумме попарных произведений атомных номеров атомов во всех связанных подграфах молекулярного графа с числом ребер i . Вершинам молекулярного графа соответствуют атомы молекулы, а ребрам графа — химические связи между этими атомами.

3. ПРЕДСКАЗАНИЕ ПЛАВЛЕНИЯ В ШИРОКОМ ДИАПАЗОНЕ ТЕМПЕРАТУР

Построена модель предсказания температуры плавления органических веществ в широком диапазоне температур. Данная модель незначительно улучшает результаты существующих аналогичных моделей и служит для подтверждения методологии и сравнения с более точными моделями в разделах 4, 5.

Модель обучена на выборке Bradley et al. [15]. Размер тестовой выборки составил 20% от размера всей выборки. Общий размер выборки 25 000 молекул. Среднее значение температуры плавления в выборке $+90.6\text{ }^{\circ}\text{C}$. Диапазон температур от $-199\text{ }^{\circ}\text{C}$ до $+562\text{ }^{\circ}\text{C}$.

Результаты в сравнении с работами других авторов представлены в таблице 1. Видно, что разработанная модель незначительно опережает по точности работы других авторов. Это подтверждает корректность методологии обучения модели. Кроме того, одинаковая точность в предлагаемой модели и в работах других авторов может свидетельствовать, что точность ограничена не свойствами модели, а свойствами обучающей

Таблица 1. Точность модели, предложенной в настоящей работе, и моделей из литературы. R^2 — коэффициент детерминации.

Работа	среднеквадратичная ошибка ($^{\circ}\text{C}$)	R^2
Эта работа	36.26	0.84
Tetko et al. [11]	36.8	-
Weiming et al. [9]	36.88	0.83

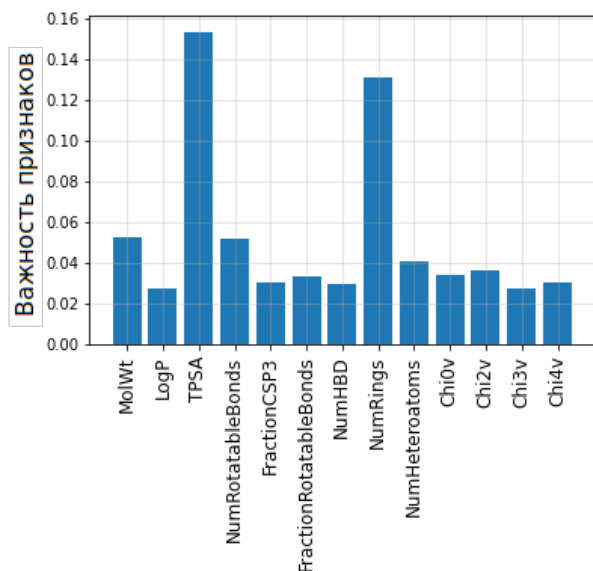


Рисунок 1. Оценка значимости признаков в задаче предсказания температуры плавления во всем температурном диапазоне (от -199°C до $+562^{\circ}\text{C}$). Показаны 13 наиболее значимых (значимость > 0.02) признаков из 93.

выборки, которая не является достаточной для получения более высокой точности.

Результаты анализа значимости признаков показаны на рисунке 1. Проведенный анализ показал наибольшую значимость признаков TPSA и NumRings. TPSA (topological polar surface area) — суммарная площадь поверхности полярных атомов (кислород/азот).

4. КЛАССИФИКАТОР ФАЗОВОГО СОСТОЯНИЯ ПРИ 25°C

Разработан метод классификации веществ на жидкости и твердые тела. Жидкостями считаются вещества с температурой плавления $T_{melt} < 25^{\circ}\text{C}$. Твердыми телами считаются вещества с температурой плавления $T_{melt} > 25^{\circ}\text{C}$.

Очевидным вариантом такого классификатора мог бы являться подход, при котором вначале определяется T_{melt} (моделью из пункта 3) и затем жидкости делятся на основе предсказанной температуры (простым сравнением с $T_{melt} = 25^{\circ}\text{C}$). Вторым вариантом может быть построение модели, которая возвращает не температуру плавления, а только два варианта значений: 0 — если $T_{melt} < 25^{\circ}\text{C}$ и 1 — если $T_{melt} > 25^{\circ}\text{C}$. Назовем такой метод прямым классификатором.

Проведено сравнение построенного прямого классификатора с алгоритмом регрессии, который определяет

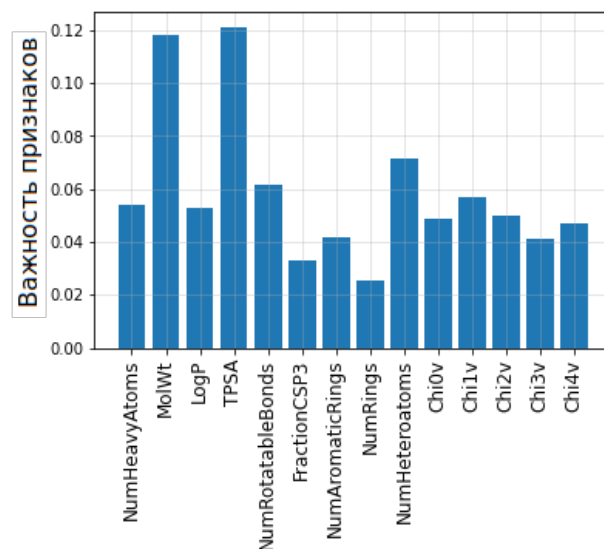


Рисунок 2. Оценка значимости признаков в задаче предсказания температуры плавления при $T_{melt} < 25^{\circ}\text{C}$. Показаны 14 наиболее значимых (значимость > 0.02) признаков из 93.

температуру плавления, а затем эта температура сравнивается с величиной $T_{melt} = 25^{\circ}\text{C}$. Показано, что построенный прямой классификатор существенно точнее такого метода.

Количество твердых тел, некорректно классифицированных как жидкости, составляет 4% для предлагаемого прямого классификатора и 8% для классификатора на основе температуры. Коэффициент детерминации R^2 составляет 0.8 и 0.7 соответственно. Таким образом, в предлагаемом методе в 2 раза снижается количество ложно-положительных результатов по предсказанию жидкости по сравнению с классификатором на основе температуры.

Повышение точности при использовании специализированного классификатора связано с тем, что алгоритм регрессии температуры минимизирует ошибку в широком температурном диапазоне. При этом для классификации более важное значение имеет именно предсказание граничных молекул с температурой плавления близкой к $T_{melt} = 25^{\circ}\text{C}$. Вещества с очень высокой и очень низкой температурой плавления достаточно просто классифицировать и поэтому предлагаемый классификатор не фокусируется на предсказании их свойств, в отличие от регрессионного алгоритма.

Выбор температуры $T_{melt} = 25^{\circ}\text{C}$ в качестве температуры разделения веществ на жидкости и твердые тела исходит не только из общих соображений, но и из классификации веществ, распространенной в научной литературе. Для анализа массива литературы использовалась база молекул PubChem [14]. По данным PubChem из всех веществ, которые отмечены в литературе как жидкости, 82% имеют температуру плавления менее 25°C . При этом 99% веществ, отмеченных как твердые тела, имеют температуру плавления более 25°C . При критерии $> 35^{\circ}\text{C}$ число некорректно определенных твердых тел становится равным 4% (вместо 1% при 25°C).

5. ПРЕДСКАЗАНИЕ ТЕМПЕРАТУРЫ ПЛАВЛЕНИЯ ДЛЯ ВЕЩЕСТВ С ТЕМПЕРАТУРОЙ ПЛАВЛЕНИЯ ДО 25 °C

Основным результатом данного раздела является построение модели предсказания температуры плавления, специализированной для веществ с $T_{melt} < 25$ °C. Показано, что априорное знание о жидком состоянии вещества при комнатной температуре ($T_{melt} < 25$ °C) позволяет существенно повысить точность предсказания температуры плавления.

Увеличение точности связано не только с уменьшением разброса точек в тестовой выборке. Улучшение точности модели, обученной на всем диапазоне температур (раздел 3), при тестировании на ограниченном диапазоне приводит к среднеквадратичной ошибке $RMSE = 27.6$ °C вместо $RMSE = 36.26$ °C. Таким образом повышение точности составляет 25%.

При этом точность метода, обученного и примененного для данных при $T_{melt} < 25$ °C, составляет $RMSE = 18.8$ °C. Таким образом, снижение погрешности по сравнению с универсальной моделью, примененной к узкому диапазону температур, составляет 32%. Относительно общего метода, примененного к широкому диапазону температур, достигается снижение погрешности в два раза.

Проведена оценка значимости признаков для данной задачи. Результаты приведены на рисунке 2. Данные результаты уместно рассматривать в сравнении со значимостью признаков для решения общей задачи в широком температурном диапазоне (рисунок 1). Видно, что модель, специализированная для жидкостей, имеет меньший разброс значимости признаков. Кроме того, большое значение начинает играть вес молекулы.

Визуализация сравнения точности методов, обученных на всем температурном диапазоне и только на жидкостях, представлена на рисунке 3. Видно, что разброс точек предлагаемого метода существенно меньше разброса точек универсального метода.

6. ВЫВОДЫ

На основе имеющегося массива экспериментальных данных построены три модели машинного обучения для температуры плавления веществ.

- Модель 1: предсказание температуры плавления веществ в широком (−199 °C +562 °C) диапазоне температур. Результаты данной модели хорошо совпадают с данными литературы. Это подтверждает корректность методологии и позволяет использовать эту модель для сравнения.
- Модель 2: классификация веществ на жидкости ($T_{melt} < 25$ °C) и твердые тела ($T_{melt} > 25$ °C). Показано, что разработанный подход является существенно более точным (в 2 раза меньше твердых веществ, ошибочно определенных как жидкости) по сравнению со стандартным предсказанием температуры плавления (моделью 1) и сравнением полученной величины с $T_{melt} = 25$ °C.
- Модель 3: предсказание температуры плавления в диапазоне температур $T_{melt} < 25$ °C. Данная модель обладает существенно более высокой (на 32%) точностью по сравнению с применением модели 1 к данному диапазону температур.

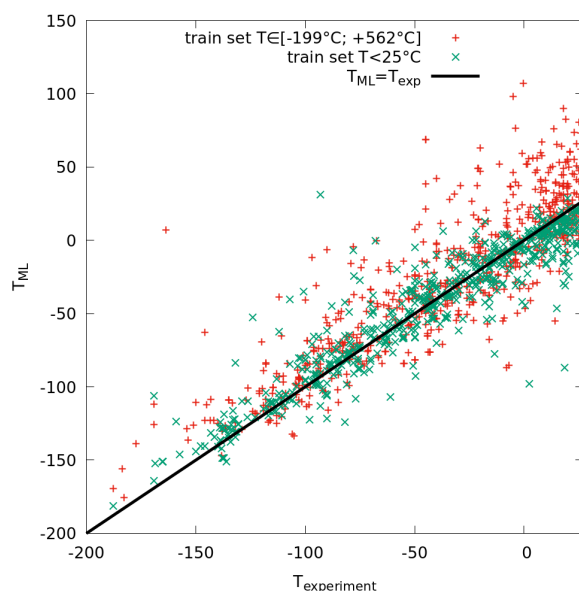


Рисунок 3. Сравнение экспериментальных данных T_{exp} и данных модели машинного обучения T_{ML} . Красные точки — температура плавления, полученная методом, обученным на широком диапазоне температур (раздел 3). Зеленые точки — методом, обученным специализированно на диапазоне температур $T_{melt} < 25$ °C. Диагональная прямая соответствует идеальному согласию предсказаний модели и экспериментальных данных.

Показано, что специализация модели для работы с жидкостями позволяет существенно повысить ее точность.

БЛАГОДАРНОСТИ

Работа была поддержана Министерством образования и науки Российской Федерации (Государственное задание № 075-00270-24-00).

СПИСОК ЛИТЕРАТУРЫ

1. Atkins P and Paula J 2006 *Physical chemistry* (Oxford University Press)
2. Jain A, Yang G and Yalkowsky S H 2004 *Ind. Eng. Chem. Res.* **43** 7618–7621
3. Ali J, Camilleri P, Brown M B, Hutt A J and Kirton S B 2012 *J. Chem. Inf. Model.* **52** 420–428
4. Yalkowsky S H and Valvani S C 1980 *J. Pharm. Sci.* **69** 912–922
5. Xu J et al 2023 *Nature* **614** 694
6. Liu Y, Lai W, Yu T, Ma Y, Guo W and Ge Z 2019 *ACS Omega* **4** 4320–4324
7. Zhang Y and Maginn E J 2013 *J. Chem. Theory Comput.* **9** 1592–1599
8. Zhang Y and Maginn E J 2012 *J. Chem. Phys.* **136** 144116–144128
9. Mi W, Chen H, Zhu D, Zhang T and Qian F 2021 *Chem. Commun.* **57** 2633–2636
10. Galeazzo T and Shiraiwa M 2022 *Environ. Sci.: Atmos.* 362–374
11. Tetko I V, M Lowe D and Williams A J 2016 *J. Cheminf.* **8** 1–18
12. Lavrinenko A K, Chernyshov I Y and Pidko E A 2023 *ACS Sustainable Chem. Eng.* **11** 15492–15502

13. Karthikeyan M, Glen R C and Bender A 2005 *J. Chem. Inf. Model.* **45** 581–590
14. Kim S *et al* 2023 *Nucleic Acids Res.* **51** D1373–D1380
15. Bradley J C, Williams A and Lang A 2014 Jean-Claude Bradley Open Melting Point Dataset URL https://figshare.com/articles/dataset/Jean_Claude_Bradley_Open_Melting_Point_Dataset/1031637
16. Friedman J H 2001 *Ann. Stat.* **29** 1189–1232
17. Кугаевских А В, Муромцев Д И и Кирсанова О В 2022 *Классические методы машинного обучения* (Санкт-Петербург: Университет ИТМО)
18. Breiman L 1996 *Mach. Learn.* **24** 123–140
19. Breiman L, Friedman J, Olshen R and Stone C 1984 *Classification and Regression Trees* (Belmont, CA: Wadsworth International Group)
20. Dorogush A V, Ershova M and Gulin A 2018 *arXiv preprint arXiv:1810.11363*
21. Breiman L 2001 *Mach. Learn.* **45** 5–32
22. Pedregosa F *et al* 2011 *J. Mach. Learn. Res.* **12** 2825–2830
23. Landrum G *et al* 2006 RDKit: Open-source cheminformatics
24. Hall L H and Kier L B 1991 *Rev. Comput. Chem.* **2** 367–422