МГУ имени М. В. Ломоносова,
химический факультет,
кафедра химической энзимологии

Курсовая работа
по аналитической химии

# Интерпретации ИК-спектров лекарственно-подобных молекул в мицеллярных системах методами хемоинформатики

Выполнил:
студент 210 группы
Малашкеевич Станислав Михайлович

Научные руководители:
- асп. 2 г/о, м.н.с. Щеголев В. О.
- д.х.н., проф. Кудряшова Е. В.

Москва, 2024

# Оглавление

# 1 Введение

Современная фармакология и химия во многом полагаются на спектральные методы анализа, один из таких — это инфракрасная спектроскопия. Однако, интерпретация ИК-спектров, особенно в сложных системах, таких как мицеллы, может вызвать затруднения. Из-за разнообразного микроокружения. В свою очередь мицеллы активно используются в фармакологии для доставки лекарств, так как они могут инкапсулировать лекарственные вещества, улучшая их доставку. Понимание взаимодействия лекарственных молекул с мицеллами на молекулярном уровне необходимо для оптимизации повышения эффективности лекарственных препаратов.

Методы хемоинформатики позволяют не только избежать трудностей, которые возникают при проведении эксперимента, начиная от несоблюдения методики и заканчивая дороговизной реагентов, но и позволяют значительно ускорить процесс подбора молекул-кандидатов при разработке лекарственных средств.

Целью работы является исследование возможностей интерпретации ИК-спектров лекарственно-подобных молекул в мицеллярных системах методами хемоинформатики, путем построения различных типов моделей способных прогнозировать $\log P$ и сопоставление результата их работы с экспериментальными данными.

# 2 Обзор литературы

## 2.1 ИК спектроскопия

Инфракрасная спектроскопия (ИК-спектроскопия) — это физико-химический метод анализа, основанный на исследовании поглощения инфракрасного излучения веществами. В ИК-спектроскопии часто используются два основных параметра — поглощение ($A$) и пропускание ($T$).

Пропускание $T$ определяется как отношение интенсивности света, прошедшего через вещество $I$, к исходной интенсивности $I_0$:

$$T = \frac{I}{I_0}. \tag{1}$$

Поглощение $A$ выражается через пропускание с помощью логарифмирования:

$$A = -\lg T, \tag{2}$$

Экспериментально поглощение определяют согласно закону Бугера-Ламберта-Бера, который связывает поглощение $A$ с концентрацией $c$ и длинной пути света $l$ через молярный коэффициент поглощения $\varepsilon$:

$$A = c\varepsilon l. \tag{3}$$

### 2.1.1 Теоретические основы метода

В простейшем варианте рассматривается двухатомная молекула в виде гармонического осциллятора. При изменения длинны связи от равновесного положения $r_0$ на величину $\delta = r - r_0$, потенциальная энергия $U$ описывается законом Гука:

$$U = K\frac{(r - r_0)^2}{2}, \tag{4}$$

где $K$ — константа упругости связи.

Полная энергия молекулы включает в себя энергию электронов $E_e$, колебательную энергию $E_{\text{кол.}}$, вращательную $E_{\text{вр.}}$ и энергию поступательного движения $E_{\text{поступ.}}$[1]:

$$E = E_e + E_{\text{кол.}} + E_{\text{вр.}} + E_{\text{поступ.}}, \tag{5}$$

причем $E_e < E_{\text{кол.}} < E_{\text{вр.}}$. ИК-излучение взаимодействует с колебательной составляющей, и излучение должно обладать соответствующей энергией достаточной только для перехода с одного колебательного уровня на другой, которые отстают друг от дру-

га на равную по энергии величину $E_0$. Она описывается частотой $\nu_0$, которая зависит от приведенной массы $\mu = \frac{m_1 m_2}{m_1+m_2}$ и константы упругости $K$:

$$\nu_0 = \frac{1}{2\pi}\sqrt{\frac{K}{\mu}}. \tag{6}$$

При переходе к многоатомной молекуле, её рассматривают, как систему осцилляторов с колебательными энергиями $E_i$. Полная колебательная энергия:

$$E = \sum_{i=1}^{N} E_i. \tag{7}$$

Однако в уравнении (7) не учитывается взаимодействия между осцилляторами. Кроме того, (7) справедливо только для гармонических колебаний. Реальные колебания зачастую ангармоничные и порождают обертоны[1], но в колебательном спектре они имеют низкую интенсивность и существенного влияния не оказывают.

Возможен случай, когда при колебании амплитуда колебаний ядер одной из связей значительно превосходит амплитуды колебаний всех остальных ядер молекулы. Тогда частоту данного колебания условно приписывают колебанию именно этой связи.

В некоторых случаях амплитуда колебаний ядер одной из связей в молекуле значительно превышает амплитуды колебаний остальных ядер. В таком случае соответствующая частота колебания может быть приписана именно этой связи. Если частота, соответствующая колебанию определенной связи, сохраняется в спектрах различных молекул, содержащих эту связь, то такую частоту называют *характеристической*. Например, для $C-H$ группы характеристическая частота 2975-2950 см$^{-1}$, для $O-H$ — 3670-3589 см$^{-1}$, а для альдегидной группы ($-CH_2-CHO$) — 1740-1720 см$^{-1}$ [2] .

## 2.2 Инфракрасные спектры

При пропускании инфракрасного излучения через образец происходит возбуждение колебательных степеней свободы молекул или их отдельных фрагментов. В результате наблюдается ослабление интенсивности излучения, прошедшего через образец, при тех длинах волн, энергия которых соответствует энергиям возбуждения колебаний в

[1] полоса обертонов — это спектральная полоса, которая возникает в колебательном спектре молекулы, когда совершается переход из основного состояния в любое помимо первого

изучаемых молекулах. Однако, в ИК спектрах проявляются только те колебания, которые изменяют дипольный момент молекулы.

Спектр в ИК-спектроскопии (ИК спектр) — это зависимость интенсивности не поглощенного ИК излучения от его частоты. ИК спектр содержит полосы поглощения, по положению и относительной интенсивности которых судят о наличии определенных типов связей и функциональных групп в молекуле вещества.

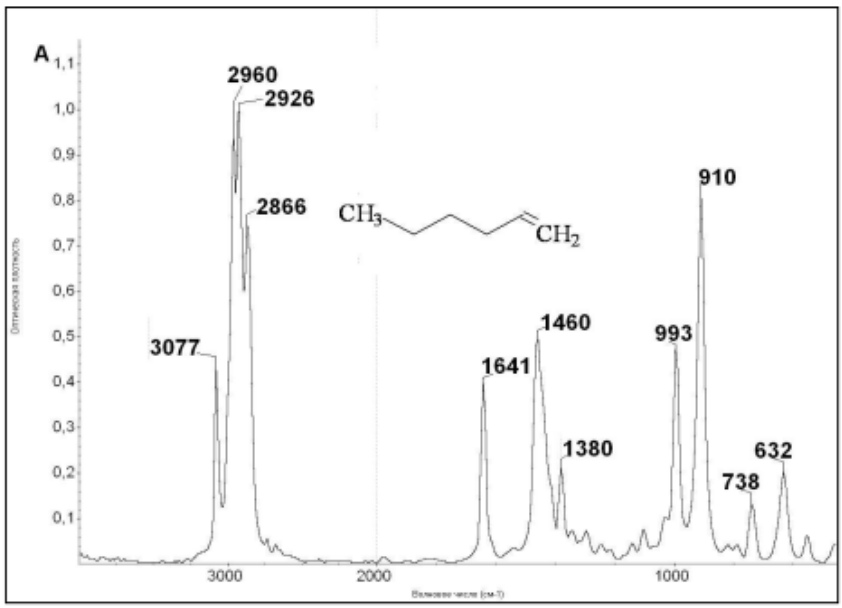


Рис. 1. ИК спектр линейного алкена. [2]

*Благодаря возможности идентификации отдельных функциональных групп и связей с помощью ИК спектроскопии, можно идентифицировать промежуточные и конечные продукты реакции.* Также следует отметить высокую специфичность метода и возможность проводить анализ твердых, жидких и газообразных веществ.

К ограничениям можно отнести неэффективность при анализе веществ, не обладающих дипольным моментом (гомоядерные двухатомные молекулы), а также затрудненность анализа водных растворов из-за поглощения водой инфракрасного излучения.

## 2.3 Коэффициент распределения ($\log D$ и $\log P$)

Липофильность вещества — характеристика, важная при разработке лекарственных средств, т.к. распределение вещества по фазам зависит от степени его гидрофобности.

В русскоязычной литературе под «коэффициентом распределения» понимается и *Distribution coefficient* (коэффициент распределения), и *Partition coefficient* (коэффициент разделения). Т.к. коэффициент *разделения* ($\alpha$) показывает отношение коэффициентов распределения двух различных веществ: $\alpha = \frac{D(A)}{D(B)}$[3].

$\log D$ (логарифм коэффициента распределения) определяется как логарифм отношения сумм равновесных концентраций $i$-ых форм вещества $X$ в *н*-октаноле $\sum_i [X]_i^{\text{окт.}}$ и воде $\sum_i [X]_i^{\text{водн.}}$ :

$$\log D = \log \frac{\sum_i [X]_i^{\text{окт.}}}{\sum_i [X]_i^{\text{водн.}}}. \tag{8}$$

Для ионизируемой молекулы соотношение равновесных концентрации меняются в зависимости от pH среды, как и значение $\log D$. Поэтому используется характеристика вещества — $\log P$ (коэффициент разделения):

$$\log P = \log \frac{[X^N_{\text{окт.}}]}{[X^N_{\text{водн.}}]}, \tag{9}$$

где $[X^N_{\text{окт.}}]$и$[X^N_{\text{водн.}}]$ — равновесные концентрации нейтрально формы в *н*-октаноле и воде. Этот параметр показывает отношение равновесных концентраций вещества в *н*-октаноле и воде.

Однако, на практике от табличного $\log P$ переходят к $\log D$, т.к. многие молекулы при нормальном pH организма находятся в разнообразных формах, которые в совокупности в $\log P$ не учитываются.

$$\log D = \log P - pK_a + pH, \quad pK_a - pH > 1, \tag{10}$$

и для кислот:

$$\log D = \log P - pK_a - pH, \quad pH - pK_a > 1. \tag{11}$$

### 2.3.1 Методы определения

Прямой метод определения $\log P$ заключается в растворении навески вещества в колбе с *н*-октанолом и водой, колбу встряхивают и после установления равновесия определяют концентрацию вещества в каждой фазе.

Экспериментальное определение сопряжено с рядом трудностей. Например, сложно определить концентраций разных форм , т.к. они имеют похожие спектры, что мешает

провести количественный анализ. Особенное это характерно для лекарственно-подобных молекул[2], которые содержат множество нетривиальных структурных фрагментов. Качественные модели предсказания $\log P$ для простых молекул, показывают большую ошибку на лекарственно-подобных. Поэтому важно быстрый и надежный метод определения $\log P$. Например, с помощью высокоэффективной жидкостной хроматографии (ВЭЖХ) с последующей детекцией УФ спектрометрией или масс-спектроскопией [4]. Изначально определяется коэффициентом удерживания вещества $k$ через фактор удерживания $k'$:

$$k' = k - 1 = \frac{t_r - t_0}{t_0} - 1 \tag{12}$$

или через время удерживания $t_r$:

$$k = \frac{t_r - t_d - \left(\frac{V_{\text{внекол.}}}{F}\right)}{t_0 - t_d - \left(\frac{V_{\text{внекол.}}}{F}\right)} - 1, \tag{13}$$

где $t_d$ — время задержки инъекции, $V_s$ — внеколоночный объем: капилляров между детектором и колонкой, между колонкой и устройством ввода, $F$ — скорость потока. По зависимости $K$ от подвижной и неподвижной фазы, можно найти $\log P$ по уравнению [5]:

$$\log P = a \cdot \log k + b \tag{14}$$

Здесь $a, b$ — эмпирические константы.

Один из первых способов теоретического определения основан на предположении об аддитивности $\log P$:

$$P = \sum_i n_i a_i, \tag{15}$$

где $a_i$ — это теоретически оцененный вклад в $\log P$ $i$-ым типом атома или фрагментом, а $a_i$ — количество атомов/фрагментов типа $i$. Т.е. есть таблица, с численный значением $\log P$ для каждого атома в молекуле с определенным окружением.

[2]Определить является вещество лекарство-подобным или нет можно по правилам Липински:
- не более 5 доноров водородной связи
- не более 10 акцепторов водородной связи
- молекулярная масса меньше 500 Дальтон
- коэффициент разделения не больше 5

## 2.4 Модели машинного обучения

### 2.4.1 Математический формализм

Машинное обучение — это область прикладной статистики, в которой применяются методы компьютерного анализа и обработки данных для обнаружение закономерностей в данных, с последующим использованием этих закономерностей для прогнозирования, например, для оценки значения физико-химических свойств молекулы.

Модель в контексте машинного обучения представляет собой математическое выражение или алгоритм, способный решать задачу аппроксимации искомой зависимости. Конкретный вид модели определяется не только типом алгоритма, но также и набором его параметров $\vec{p} = (p_1, \cdots, p_m)$, которые необходимо настраивать для нахождения достаточно точной аппроксимации на известных данных.

Данные, используемые для обучения модели, представлены в виде набора чисел, а более формально — в виде векторов $\vec{x} = (x_1, x_2, \cdots, x_n)$, где каждая компонента $x_i$ — это конкретное значение некоторой переменной $f_i$, от которой зависит искомая зависимость:

$$y = M\Big(\vec{f}, \vec{p}\Big), \tag{16}$$

где $M$ — модель, $y$ — целевая переменная, или искомая зависимость; $\vec{f} = (f_1, \cdots, f_k)$ — переменные, объясняющие искомую зависимость, или *признаки* (англ. *features*); $\vec{p} = (p_1, \cdots, p_m)$ — набор параметров модели. Так, например, атомная масса элемента зависит от двух признаков: количества протонов $f_p$ и количества нейтронов $f_n$ в ядре: для атома $^7$Li атомная масса равна:

$$\underbrace{7.016}_{y} = \underbrace{1.007}_{m_{p^+}} \cdot \big[f_p = 3\big] + \underbrace{1.009}_{m_{n^0}} \cdot \big[f_p = 4\big] + \varepsilon, \tag{17}$$

где $\varepsilon$ — случайная ошибка модели, возникающая в следствие дефекта массы. Отличие между $\vec{f}$ и $\vec{x}$ заключается в том, что $f_i$ — представляет собой абстрактную величину, в то время как $x_i$ — это её конкретное значение, получаемое, например, из эксперимента[3].

[3] в терминах случайных величин: $x_i \sim f_i$, т.е. $x_i$ имеет распределение $f_i$

Обучение модели — настройка её параметров на обучающих данных, в процессе обучения увеличивается точность предсказаний модели.

Вся совокупность обучающих данных называется *обучающей выборкой* и организуется в виде *матрицы признаков* $X$, и вектора целевых значений $\vec{y}$:

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \tag{18}$$

что удобно в вычислительном плане и для компактного представления данных.

Каждый столбец матрицы $X$ соответствует объясняющей переменной $f_j = \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$, каждая строка соответствует набору всех признаков для одного наблюдения $\vec{x}_i = \left(x_{i,1}, \cdots, x_{i,k}\right)$, например, для одной молекулы $\vec{x}_i$ — молярной массе, количеству гидроксильных групп и т.д. Вектор целевых значений содержит в себе все известные значения искомой зависимости, например, экспериментальное измеренные $\log P$.

#### 2.4.1.1 Параметры и гиперпараметры

Обучение состоит в настройке параметров модели $\vec{p}$. Контроль обучения — это процесс оценки качества модели на известных примерах из обучающей выборки. Это необходимо для оценки эффективности модели и для принятия решений о стратегии обучения.

Модели могут включать в себя параметры, которые не могут быть настроены вместе с другими в процессе обучения, их называют гиперпараметрами и они устанавливаются до начала обучения модели. Например, наличие свободного члена $\beta_0$ в простой линейной модели можно считать гиперпараметром, однако само значение $\beta_0$ настраивается в процессе обучения.

#### 2.4.1.2 Обучающая, валидационная и контрольная выборки

Для эффективной оценки качества модели и настройки её параметров данные разбивают на обучающую, валидационную и контрольную (тестовую) выборки.

$$X = X_{\text{обуч}} \sqcup X_{\text{вал}} \sqcup X_{\text{контр}}, \tag{19}$$

где знаком «$\sqcup$» обозначено объединение двух не пересекающихся наборов (множеств).

1. На $X_{\text{обуч}}$ настраиваются основные параметры модели, например, коэффициенты в линейной регрессии.
2. На $X_{\text{обуч}}$ оценивается качество модели[4], а также могут оцениваться гиперпараметры, например, можно сравнить между собой две линейные модели со свободным членом и без.
3. Контрольная выборка ($X_{\text{контр}}$) служит для конечной оценки качества после всех этапов настройки параметров и гиперпараметров, поскольку эти данные никогда не использовались для настройки модели, на них можно получить объективную оценку результата.

#### 2.4.1.3 Переобучение

Переобучение — это общее явление, когда модель машинного обучения слишком агрессивно подстраивается под обучающие данные и обладает плохой обобщающей способностью на новых, ранее не виденных данных. Это может произойти из-за слишком сложной модели, включающей в себя неоправданно большое число параметров, или недостаточного размера обучающей выборки.

Для борьбы с переобучением используются разные стратегии:

- увеличивается объем обучающих данных;
- применяется метод *регуляризации* т.е. наложение ограничения на параметры модели, например, добавляются штрафы за слишком большие или малые значения параметров;
- проводится отбор признаков — выбор наиболее значимых и полезных характеристик из набора данных, что позволяет упростить модель и повысить точность прогнозирования. Это особенно важно, когда набор данных содержит большое количество признаков, многие из которых могут быть избыточными или даже шумом.

### 2.4.2 Векторное представление молекул

Алгоритмические методы работы с данными (в частности машинное обучение) требуют, чтобы исходная информация была представлена в единообразном виде — в виде векторов и матриц. С другой стороны, молекулы существуют во множестве конформа-

[4]для оценки качества модели можно использовать, например, $R^2$, о чём будет сказано далее

ций, одно вещество может описываться равновесной смесью изомеров (циклическая и линейная глюкоза), их свойства зависят от внешних условий (давления, температуры) и т.д. Из-за этих особенностей формальные представления молекул в виде набора чисел (векторов) всегда опускают большу́ю часть их свойств.

На практике не существует одного метода представления молекул; качество методов машинного обучения зависит, в частности, от выбора способа формального представления молекулы. Отдельной задачей является выбор метода векторизации молекул, который сохраняет свойства наиболее важные для работы выбранного алгоритма.

Самым простым представлением молекулы можно считать набор её физически измеряемых свойств, например, для линейных алканов — число атомов углерода $n$ и температуры кипения $T_{\text{кип}}$:

$$\text{н-C}_n\text{H}_{2n} \quad \Leftrightarrow \quad (n, T_{\text{кип}}) \tag{20}$$

С описанным представлением можно обучить линейную модель $T_{\text{кип}} = a \cdot n + b$, обладающую предсказательной силой (Рис. 2).

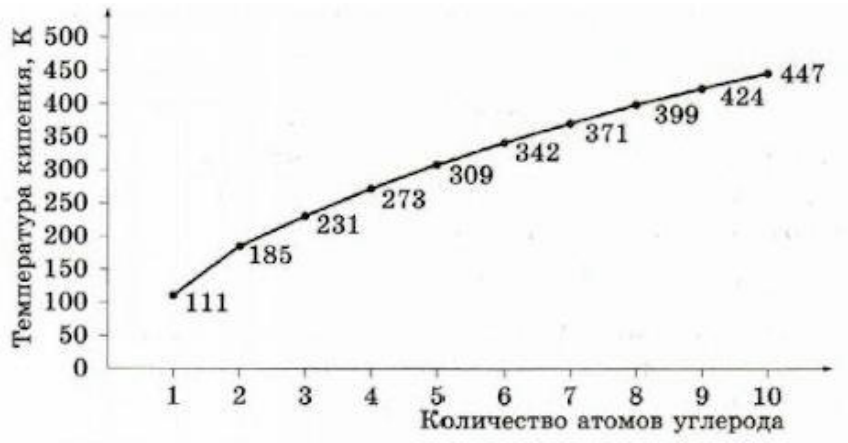


Рис. 2. Зависимость температуры кипения алканов от количества атомов углерода

Приведём часто используемые способы векторизации молекул.

#### 2.4.2.1 Молекулярные дескрипторы

Молекулярные дескрипторы — это физико-химические характеристики молекулы, которые могут быть вычислены по её молекулярному графу (структурной формуле или пространственной конформации). Приведём примеры дескрипторов.

- **Физические**: молярная масса, площадь молекулярной поверхности, заряд.

- **Функциональные**: наличие функциональных групп в структуре (−COOH, −OH, $-NH_2$).
- **Структурные**: описывают более общие свойства молекулярного графа (например, суммарная площадь полярных атомов TPSA (англ. *topological polar surface area*)).

#### 2.4.2.2 Структурные ключи

Структурные ключи представляют собой бинарные векторы[5], где каждая компонента вектора (бит) отражает наличие (1) или отсутствие (0) определенного структурного фрагмента в молекуле.

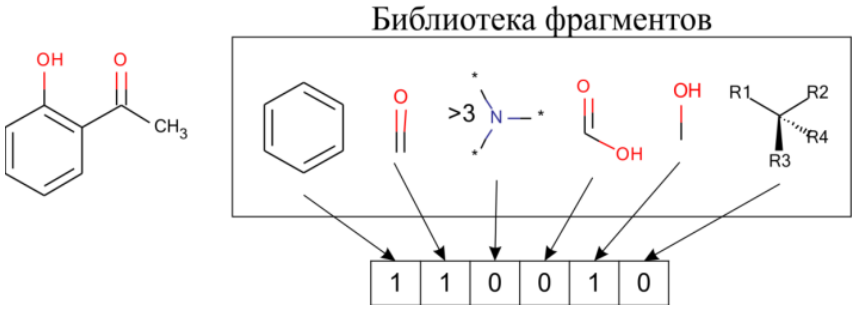


Рис. 3. Создание структурного ключа [6]

Процесс создания структурных ключей включает в себя разбиение молекул на фрагменты и их кодирование в бинарные векторы. Каждый бит в векторе соответствует конкретному структурному фрагменту (атому, хим. связи или функциональной группе). Набор всех фрагментов называют *библиотекой*.

Структурные ключи широко применяются для поиска похожих молекул и оценки структурного подобия. Однако, у структурных ключей есть ограничения. Создание библиотеки фрагментов может быть трудоемким процессом. Представление молекулы в виде структурных ключей зависит от качества сконструированной библиотеки.

#### 2.4.2.3 Молекулярные отпечатки

Молекулярные отпечатки — это тип структурных ключей, библиотека фрагментов для которых генерируется алгоритмически, а не создаётся заранее; принцип формирования векторного представления сохраняется: бинарный вектор (из 0 и 1) кодирует наличие определённых структурных фрагментов.

[5]бинарный вектор — это вектор, состоящий только из 0 или 1, называемых битами

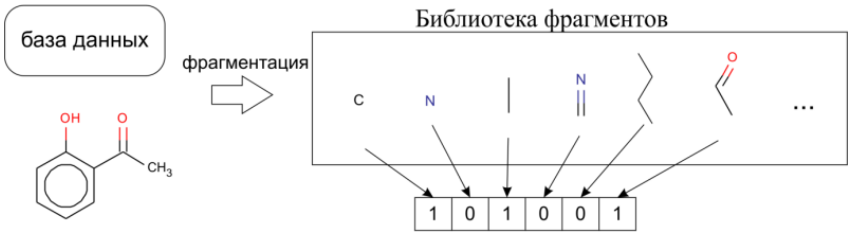


Рис. 4. Образование молекулярных отпечатков [6]

Использование молекулярных отпечатков предпочтительно, поскольку они позволяют избежать ситуации, когда для исходной библиотеки были выбраны неудачные структурные фрагменты.

Для задания соответствия между структурным фрагментом и номером компоненты (индексом) в векторе молекулярного отпечатка используют *хеш-функции* $h$, которые отображают структурные фрагменты $f \in F$ в натуральное число:

$$h : F \to \mathbb{N}. \tag{21}$$

Преобразование структурного фрагмента в число (*хеш-код*) с помощью хэш-функции $h$ называется *хешированием*.

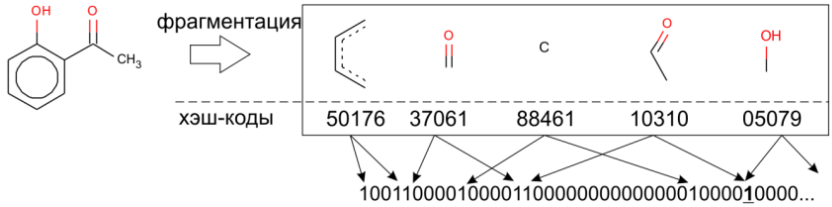


Рис. 5. Создание хешированных молекулярных отпечатков [6]

При преобразовании (хешировании) структурных фрагментов в числа неизбежно теряется много информации, поэтому существует вероятность совпадения хеш-кодов двух фрагментов: эта ситуация называется *коллизией*. У качественных хэш-функций похожие фрагменты $f \in F$ отображаются в сильно различающиеся хэш-коды, кто позволяет избегать коллизий[6].

[6] в базе данных из $M = 10^6$ молекул при кодировании каждой молекулы бинарным вектором (молекулярным отпечатком) длины $n = 64$ вероятность коллизии:

$$\mathbb{P}[\text{коллизия}] = 1 - \prod_{i=1}^{M}\left(1 - \frac{i-1}{2^n}\right) = 0.000002\% \tag{22}$$

### 2.4.3 Линейные модели

Линейные модели — это наиболее широко используемый инструмент в статистике. Они просты, значительная часть теоретических результатов получена именно на них. Линейная модель задаёт зависимость между целевым значением $\vec{y}$ и набором объясняющих переменных $X$ (представленных в виде матрицы признаков) в виде:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k} + \varepsilon_i \tag{23}$$

где $\varepsilon_i$ — случайная не прогнозируемая ошибка, имеющая нормальное распределение ($\varepsilon_i \sim N(\mu = 0, \sigma^2)$), $\beta_j$ — коэффициенты (параметры) линейной модели.

В векторном виде:

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}. \tag{24}$$

Для настройки параметров модели $\vec{\beta}$ минимизируется *сумма квадратов остатков* RSS (англ. *residual sum of squares*) между целевыми значениями $y_i$ и их оценками $\hat{y_i}$:

$$\mathrm{RSS}\left(\vec{\beta}\right) := \sum_i \left(\hat{y}_i - y_i\right)^2 = \sum_i {\hat{\varepsilon}_i}^2 \to \min_{\vec{\beta}}, \tag{25}$$

поэтому этот метод так же называется *методом наименьших квадратов* (МНК).

Простые линейные модели интерпретируемы и вычислительно эффективны, однако обладают недостатками: они не способны описывать сложные нелинейные зависимости и подвержены переобучению.

### 2.4.4 Функция потерь и эмпирический риск

МНК можно переформулировать в более общем виде через введение *функции потерь* $L$ и *эмпирического риска* $Q$:

- Функция потерь $L(\hat{y}, y)$ (англ. *loss function*) характеризует различие между истинным значением целевой переменной $y$ и её предсказанной оценкой $\hat{y}$; для линейной регрессии функция потерь квадратична: $L(\hat{y}_i, y_i) = \left(\hat{y}_i - y_i\right)^2$.
- Эмпирический риск $Q$ — это средняя ошибка на обучающей выборке, которая минимизируется в процессе настройки модели (обучения):

$$Q = \frac{1}{n} \sum_{i=1}^{n} L(\hat{y_i}, y_i) \to \min, \tag{26}$$

где $n$ — размер обучающей выборки; для линейной регрессии $Q = \mathrm{RSS}$ (25).

В описанном формализме задача оптимизации произвольной (линейной или нелинейной) модели сводится к минимизации эмпирического риска $Q$.

### 2.4.5 Регуляризация

Регуляризация — это метод борьбы с переобучением путём введения ограничений на параметры модели:

- $L_1$-регуляризация (также называемая Lasso) добавляет штраф за **сумму абсолютных значений** параметров модели, в результате оптимальные значения параметров $\vec{\beta}$ лежает близко к 0, т.е. оказываются в компактной области пространства признаков и имеют небольшие интерпретируемые значения:

$$Q_1 = \text{RSS} + \lambda \cdot \sum_i |\beta_i| \to \min_{\vec{\beta}}. \tag{27}$$

- $L_2$-регуляризация (также называемая Ridge) добавляет штраф на **сумму квадратов** параметров модели, этот подход работает аналогично $L_1$:

$$Q_2 = \text{RSS} + \lambda \cdot \sum_i \beta_i^2 \to \min_{\vec{\beta}}. \tag{28}$$

Регуляризация позволяет достичь баланса между точностью модели и её сложность. На практике $L_2$-регуляризация менее агрессивно устремляет параметры к 0 из-за гладкости квадратичного штрафа (параболы) при $\beta_j \to 0$, поэтому она используется в ситуации, когда в модели присутствуют сильно коррелированные признаки, которые должны

### 2.4.6 Нелинейные модели

В реальности не всегда оказывается, что данные линейно зависимы. Для подобных случаев существуют нелинейные модели, позволяющие учитывать более сложные взаимосвязи между данными.

Типичными примерами нелинейных моделей являются полиномиальные модели, логистические регрессии, методы ближайших соседей и опорных векторов. Эти модели находят применение там, где линейные модели недостаточно эффективны в силу сложности или нелинейной структуры данных.

Однако нелинейные модели могут оказаться более сложными в интерпретации и требовать более тщательного анализа результатов из-за их особенностей. Тем не менее, с

правильным подходом они представляют собой мощный инструмент для анализа и прогнозирования данных.

#### 2.4.6.1 Метод $k$ ближайших соседей ($k$NN)

Метод $k$ ближайших соседей $k$NN (англ. *$k$ nearest neighbours*) — один из простейших алгоритмов в машинном обучении, основанный на *гипотезе непрерывности*, утверждающей, что близким объектам $\vec{x}$ соответствуют близкие ответы $y$.

Алгоритмически, предсказание значения целевой переменной $y$ для нового объекта $\vec{x}$ вычисляется усреднением известных ответов $y$ для $k$ штук ближайших к $\vec{x}$ объектов из обучающей выборки[7]:

$$\hat{y}(\vec{x}) = \frac{1}{k} \sum_{\vec{x}' \in N_k(\vec{x})} y(\vec{x}'), \tag{29}$$

где $\vec{x}' \in N_{k(\vec{x})}$ — это множество $k$ ближайших к $\vec{x}$ соседей.

Для лучшего учёта гипотезы непрерывности, известные ответы на соседях усредняются с весом $w$, обратно пропорциональным расстонию $\rho(\vec{x}, \vec{x}')$ от нового объекта $\vec{x}$ до его соседа $\vec{x}'$:

$$\hat{y}(\vec{x}) = \frac{1}{W} \sum_{\vec{x}' \in N_k(\vec{x})} w(\rho(\vec{x}, \vec{x}')) \cdot y(\vec{x}'), \tag{30}$$

где $W$ — сумма весов; вклад каждого соседа в итоговый ответ определяется весом $w(\rho(\vec{x}, \vec{x}')$, который убывает с ростом расстояния.

$k$NN не лишён недостатков:

1. В реальных задачах, когда объекты $\vec{x}$ описываются большим набором разнородных признаков $\vec{x} = (x_1, \cdots, x_m)$, гипотеза непрерывности обычно ошибочка, т.е. пространство, в котором лежат векторы объектов $\vec{x}$ неоднородно, расстояние между объектами $\rho(\vec{x}, \vec{x}')$ неинтерпретируемо, веса соседей $x'$ не отражают их вклад в значение целевой переменной $y$. Например, при описании молекулы набором разнородных дескрипторов (молярная масса, заряд, площадь поверхности) $m = (M, Q, S)$, простое евклидово расстояние $\rho(m_1, m_2) = \sqrt{(M_1 - M_2)^2 + (Q_1 - Q_2)^2 + (S_1 - S_2)^2} \sim \sqrt{\text{масса}^2 + \text{заряд}^2 + \text{длина}^4}$ лишено

[7] фактически, в $k$NN отсутствует этап обучения

смысла. С этим ограничением можно бороться путём введения функций расстояния, учитывающих особенности описанного пространства.

2. Поиск ближайших соседей в больших обучающих выборках ресурсозатратен, однако, эта проблема решается эффективными алгоритмами поиска.

#### 2.4.6.2 Метод опорных векторов

В 60-е годы группа советских математиков из Института Проблем Управления (ИПУ РАН) под руководством В. Н. Вапника и А. Я. Червоненкиса разработали метод обобщенного портрета[8]. В 90-е годы метод был развит под руководством В. Н. Вапника и после некоторых улучшений получил современное название — метод опорных векторов (англ. *support vector machine*, SVM). Изначально разработанный для классификации метод был перенесен на задачу регрессии, алгоритм получил название регрессия на опорных векторах (англ. *support vector regression*, SVR).

Фактически, в простейшем случае работа алгоритма SVR сводится к нахождению зависимости в данных вида $a = f(\vec{x})$ такой, чтобы все её значения на объектах обучающей выборки $f(\vec{x}_i)$ отличались от истинных $y_i$ не более, чем на $\varepsilon$, т.е.

$$f: \quad |y_i - f(\vec{x}_i)| \leq \varepsilon. \tag{31}$$

В простейшем случае ищется линейная зависимость:

$$f(\vec{x}) = \sum_{i=1}^{n} w_i \cdot x_i + w_0 = \langle \vec{w}, \vec{x} \rangle + w_0, \tag{32}$$

где $\vec{w} = \left(w_1, ..., w_j\right)$ — вектор коэффициентов линейной модели (интерпретируется как вес признака), а $w_0$ — свободный член.

При наличии в данных выбросов условие (31) слишком ограничивающее и на практике ослабляется, т.е. при нахождении $f(\vec{x})$ её разрешается отступать от значений $y_i$ более, чем на $\varepsilon$, но за это вводится штраф. Эмпирический риск принимает вид:

$$Q = \sum_{i=1}^{N} \max(0, |y_i - \langle \vec{w}, \vec{x}_i \rangle + w_0)| - \varepsilon) + \frac{1}{2C} \|\vec{w}\|^2 \to \min_{\vec{w}, w_0} \tag{33}$$

[8] заключается в построении плоскости, разделяющей два класса, учетываются лишь точки лежащие вблизи разделяющей плоскости

где $\langle \vec{w}, \vec{x}_i \rangle$ обозначает скалярное произведение, $\frac{1}{2C} > 0$ — коэффициент регуляризации, $\|\vec{w}\|$ — норма вектора весовых коэффициентов.

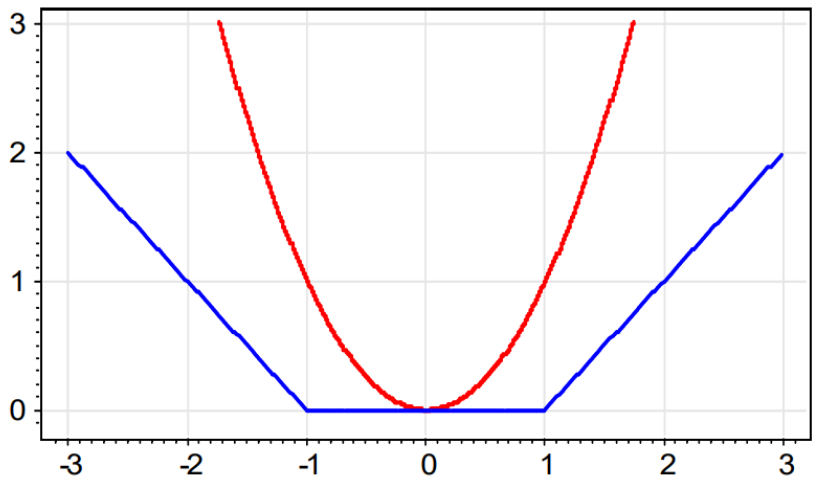


Рис. 6. Функции потерь в задаче МНК (красная парабола); в методе регрессии на опорных векторах (синяя кусочно-линейная, ε = 1)

Из Рис. 6 видно, что функция потерь в SVR учитывает только точки, вышедшие за диапазон $\varepsilon$; подобные функции называют $\varepsilon$-нечувствительными.

$\varepsilon$ задает область пространства вокруг $f(\vec{x})$ внутри, которой не накладываются штрафы, эту область называет *$\varepsilon$-трубой*. Объекты, вышедшие за пределы $\varepsilon$ — *опорные векторы*, т.е. именно ошибку на них минимизирует алгоритм.

Однако, классический метод опорных векторов пригоден только для анализа линейной зависимости в данных. Для снятия этого ограничения используют функции *ядра* $K$ — обобщения скалярного произведения на нелинейные (неевклидовы) пространства:

$$K(\vec{x}_i, \vec{x}_j) = \langle \varphi(\vec{x}_i) \varphi(\vec{x}_j) \rangle, \tag{34}$$

где $\varphi(\vec{x}_i)$ — нелинейная функция преобразования, отображающая исходные данные в новое пространство.

Примеры ядер:

- линейное ядро:

$$K(\vec{x}_i, \vec{x}_j) = \langle \vec{x}_i, \vec{x}_j \rangle. \tag{35}$$

- гауссово:

$$K(\vec{x}_i, \vec{x}_j) = e^{-\gamma\|\vec{x}_i - \vec{x}_j\|^2}, \quad \gamma \in \mathbb{R}. \tag{36}$$

- полиномиальное:

$$K(\vec{x}_i, \vec{x}_j) = \left(a \cdot \langle \vec{x}_i \vec{x}_j \rangle + b\right)^p, \quad a, b, p \in \mathbb{R}. \tag{37}$$

SVR — метод машинного обучения способный находить в данных линейные и нелинейные зависимости (за счет применения ядра), устойчивый к выбросам (за счет возможности игнорирования части наблюдений) и обладающий высокой точностью предсказаний, но требующий большого количества вычислительных ресурсов при нахождении зависимости в нелинейных пространствах.

#### 2.4.6.3 Метод случайного леса

В основе метода случайного леса (англ. *random forest*) лежит две идеи:

1. решающее дерево — набор правил, организованный в виде иерархического графа (Рис. 7), которые необходимо применить для прогнозирования целевого значения.
2. *ансамбль моделей* — множество моделей, ответы которых комбинируются путем усреднения целевых значений $\hat{y}$ (предсказанных одиночными моделями).

Случайный лес — это ансамбль (множество) решающих деревьев, где кажое дерево строится независимо на разных подмножества обучающей выборки, и результат усредняется. Такой метод построения ансамблей называется бэггинг (англ. *bagging* — ***b****ootstrap* ***agg****regat****ing***).

Дерево состоит из корня, внутренних узлов и листьев — конечных узлов. В корне и узлах содержатся условия, по которым объекты разделяются на группы; каждому листу приписывается значение — ответ для объектов дошедших до этого листа, который, например, может вычисляться как среднее арифметическое целевых значений $y_i$ попавших в данный лист объектов из обучающей выборки.

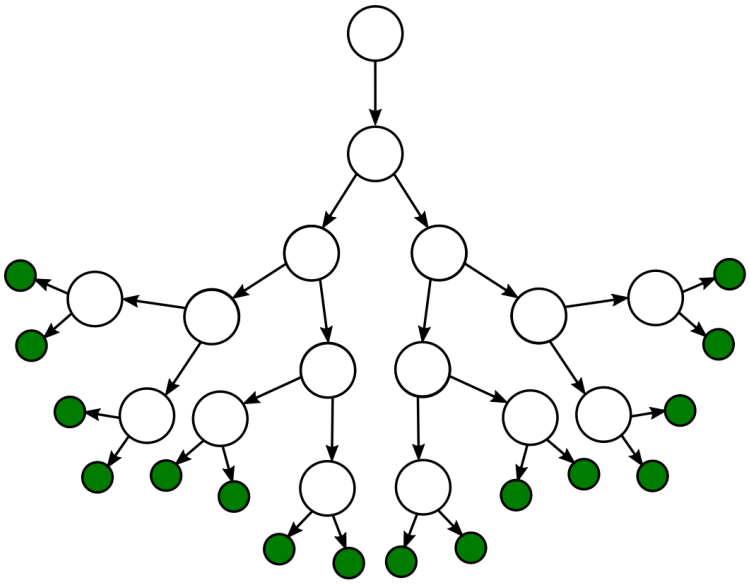

Рис. 7. Дерево принятия решений

В каждом узле объекты с одинаковым значением тестируемого признака направляются в одну дочернюю вершину. При прохождении по дереву (Рис. 7) — повышается однородность, т.е. похожие объекты концентрируются в одних вершинах. Например, молекулы при прохождении по дереву будут группироваться по схожести свойств.

Неоднородность объектов в каждой вершине $V$ может быть определена по формуле:

$$H(V) = \frac{1}{n} \sum_{\vec{x}_i \in V} \left( y_i - \overline{y}_V \right)^2, \tag{38}$$

где $n$ — количество объектов обучающей выборки, дошедших до вершины $V$, $\overline{y}_V$ — среднее значение целевых значений на этих объектах.

В (38) неоднородность измеряется через среднеквадратичное отклонение, однако могут использоваться и другие функции потерь.

При построении дерева новые листы создаются тогда, когда повышается однородность: оценивается неоднородность родительской вершины $V$ и двух потенциальных дочерних вершин $U_{\text{лев}}$ и $U_{\text{прав}}$, рассчитывается уменьшение неоднородности:

$$\Delta H = H(V) - \underbrace{\left( \frac{n_{\text{лев}}}{n} \cdot H(U_{\text{лев}}) + \frac{n_{\text{прав}}}{n} \cdot H\left(U_{\text{прав}}\right) \right)}_{\text{взвешенная неоднородность дочерних вершин}} \to \max_{U_{\text{лев}}, U_{\text{прав}}}, \tag{39}$$

где $n$, $n_{\text{лев}}$, $n_{\text{прав}}$ — число объектов в родительской и дочерних вершинах.

Уравнение (39) называется *критерием ветвления* (англ. *split criterion*), при создании дочерних вершин, он максимизируется путём перебора объектов в дочерних вершинах:

$$V = U^*_{\text{лев}} \sqcup U^*_{\text{прав}}, \tag{40}$$

здесь $U^*_{\text{лев}}$ и $U^*_{\text{прав}}$ — произвольные разбиения родительской вершины $V$ на две дочерние, «$\sqcup$» — объединение не пересекающихся множеств. Среди всех разбиений будет выбрано то, которое максимизирует критерий ветвления.

При движении по дереву уменьшается количество объектов обучающей выборки, дошедших до каждого узла. Рассчитанные по (39) уменьшения неоднородности оцениваются по малому количеству объектов и теряют надежности, т.е. возрастает погрешность оценки $\Delta H$.

Если не задавать ограничения на количество ветвлений (глубину дерева), в каждый лист попадет по одному уникальному объекту, т.е. дерево запомнит всю обучающую выборку наизусть.

Переобучение в деревьях проявляется как *а)* неверно принятые решения о ветвлении, в следствие ненадежных оценок $\Delta H$ по обучающей выборке; и как *б)* излишняя подстройка листов дерева под объекты обучающей выборки.

Для борьбы с переобучением и увеличения обобщающей способности к дереву применяется регуляризация:

- ограничение глубины дерева,
- ограничение на минимальное количество объектов необходимых для ветвления или создания листа,
- введение $\Delta H_{\min}$, т.е. минимальное изменение неоднородности, достаточное для создания дочерних вершин.

Метод случайного леса (ансамбль решающих деревьев) отличается высокой прогнозирующей способностью, простотой их построения (каждое дерево строится быстро и независимо), высокой вычислительная эффективность (разные деревья ансамбля могут рассчитываться параллельно на разных ядрах процессора), по сравнению с мето-

дом опорных векторов, но требует длительного подбора гиперпараметров (в частности параметров регуляризации).

### 2.4.7 Оценка качества моделей

Оценка качества модели — важный этап машинного обучения, он необходим для выбора модели с наибольшей предсказательной способностью. В регрессии оценка качества модели зависит от отклонения предсказываемых значений $\hat{y}$ от истинных $y_i$ и характеризует, как точно модель предсказывает целевое значение $y_i$. Существуют различные способы вычисления подобных оценок ($R^2$, MSE, RMSE, MAE и др.).

#### 2.4.7.1 Дисперсия и корреляция

Для генеральной совокупности $\mathcal{A} = \{a_1, ..., a_m\}$, содержащей все возможные значения переменной $a$, дисперсия равна:

$$D_{\mathcal{A}} := \frac{1}{m} \sum_i (a_i - \overline{a})^2 = \sigma^2, \tag{41}$$

где $\sigma$ — стандартное отклонение генеральной совокупности, а $\overline{a} = \frac{1}{n} \sum_i a_i$ — среднее значение по всем $a_i$.

Для выборки значений величины $A = \{a_1, ..., a_n\} \subset \mathcal{A}$ (где $n < m$) выборочная дисперсия определяется формулой:

$$D_A := \frac{1}{n-1} \sum_i (a_i - \overline{a})^2 = S^2, \tag{42}$$

где $S$ — стандартное отклонение выборки.

Разница в коэффициентах перед суммой в (41) и (42) обусловлена т.н. *поправкой Бесселя* ($\frac{1}{n} \to \frac{1}{n-1}$), она вводится из соображения, что выборка содержит лишь часть генеральной совокупности и, следовательно, кажущаяся дисперсия выборки $A$ может уменьшаться ($D_A < D_{\mathcal{A}}$); в (42) это компенсируют увеличением коэффициента ($\frac{1}{n-1} > \frac{1}{n}$).

Линейную зависимость двух величин $A$ и $B$ характеризуют с помощью коэффициента *корреляции*:

$$r_{A,B} := \frac{\frac{1}{n-1} \sum_i (a_i - \overline{a}) \left(b_i - \overline{b}\right)}{\sqrt{\frac{1}{n-1} \sum_i (a_i - \overline{a})^2 \cdot \frac{1}{n-1} \sum_i \left(b_i - \overline{b}\right)^2}}) = \frac{\frac{1}{n-1} \sum_i (a_i - \overline{a}) \left(b_i - \overline{b}\right)}{\sigma_A \cdot \sigma_B}, \tag{43}$$

числитель в (43) называют *ковариацией*, в знаменателе стоит произведение стандартных отклонений, которые могут быть вычислены по (42) или по (41) с поправкой на коэффициент $\frac{1}{m}$.

С помощью корреляции можно характеризовать линейные модели, если в качестве A взять множество $\hat{y}$, а B — множество $y_i$.

#### 2.4.7.2 $R$-квадрат ($R^2$)

Одна из основных метрик — это $R^2$ ($R$-квадрат), она показывает долю дисперсии $D$ зависимой переменной $y_i$, которую можно объяснить через независимые переменные $X$, построив модель $M$ вида $\hat{y} = f(\vec{x})$:

она определяется через отношение дисперсий истинных ($D_Y$) и предсказанных ($D_{\widehat{Y}}$) значений; по сути, $R^2$ — это доля дисперсии, объяснённой моделью $M$ через переменные $X = \{x_1, ..., x_k\}$:

$$R^2 = \frac{D_{\widehat{Y}}}{D_Y}, \tag{44}$$

где $Y = \{y_1, ..., y_n\}$ — набор истинных значений предсказываемой переменной, а $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_n\}$ — соответствующие им значения, предсказанные моделью $M$.

При подстановке выражений дисперсии (42) или (41) в определение $R^2$ по формуле (44), после сокращения общего множителя $\frac{1}{n-1}$ (или $\frac{1}{m}$) получают выражение состоящее из отношения сумм квадратов:

$$R^2 := \frac{\mathrm{ESS}}{\mathrm{TSS}}. \tag{45}$$

- Числитель ESS в (45) называют *объясненной суммой квадратов* (англ. *explained sum of squares*), ESS представляет собой разброс предсказанных значений $\hat{y}_i$ относительно среднего $\overline{y}$:

$$\mathrm{ESS} := \sum_i (\hat{y}_i - \overline{y})^2. \tag{46}$$

- Знаменатель TSS в (45) называют *полной суммой квадратов* (англ. total sum of squares), показывающей разброс реальных значений $y_i$ относительно их среднего значения:

$$\mathrm{TSS} := \sum_i (y_i - \overline{y})^2. \tag{47}$$

Для линейных моделей $M$ вида

$$y(\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{48}$$

можно показать, что TSS, ESS и RSS в уравнениях (47), (46) и (25) связаны соотношением:

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}, \tag{49}$$

выразив ESS из (49) и подставив в (45), получают альтернативное определение:

$$R^2 := 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \tag{50}$$

это выражение чаще всего используют на практике. Также для модели (48) можно показать, что $R^2$ равен квадрату корреляции (43):

$$R^2 = r^2. \tag{51}$$

Определения (45) и (50) эквивалентны для моделей вида (48), но не для произвольной модели — ввиду того, что соотношение (49) в общем случае неверно. Аналогично, свойство (51) выполнено только при оговоренных условиях. В общем случае определения $R^2$ задаваемые формулами (45) и (50) неэквивалентны, однако на практике выражение (50) оказывается лучше.

Для простейшей несмещённой[9] линейной модели

$$M: \quad y(\vec{x}) = \overline{y} \tag{52}$$

значение $R^2$:

$$R^2_M = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum_i \left(\overline{y} - \overline{y}\right)^2}{\mathrm{TSS}} = 0. \tag{53}$$

Для альтернативной намерено смещённой модели

$$M': \quad y(\vec{x}) = \overline{y} + \varepsilon \tag{54}$$

со систематической ошибкой $\varepsilon = \mathrm{const} \neq 0$ использование определения (45) приведёт к завышению оценки качества:

$$R^2_{M'} = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum_{i=1}^{n} \left((\overline{y} + \varepsilon) - \overline{y}\right)^2}{\mathrm{TSS}} = \frac{n \cdot \varepsilon^2}{\mathrm{TSS}} > 0, \tag{55}$$

[9] среднее оценок совпадает со средним выборки: $\overline{\hat{y}} = \overline{y}$, т.е. отсутствует систематическая погрешность

где $n$ — число наблюдений, а TSS определяется исключительно исходной выборкой и не зависит от выбора модели.

В методе МНК, где оптимизируются ошибки $\mathrm{RSS} = \hat{\varepsilon}_1^2 + \ldots + \hat{\varepsilon}_n^2 \to \min$, модель $M'$ можно получить только по оплошности, однако, полученное из определения (45) соотношение

$$R^2_M < R^2_{M'} \tag{56}$$

всё равно нельзя игнорировать: заведомо плохая модель, сдвинутая на константу $\varepsilon$, получает излишне оптимистичную оценку качества. Параметры более сложных нелинейных моделей находятся численно через оптимизацию $Q \to \min$, что не гарантирует единственности и корректности решения. Определения (45) разумно только несмещённых линейных моделей, для произвольных моделей оно не даёт точной оценки качества.

Напротив, из определения (50) получают более реалистичные заниженные оценки $R^2$:

$$R^2_{M'} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^n \left((\overline{y} + \varepsilon) - y_i\right)^2}{\mathrm{TSS}} \leq 1 - \frac{\sum_{i=1}^n \left(\overline{y} - y_i\right)^2}{\mathrm{TSS}} = 0, \tag{57}$$

при увеличении систематической ошибки, будет увеличиваться RSS, что приведёт к снижению значения $R^2$, т.е., заведомо плохие модели получат меньшую оценку качества.

| модель | $R^2 = (\cdot)(\mathrm{ESS})$ | $R^2 = (\cdot)(\mathrm{RSS})$ | $\max R^2$ | $\min R^2$ |
|---|---|---|---|---|
| $M : y(\vec{x}) = \overline{y}$ | 0 | 0 | 0 | 0 |
| $M' : y(\vec{x}) = \overline{y} + \varepsilon$ | $\frac{n \cdot \varepsilon^2}{\mathrm{TSS}}$ | $\leq 0$ | $+\infty$ | $-\infty$ |

Таким образом, вычисление $R^2$ по формуле (50) позволяет:

1. избегать завышения кажущегося качества модели;
2. использовать безопасные оценки $R^2$ за границами применимости исходного определения (45), т.е. для произвольной (линейной или нелинейной) модели, а не только модели задаваемой условием (48).

Отношение сумм квадратов $\frac{\mathrm{RSS}}{\mathrm{TSS}}$ в выражении (50) для $R^2$ всегда положительно и неограниченно сверху. Величина самого $R^2$ лежит в диапазоне от $-\infty$ до 1. Значению $R^2 = 1$ соответствует идеальная модель. Непосредственно из (50) можно показать, что

$R^2 = 0$ для простейшей линейной модели $y(\vec{x}) = \overline{y}$. При RSS $>$ TSS значение $R^2 < 0$, т.е. модель демонстрируют качество ниже, чем $y(\vec{x}) = \overline{y}$ и, следовательно, вовсе не обладают предсказательной силой. Модель может быть неограниченно плохой, поскольку для $R^2$ разрешены любые отрицательные значения. Разумные модели характеризуются $R^2 > 0$, для моделей с высокой предсказательной силой, значение $R^2 \to 1$. Но коэффициент детерминации лишь одна из характеристик модели, не гарантирующая безупречность её работы, поэтому следует использовать её в сочетании с другими метриками.

#### 2.4.7.3 Усреднённые ошибки

Помимо $R^2$ часто используются и другие метрики:

- MSE — среднеквадратичная ошибка (англ. mean squared error).
- RMSE — корень из среднеквадратичной ошибки (англ. root mean squared error).
- MAE — средняя абсолютная ошибка (англ. mean absolute error).

MSE — есть сумма квадратов отклонений истинных значений $y_i$ от их предсказанных $\hat{y}_i$, деленная на количество наблюдений $n$:

$$\text{MSE} := \frac{1}{n}\sum_i (\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_i \hat{\varepsilon_i}^2 = \frac{\text{RSS}}{n}. \tag{58}$$

Также из определения (58) следует, что MSE — это RSS деленная на количество наблюдений.

Отметим несколько свойств этой метрики, возникающих в силу квадратичной зависимости:

- метрика обладает одним глобальным минимумом, т.е. качество модели монотонно меняется (ухудшается) с увеличением значения метрики MSE;
- неустойчива к наличию выбросов в данных, поскольку слагаемых $\hat{\varepsilon}_i^2$ вносят большой вклад в значением метрики для измерений с большой ошибкой $\hat{\varepsilon}_i$;
- интерпретация значения метрики MAE затруднена из-за возведения исходных единиц измерения в квадрат.

Один из способов снижения негативного эффекта от выбросов — замена квадрата в определении на модуль:

$$\mathrm{MAE} := \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{59}$$

MAE показывает среднее абсолютное отклонение предсказанного значения $\hat{y}_i$ от реального значения $y_i$, но в отличие от (58) отклонения $\hat{\varepsilon}_i$ в квадрат не возводятся. По сравнению с MSE, MAE обладает полезными свойствами:

- лучшая интерпретируемость результата — т.к. сохраняется размерность единиц;
- бóльшая устойчивость к выбросам, ввиду отсутствия квадратичной зависимости $\hat{\varepsilon}_i^2$, что нивелирует влияние выбросов на оценку качества.

RMSE измеряет среднеквадратичное отклонение между прогнозируемыми значениями и фактическими значениями целевой переменной и является квадратным корнем из MSE:

$$\mathrm{RMSE} := \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i}\left(\hat{y}_i - y_i\right)^2}, \tag{60}$$

При использовании RMSE сохраняются единицы измерения, но из-за возведения $\hat{\varepsilon}_i$ в квадрат сохраняется особенность MSE — неустойчивость к выбросам.

Отметим общие свойства этих метрик. Все они представляют собой четные функции[10] от отклонения $\hat{\varepsilon}_i$. Поэтому не происходит компенсации ошибок, т.е. отрицательные ошибки ($\hat{\varepsilon}_i < 0$) не нейтрализуют положительные ($\hat{\varepsilon}_i > 0$). Модели с нулевой средней ошибкой (MAE, MSE, RMSE = 0) уже оптимальны, и не могут быть улучшены.

Все обозначенные в этом разделе метрики имеют свои преимущества и недостатки. Для получения корректного представления о качестве модели предпочтительно использовать их комбинацию.

[10] функция $f(x)$ называется *чётной*, если выполняется: $f(x) = f(-x)$

# 3 Экспериментальная часть

В ходе экспериментальной части было определено распределение веществ-ингибиторов (АТМБ и Диллапиол) в мицеллярной системе, для этого было зарегистрировано по три ИК-спектра каждого анализируемого вещества в обращенных мицеллах, полярной и неполярной фазах. По смещению пика функциональной группы судили о её распределении по фазам. После чего обучили интерпретируемую модель и провели теоретический анализ.

## 3.1 Материалы и методы

### 3.1.1 Инструментальные методы анализа

Для регистрации ИК-спектров в пластиковой пробирке навеску АОТ

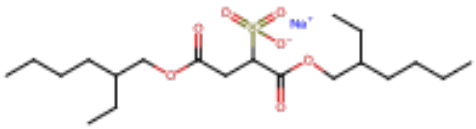

Рис. 8. АОТ (бис(2-этилгексил)сульфосукцинат натрия, Sigma-Aldrich, США) растворили в октане для получения концентрации 0.1 мг/мл, после чего медленно добавили воду с растворенным в ней ингибитором (оба синтезированы в ИОХ РАН) при перемешивании до достижения молярного отношения воды к АОТ 22:1.

Для анализа распределения вещества сравнивали его ИК-спектр в обращенных мицеллах со спектрами в полярной и неполярной фазах. Изменение микроокружения вызывало смещение пика, соответствующего конкретной функциональной группе, вправо или влево относительно его положения в мицеллах. Считалось, что группа находится в той фазе, где это смещение было минимальным.

В качестве полярной фазы использовался раствор ингибитора в смеси вода-этанол в объемном соотношении один к одному. Для неполярной фазы ингибитор растворили в системе октан-этанол в объемном соотношении 1:1. Спектры регистрировали на ИК-Фурье спектрометре Bruker Tensor 27 с терморегулируемой ячейкой с цинк-селеновым кристаллом. Диапазон измерений был от 900 до 3000 $см^{-1}$.

### 3.1.2 Построение модели $\log P$

Для предсказания $\log P$ было построено несколько моделей машинного обучения с помощью библиотеки scikit-learn для языка Python, например регрессии: линейная, Lasso, Ridge, *k*-ближайших соседей (KNN), модели на опорных векторах (SVR) и случайного леса (RF), а также использовали нейронную сеть ChemProp [7]. Обучение велось либо на основе дескрипторов, полученных с помощью библиотеки RDKit, либо на основе хешированных молекулярных отпечатках, полученных с помощью надстройки Datamol.[11] Для обучения было отобрано два набора данных [8], [9],состоящих из молекул и их экспериментальных значений $\log P$, оценку предполагалось производить с помощью 707 лекарственно-подобных молекул, для которых $\log P$ определяли с помощью обращенно-фазовой абсорбционной хроматографии [4].

Для повышения точности модели дескрипторы были предварительно стандартизованы путем исключения признаков, имеющих малую дисперсию и приведения оставшихся к одной шкале от 0 до 1 или от −1 до 1. Признаков с только отрицательными значениями не было. После стандартизации произвели отбор признаков с помощью алгоритмов SFS и RFECV.

Оценку, обучение и подбор параметров производился с помощью кросс-валидации со случайным перемешиванием молекул и последующим разделением обучающей выборки на 5 частей, 4 из которых передавались на обучение и одна на валидацию. Метрики качества включали среднеквадратичную ошибку (MSE), коэффициент детерминации ($R^2$) и среднюю абсолютную ошибку (MAE).

## 3.2 Результаты и обсуждения

В результате отбора признаков оказалось, что их оптимальное количество для разных моделей лежит в диапазоне 40-80 (Рис. 9). Для моделей, использующих молекулярные отпечатки, отбор признаков не проводили, т.к. он приводил лишь к незначительному повышению метрик качества (Рис. 10). В итоге модели, обученные на молекулярных отпечатках показывали $R^2$ лучше, чем обученные на дескрипторах.

[11] т.е. матрица признаков состояла либо из дескрипторов, либо из хешированных молекулярных отпечатков

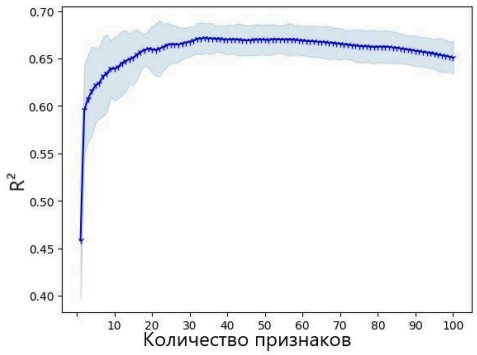


Рис. 9. График зависимости $R^2$ от количества параметров для модели SVR, обученной на дескрипторах

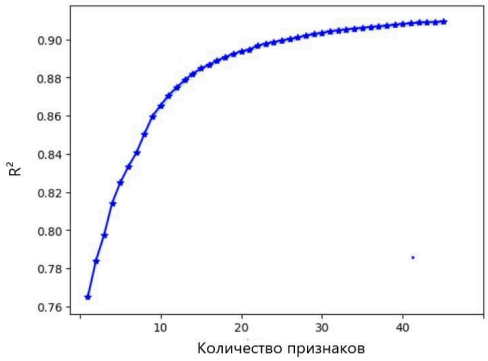


Рис. 10. График зависимости $R^2$ от количества параметров для модели SVR, обученной на молекулярных отпечатках

Наиболее точными моделями оказались SVR, RF и нейронная сеть ChemProp, иные модели показали $R^2$ менее 0.6.

| | **Случайный лес** | **SVR** | **Нейронная сеть** |
|---|---|---|---|
| $R^2$ | 0.87 | 0.91 | 0.95 |
| MAE | 0.30 | 0.28 | 0.16 |
| MSE | 0.34 | 0.30 | 0.17 |

Однако эти результаты были показаны на валидационных данных, а на тестовых все модели показывали $R^2$ меньше нуля, что характерно для систематических ошибок. Также градуировка, по которой находились значения $\log P$ из коэффициента удерживания, была построена для малого числа молекул (55) с молекулярной массой менее 100 Да [5] и корректность её использования в данном случае ($Mr_{ср.}$~400 Да) находится под вопросом.

Для коррекции градуировочной зависимости на молекулы с большей молекулярной массой, было предложено взять за верные данные значения $\log P$, рассчитанные нейронной сетью и аппроксимировать к ним экспериментальные значения $\log P$. Однако, такой метод имеет ряд недостатков: измерения осуществлялись на трех различных колонках с разной градуировочной зависимостью, но данных какие молекулы измерялись на какой колонке нет, также модель вносит неточность, т.к. не является идеальной.

В результате регистрации ИК-спектра АТМБ пик соответствующий колебанию метокси-группам (2890) оказался ближе к полярной фазе (2901), чем к неполярной (2924), а пик относящийся к колебаниям ароматической C-C связи занял промежуточное значение: в мицеллах (1491-1488 и 1454), в полярной фазе (1488,5 и 1449-1456) и неполярной фазе (1492 и 1466). Следовательно АТМБ преимущественно находится внутр мицеллы, но его гидрофобный хвост расположен между молекулами АОТ. Аналогичные результаты показала и интерпретируемая модель: гидрофобный хвост направлен в менее полярную среду, а метокси-группы в более полярную.

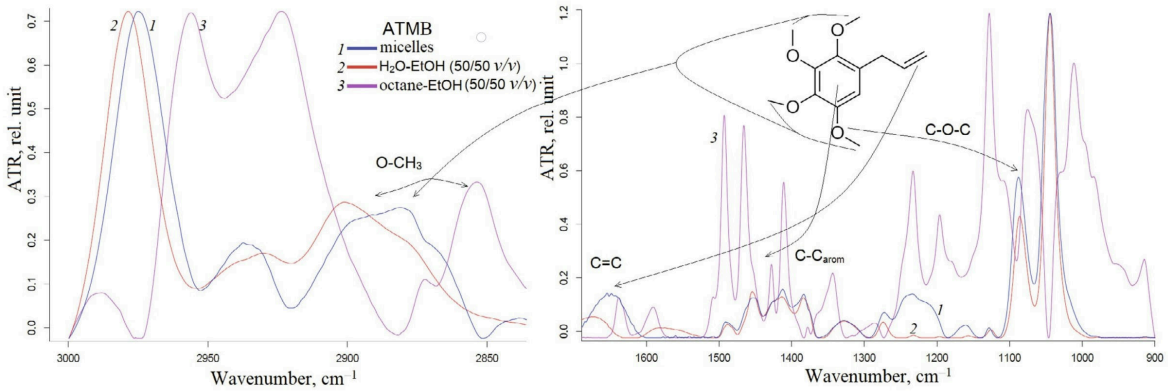


Рис. 11. ИК-спектр АТМБ в мицеллах, полярной и неполярной фазе

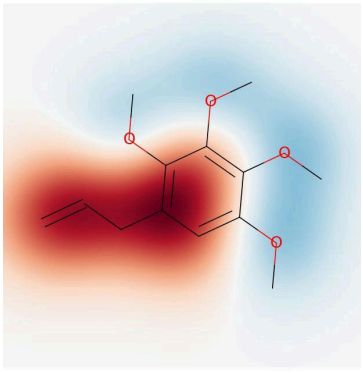

Рис. 12. Структурная формула АТМБ, где красным отмечена гидрофобная часть, а синим гидрофильная

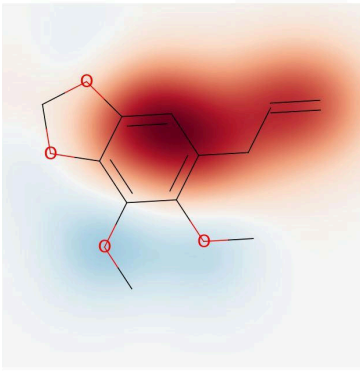

Рис. 13. Структурная формула Диллапиола, где красным отмечена гидрофобная часть, а синим гидрофильная

В случае Диллапиола аналогично группы O-$CH_2$-O, =C-O-C и -O-C$H_3$ ориентированы внутрь мицеллы, а ароматическая часть находится внутри интерфейса. Результаты модели совпали с экспериментальным анализом, кроме части O-$CH_2$-O, которую модель отнести к конкретной фазе не смогла.

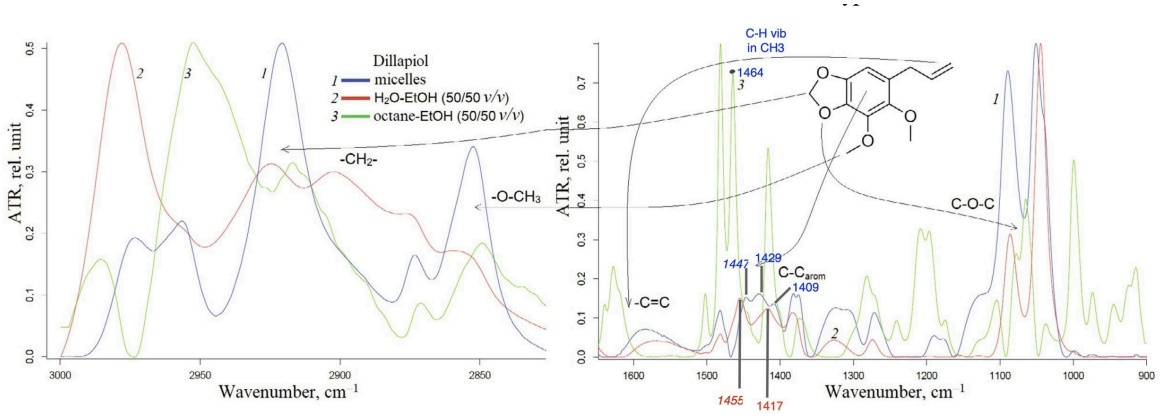


Рис. 14. ИК-спектр Диллапиола в мицеллах, полярной и неполярной фазе

С помощью точной точной модели были получены значения $\log P$: 2.12 и 2.56 для АТМБ и диллапиола, что говорит о их высокой гидрофобности, однако, судя по экспериментальным данным оба вещества находятся в интерфейсе мицеллы и «общая гидрофобность» молекулы не информативна для анализа мицеллярных систем.

# 4 Выводы

1. Абсолютная ошибка предсказания  0.3 ед. LogP, что сопоставимо с экспериментальной.
2. В наборе экспериментальных данных по LogP, полученных методом обращенно-фазовой хроматографии [4], обнаружена систематическая ошибка = 0.739 , предложен метод компенсирования этой ошибки.
3. Данные по распределению ингибиторов в мицеллярной системе, полученные методом ИК совпадают с предсказаниями интерпретируемой модели.
4. АТМБ и диллапиол преимущественно находятся в неполярной фазе (LogP равен 2.12 и 2.56).
5. Из-за внутримолекулярного различия LogP (0.52 и 0.73 для АТМБ и Диллапиола) ингибиторы находятся в динамическом межфазном равновесии, что проявляется на ИК в виде смещения максимума характеристических частот.

# Источники


[1] J. Coates и others, «Interpretation of infrared spectra, a practical approach», *Encyclopedia of analytical chemistry*, т. 12, сс. 10815–10837, 2000.

[2] Б. Тарасевич, «ИК спектры основных классов органических соединений», 2012.

[3] Ю. А. Золотов, Е. Н. Дорохова, В. И. Фадеева, и others, *Основы аналитической химии. Кн. 2. Методы химического анализа*. 1999.

[4] S. Martel *и др.*, «Large, chemically diverse dataset of log P measurements for benchmarking studies», *European Journal of Pharmaceutical Sciences*, т. 48, вып. 1–2, сс. 21–29, 2013.

[5] A. Guillot *и др.*, «Lipophilicity determination of highly lipophilic compounds by liquid chromatography», *Chemistry & Biodiversity*, т. 6, вып. 11, сс. 1828–1836, 2009.

[6] Т. Маджидов, И. Баскин, И. Антипин, и А. Варнек, «Введение в хемоинформатику (серия пособий)», *Международный журнал экспериментального образования*, вып. 10–2, сс. 198–200, 2015.

[7] E. Heid *и др.*, «Chemprop: A machine learning package for chemical property prediction», *Journal of Chemical Information and Modeling*, т. 64, вып. 1, сс. 9–17, 2023.

[8] M. Popova, O. Isayev, и A. Tropsha, «Deep reinforcement learning for de novo drug design», *Science advances*, т. 4, вып. 7, с. eaap7885, 2018.

[9] N. Ulrich, K.-U. Goss, и A. Ebert, «Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation», *Communications Chemistry*, т. 4, вып. 1, с. 90–91, 2021.