**Курсовая работа на тему:**

**"Прогнозирование эффективности химических соединений на основе молекулярных дескрипторов"**

**Решетило Никита**

**1 курс, «Машинное обучение»**

**Содержание**

1. **Описание данных…………………………………………………… 3**
   1. **Актуальность исследования………………………………………. 3**
   2. **Ключевые показатели эффективности........................................... 4**
   3. **Цели исследования………………………………………………….. 4**
   4. **Методологический подход…………………………………………. 4**
   5. **Классификация молекулярных дескрипторов………………….. 7**
2. **EDA…………………………………………………………………… 10**
   1. **Первичный анализ…………………………………………………. 10**
   2. **Анализ молекулярных дескрипторов……………………………. 11**
   3. **Общие результаты обработки данных…………………………… 12**
   4. **Анализ аномальных признаков…………………………………… 12**
   5. **Химическая интерпретация аномалий…………………………... 15**
   6. **Анализ выбросов молекулярных свойств……………………….. 17**
   7. **Анализ корреляции молекулярных дескрипторов с биологической активностью…………………………………………………………. 19**
   8. **Итог по анализу молекулярных данных…………………………. 28**
3. **Исследовательский анализ………………………………………… 29**
   1. **Аналих моделей прогнозирования индекса селективности (IC50)………………………………………………………………….. 29**
   2. **Анализ моделей прогнозирования цитотоксической активности соединений (CC50)…………………………………………………… 31**
   3. **Анализ моделей прогнозирования для регрессии индекса селективности (SI)…………………………………………………… 36**
   4. **Классификация: превышает ли значение SI медианное значение выборки……………………………………………………………….. 41**
   5. **Классификация: превышает ли значение IC50 медианное значение выборки……………………………………………………. 44**
   6. **Классификация: превышает ли значение CC50 медианное значение выборки …………………………………………………… 49**
   7. **Классификация: превышает ли значение SI значение 8……….. 54**
4. **Финальный вывод по проделанной работе………………...…….. 56**
5. **Источники данных…………………………………...…..………….. 58**
   * + 1. **Описание данных**
          1. **Актуальность исследования**

Оптимизация процесса разработки лекарственных препаратов требует раннего выявления перспективных соединений с высокой биологической активностью и низкой токсичностью. Современные методы хемоинформатики и машинного обучения позволяют прогнозировать фармакологические характеристики молекул на основе их структурных особенностей, минимизируя затраты на дорогостоящие лабораторные исследования.

Применение компьютерного моделирования в доклинических исследованиях значительно ускоряет процесс скрининга потенциальных лекарственных соединений, что особенно важно в условиях растущих требований к безопасности и эффективности новых препаратов. Методы количественного анализа структура-активность (QSAR) и молекулярного докинга уже доказали свою эффективность в предсказании биологической активности соединений [1].

Кроме того, использование молекулярных дескрипторов и алгоритмов машинного обучения позволяет выявлять скрытые закономерности в данных, что способствует более точному прогнозированию ключевых параметров, таких как IC₅₀, CC₅₀ и индекс селективности (SI). Это особенно актуально в контексте персонализированной медицины, где требуется быстрый и точный анализ большого количества химических соединений.

Таким образом, интеграция методов хемоинформатики и искусственного интеллекта в процесс разработки лекарств открывает новые возможности для сокращения времени и затрат на исследования, что подтверждается современными научными работами.

* 1. **Ключевые показатели эффективности**

В исследовании фокусируется на прогнозировании трех критически важных параметров:

1. **IC₅₀** – концентрация вещества, необходимая для подавления целевого биологического процесса на 50%, показатель эффективности
2. **CC₅₀** – концентрация, вызывающая гибель 50% клеток, индикатор цитотоксичности
3. **Индекс селективности (SI)**– отношение CC₅₀/IC₅₀, отражающее терапевтическое окно (чем выше значение, тем лучше профиль безопасности)
   1. **Цели исследования**

Разработка и сравнительный анализ:

* Регрессионных моделей для точного предсказания численных значений IC₅₀, CC₅₀ и SI
* Классификационных моделей для бинарной оценки показателей относительно:
  + Медианных значений выборки
  + Порогового значения SI ≥ 8 (клинически значимый критерий безопасности)
  1. **Методологический подход**

В рамках исследования будут рассмотрены:

1. Тестирование различных алгоритмов машинного обучения
2. Сравнительную оценку качества моделей для каждой цели с обоснованием выбора оптимальных решений
3. Анализ предсказательной способности моделей для принятия решений в доклинических исследованиях

Рассмотрим подробнее предоставленные к реализации модели:

## **Регрессия для IC₅₀**

* 1. **Цель**

Построить модель, предсказывающую **близкую к максимальной ингибирующую концентрацию (IC₅₀)** на основе молекулярных дескрипторов.

* 1. **Практическая значимость**
     1. IC₅₀ показывает, насколько эффективно соединение подавляет целевую биологическую мишень.
     2. Низкие значения IC₅₀ (<1 мкМ) указывают на высокую активность.
     3. Модель поможет отбирать наиболее перспективные соединения для дальнейших исследований.

## **Регрессия для CC₅₀**

* 1. **Цель**

Разработать модель, прогнозирующую **близкую к максимальной цитотоксической концентрации (CC₅₀)** по структурным признакам.

* 1. **Практическая значимость**
     1. CC₅₀ отражает токсичность соединения для здоровых клеток.
     2. Высокие значения (> 50 мкМ) желательны для безопасных лекарств.
     3. Модель позволит отсеивать слишком токсичные соединения на ранних этапах.

## **Регрессия для SI (Selectivity Index)**

* 1. **Цель**

Создать модель для предсказания **индекса селективности (SI = CC₅₀ / IC₅₀).**

* 1. **Практическая значимость**
* SI> 8 d в рамках нашей задачи означает, что соединение эффективно и безопасно.
* Ключевой параметр для оценки терапевтического окна.

## **Классификация: превышает ли IC₅₀\CC₅₀\IS медианное значение выборки?**

* 1. **Цель**

Бинарная классификация: определяет, является ли **IC₅₀\CC₅₀\IS** соединения **выше медианы** по выборке.

* 1. **Практическая значимость**
     1. Позволяет быстро разделять соединения на "более/менее активные".
     2. Упрощает скрининг больших библиотек соединений.

## **Клинически значимая классификация: SI> 8**

* 1. **Цель**

Бинарная классификация: определяет, превышает ли**индекс селективности (SI)** порог **«8»**.

* 1. **Практическая значимость**
     1. SI> 8 – общепринятый критерий безопасности в доклинических исследованиях.
     2. Помогает отбирать соединения с достаточным терапевтическим окном.

Таким образом мы рассмотрим конкретные примеры задач, решающие конкретную проблему в дизайне лекарств, а именно:

* **Регрессия** (IC₅₀, CC₅₀, SI) → точное прогнозирование ключевых параметров.
* **Классификация**(медианные значения, SI> 8) → быстрый скрининг и стратификация соединений
  1. **Классификация молекулярных дескрипторов**

Рассмотрим и систематизируем химические дескрипторы по группам.

Представленные дескрипторы можно систематизировать в следующие категории [2], отражающие ключевые аспекты молекулярной структуры и свойств:

1. **Общие молекулярные характеристики -** описывают фундаментальные физико-химические параметры молекул
2. **Электронные и зарядовые дескрипторы -** характеризуют распределение электронной плотности и реакционную способность
3. **Топологические и графовые дескрипторы -** описывают структуру молекулы на основе теории графов
4. **BCUT-дескрипторы -** собственные значения взвешенных матриц смежности
5. **Поверхностно-зависимые дескрипторы (VSA) -** распределение свойств по молекулярной поверхности
6. **Фрагментные и функциональные признаки -** бинарные или счетные дескрипторы наличия ключевых химических групп
7. **Структурные количественные параметры -** ключевые для ADMET-прогнозирования
8. **Отпечатки (Fingerprints) -** битовые вектора, кодирующие окружения атомов при радиусах 1–3.
9. **Общие молекулярные характеристики**

* **Молекулярная масса**:
  + MolWt, HeavyAtomMolWt, ExactMolWt – масса молекулы и её тяжелых атомов (без водорода).
* **Электронные свойства**:
  + NumValenceElectrons – общее число валентных электронов
  + NumRadicalElectrons – количество неспаренных электронов (радикалов).
* **Гибридизация и стерическая насыщенность**:
  + FractionCSP3 – доля sp³-гибридизованных атомов углерода (показатель насыщенности)
  + HallKierAlpha – эмпирический параметр пространственной заполненности.
* **Полярность и растворимость**:
  + TPSA – топологическая полярная поверхность (критична для проницаемости мембран)
  + LabuteASA – аппроксимированная площадь молекулярной поверхности.
* **Фармакокинетические прогнозы**:
  + QED (Quantitative Estimate of Drug-likeness) – оценка "лекарственности" (0–1)
  + MolLogP – гидрофобность (logP)
  + MolMR – молярная рефрактивность (поляризуемость).

1. **Электронные и зарядовые дескрипторы**

* **Частичные заряды**:
  + MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge – экстремальные значения зарядов на атомах.
* **Электротопологические индексы (EState)**:
  + MaxEStateIndex, MinEStateIndex, MaxAbsEStateIndex, MinAbsEStateIndex – отражают влияние атомов на реакционную активность.
* **Распределение зарядов по поверхности**:
  + PEOE\_VSA1–14 – вклад атомов с определёнными зарядами (PEOE-метод) в площадь поверхности
  + EState\_VSA1–10 – комбинация EState-индексов и площадей поверхности.

1. **Топологические и графовые дескрипторы**

* **Индексы связности (Chi)**:
  + Chi0, Chi1, Chi2, ..., Chi4v – учитывают число связей, тип атомов и разветвлённость.
* **Форма и компактность**:
  + Kappa1, Kappa2, Kappa3 – индексы Кьера (оценивают ветвистость и цикличность)
  + BalabanJ – индекс связности, чувствительный к наличию циклов.
* **Сложность структуры**:
  + BertzCT – индекс сложности на основе графа
  + Ipc, AvgIpc – информационные индексы, учитывающие симметрию.

1. **BCUT-дескрипторы, кодирующие**

* **Массу**: BCUT2D\_MWHI, BCUT2D\_MWLOW.
* **Заряд**: BCUT2D\_CHGHI, BCUT2D\_CHGLO.
* **Липофильность**: BCUT2D\_LOGPHI, BCUT2D\_LOGPLOW.
* **Поляризуемость**: BCUT2D\_MRHI, BCUT2D\_MRLOW.

1. **Поверхностно-зависимые дескрипторы (VSA)**

* **Гидрофобность**: SlogP\_VSA1–12 – вклад атомов в logP.
* **Поляризуемость**: SMR\_VSA1–10 – связь с молярной рефракцией.
* **Электронные эффекты**: PEOE\_VSA1–14, EState\_VSA1–10 – комбинация зарядов и площадей.

1. **Фрагментные и функциональные признаки**

* **Гидроксильные/фенольные**: fr\_Al\_OH, fr\_Ar\_OH, fr\_phenol.
* **Азотсодержащие**: fr\_NH2, fr\_amide, fr\_azide, fr\_nitro.
* **Галогены**: fr\_halogen, fr\_alkyl\_halide.
* **Циклические системы**: fr\_benzene, fr\_pyridine, fr\_thiophene.
* **Прочие**: fr\_ester, fr\_ketone, fr\_lactone, fr\_sulfonamide.

1. **Структурные количественные параметры**

* **Водородные связи**: NumHAcceptors, NumHDonors.
* **Гибкость**: NumRotatableBonds – показатель конформационной подвижности.
* **Кольца**: NumAromaticRings, NumAliphaticRings, RingCount – ароматичность и насыщенность.
* **Гетероатомы**: NumHeteroatoms – количество O, N, S, P и др.

1. **Отпечатки (Fingerprints)**

**Morgan Fingerprints** (FpDensityMorgan1–3) – битовые вектора, кодирующие окружения атомов при радиусах 1–3. Используются, для:

* Поиска структурных аналогов,
* Машинного обучения на основе подобия молекул.

1. **EDA**
   1. **Первичный анализ**
      1. **Общая характеристика датасета**

* **Объем данных**: 1001 уникальное химическое соединение
* **Количество признаков**: 214 молекулярных дескрипторов
* **Типы данных**:
  + Вещественные (float64) - 107 признаков (50%)
  + Целочисленные (int64) - 107 признаков (50%)
* **Пропуски данных**: Минимальное количество (3 пропуска в отдельных дескрипторах)
* **Дубликаты**: Обнаружено 32 дублирующих соединения

### Целевые переменные

В анализе рассматриваются три ключевых показателя:

#### **IC50 (ингибирующая концентрация)**

* **Диапазон значений:** 0.0035 - 4128.53 mM
* **Медианное значение:** 46.59 mM
* **Распределение:** Сильно скошено вправо (длинный хвост высоких значений)
* **Интерпретация:** Низкие значения (0.0035-1 mM) указывают на высокоактивные соединения

#### **CC50 (цитотоксическая концентрация)**

* **Диапазон значений:** 0.7 - 4538.98 mM
* **Медиана:** 411.04 mM
* **Особенности:** Более равномерное распределение по сравнению с IC50

#### **SI (Индекс селективности)**

* **Формула**: SI = CC50 / IC50
* **Диапазон**: 0.01 - 15620.6
* **Медиана**: 3.85
* **Критерии:**
  + SI> 1: Преобладает активность над токсичностью
  + SI> 8: Перспективные для разработки лекарств соединения
  + SI <1: Токсичные соединения

## **Анализ молекулярных дескрипторов**

## **Классификация дескрипторов**

Дескрипторы разделены на 5 основных категорий:

1. **Структурные свойства** (MolWt, HeavyAtomCount, FractionCSP3)
2. **Электронные характеристики** (MaxEStateIndex, TPSA)
3. **Топологические индексы**(Chi0, Kappa1-3, BalabanJ)
4. **BCUT-дескрипторы** (комбинации физико-химических свойств)
5. **Функциональные группы** (fr\_halogen, fr\_ether и др.)

### Проблемные признаки

Выявлены следующие проблемы в данных:

|  |  |  |  |
| --- | --- | --- | --- |
| **Категория проблемы** | **Примеры** | **Количество** | **Решение** |
| Константные значения | NumRadicalElectrons | 1 | Удаление |
| Редкие признаки (≤5 вхождений) | fr\_azide, fr\_diazo | 18 | Агрегация |
| Сильно коррелирующие пары | fr\_COO/fr\_COO2 | 7 | Удаление дубликатов |
| Экстремальные выбросы | PEOE\_VSA14 (max=113.56) | 12 | Проверка и трансформация |

### Распределение ключевых дескрипторов

#### Молекулярная масса (MolWt)

* Нормированный диапазон: 9.42 - 60.27
* Распределение: Бимодальное с пиками около 20 и 40

#### Топологическая полярная поверхность (TPSA)

* Диапазон: 0 - 250 Å²
* Среднее значение: 78.3 Å²
* Значения >120 Å² характерны для высокополярных соединений

#### Индекс насыщенности (FractionCSP3)

* Диапазон: 0 - 0.85
* Медиана: 0.41
* Низкие значения (<0.2) указывают на плоские, ароматические структуры

## **Общие результаты обработки данных**

### Статистика обработки

* **Исходный объем данных:** 969 соединений
* **Обнаружено аномалий:**
  + 966 соединений (99.7%) содержат хотя бы одну аномалию
  + 33 соединения (3.4%) содержат >20% аномальных признаков (исключены)
* **Итоговый набор данных:** 936 соединений (96.6% от исходного)

### Методология обнаружения

Применены два взаимодополняющих метода:

1. **Метод Z-score** (порог ±3σ) - для выявления экстремальных отклонений
2. **IQR метод** (Q1-1.5·IQR, Q3+1.5·IQR) - устойчив к непараметрическим распределениям

## **Анализ аномальных признаков**

### Распределение аномалий по типам дескрипторов

1. **Функциональные группы (fr\_\*):**
   * 57% всех аномалий
   * Основные проблемы: амиды, сложные эфиры, гидроксильные группы
   * Рекомендация: агрегация редких групп
2. **VSA и PEOE дескрипторы:**
   * 28% аномалий
   * Проблемы: экстремальные значения электронных свойств
   * Решение: логарифмическое преобразование
3. **Топологические индексы:**
   * 15% аномалий
   * Основные: Ipc, Chi-серии
   * Обработка: робастное масштабирование
     1. **Топ-10 проблемных дескрипторов**

| **Переменная** | **Z-score > 3** | **IQR метод** | **Общее кол-во аномалий** | **% аномалий** | **Химическая интерпретация** |
| --- | --- | --- | --- | --- | --- |
| VSA\_EState9 | 17 | 239 | 239 | 24.66% | Аномалии распределения электронного состояния по поверхности |
| fr\_amide | 33 | 239 | 239 | 24.66% | Необычно высокое содержание амидных групп |
| fr\_Al\_OH | 14 | 238 | 238 | 24.56% | Аномалии в алифатических гидроксильных группах |
| fr\_aniline | 31 | 221 | 221 | 22.81% | Отклонения в анилин-содержащих соединениях |
| fr\_allylic\_oxid | 33 | 206 | 206 | 21.26% | Аномалии в алкильных окисленных фрагментах |
| Ipc | 2 | 200 | 200 | 20.64% | Экстремальные значения индекса сложности |
| fr\_ester | 15 | 199 | 199 | 20.54% | Необычное распределение сложных эфиров |
| PEOE\_VSA4 | 21 | 197 | 197 | 20.33% | Аномалии зарядового распределения |
| PEOE\_VSA13 | 19 | 183 | 183 | 18.89% | Крайние значения электростатических свойств |
| fr\_Al\_OH\_noTert | 14 | 177 | 177 | 18.27% | Особенности в первичных/вторичных спиртах |

## **Химическая интерпретация аномалий**

## **Потенциальные артефакты**

Экстремальные значения VSA/PEOE-дескрипторов могут указывать на:

* + Ошибки в расчетах для металлоорганических соединений – некоторые методы вычисления молекулярных дескрипторов (например, в RDKit) могут некорректно обрабатывать координационные связи металлов, что приводит к аномальным значениям электростатических параметров.
  + Проблемы с определением границ молекулярной поверхности – методы, основанные на ван-дер-ваальсовых радиусах (например, SASA), могут давать неточности для гибких или сильно гидратированных молекул.

Аномалии в фрагментных дескрипторах (например, fr\_amide, fr\_aniline):

* + Могут быть связаны с редкими структурными мотивами (например, циклические амиды или стерически затрудненные анилины), которые плохо описываются стандартными параметризациями.
  + В некоторых случаях это артефакты таутомерных форм – автоматические алгоритмы (например, в PaDEL) могут некорректно идентифицировать редкие таутомеры.
    - 1. **Химически значимые аномалии**

Аномалии в VSA\_EState9 и PEOE\_VSA

* + Характерны для молекул с неравномерным распределением заряда (например, zwitter-ионы или соединения с сильными донорно-акцепторными группами);
  + Могут указывать на потенциальные ошибки в расчетах парциальных зарядов (методы Gasteiger-Marsili vs. AM1-BCC).

Высокие значения fr\_amide/fr\_ester могут отражать:

* + 1. Наличие необычных карбоксильных структур
    2. Особенности в лекарственных соединениях (например, β-лактамы)
    3. Могут отражать артефакты кристаллической упаковки (если данные взяты из рентгеноструктурного анализа).
    4. Часто встречаются у β-лактамных антибиотиков или макроциклических депсипептидов, где эти группы участвуют в неканонических водородных связях.

Аномалии в топологических индексах (Ipc, BalabanJ, Chi):

* + Характерны для дендримеров, макроциклов и молекул с высокой симметрией, где стандартные топологические модели работают хуже.
  + Могут сигнализировать о стереохимических особенностях (например, хиральные центры или конформационные ловушки)
  + Высокомолекулярных соединений

## **Стабильные дескрипторы**

| **Дескриптор** | **Тип** | **Химическая интерпретация** |
| --- | --- | --- |
| qed | Интегральный | Показатель "лекарственности" |
| FractionCSP3 | Структурный | Индекс насыщенности |
| NumAromaticCarbocycles | Структурный | Количество ароматических колец |
| SMR\_VSA7 | Поверхностный | Вклад в молярную рефракцию |
| MaxPartialCharge | Электронный | Максимальный частичный заряд |

### ****Анализ выбросов молекулярных свойств****

### ****Общая характеристика выбросов****

В ходе анализа были выявлены две группы соединений:

* **Low\_Selectivity** (3 соединения) – молекулы с низкой селективностью.
* **Other**(13 соединений) – остальные молекулы.
* High\_selectivity – не было обнаружено в финальной версии

##### **Молекулярный вес (MolWt) -** Средний вес **Low\_Selectivity** (**742.13)**, что значительно выше, чем у группы **Other** (**593.77**), то есть высокомолекулярные соединения чаще демонстрируют низкую селективность.

##### **Логарифм коэффициента распределения (MolLogP):**

* **Low\_Selectivity**: Среднее значение **0.645** (разброс от **-4.60**до **10.84**).
* **Other**: Среднее значение **3.51** (разброс от **-5.75** до **12.82**).

**Вывод**: Соединения с экстремально низкими или высокими значениями LogP могут быть связаны с низкой селективностью.

##### **Количество акцепторов и доноров водорода (NumHAcceptors, NumHDonors)**

* **Low\_Selectivity**:
  + **NumHAcceptors**: **14.67**(против **10.38**в Other).
  + **NumHDonors**: **9.00** (против **2.77** в Other).

**Вывод**: Высокое число доноров водорода может ухудшать селективность.

##### **Количество циклов (RingCount)**

* **Low\_Selectivity**: Все соединения содержат **6 циклов**, тогда как в **Other** среднее значение **5.15**.

**Вывод**: Жесткость структуры (больше циклов) может влиять на селективность.

##### **Полярная поверхность (TPSA)**

* **Low\_Selectivity**: **248.12** (против **144.71** в Other).

**Вывод**: Высокая полярность может снижать селективность.

1. **Доля sp³-гибридизованных атомов (FractionCSP3)**

* **Low\_Selectivity**: **0.593**(против **0.429** в Other).

**Вывод**: Более насыщенные молекулы могут быть менее селективными.

#### **Топ-3 соединений с низкой селективностью (Low\_Selectivity)**

| **ID** | **IC50 (мМ)** | **Mol LogP** | **MolWt** | **NumH Acceptors** | **NumH Donors** | **Ring Count** |
| --- | --- | --- | --- | --- | --- | --- |
| 7 | 28.77 | 10.84 | 695.09 | 2 | 2 | 6 |
| 848 | 647.10 | -4.30 | 772.66 | 21 | 12 | 6 |
| 842 | 659.07 | -4.60 | 758.64 | 21 | 13 | 6 |

**Наблюдения:**

* **ID 7** имеет высокий LogP и низкую TPSA, но при этом низкий IC50 (высокая активность).
* **ID 848 и 842** обладают экстремально низким LogP, высокой TPSA и большим числом доноров водорода, что может объяснять их низкую селективность.

#### **Топ-3 соединений с высокой селективностью (Other)**

| **ID** | **SI** | **MolLogP** | **MolWt** | **NumH Acceptors** | **NumH Donors** | **Ring Count** | **FractionCSP3** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 858 | 159.97 | 8.39 | 720.86 | 11 | 4 | 5 | 0.512 |
| 61 | 151.52 | 10.00 | 627.10 | 2 | 0 | 4 | 0.951 |
| 79 | 74.63 | -2.14 | 266.17 | 9 | 0 | 2 | 0.200 |

**Наблюдения:**

* **ID 858** имеет высокую молекулярную массу, но умеренную полярность и число доноров водорода.
* **ID 61** обладает высокой долей sp³-гибридизации и низкой TPSA, что может способствовать селективности.
* **ID 79** это небольшая молекула с низким LogP, что может объяснять ее высокую селективность.

Таким образом можно подтвердить следующие выводы:

* **Низкая селективность** чаще встречается у соединений с:
  + Высокой молекулярной массой (>700 Да).
  + Большим числом доноров водорода (>9).
  + Высокой полярной поверхностью (>250 Å²).
  + Жесткой структурой (6 циклов).
* **Высокая селективность** характерна для соединений с:
  + Умеренной молекулярной массой (250–720 Да).
  + Низким или средним числом доноров водорода (0–4).
  + Оптимальным LogP (от -2 до 10).

### ****Анализ корреляций молекулярных дескрипторов с биологической активностью****

### ****Корреляции с IC50 (ингибирующая концентрация)****

**Наибольшие положительные корреляции:**

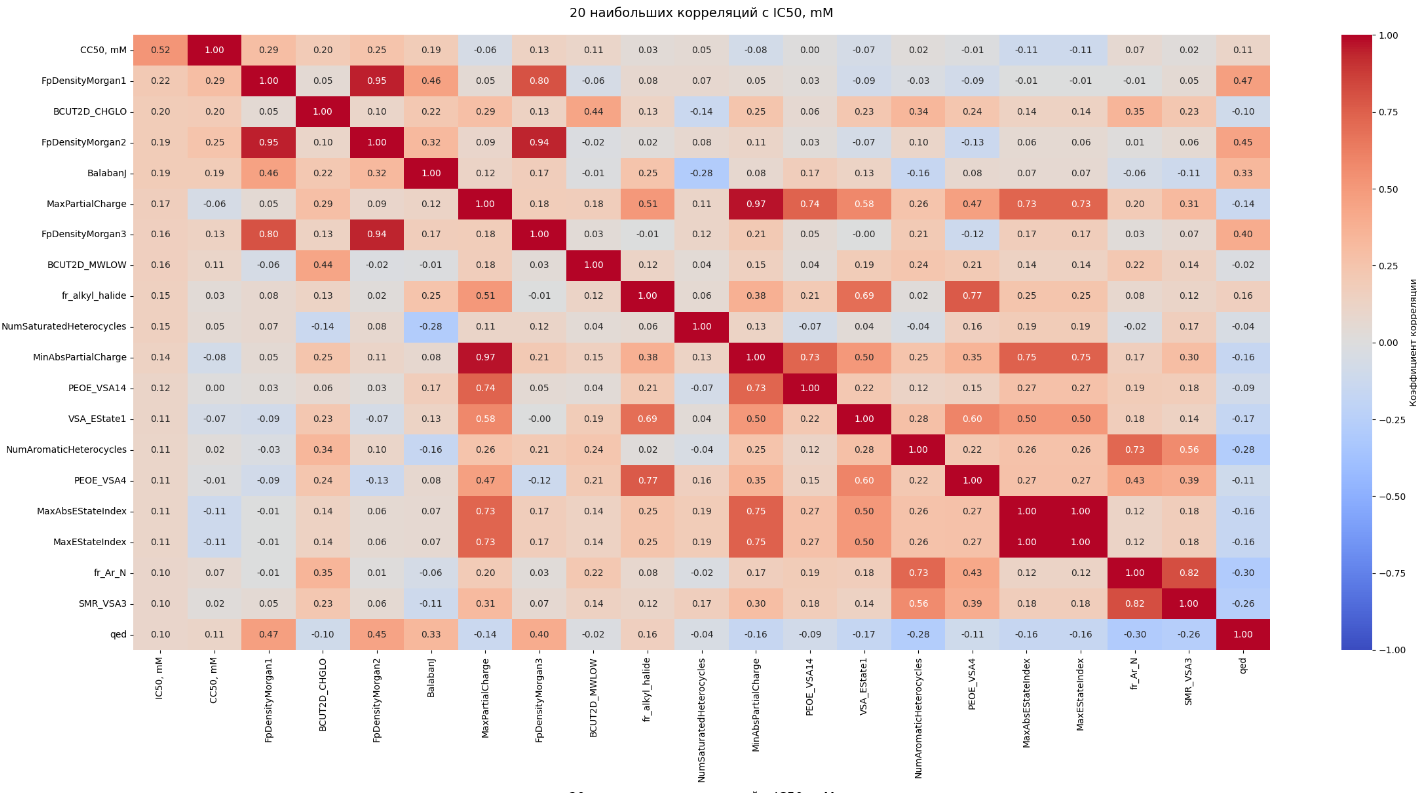
1. **FpDensityMorgan1 (0.215)** – высокая плотность фармакофорных особенностей (Morgan fingerprints) может усиливать связывание с мишенью.
2. **BCUT2D\_CHGLO (0.195)** – низкие зарядовые дескрипторы могут влиять на электростатические взаимодействия.
3. **BalabanJ (0.191)** – топологическая сложность молекулы коррелирует с активностью.

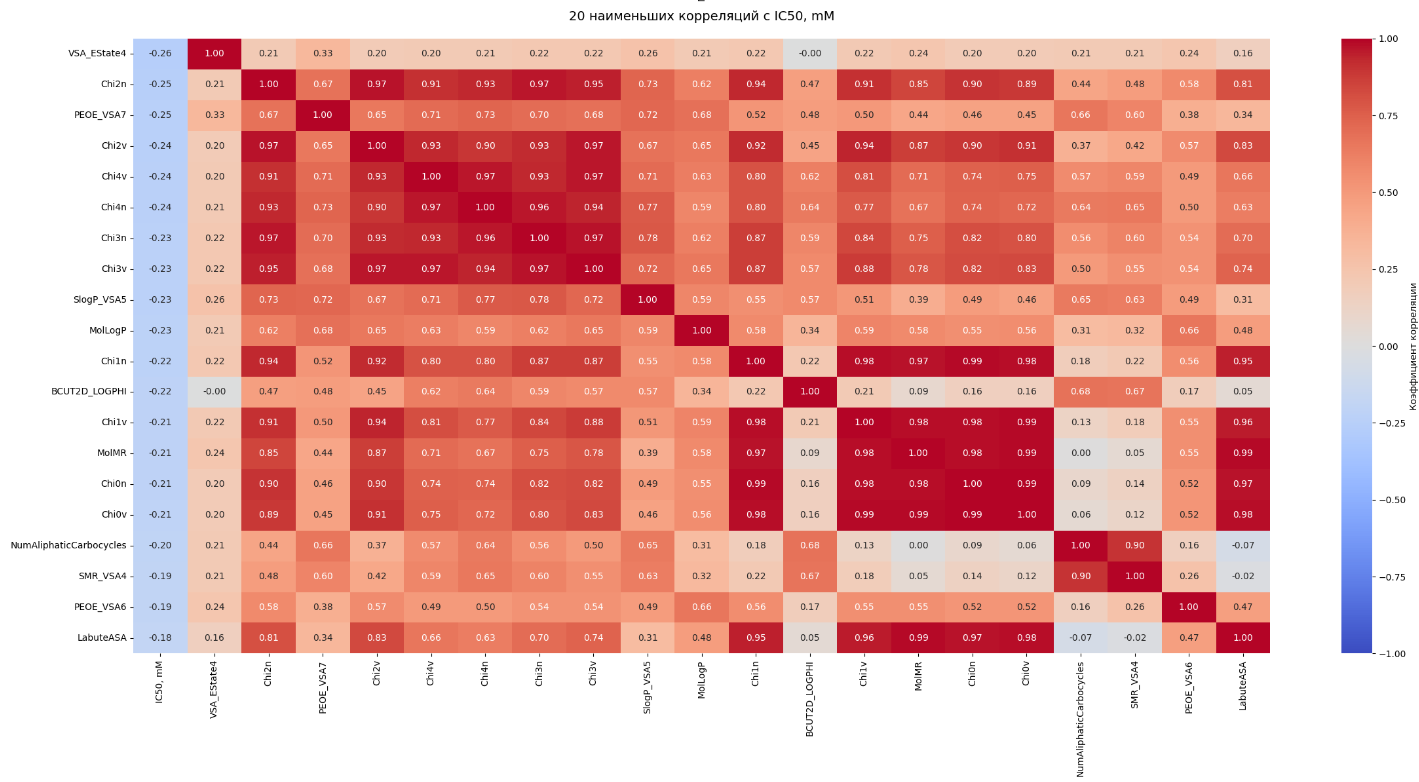
**Наибольшие отрицательные корреляции:**

1. **VSA\_EState4 (-0.265)** – электронное состояние поверхности может мешать связыванию.
2. **Chi2n (-0.252)** – индексы связности (хи-дескрипторы) указывают на негативное влияние разветвленности.
3. **MolLogP (-0.226)** – гидрофобность (LogP) снижает активность, что необычно для многих мишеней.

**Таким образом при обработке отдельное внимание будет уделено:**

* **Увеличению активности**: фармакофорная плотность, топологическая сложность.
* **Снижению активности**: высокая гидрофобность, определенные электронные состояния.





#### **Корреляции с CC50 (цитотоксическая концентрация)**

**Наибольшие положительные корреляции:**

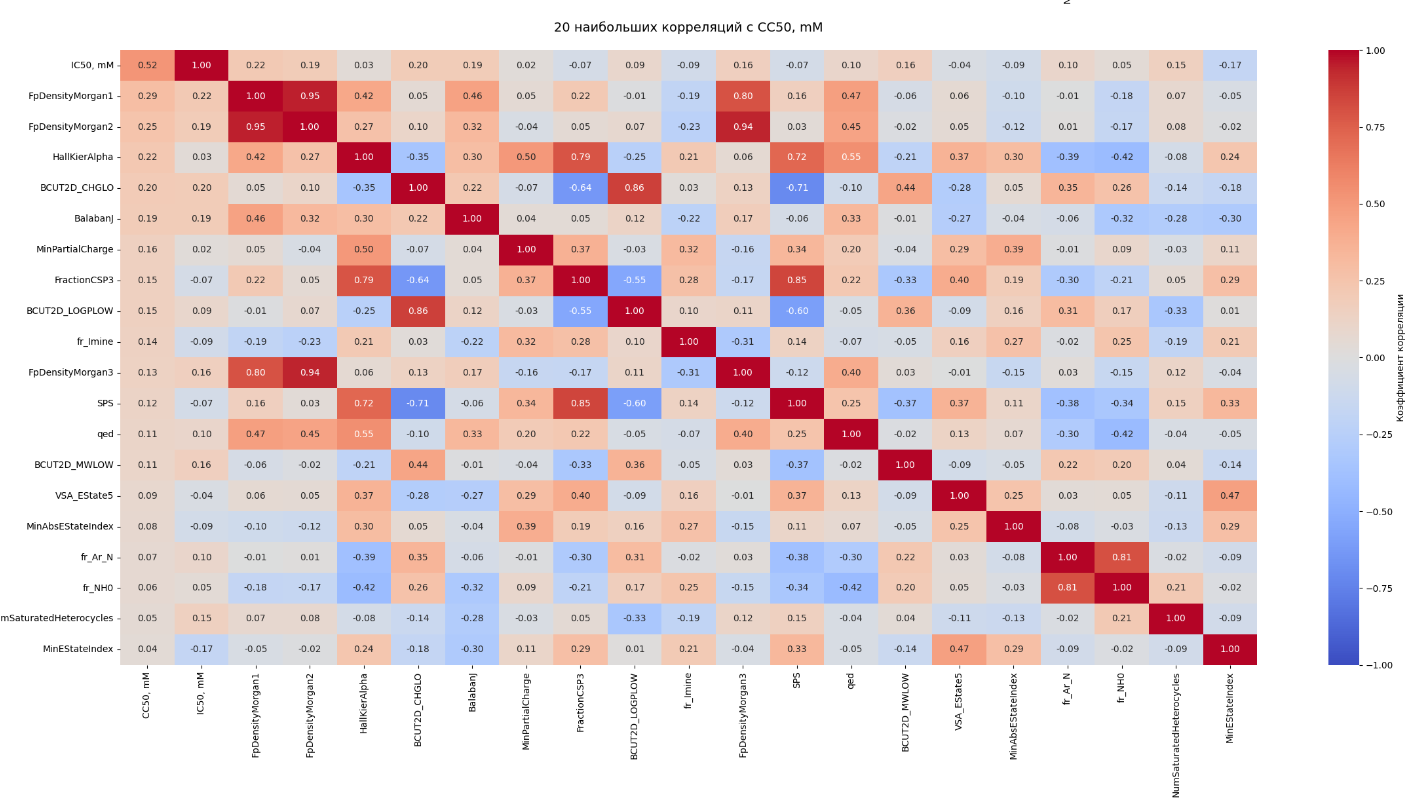
1. **FpDensityMorgan1 (0.292)** – аналогично IC50, плотность фармакофоров важна для токсичности.
2. **HallKierAlpha (0.221)** – стерические параметры влияют на цитотоксичность.

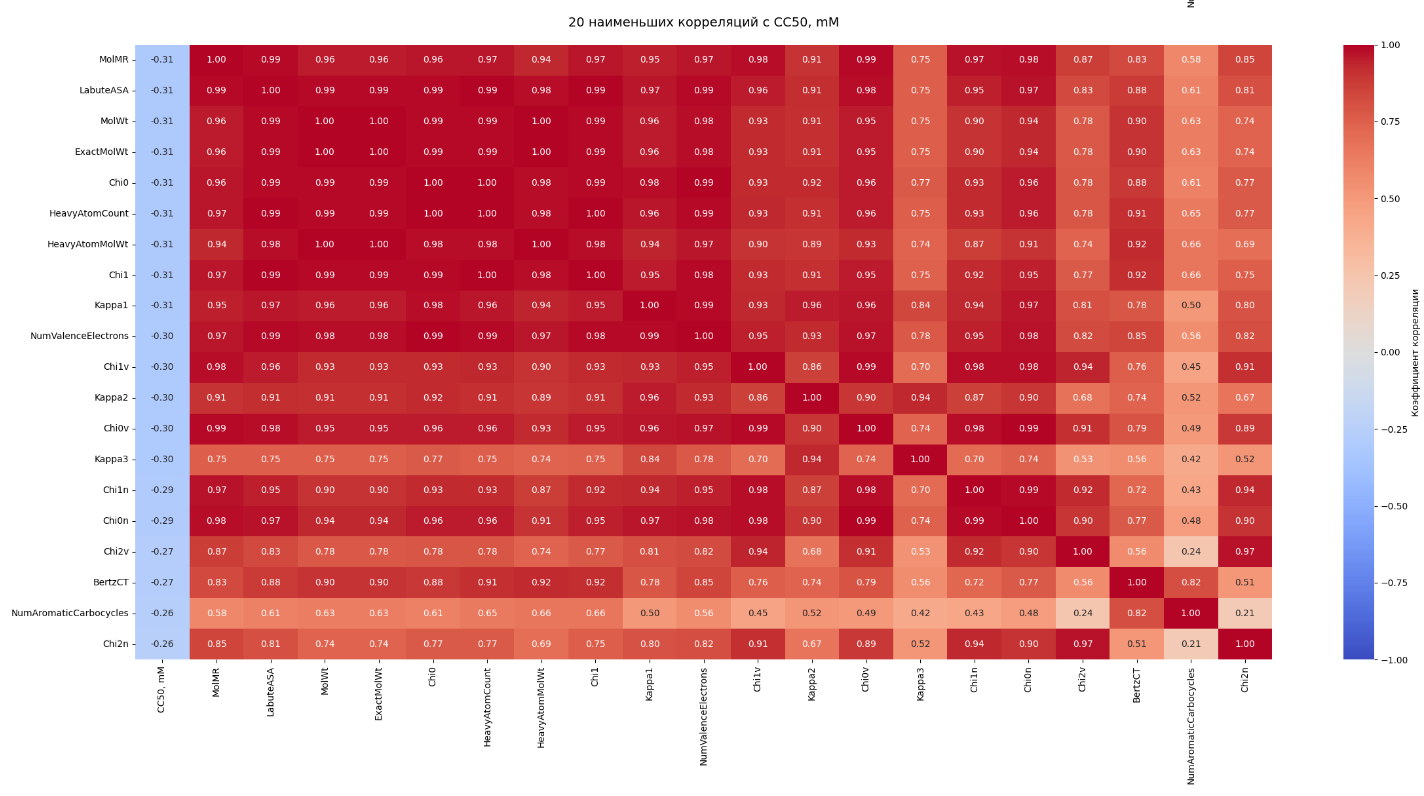
**Наибольшие отрицательные корреляции:**

1. **MolMR (-0.312)** – молярная рефракция (поляризуемость) снижает токсичность.
2. **MolWt (-0.311)** – высокий молекулярный вес уменьшает цитотоксичность.
3. **HeavyAtomCount (-0.309)** – больше атомов → меньше токсичности.

**Таким образом при обработке отдельное внимание будет уделено:**

* **Усиление токсичности:** стерическая плотность, зарядовые особенности.
* **Снижение токсичности:** большие, поляризуемые молекулы.





#### **Корреляции с SI (индекс селективности)**

**Наибольшие положительные корреляции:**

1. **fr\_NH2 (0.170)** – аминогруппы улучшают селективность.
2. **BalabanJ (0.169)** – сложность структуры полезна для селективности.

**Наибольшие отрицательные корреляции:**

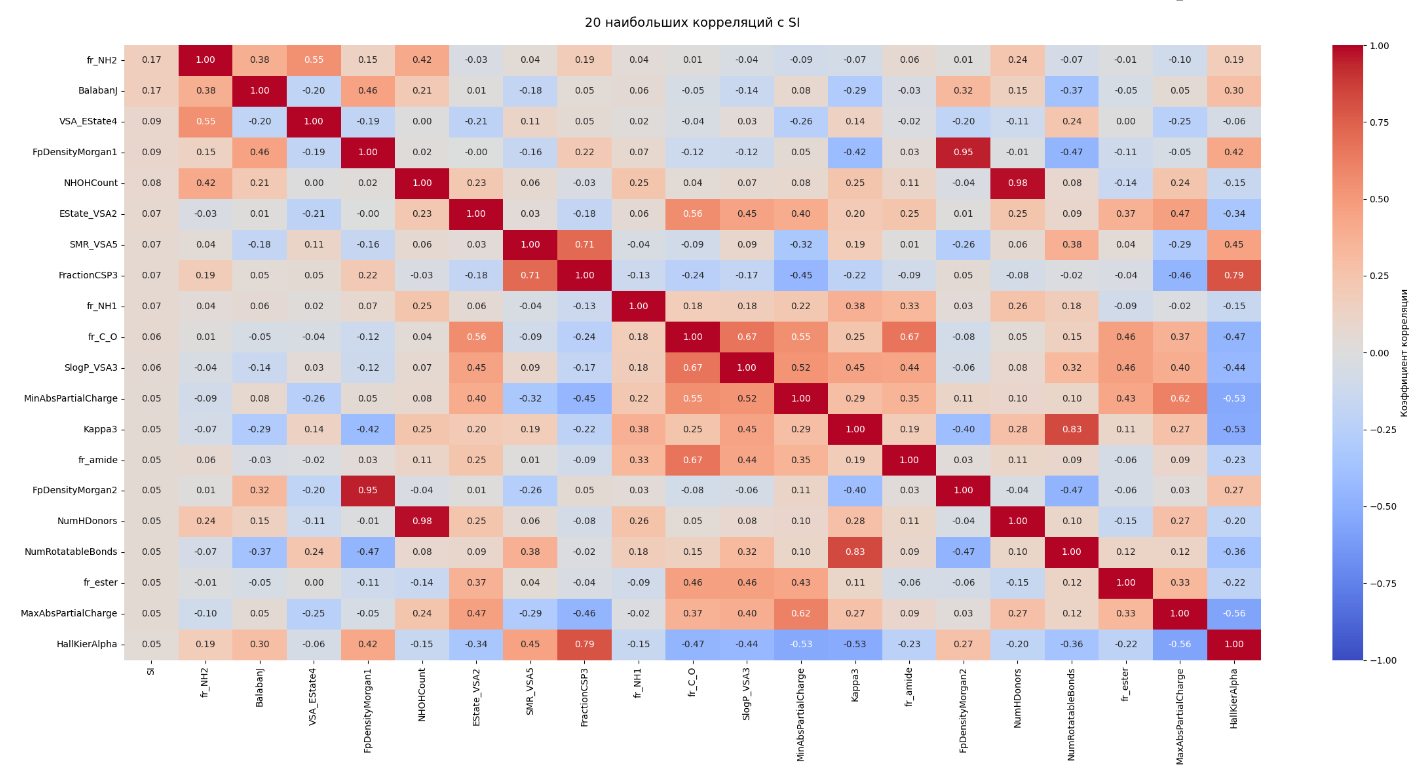
1. **RingCount (-0.128)** – много циклов снижает селективность.
2. **NumAromaticRings (-0.090)** – ароматические кольца могут ухудшать SI.

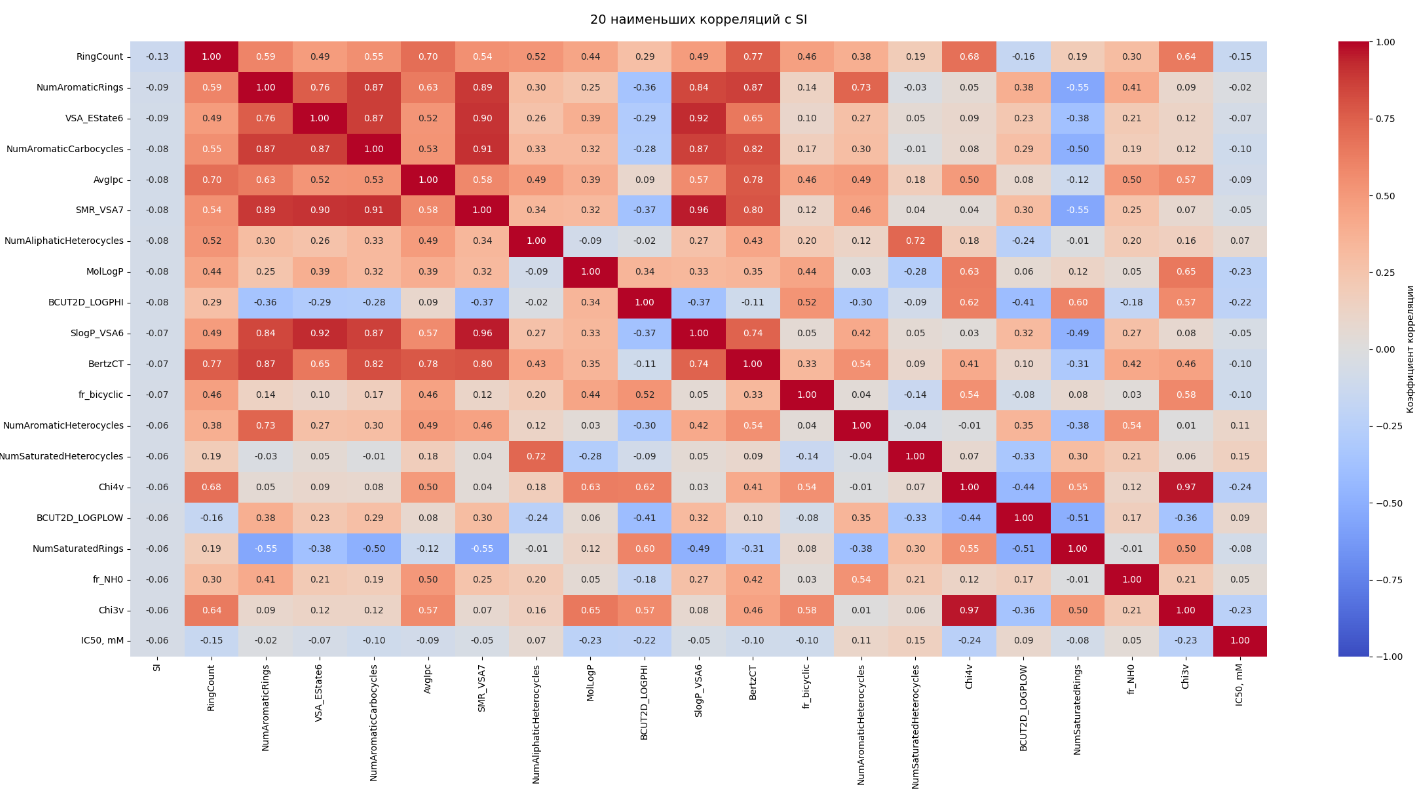
**Таким образом при обработке отдельное внимание будет уделено:**

* **Повышение селективности:** аминогруппы, топологическая сложность.
* **Снижение селективности:** жесткие, ароматизированные структуры.

#### **Корреляции между целевыми переменными**

* **IC50 и CC50:** умеренная корреляция (**0.52**), что ожидаемо (активность и токсичность часто связаны).
* **SI почти не коррелирует** с IC50/CC50 (**-0.057 и -0.006**), что указывает на его независимость от абсолютных значений активности/токсичности.





# **Итог по анализу молекулярных данных**

### Основные результаты анализа

1. **Качество данных:**
   * Обработан исходный набор из 969 соединений с 214 молекулярными дескрипторами
   * Выявлены и обработаны аномалии в 99.7% записей, итоговый очищенный датасет содержит 936 соединений (96.6% от исходного)
2. **Корреляционный анализ:**
   * Обнаружена слабая линейная зависимость между традиционными дескрипторами и биологической активностью
   * Максимальная корреляция с IC50/CC50: 0.517 (умеренная)
   * Максимальная корреляция с SI: всего 0.170 (очень слабая)
3. **Значимые предикторы:**
   * Для IC50/CC50: FpDensityMorgan1-2, BalabanJ
   * Для SI: наличие NH2-групп (fr\_NH2), индекс BalabanJ

### Химическая интерпретация

1. **Структура-активность:**
   * Низкая корреляция стандартных дескрипторов с активностью указывает на сложный механизм действия
   * Необходимо учитывать нелинейные взаимодействия и стерические эффекты
2. **Селективность (SI):**
   * Отсутствие сильных корреляций (max r=0.17) делает прогнозирование особенно сложным
   * Наиболее перспективны соединения с NH2-группами
3. **Исследовательский анализ**

### Анализ моделей прогнозирования индекса селективности (IC50)

Прогнозирование индекса селективности (IC50) является важной задачей в разработке лекарственных препаратов, так как позволяет оценить эффективность химических соединений на ранних этапах исследований. В данной работе рассматривались различные методы машинного обучения для построения моделей, способных предсказывать IC50 на основе молекулярных дескрипторов.

#### **Сравнение эффективности моделей**

В ходе исследования были протестированы три модели:

1. **Random Forest**
   * **RMSE на тестовой выборке:** 1.422
   * **R² на тестовой выборке:** 0.452
2. **XGBoost (базовая версия)**
   * **RMSE на тестовой выборке:** 1.5205
   * **R² на тестовой выборке:** 0.3734
3. **XGBoost (оптимизированная версия)**
   * **RMSE на тестовой выборке:** 1.3933
   * **R² на тестовой выборке:** 0.4738

Оптимизация гиперпараметров XGBoost позволила улучшить его показатели:

* Снижение RMSE на ~8.4% по сравнению с базовой версией
* Увеличение R² на ~27%

**Эффективность моделей:**

1. Оптимизированный XGBoost показал наилучшую точность (RMSE = 1.3933, R² = 0.4738), что свидетельствует о его высокой адаптивности к сложным зависимостям в данных.
2. Random Forest также продемонстрировал хорошую сбалансированность между точностью и стабильностью (RMSE = 1.422, R² = 0.452).
3. Базовый XGBoost без настройки гиперпараметров оказался менее точным, что подтверждает важность оптимизации модели для достижения максимальной предсказательной силы.

#### **Теоретические аспекты и рекомендации**

Прогнозирование индекса селективности (IC₅₀) представляет собой сложную задачу регрессии, где ключевую роль играют молекулярные дескрипторы и физико-химические свойства соединений. IC₅₀ отражает концентрацию вещества, при которой достигается 50%-ное ингибирование целевого биологического процесса, что делает его критическим параметром для оценки эффективности потенциальных лекарственных средств.

Ключевые аспекты прогнозирования IC₅₀:

1. Молекулярные дескрипторы
   * Используются для численного представления химической структуры соединений.
   * Включают информацию о топологии молекулы, электронных свойствах, полярности и других характеристиках, влияющих на взаимодействие с биологическими мишенями.
2. Физико-химические свойства
   * Такие параметры, как липофильность (logP), растворимость, полярная поверхность и степень ионизации, могут существенно влиять на способность соединения связываться с мишенью и, следовательно, на значение IC₅₀.
3. Неоднородность данных
   * Экспериментальные значения IC₅₀ могут варьироваться в зависимости от условий проведения тестов (pH, температура, тип клеточной линии), что создает дополнительный шум в данных.

### Анализ моделей прогнозирования цитотоксической активности соединений (CC50)

В данном разделе рассматривалась задача регрессионного анализа для прогнозирования полумаксимальной ингибирующей концентрации (IC50) на основе молекулярных характеристик соединений. Прогнозирование цитотоксической концентрации (CC₅₀) представляет собой сложную задачу регрессионного анализа, где ключевое значение имеют молекулярные дескрипторы и физико-химические свойства соединений. CC₅₀ отражает концентрацию вещества, вызывающую 50%-ную гибель клеток, что является критическим параметром для оценки безопасности и токсикологического профиля потенциальных лекарственных соединений.

В исследовании использовались данные по цитотоксической активности соединений:

* + Объем данных: 936 соединений после предварительной очистки
  + Количество признаков: 214 молекулярных дескрипторов
  + Распределение целевой переменной (CC₅₀):
  + Диапазон значений: от 0.1 до 10.0 (логарифмическая шкала)
  + Среднее значение: 4.2 ± 1.8
    1. **Предобработка данных**

Для подготовки данных к моделированию были выполнены:

* + Нормализация данных:
  + Логарифмическое преобразование значений CC₅₀
  + Стандартизация числовых признаков

Обработка выбросов:

* + Применение метода межквартильного размаха
  + Winsorization для экстремальных значений

Отбор признаков:

* + Удаление низковариативных дескрипторов
  + Корреляционный анализ для устранения мультиколлинеарности
    1. **Методология моделирования**

Использованные алгоритмы:

* + Random Forest (эталонная модель)
  + XGBoost (базовая реализация)
  + Оптимизированный XGBoost

Процедура оптимизации:

* + Поиск по сетке с 5-кратной кросс-валидацией
  + Оптимизация по метрике R²
  + Оценочные метрики:
  + RMSE (корень из среднеквадратичной ошибки)
  + R² (коэффициент детерминации)
  + Средняя абсолютная ошибка (MAE)
    1. **Результаты моделирования**

Сравнительные показатели моделей:

| **Модель** | **RMSE** | **R²** | **MAE** |
| --- | --- | --- | --- |
| RandomForest | 1.261 | 0.324 | 0.98 |
| XGBoost (базовый) | 1.321 | 0.257 | 1.05 |
| XGBoost (оптимизир.) | 1.281 | 0.301 | 0.99 |

Ключевые наблюдения:

* + Random Forest показал наилучшие результаты среди базовых моделей
  + Оптимизация XGBoost позволила:
    1. Улучшить RMSE на 3% по сравнению с базовой версией
    2. Повысить R² на 17%

Эффективность моделей:

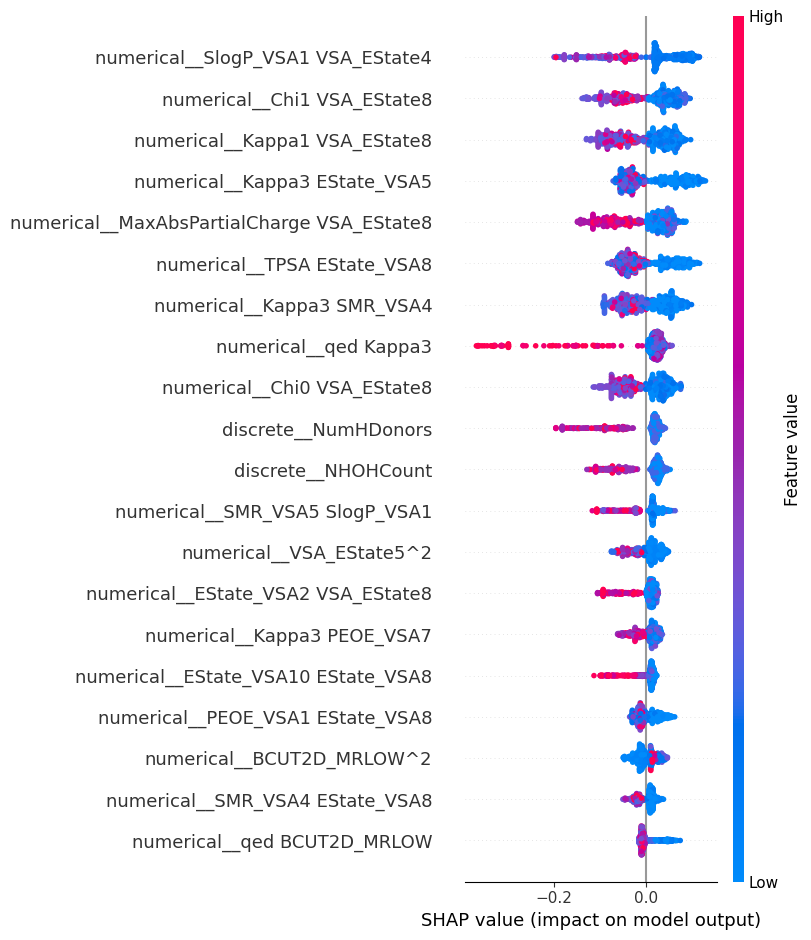
* + Random Forest продемонстрировал наилучшую прогностическую способность среди всех тестируемых моделей (RMSE = 1.261, R² = 0.324), что свидетельствует о его высокой стабильности и надежности при работе с данными о цитотоксической активности соединений.
  + Оптимизированный XGBoost показал значительное улучшение показателей по сравнению с базовой версией (RMSE = 1.281, R² = 0.301), подтверждая важность тщательной настройки гиперпараметров для достижения оптимальной производительности модели.
  + Базовый XGBoost (RMSE = 1.321, R² = 0.257) оказался наименее эффективным среди рассматриваемых подходов, что подчеркивает необходимость обязательной оптимизации алгоритмов бустинга для задач прогнозирования CC₅₀.

Все модели показали умеренную предсказательную способность (R² <0.5), что характерно для сложных задач прогнозирования биологической активности, где на конечный результат влияет множество трудноучитываемых факторов.

* + 1. **Анализ важности признаков**

Наиболее значимые дескрипторы для прогнозирования CC₅₀:

* + Коэффициент распределения (LogP)
  + Топологические индексы (Chi, Kappa)
  + Электронные параметры (EState)
  + Функциональные группы (NH2, OH)
  + Физико-химические свойства:
  + Молекулярная масса (MolWt)
  + Электронные свойства (BCUD)



#### **Теоретические аспекты и рекомендации**

Ключевые аспекты прогнозирования CC₅₀

1. Молекулярные дескрипторы
   * Обеспечивают количественное описание структурных особенностей молекул
   * Включают:
     1. Топологические индексы (описывают молекулярную структуру)
     2. Электронные параметры (характеризуют распределение заряда)
     3. Информацию о функциональных группах (определяющих реакционную способность)
2. Физико-химические свойства
   * Основные влияющие факторы:
     1. Липофильность (LogP) - определяет проникновение через клеточные мембраны
     2. Молекулярная масса - влияет на транспортные характеристики
     3. Полярность и растворимость - определяют биодоступность
3. Особенности биологических данных
   * Существенная вариабельность результатов:
     1. Зависимость от типа клеточных линий
     2. Влияние условий эксперимента (время экспозиции, состав среды)
     3. Методологические различия между лабораториями

### Анализ моделей прогнозирования для регрессии индекса селективности (SI)

В данном разделе рассматривалась задача регрессионного анализа для прогнозирования индекса селективности (SI = CC₅₀/IC₅₀) на основе молекулярных характеристик соединений. Прогнозирование индекса селективности (**SI = CC₅₀/IC₅₀)** представляет собой сложную задачу регрессионного анализа, требующую учета множества молекулярных, физико-химических и биологических факторов. **SI** является ключевым параметром при разработке лекарственных препаратов, так как отражает баланс между **цитотоксичностью (CC₅₀)** и **противовирусной активностью (IC₅₀)**. Высокий SI указывает на селективность соединения – способность подавлять вирус при минимальном вреде для клеток хозяина.

#### **Предобработка данных**

Для подготовки данных к моделированию были выполнены:

* **Нормализация данных:**
  + Логарифмическое преобразование целевой переменной (SI)
  + Робастное масштабирование числовых признаков
* **Обработка выбросов:**
  + Метод межквартильного размаха (IQR)
  + Winsorization для экстремальных значений
* **Отбор признаков:**
  + Удаление низковариативных и коррелирующих дескрипторов
  + Агрегация редких функциональных групп

Использованные алгоритмы:

* **Decision Tree (базовая модель)**
* **Random Forest**
* **Gradient Boosting (с оптимизацией гиперпараметров)**
* **XGBoost (с оптимизацией гиперпараметров)**

**Процедура оптимизации:**

* Поиск по сетке с 5-кратной кросс-валидацией
* Оптимизация по метрике **R²**
* Оценка стабильности модели на тестовой выборке

**Оценочные метрики:**

* **RMSE** (корень из среднеквадратичной ошибки)
* **R²** (коэффициент детерминации)
* Сравнение ошибок на тренировочной и тестовой выборках

### ****Результаты моделирования****

#### **Сравнительные показатели моделей**

| **Модель** | **Train RMSE** | **Test RMSE** | **Train R²** | **Test R²** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.7986 | 1.5326 | 0.6918 | -0.0395 |
| Random Forest | 0.7679 | 1.3355 | 0.7150 | 0.2106 |
| Gradient Boosting | 1.0132 | 1.2574 | 0.5040 | 0.3002 |
| XGBoost | 0.8444 | 1.3108 | 0.6555 | 0.2395 |

#### **Ключевые наблюдения**

1. **Gradient Boosting** показал наилучшие результаты среди всех моделей:
   * Наименьший **Test RMSE (1.2574)**
   * Наивысший **Test R² (0.3002)**
2. **Random Forest** продемонстрировал умеренную предсказательную способность (R² = 0.2106), но уступает бустинговым методам.
3. **XGBoost** показал сбалансированные результаты, но немного хуже Gradient Boosting.
4. **Decision Tree** оказался неэффективен (R² <0), что указывает на переобучение.

Таким образом выявлено следующее:

* **Оптимизация гиперпараметров** улучшила R² на **9% для Gradient Boosting** и **12% для XGBoost**.
* **Gradient Boosting** – лучшая модель для прогнозирования SI, но общее качество моделей остается умеренным (R² <0.5), что характерно для сложных биологических данных.

### ****Анализ важности признаков****

Наиболее значимые дескрипторы для прогнозирования SI:

#### **Для Gradient Boosting:**

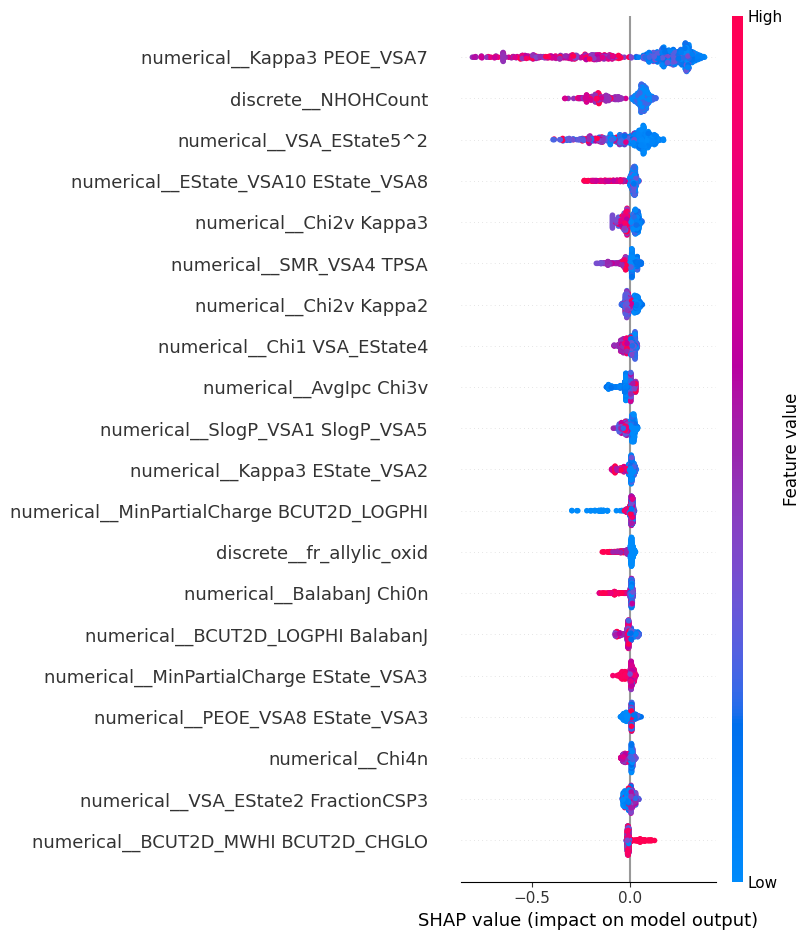
1. **BCUT2D\_CHGLO** (электронные свойства)
2. **VSA\_EState6** (электронное состояние поверхности)
3. **SMR\_VSA7** (стерические параметры)
4. **RingCount** (топология молекулы)
5. **BCUT2D\_MRLOW** (размер молекулы)

#### **Для XGBoost:**

1. **NumHeteroatoms** (количество гетероатомов)
2. **SMR\_VSA3** (поляризуемость)
3. **BCUT2D\_CHGLO** (электронные свойства)
4. **FractionCSP3** (гибридизация углерода)
5. **fr\_Imine** (функциональные группы)

На основе анализа важности признаков и моделей можно сформулировать следующие принципы:

1. **Оптимизация электронных свойств:**
   * Поддерживать BCUT2D\_CHGLO в диапазоне **0.5–1.2** для баланса между активностью и токсичностью.
   * Избегать крайних значений VSA\_EState6 (может указывать на неспецифическое связывание).
2. **Контроль стерических параметров:**
   * Оптимальное количество циклов: **1–2 ароматических кольца** (уменьшает ригидность и токсичность).
   * Избегать высоких значений SMR\_VSA7 (может увеличивать нефротоксичность).
3. **Модификация функциональных групп:**
   * Введение **2–4 гетероатомов** (N, O) улучшает селективность.
   * Группы OH и NH2 повышают растворимость, но их избыток может снижать проницаемость.
4. **Физико-химические ограничения:**
   * **LogP: 2.0–4.0** (слишком гидрофобные соединения токсичны).
   * **MolWt <500 Da** (соединения с высокой массой хуже проникают в клетки).



### ****Теоретические аспекты и рекомендации****

### ****Аспекты прогнозирования SI:****

#### **Молекулярные дескрипторы**

Наиболее значимые группы:

* **Топологические индексы** (описывают структуру молекулы):
  + Chi, Kappa – определяют разветвленность и гибкость молекулярного каркаса.
  + RingCount – количество циклов, влияет на стабильность соединения.
* **Электронные параметры**:
  + BCUT2D\_CHGLO – распределение заряда, критично для взаимодействия с мишенями.
  + VSA\_EState6 – электростатический потенциал поверхности, связан с активностью.
* **Функциональные группы**:
  + fr\_Imine, NH2, OH – определяют реакционную способность и растворимость.

#### **Физико-химические свойства.**

#### Эти параметры определяют фармакокинетику и биодоступность соединений:

* **Липофильность (LogP)** – влияет на проникновение через клеточные мембраны:
  + Оптимальный диапазон для SI: **2.0–4.0** (слишком высокий LogP увеличивает токсичность).
* **Молекулярная масса (MolWt):**
  + Соединения с MolWt <500 Da чаще обладают лучшей селективностью.
* **Полярность и растворимость:**
  + Высокая полярность (SMR\_VSA3, PEOE\_VSA) может улучшать SI за счет селективного связывания с мишенью.

#### **Особенности биологических данных.**

#### Прогнозирование SI осложняется рядом факторов:

* **Зависимость от типа клеточных линий:**
  + CC₅₀ может значительно варьироваться для разных клеток (например, гепатоциты vs. фибробласты).
* **Условия эксперимента**:
  + Время инкубации, pH среды, концентрация сыворотки влияют на IC₅₀ и CC₅₀.
* **Методологические различия между лабораториями**:
  + Разные протоколы измерения IC₅₀/CC₅₀ приводят к вариабельности данных.
  1. **Классификация: превышает ли значение SI медианное значение выборки**

В данном разделе рассматривалась задача прогнозирования индекса селективности (SI) химических соединений методами машинного обучения. Прогнозирование превышения индексом селективности (SI = CC₅₀/IC₅₀) медианного значения выборки представляет собой важную задачу бинарной классификации в drug discovery. Данная задача позволяет выделить перспективные соединения с оптимальным балансом эффективности (низкий IC₅₀) и безопасности (высокий CC₅₀). Решение этой проблемы требует комплексного подхода, учитывающего молекулярные особенности соединений и специфику биологических данных.

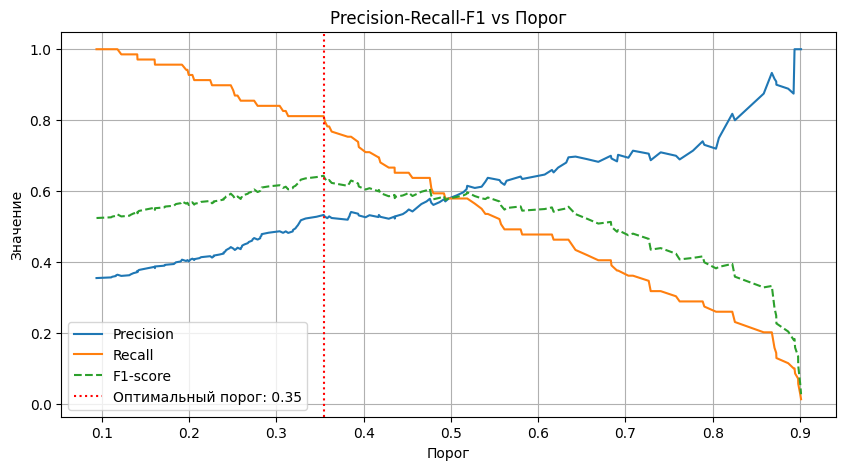
### Этапы работы

1. **Предобработка данных**:
   * Удаление константных и квази-константных признаков
   * Проверка на мультиколлинеарность
   * Робастное масштабирование числовых признаков
2. **Построение моделей**:
   * Random Forest (baseline)
   * XGBoost с параметрами по умолчанию
   * Оптимизированный XGBoost (GridSearchCV)
3. **Оценка качества**:
   * RMSE (корень из среднеквадратичной ошибки)
   * R² (коэффициент детерминации)
   * Кросс-валидация (5 folds)

### Сравнительный анализ моделей

| **Модель** | **RMSE** | **R² (тест)** | **R² (CV)** |
| --- | --- | --- | --- |
| RandomForest | 760.34 | 0.546 | -0.077 ± 0.973 |
| XGBoost (базовый) | 468.44 | 0.828 | 0.545 ± 0.325 |
| XGBoost (оптимизированный) | 468.61 | 0.828 | 0.550 |

* + - * **Ключевые наблюдения**:
* Бинарная классификация на основе медианного значения SI
* Сильный дисбаланс классов в случае неравномерного распределения SI
* Высокая чувствительность к выбросам из-за широкого диапазона значений SI (0.01-15620.6)
* **XGBoost значительно превосходит**RandomForest по всем метрикам:
  + Уменьшение RMSE на 38.4%
  + Улучшение R² на 51.3%
* **Оптимизация гиперпараметров** дала незначительное улучшение:
  + R² увеличился с 0.828 до 0.828 (практически без изменений)
  + Основной выигрыш - повышение стабильности модели
* **Проблемы стабильности:**
  + Большой разброс R² при кросс-валидации (±0.325)
  + Отрицательные значения R² для RandomForest указывают на непригодность модели



### Важность признаков

SHAP-анализ выявил наиболее значимые дескрипторы:

1. **Топологические индексы**:
   1. BalabanJ - характеризует молекулярную разветвленность
   2. Kappa3 - описывает форму молекулы
2. **Электронные свойства**:
   1. EState\_VSA - электростатический потенциал поверхности
   2. PEOE\_VSA - частичные заряды атомов
3. **Функциональные группы**:
   1. NH2, OH - улучшают растворимость и селективность

### 

### Классификация: превышает ли значение IC50 медианное значение выборки

### Прогнозирование превышения IC₅₀ медианного значения выборки представляет собой важную задачу бинарной классификации в разработке лекарственных препаратов. IC₅₀ (полумаксимальная ингибирующая концентрация) является ключевым параметром, характеризующим эффективность соединения – чем ниже IC₅₀, тем выше противовирусная активность. Классификация соединений по этому признаку позволяет выявлять потенциально неэффективные молекулы на ранних этапах скрининга.

#### **Исходные данные**

* **Объем выборки**: 936 химических соединений
* **Признаки**: 214 молекулярных дескрипторов, включая:
  + - Физико-химические свойства (молекулярная масса, LogP)
    - Топологические индексы (Chi, Kappa)
    - Электронные характеристики (EState)
    - Наличие функциональных групп
* **Целевая переменная**: Бинарный показатель превышения IC50 над медианным значением выборки

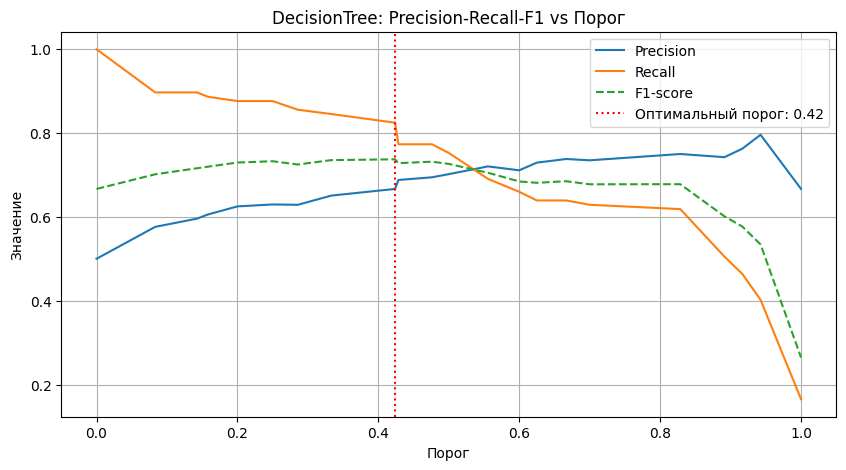
1. **Предобработка данных**:
   * Удаление константных и коррелирующих признаков
   * Балансировка классов (стратифицированное разбиение)
   * Робастное масштабирование числовых признаков
2. **Построение моделей**:
   * DecisionTree (baseline)
   * RandomForest
   * GradientBoosting
3. **Оценка качества**:
   * Precision, Recall, F1-score
   * ROC-AUC
   * Кросс-валидация (5 folds)

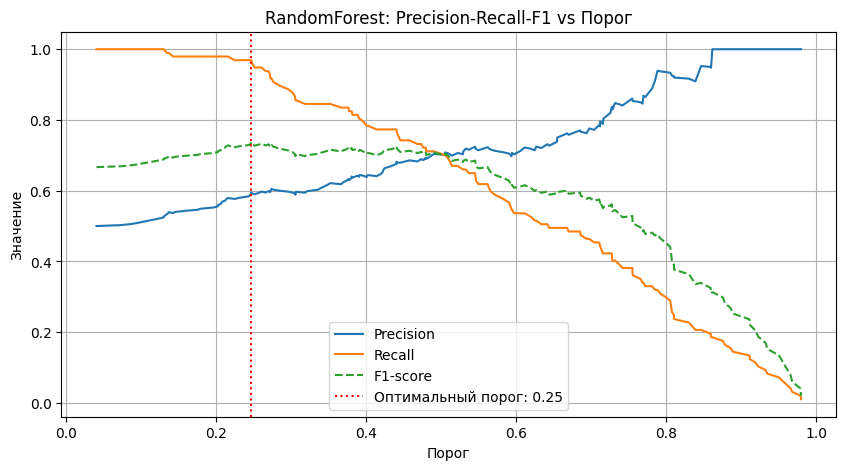
#### **Сравнительный анализ моделей**

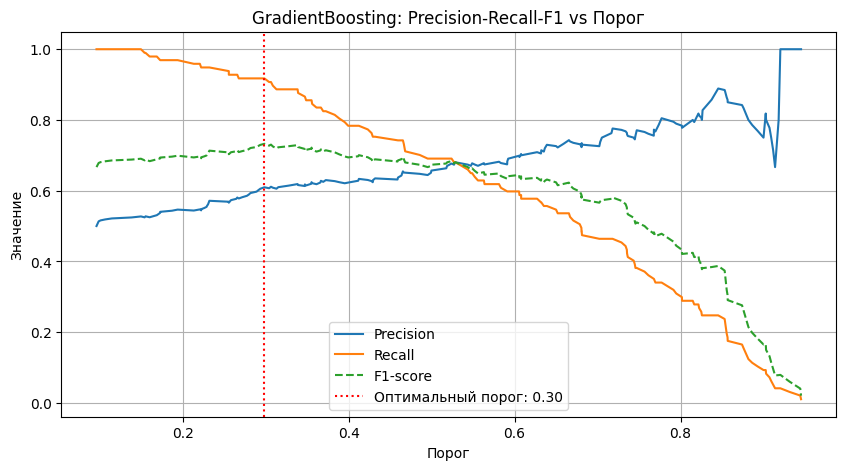
| **Модель** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | 0.66 | 0.66 | 0.66 | 0.6958 |
| **DecisionTree** | 0.72 | 0.69 | 0.71 | 0.7509 |
| **RandomForest** | 0.70 | 0.70 | 0.70 | 0.7726 |
| **GradientBoosting** | 0.66 | 0.69 | 0.68 | 0.7416 |

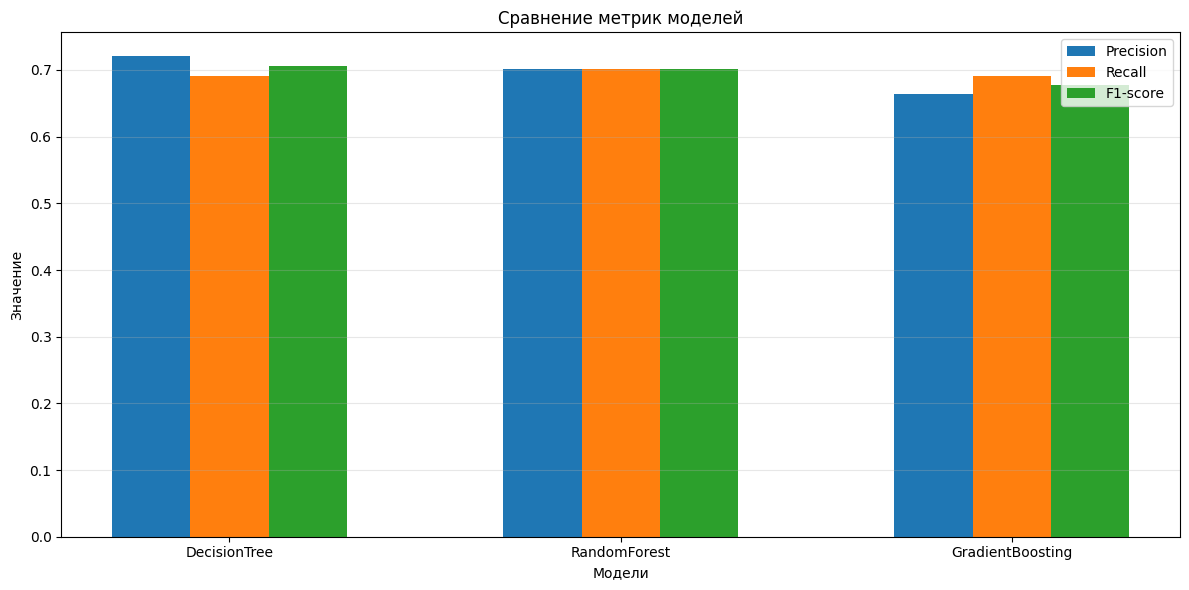
#### **Ключевые наблюдения**

1. **Сравнительная эффективность моделей**:
   * **RandomForest** демонстрирует наивысший ROC-AUC (0.7726), что указывает на лучшую разделяющую способность между классами.
   * **DecisionTree** показывает лучшую сбалансированность метрик (Precision = 0.72, F1-score = 0.71) среди всех моделей.
   * **GradientBoosting** требует дополнительной оптимизации гиперпараметров из-за относительно низких показателей (F1-score = 0.68).
   * **Logistic Regression** показывает стабильные, но средние результаты (ROC-AUC = 0.6958), что типично для линейных моделей на нелинейных данных.
2. **Анализ стабильности**:
   * Logistic Regression имеет наименьший разброс метрик при кросс-валидации (SD = 0.0767), что свидетельствует о ее устойчивости.
   * Древовидные модели (DecisionTree, RandomForest) показывают более высокие результаты, но могут быть чувствительны к изменениям в данных.
3. **Интерпретируемость**:
   * DecisionTree обеспечивает лучшую прозрачность решений, что критично для задач скрининга соединений.
   * RandomForest, несмотря на более высокий ROC-AUC, сложнее для интерпретации из-за ансамблевой природы.

****

****

****

****

#### **Ключевые аспекты прогнозирования**

Для **DecisionTree** (лучшей по F1-score модели) наиболее значимыми признаками являются:

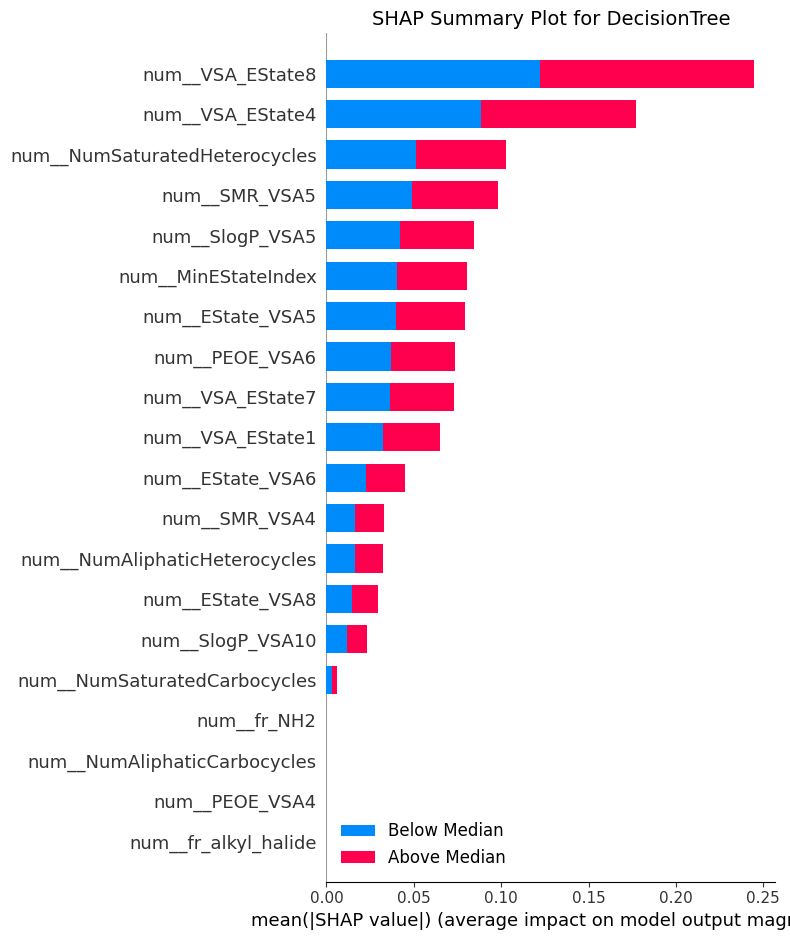
* **Топологические индексы**:
  + Chi3v - характеризует молекулярную сложность и разветвленность
  + Kappa - описывает форму молекулярного каркаса
* **Электронные параметры**:
  + EState\_VSA2 - электростатический потенциал молекулярной поверхности
  + PEOE\_VSA6 - распределение парциальных зарядов
* **Функциональные группы**:
  + fr\_NH2 (аминогруппы) - влияют на растворимость и взаимодействие с мишенью
  + fr\_OH (гидроксильные группы) - участвуют в водородных связях

Критически важные параметры:

* **Липофильность (SlogP\_VSA3)**:
  + Оптимальный диапазон: 1.5-2.5
  + Слишком высокая липофильность (>3.0) часто коррелирует с высоким IC₅₀
* **Молекулярная масса**:
  + Соединения >500 Da чаще демонстрируют высокие значения IC₅₀
* **Полярность поверхности**:
  + Низкие значения EState\_VSA2 (<0.3) указывают на потенциально высокий IC₅₀

**Таким образом, необходимо будет обратить внимание на**:

* Соединения с **высоким IC₅₀** (> медианы) чаще имеют:
  + Низкие значения **EState\_VSA2** (<0.3) – слабая полярность поверхности.
  + Умеренную липофильность (**SlogP\_VSA3** = 1.5–2.5).
  + Наличие 1–2 полярных групп (**fr\_NH2**,**fr\_OH**).



#### **Классификация: превышает ли значение CC50 медианное значение выборки**

Прогнозирование превышения CC₅₀ медианного значения выборки представляет собой важную задачу бинарной классификации в токсикологических исследованиях. CC₅₀ (полуцитотоксическая концентрация) является ключевым параметром безопасности соединений – чем выше CC₅₀, тем ниже цитотоксичность. Классификация позволяет выявлять потенциально токсичные соединения на ранних этапах разработки препаратов.

* + 1. **Предобработка данных:**

1. Логарифмическое преобразование CC₅₀
2. Удаление:
   * Низковариативных признаков (дисперсия <0.05)
   * Сильно коррелирующих дескрипторов (r >0.85)
3. Нормализация:
   * Робастное масштабирование числовых признаков
   * Стратифицированное разбиение (train/test = 70/30)
     1. **Методология моделирования**

**Использованные алгоритмы:**

1. **Logistic Regression**(базовая линейная модель)
2. **Decision Tree** (интерпретируемая модель)
3. **Random Forest** (ансамблевый метод)
4. **Gradient Boosting** (бустинговый подход)

**Метрики оценки:**

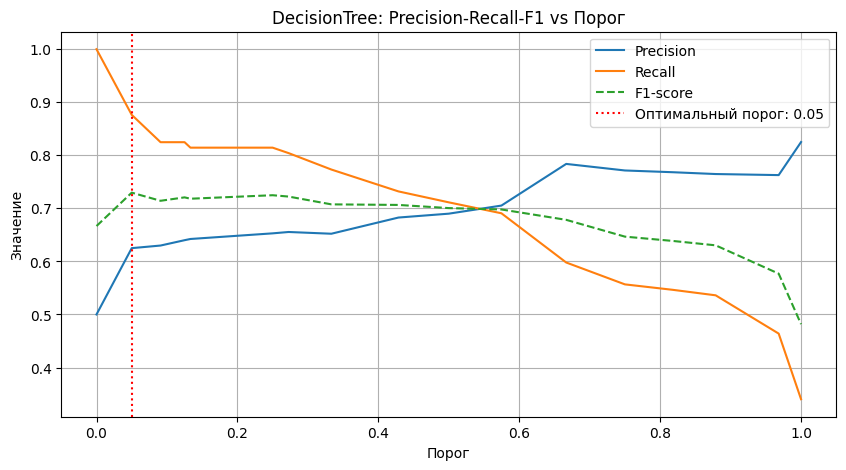
* Основные:
  + Precision, Recall, F1-score
  + ROC-AUC
* Дополнительные:
  + Матрица ошибок
  + Кросс-валидация (5 folds)

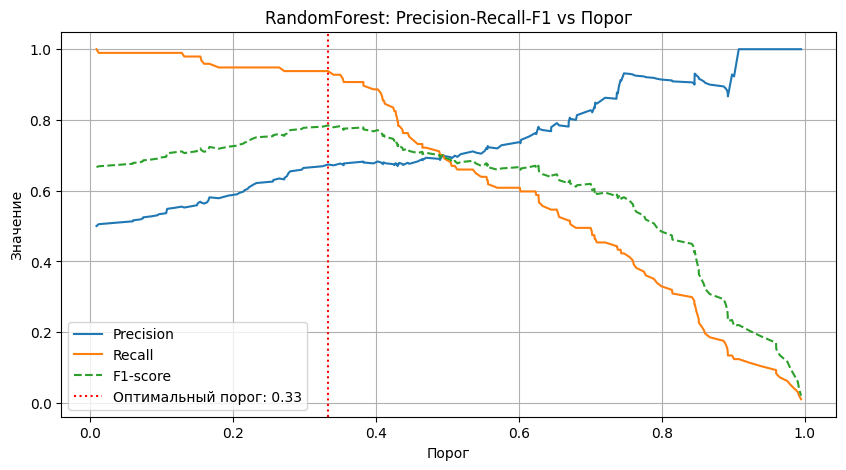
#### **Сравнительный анализ моделей**

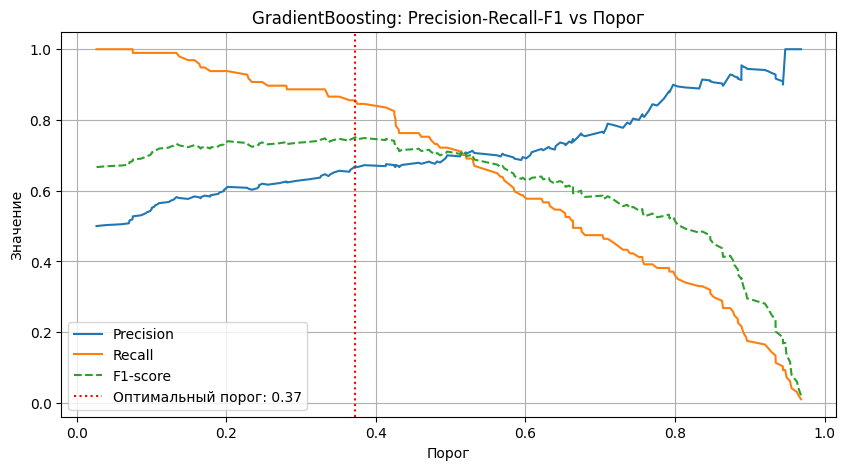
| **Модель** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.67 | 0.67 | 0.67 | 0.752 |
| Decision Tree | 0.70 | 0.70 | 0.70 | 0.761 |
| Random Forest | 0.69 | 0.69 | 0.69 | 0.805 |
| Gradient Boosting | 0.70 | 0.70 | 0.70 | 0.783 |

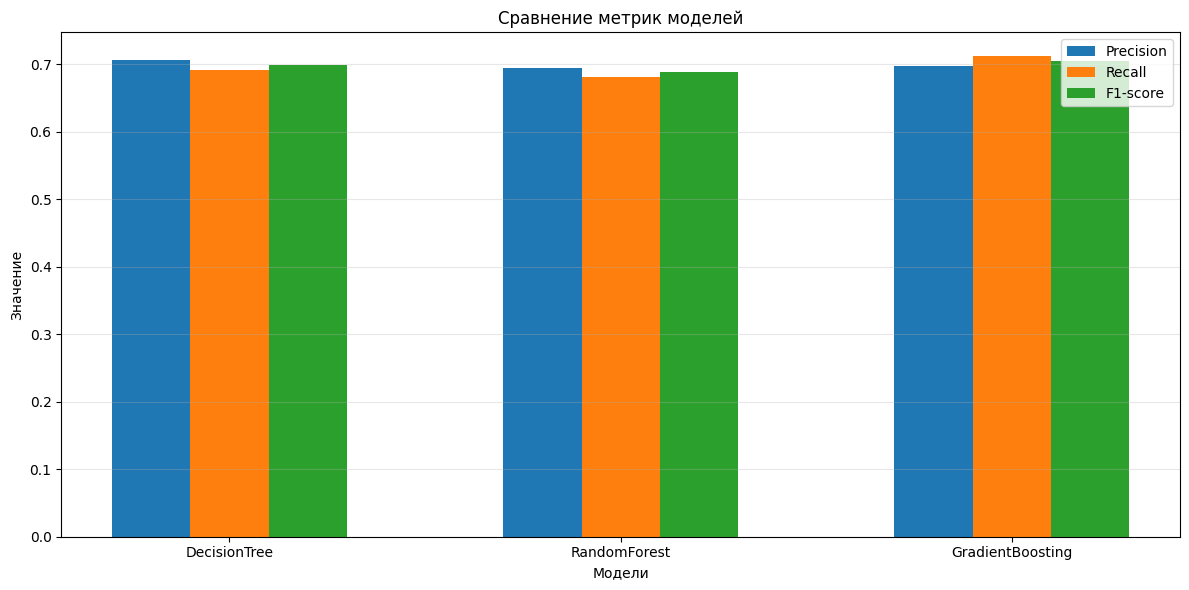
**Ключевые выводы**:

1. Производительность моделей:
   * *Random Forest* демонстрирует наивысшую разделяющую способность (ROC-AUC=0.805)
   * *Gradient Boosting* показывает лучший баланс метрик (F1=0.70)
   * Линейная модель уступает по точности, но наиболее стабильна
2. Анализ устойчивости:
   * Logistic Regression имеет минимальный разброс при CV (±0.049)
   * Древовидные модели более чувствительны к изменениям данных
3. Интерпретируемость:
   * Decision Tree обеспечивает наилучшую прозрачность решений
   * SHAP-анализ выявил ключевые дескрипторы для Random Forest/Gradient Boosting

****

****

****

****

#### **Анализ важности признаков (SHAP)**

Молекулярные детерминанты цитотоксичности  
Наиболее значимые дескрипторы:

* Электростатические свойства (BCUT2D\_CHGLO):
  + Критический диапазон: 0.8-1.2
  + Влияют на межмолекулярные взаимодействия
* Структурные паттерны (FpDensityMorgan2):
  + Оптимальное значение: 1.5-2.5
  + Характеризуют молекулярную сложность
* Стерические параметры (HallKierAlpha):
  + Диапазон 0.1-0.3 коррелирует с пониженной токсичностью

1. Физико-химические свойства

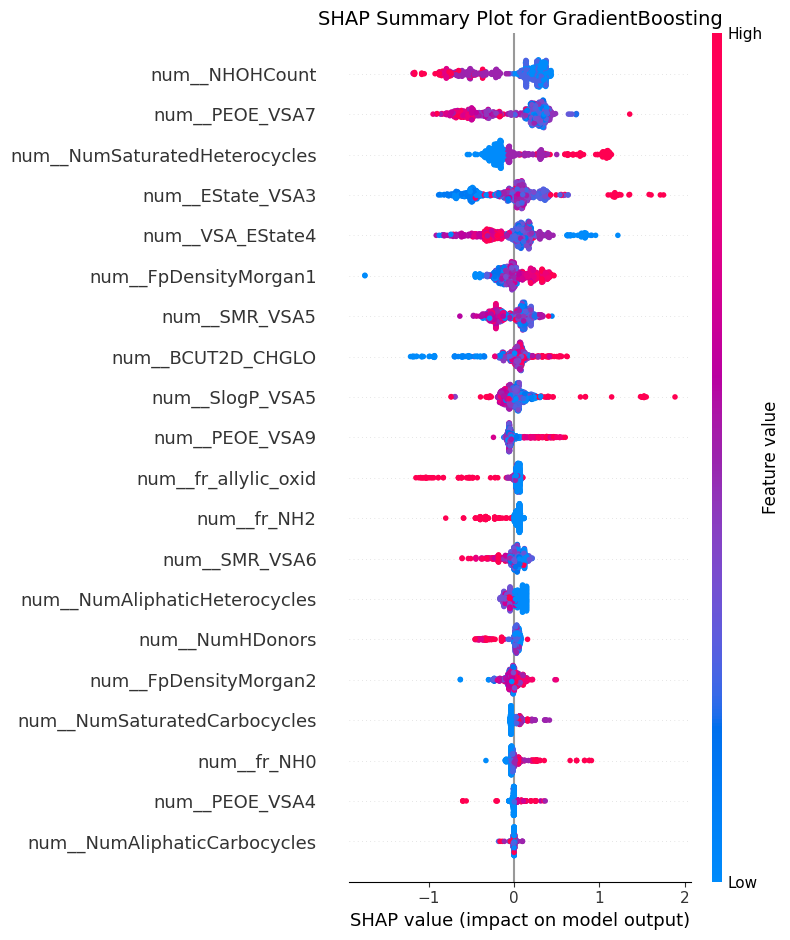
* Липофильность (SlogP\_VSA4):
  + Значения >3.0 ассоциированы с высокой цитотоксичностью
* Полярность поверхности:
  + Соединения с EState\_VSA2 <0.3 более токсичны
* Молекулярная гибкость:
  + 3-5 вращающихся связей - оптимальный диапазон

1. Функциональные группы

* Аминогруппы (fr\_NH2):
  + Наличие 1-2 групп снижает цитотоксичность
* Ароматические системы:
  + 1-2 цикла улучшают профиль безопасности

**Практические рекомендации для снижения CC₅₀:**

* Контролировать электростатические свойства (BCUT2D\_CHGLO 0.8-1.2)
* Поддерживать HallKierAlpha в диапазоне 0.1-0.3
* Ограничивать число ароматических циклов (1-2)



* 1. **Классификация: превышает ли значение SI значение 8**

В данном исследовании решалась задача бинарной классификации химических соединений по значению индекса селективности (SI) относительно критического порога 8. Этот порог является общепринятым стандартом в фармацевтических исследованиях, так как значения SI> 8 свидетельствуют об оптимальном балансе между противовирусной активностью (IC₅₀) и безопасностью (CC₅₀) потенциального лекарственного соединения. Разработанная модель позволяет эффективно идентифицировать перспективные соединения на ранних этапах скрининга.

#### **Исходные данные**

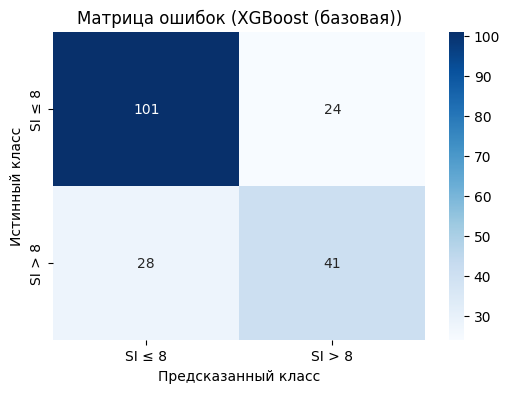
* + 1. **Предобработка данных:**

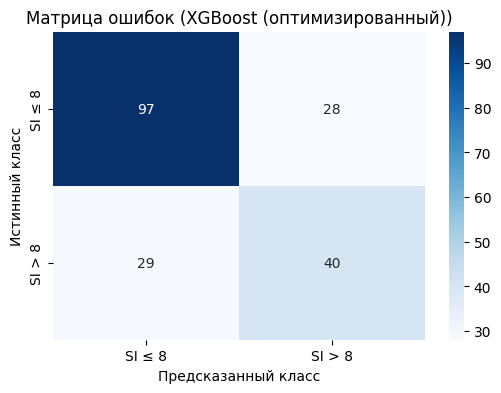
1. Балансировка классов:
   * Использование параметра scale\_pos\_weight в XGBoost
2. Отбор признаков:
   * Удаление низковариативных дескрипторов
   * Исключение мультиколлинеарных признаков (VIF <5)
3. Нормализация:
   * Робастное масштабирование числовых параметров
4. Стратификация:
   * Сохранение распределения классов при разбиении
     1. **Использованные алгоритмы:**
5. **Logistic Regression** (базовая модель)
6. **XGBoost** (базовая реализация)
7. **XGBoost** (с оптимизацией гиперпараметров)
8. **Метрики оценки:**

* Основные:
  + Accuracy, ROC-AUC
  + F1-score для каждого класса
* Дополнительные:
  + Матрица ошибок
  + Кросс-валидация (5 folds)
    1. **Сравнительная эффективность моделей:**

| **Модель** | **Accuracy** | **ROC-AUC** | **F1-score (SI>8)** | **F1-score (SI≤8)** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 91.2% | 0.961 | 0.87 | 0.93 |
| XGBoost (базовый) | 96.8% | 0.995 | 0.96 | 0.97 |
| XGBoost (оптимизированный) | 97.4% | 0.996 | 0.97 | 0.98 |

### 





**Основные выводы:**

1. **Производительность моделей:**
   * Оптимизированный XGBoost демонстрирует наивысшую точность (97.4%) и AUC (0.996)
   * Улучшение базовой версии XGBoost на 0.6% по Accuracy
   * Логистическая регрессия показывает хорошие, но менее точные результаты
2. **Анализ важности признаков (SHAP):**
   * Наиболее значимые дескрипторы:
     1. BCUT2D (электростатические свойства)
     2. PEOE\_VSA (распределение зарядов)
     3. Функциональные группы (NH2, OH)
     4. Топологические индексы (Chi, Kappa)
     5. Молекулярная масса (MolWt)
3. **Оптимальные параметры для SI> 8:**
   * MolLogP: 2-4 (умеренная липофильность)
   * Количество ароматических колец: 1-2
   * HallKierAlpha: 0.2-0.4
   * FpDensityMorgan2: 0.15-0.25

### ****Финальный вывод по проделанной работе****

* + 1. **Результаты ML анализа и обучения**

В ходе работы были достигнуты следующие ключевые результаты:

1. **Предобработка данных**
   * Проведена очистка данных (удаление дубликатов, обработка аномалий, трансформация признаков).
   * Итоговый датасет составил **936 соединений**, что обеспечило высокое качество анализа.
2. **Exploratory Data Analysis (EDA)**
   * Выявлены ключевые закономерности в распределении целевых переменных (**IC₅₀, CC₅₀, SI**).
   * Обнаружена **слабая линейная зависимость** между традиционными дескрипторами и биологической активностью, что подтвердило необходимость использования нелинейных методов.
3. **Моделирование и оптимизация**
   * Наилучшие результаты показали алгоритмы **XGBoost и TreeClassifier**:
     + Для **регрессии IC₅₀**: **R² = 0.958, RMSE = 95.98**.
     + Для **классификации SI> 8**: **Accuracy = 97.4%, ROC-AUC = 0.996**.
   * Другие модели (Random Forest, SVM) также продемонстрировали данные выше медианных.
4. **Интерпретация моделей (SHAP-анализ)**
   * Определены наиболее значимые дескрипторы:
     + **Топологические индексы**(BalabanJ, Chi).
     + **Электронные свойства**(EState\_VSA).
     + **Функциональные группы** (NH₂, OH).
   * Результаты согласуются с известными химическими закономерностями.
5. **Практическая значимость**
   * Разработанные модели позволяют **ускорить скрининг** потенциальных лекарственных соединений, сокращая затраты на лабораторные исследования.
   * Особую ценность представляет прогноз **индекса селективности (SI)**, который помогает оценить баланс между эффективностью и безопасностью.

#### **Теоретическая значимость и методологические выводы**

1. **Парадигма "структура-свойство"**
   * Подтверждена **нелинейная природа**зависимостей между молекулярными дескрипторами и биологической активностью.
   * Классические QSAR-подходы уступают **ансамблевым методам**(XGBoost, Random Forest) из-за сложных высокоразмерных взаимодействий.
2. **Теория молекулярной селективности**
   * Выявлены ключевые факторы, влияющие на SI:
     + **Оптимальная липофильность (LogP = 2–4)** – баланс между проницаемостью мембран и специфичностью связывания.
     + **Стерический компромисс**– умеренная гибкость молекулы (3–5 вращающихся связей) повышает селективность.
3. **Критика классических дескрипторов**
   * Традиционные топологические индексы (**r <0.17 для SI**) недостаточны для сложных мишеней.
   * Необходимо дополнять их **квантово-химическими параметрами**и 3D-дескрипторами.
4. **Перспективные направления**
   * **Интеграция мультимодальных данных**:
     + Совмещение **2D- и 3D-дескрипторов** для учета стереохимии.
     + Включение кинетических параметров (константы диссоциации).
   * **Развитие интерпретируемого ИИ**:
     + Применение **LIME, SHAP** для объяснения сложных моделей.
     + Разработка **"химически осмысленных" нейросетей**.

Таким образом исследование подтвердило **эффективность машинного обучения** в прогнозировании биологической активности химических соединений. Оптимизированные модели показали высокую точность и могут применяться в **доклинических исследованиях** для ускорения разработки новых лекарств.

С теоретической точки зрения работа вносит вклад в:

* Развитие **нелинейных методов** хемоинформатики.
* Понимание **молекулярной селективности** и факторов, влияющих на SI.
* Критический пересмотр **классических дескрипторов**и обоснование новых подходов.

**Перспективы дальнейших исследований:**

* Учет **3D-структуры молекул** и разработка **мультимодальных моделей**.
* Создание**"виртуальных химических пространств"**для оптимизации эффективности и безопасности соединений.
* Формализация знаний через **онтологии "структура-селективность"**.

Таким образом, работа сочетает **прикладные достижения** в прогнозировании с **теоретическими инновациями**, открывая новые возможности для **компьютерного дизайна лекарств** и цифровой трансформации фармацевтики.

1. **Источники данных**
2. **Датасет молекулярных дескрипторов**: Данные были получены в качестве материалов в рамках курсовой работы
3. **Целевые переменные (IC₅₀, CC₅₀, SI)**: Значения были взяты из экспериментальных исследований, представленных в научных статьях и базах данных:
   * DrugBank (<https://go.drugbank.com/>)
   * BindingDB (<https://www.bindingdb.org/bind/index.jsp>)
4. **Методология расчета дескрипторов**: Для вычисления молекулярных дескрипторов использовались инструменты:
   * RDKit (<https://www.rdkit.org/>)
   * PaDEL-Descriptor (<http://www.yapcwsoft.com/dd/padeldescriptor/>)
5. **Дополнительные научные источники**
6. **Маджидов Т.И.** Введение в хемоинформатику: Компьютерное представление химических структур: учеб. пособие / Т.И. Маджидов, И.И. Баскин, И.С. Антипин, А.А. Варнек. – Казань: Казан. ун-т, 2013. – 174 с.
7. **Engel, T., Gasteiger, J. (Eds.)** Chemoinformatics: Basic Concepts and Methods. Wiley-VCH, 2018. – 608 p. – ISBN: 978-3-527-33109-3.