Курсовая работа на тему:

"Прогнозирование эффективности химических соединений на основе молекулярных дескрипторов"

Решетило Никита

1 курс, «Машинное обучение»

Содержание

I. Описание данных

a. 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 IC_{50}, CC_{50} 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.

b. Ключевые показатели эффективности

В исследовании фокусируется на прогнозировании трех критически важных параметров:

1. IC_{50} – концентрация вещества, необходимая для подавления целевого биологического процесса на 50%, показатель эффективности

2. CC_{50} – концентрация, вызывающая гибель 50% клеток, индикатор цитотоксичности
3. Индекс селективности (SI) – отношение CC_{50}/IC_{50}, отражающее терапевтическое окно (чем выше значение, тем лучше профиль безопасности)

c. Цели исследования

Разработка и сравнительный анализ:

- Регрессионных моделей для точного предсказания численных значений IC_{50}, CC_{50} и SI
- Классификационных моделей для бинарной оценки показателей относительно:
  - Медианных значений выборки
  - Порогового значения $SI \geq 8$ (клинически значимый критерий безопасности)

d. 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:

1. Регрессия для IC_{50}

a. **Цель**

Построить модель, предсказывающую близкую к максимальной ингибирующую концентрацию (IC_{50}) на основе молекулярных дескрипторов.

b. **Практическая значимость**

i. IC_{50} показывает, насколько эффективно соединение подавляет целевую биологическую мишень.
ii. Низкие значения IC_{50} (<1 мкМ) указывают на высокую активность.
iii. Модель поможет отбирать наиболее перспективные соединения для дальнейших исследований.

---

2. Регрессия для CC_{50}

a. **Цель**

Разработать модель, прогнозирующую близкую к максимальной цитотоксической концентрации (CC_{50}) по структурным признакам.

b. **Практическая значимость**

i. CC_{50} отражает токсичность соединения для здоровых клеток.
ii. Высокие значения (> 50 мкМ) желательны для безопасных лекарств.
iii. Модель позволит отсеивать слишком токсичные соединения на ранних этапах.

---

3. Регрессия для SI (Selectivity Index)

a. **Цель**

Создать модель для предсказания индекса селективности (SI = CC_{50} / IC_{50})**.**

b. **Практическая значимость**

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

---

4. **Классификация: превышает ли IC_{50}\CC_{50}\IS медианное значение выборки?**

    a. **Цель**

    Бинарная классификация: определяет, является ли IC_{50}\CC_{50}\IS соединения выше медианы по выборке.

    b. **Практическая значимость**

        i. Позволяет быстро разделять соединения на "более/менее активные".
        ii. Упрощает скрининг больших библиотек соединений.

---

5. **Клинически значимая классификация: SI> 8**

    a. **Цель**

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

    b. **Практическая значимость**

        i. SI> 8 – общепринятый критерий безопасности в доклинических исследованиях.
        ii. Помогает отбирать соединения с достаточным терапевтическим окном.

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

- Регрессия (IC_{50}, CC_{50}, SI) → точное прогнозирование ключевых параметров.
- Классификация (медианные значения, SI> 8) → быстрый скрининг и стратификация соединений

е. 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.

---

1. 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 sp^3-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).

---

2. 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.

---

3. 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.

---

4. **BCUT-дескрипторы, кодирующие**
   - **Массу**: BCUT2D_MWHI, BCUT2D_MWLOW.
   - **Заряд**: BCUT2D_CHGHI, BCUT2D_CHGLO.
   - **Липофильность**: BCUT2D_LOGPHI, BCUT2D_LOGPLOW.
   - **Поляризуемость**: BCUT2D_MRHI, BCUT2D_MRLOW.

---

5. **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.

---

6. **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.

---

7. **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.

---

9. **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.

II. EDA

a. Первичный анализ

i. 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

ii. 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

b. Анализ молекулярных дескрипторов

i. 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.)

ii. 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

iii. 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

c. Общие результаты обработки данных

i. 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)

ii. 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

d. Анализ аномальных признаков

i. 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

ii. 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

Переменная	Z-score > 3	IQR метод	Общее кол-во аномалий	% аномалий	Химическая интерпретация
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

e. Химическая интерпретация аномалий

I. 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.

II. 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

f. Анализ выбросов молекулярных свойств

i. 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

1. 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.
2. 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.

3. 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.

4. 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.

5. Полярная поверхность (TPSA)

- Low_Selectivity: 248.12 (против 144.71 в Other).

Вывод: Высокая полярность может снижать селективность.

6. Доля sp^3-гибридизованных атомов (FractionCSP3)

- Low_Selectivity: 0.593 (против 0.429 в Other).

Вывод: Более насыщенные молекулы могут быть менее селективными.

ii. 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.

iii. Топ-3 соединений с высокой селективностью (Other)

ID	SI	MolLogP	MolWt	NumH Acceptors	NumH Donors	Ring Count	Fraction CSP3
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 sp^3-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).

g. Анализ корреляций молекулярных дескрипторов с биологической активностью

i. Корреляции с IC50 (ингибирующая концентрация)

Наибольшие положительные корреляции:

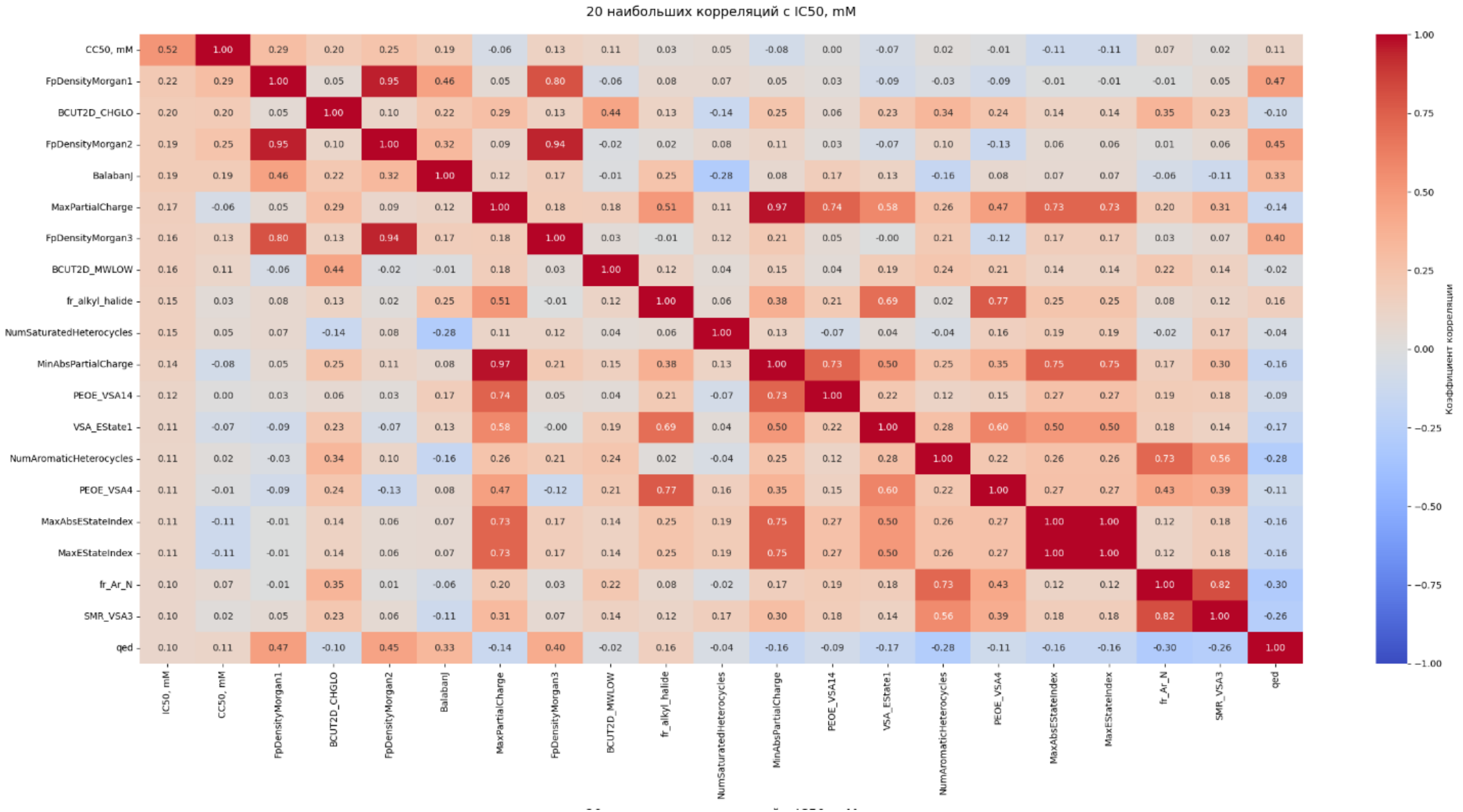
1. FpDensityMorgan1 (0.215) – высокая плотность фармакофорных особенностей (Morgan fingerprints) может усиливать связывание с мишенью.
2. BCUT2D_CHGLO (0.195) – низкие зарядовые дескрипторы могут влиять на электростатические взаимодействия.
3. BalabanJ (0.191**)** – топологическая сложность молекулы коррелирует с активностью.

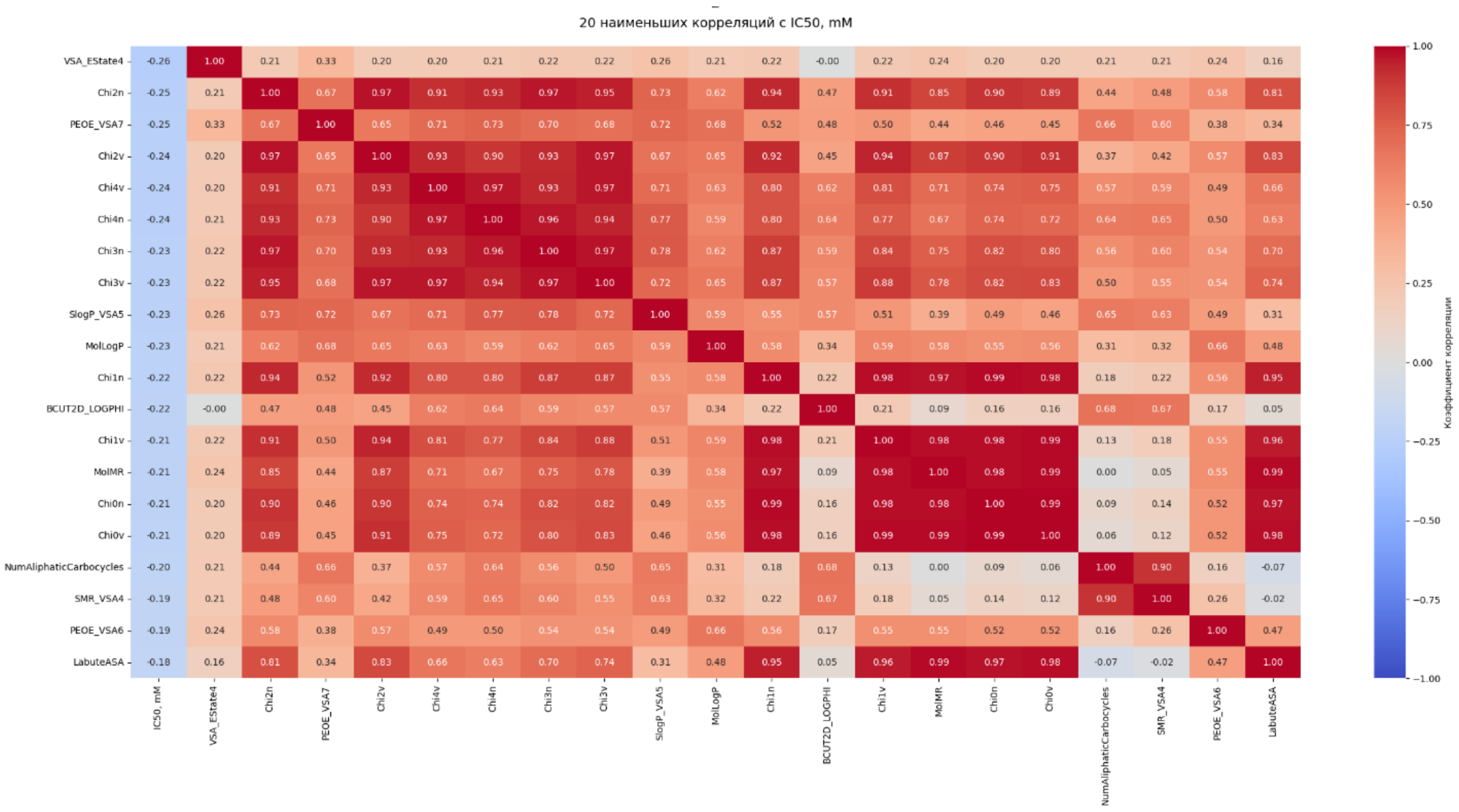
Наибольшие отрицательные корреляции:

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.


20 наибольших корреляций с IC50, mM
CC50, mM 0.52 1.00 0.29 0.20 0.25 0.19 -0.06 0.13 0.11 0.03 0.05 -0.08 0.00 -0.07 0.02 -0.01 -0.11 -0.11 0.07 0.02 0.11
FpDensityMorgan1 0.22 0.29 1.00 0.05 0.95 0.46 0.05 0.80 -0.06 0.08 0.07 0.05 0.03 -0.09 -0.03 -0.09 -0.01 -0.01 -0.01 0.05 0.47
BCUT2D_CHGLO 0.20 0.20 0.05 1.00 0.10 0.22 0.29 0.13 0.44 0.13 -0.14 0.25 0.06 0.23 0.34 0.24 0.14 0.14 0.35 0.23 -0.10
FpDensityMorgan2 0.19 0.25 0.95 0.10 1.00 0.32 0.09 0.94 -0.02 0.02 0.08 0.11 0.03 -0.07 0.10 -0.13 0.06 0.06 0.01 0.06 0.45
BalabanJ 0.19 0.19 0.46 0.22 0.32 1.00 0.12 0.17 -0.01 0.25 -0.28 0.08 0.17 0.13 -0.16 0.08 0.07 0.07 -0.06 -0.11 0.33
MaxPartialCharge 0.17 -0.06 0.05 0.29 0.09 0.12 1.00 0.18 0.18 0.51 0.11 0.97 0.74 0.58 0.26 0.47 0.73 0.73 0.20 0.31 -0.14
FpDensityMorgan3 0.16 0.13 0.80 0.13 0.94 0.17 0.18 1.00 0.03 -0.01 0.12 0.21 0.05 -0.00 0.21 -0.12 0.17 0.17 0.03 0.07 0.40
BCUT2D_MWLOW 0.16 0.11 -0.06 0.44 -0.02 -0.01 0.18 0.03 1.00 0.12 0.04 0.15 0.04 0.19 0.24 0.21 0.14 0.14 0.22 0.14 -0.02
fr_alkyl_halide 0.15 0.03 0.08 0.13 0.02 0.25 0.51 -0.01 0.12 1.00 0.06 0.38 0.21 0.69 0.02 0.77 0.25 0.25 0.08 0.12 0.16
NumSaturatedHeterocycles 0.15 0.05 0.07 -0.14 0.08 -0.28 0.11 0.12 0.04 0.06 1.00 0.13 -0.07 0.04 -0.04 0.16 0.19 0.19 -0.02 0.17 -0.04
MinAbsPartialCharge 0.14 -0.08 0.05 0.25 0.11 0.08 0.97 0.21 0.15 0.38 0.13 1.00 0.73 0.50 0.25 0.35 0.75 0.75 0.17 0.30 -0.16
PEOE_VSA14 0.12 0.00 0.03 0.06 0.03 0.17 0.74 0.05 0.04 0.21 -0.07 0.73 1.00 0.22 0.12 0.15 0.27 0.27 0.19 0.18 -0.09
VSA_EState1 0.11 -0.07 -0.09 0.23 -0.07 0.13 0.58 -0.00 0.19 0.69 0.04 0.50 0.22 1.00 0.28 0.60 0.50 0.50 0.18 0.14 -0.17
NumAromaticHeterocycles 0.11 0.02 -0.03 0.34 0.10 -0.16 0.26 0.21 0.24 0.02 -0.04 0.25 0.12 0.28 1.00 0.22 0.26 0.26 0.73 0.56 -0.28
PEOE_VSA4 0.11 -0.01 -0.09 0.24 -0.13 0.08 0.47 -0.12 0.21 0.77 0.16 0.35 0.15 0.60 0.22 1.00 0.27 0.27 0.43 0.39 -0.11
MaxAbsEStateIndex 0.11 -0.11 -0.01 0.14 0.06 0.07 0.73 0.17 0.14 0.25 0.19 0.75 0.27 0.50 0.26 0.27 1.00 1.00 0.12 0.18 -0.16
MaxEStateIndex 0.11 -0.11 -0.01 0.14 0.06 0.07 0.73 0.17 0.14 0.25 0.19 0.75 0.27 0.50 0.26 0.27 1.00 1.00 0.12 0.18 -0.16
fr_Ar_N 0.10 0.07 -0.01 0.35 0.01 -0.06 0.20 0.03 0.22 0.08 -0.02 0.17 0.19 0.18 0.73 0.43 0.12 0.12 1.00 0.82 -0.30
SMR_VSA3 0.10 0.02 0.05 0.23 0.06 -0.11 0.31 0.07 0.14 0.12 0.17 0.30 0.18 0.14 0.56 0.39 0.18 0.18 0.82 1.00 -0.26
qed 0.10 0.11 0.47 -0.10 0.45 0.33 -0.14 0.40 -0.02 0.16 -0.04 -0.16 -0.09 -0.17 -0.28 -0.11 -0.16 -0.16 -0.30 -0.26 1.00
IC50, mM
CC50, mM
FpDensityMorgan1
BCUT2D_CHGLO
FpDensityMorgan2
BalabanJ
MaxPartialCharge
FpDensityMorgan3
BCUT2D_MWLOW
fr_alkyl_halide
NumSaturatedHeterocycles
MinAbsPartialCharge
PEOE_VSA14
VSA_EState1
NumAromaticHeterocycles
PEOE_VSA4
MaxAbsEStateIndex
MaxEStateIndex
fr_Ar_N
SMR_VSA3
qed
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Коэффициент корреляции


20 наименьших корреляций с IC50, mM
VSA_EState4 -0.26 1.00 0.21 0.33 0.20 0.20 0.21 0.22 0.22 0.26 0.21 0.22 -0.00 0.22 0.24 0.20 0.20 0.21 0.21 0.24 0.16
Chi2n -0.25 0.21 1.00 0.67 0.97 0.91 0.93 0.97 0.95 0.73 0.62 0.94 0.47 0.91 0.85 0.90 0.89 0.44 0.48 0.58 0.81
PEOE_VSA7 -0.25 0.33 0.67 1.00 0.65 0.71 0.73 0.70 0.68 0.72 0.68 0.52 0.48 0.50 0.44 0.46 0.45 0.66 0.60 0.38 0.34
Chi2v -0.24 0.20 0.97 0.65 1.00 0.93 0.90 0.93 0.97 0.67 0.65 0.92 0.45 0.94 0.87 0.90 0.91 0.37 0.42 0.57 0.83
Chi4v -0.24 0.20 0.91 0.71 0.93 1.00 0.97 0.93 0.97 0.71 0.63 0.80 0.62 0.81 0.71 0.74 0.75 0.57 0.59 0.49 0.66
Chi4n -0.24 0.21 0.93 0.73 0.90 0.97 1.00 0.96 0.94 0.77 0.59 0.80 0.64 0.77 0.67 0.74 0.72 0.64 0.65 0.50 0.63
Chi3n -0.23 0.22 0.97 0.70 0.93 0.93 0.96 1.00 0.97 0.78 0.62 0.87 0.59 0.84 0.75 0.82 0.80 0.56 0.60 0.54 0.70
Chi3v -0.23 0.22 0.95 0.68 0.97 0.97 0.94 0.97 1.00 0.72 0.65 0.87 0.57 0.88 0.78 0.82 0.83 0.50 0.55 0.54 0.74
SlogP_VSA5 -0.23 0.26 0.73 0.72 0.67 0.71 0.77 0.78 0.72 1.00 0.59 0.55 0.57 0.51 0.39 0.49 0.46 0.65 0.63 0.49 0.31
MolLogP -0.23 0.21 0.62 0.68 0.65 0.63 0.59 0.62 0.65 0.59 1.00 0.58 0.34 0.59 0.58 0.55 0.56 0.31 0.32 0.66 0.48
Chi1n -0.22 0.22 0.94 0.52 0.92 0.80 0.80 0.87 0.87 0.55 0.58 1.00 0.22 0.98 0.97 0.99 0.98 0.18 0.22 0.56 0.95
BCUT2D_LOGPHI -0.22 -0.00 0.47 0.48 0.45 0.62 0.64 0.59 0.57 0.57 0.34 0.22 1.00 0.21 0.09 0.16 0.16 0.68 0.67 0.17 0.05
Chi1v -0.21 0.22 0.91 0.50 0.94 0.81 0.77 0.84 0.88 0.51 0.59 0.98 0.21 1.00 0.98 0.98 0.99 0.13 0.18 0.55 0.96
MolMR -0.21 0.24 0.85 0.44 0.87 0.71 0.67 0.75 0.78 0.39 0.58 0.97 0.09 0.98 1.00 0.98 0.99 0.00 0.05 0.55 0.99
Chi0n -0.21 0.20 0.90 0.46 0.90 0.74 0.74 0.82 0.82 0.49 0.55 0.99 0.16 0.98 0.98 1.00 0.99 0.09 0.14 0.52 0.97
Chi0v -0.21 0.20 0.89 0.45 0.91 0.75 0.72 0.80 0.83 0.46 0.56 0.98 0.16 0.99 0.99 0.99 1.00 0.06 0.12 0.52 0.98
NumAliphaticCarbocycles -0.20 0.21 0.44 0.66 0.37 0.57 0.64 0.56 0.50 0.65 0.31 0.18 0.68 0.13 0.00 0.09 0.06 1.00 0.90 0.16 -0.07
SMR_VSA4 -0.19 0.21 0.48 0.60 0.42 0.59 0.65 0.60 0.55 0.63 0.32 0.22 0.67 0.18 0.05 0.14 0.12 0.90 1.00 0.26 -0.02
PEOE_VSA6 -0.19 0.24 0.58 0.38 0.57 0.49 0.50 0.54 0.54 0.49 0.66 0.56 0.17 0.55 0.55 0.52 0.52 0.16 0.26 1.00 0.47
LabuteASA -0.18 0.16 0.81 0.34 0.83 0.66 0.63 0.70 0.74 0.31 0.48 0.95 0.05 0.96 0.99 0.97 0.98 -0.07 -0.02 0.47 1.00
IC50, mM
VSA_EState4
Chi2n
PEOE_VSA7
Chi2v
Chi4v
Chi4n
Chi3n
Chi3v
SlogP_VSA5
MolLogP
Chi1n
BCUT2D_LOGPHI
Chi1v
MolMR
Chi0n
Chi0v
NumAliphaticCarbocycles
SMR_VSA4
PEOE_VSA6
LabuteASA
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Коэффициент корреляции

ii. **Корреляции с CC50 (цитотоксическая концентрация)**

Наибольшие положительные корреляции:

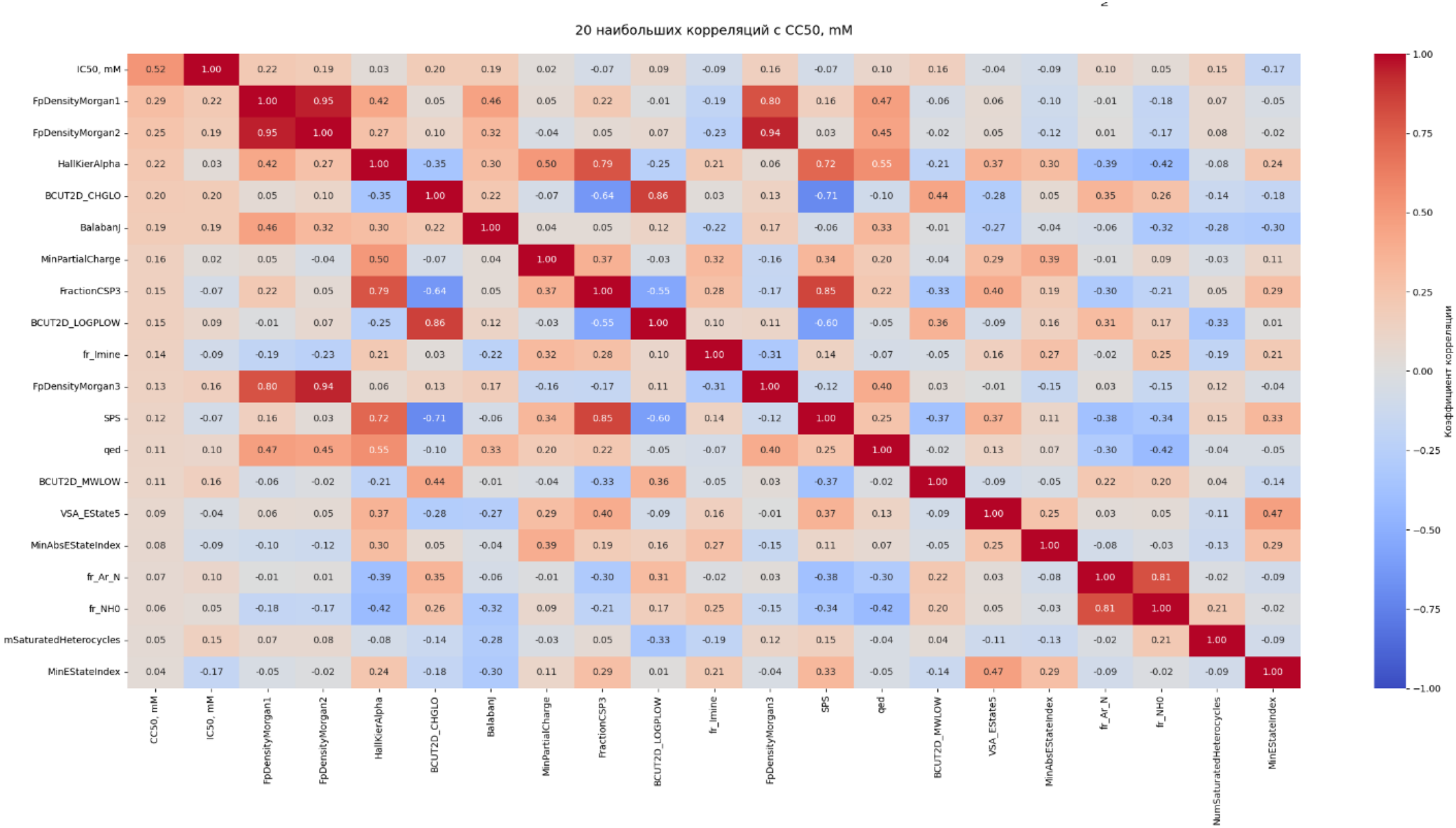
1. FpDensityMorgan1 (0.292) – аналогично IC50, плотность фармакофоров важна для токсичности.
2. HallKierAlpha (0.221) – стерические параметры влияют на цитотоксичность.

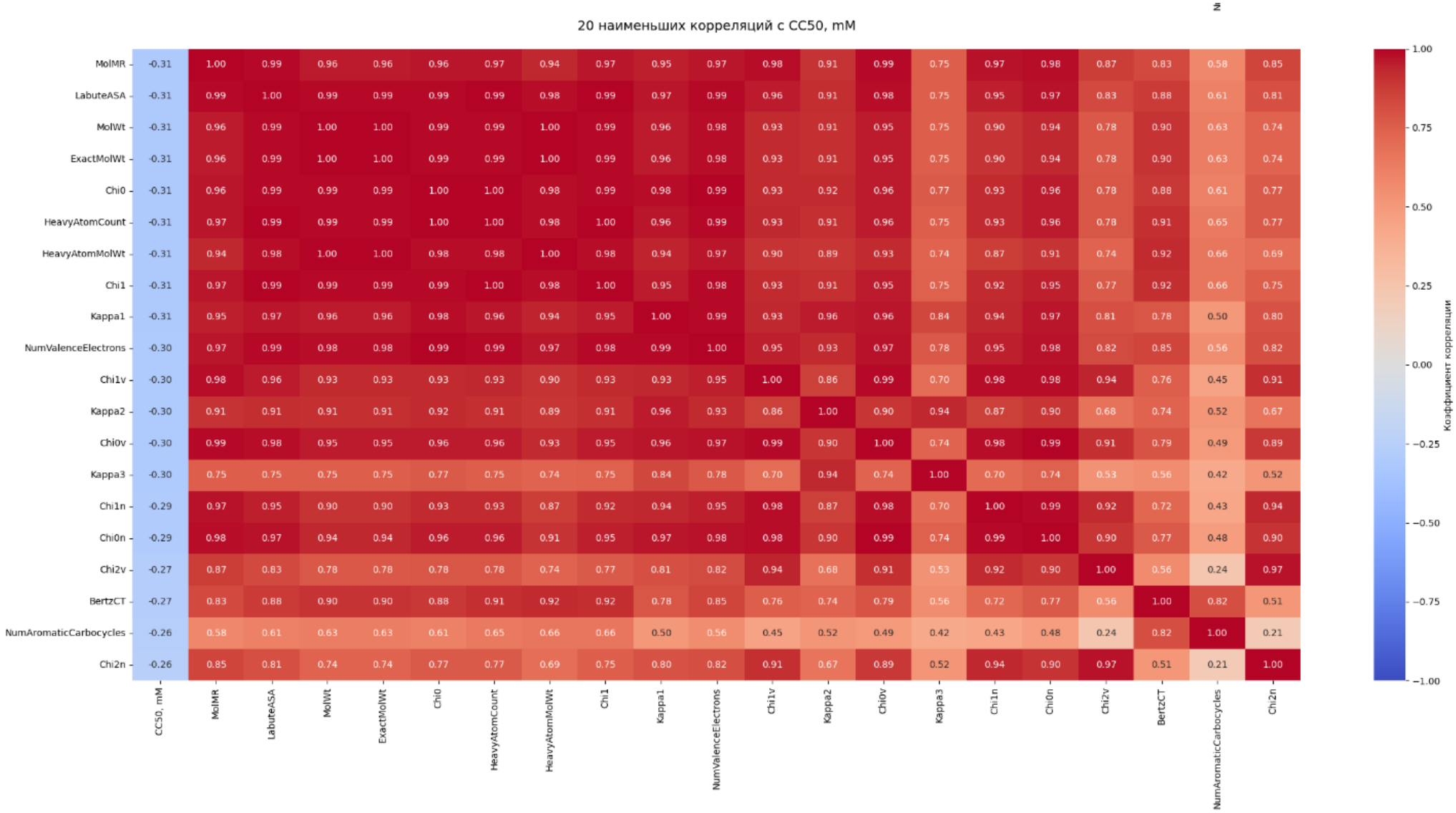
Наибольшие отрицательные корреляции:

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.


20 наибольших корреляций с CC50, mM
IC50, mM 0.52 1.00 0.22 0.19 0.03 0.20 0.19 0.02 -0.07 0.09 -0.09 0.16 -0.07 0.10 0.16 -0.04 -0.09 0.10 0.05 0.15 -0.17
FpDensityMorgan1 0.29 0.22 1.00 0.95 0.42 0.05 0.46 0.05 0.22 -0.01 -0.19 0.80 0.16 0.47 -0.06 0.06 -0.10 -0.01 -0.18 0.07 -0.05
FpDensityMorgan2 0.25 0.19 0.95 1.00 0.27 0.10 0.32 -0.04 0.05 0.07 -0.23 0.94 0.03 0.45 -0.02 0.05 -0.12 0.01 -0.17 0.08 -0.02
HallKierAlpha 0.22 0.03 0.42 0.27 1.00 -0.35 0.30 0.50 0.79 -0.25 0.21 0.06 0.72 0.55 -0.21 0.37 0.30 -0.39 -0.42 -0.08 0.24
BCUT2D_CHGLO 0.20 0.20 0.05 0.10 -0.35 1.00 0.22 -0.07 -0.64 0.86 0.03 0.13 -0.71 -0.10 0.44 -0.28 0.05 0.35 0.26 -0.14 -0.18
BalabanJ 0.19 0.19 0.46 0.32 0.30 0.22 1.00 0.04 0.05 0.12 -0.22 0.17 -0.06 0.33 -0.01 -0.27 -0.04 -0.06 -0.32 -0.28 -0.30
MinPartialCharge 0.16 0.02 0.05 -0.04 0.50 -0.07 0.04 1.00 0.37 -0.03 0.32 -0.16 0.34 0.20 -0.04 0.29 0.39 -0.01 0.09 -0.03 0.11
FractionCSP3 0.15 -0.07 0.22 0.05 0.79 -0.64 0.05 0.37 1.00 -0.55 0.28 -0.17 0.85 0.22 -0.33 0.40 0.19 -0.30 -0.21 0.05 0.29
BCUT2D_LOGPLOW 0.15 0.09 -0.01 0.07 -0.25 0.86 0.12 -0.03 -0.55 1.00 0.10 0.11 -0.60 -0.05 0.36 -0.09 0.16 0.31 0.17 -0.33 0.01
fr_Imine 0.14 -0.09 -0.19 -0.23 0.21 0.03 -0.22 0.32 0.28 0.10 1.00 -0.31 0.14 -0.07 -0.05 0.16 0.27 -0.02 0.25 -0.19 0.21
FpDensityMorgan3 0.13 0.16 0.80 0.94 0.06 0.13 0.17 -0.16 -0.17 0.11 -0.31 1.00 -0.12 0.40 0.03 -0.01 -0.15 0.03 -0.15 0.12 -0.04
SPS 0.12 -0.07 0.16 0.03 0.72 -0.71 -0.06 0.34 0.85 -0.60 0.14 -0.12 1.00 0.25 -0.37 0.37 0.11 -0.38 -0.34 0.15 0.33
qed 0.11 0.10 0.47 0.45 0.55 -0.10 0.33 0.20 0.22 -0.05 -0.07 0.40 0.25 1.00 -0.02 0.13 0.07 -0.30 -0.42 -0.04 -0.05
BCUT2D_MWLOW 0.11 0.16 -0.06 -0.02 -0.21 0.44 -0.01 -0.04 -0.33 0.36 -0.05 0.03 -0.37 -0.02 1.00 -0.09 -0.05 0.22 0.20 0.04 -0.14
VSA_EState5 0.09 -0.04 0.06 0.05 0.37 -0.28 -0.27 0.29 0.40 -0.09 0.16 -0.01 0.37 0.13 -0.09 1.00 0.25 0.03 0.05 -0.11 0.47
MinAbsEStateIndex 0.08 -0.09 -0.10 -0.12 0.30 0.05 -0.04 0.39 0.19 0.16 0.27 -0.15 0.11 0.07 -0.05 0.25 1.00 -0.08 -0.03 -0.13 0.29
fr_Ar_N 0.07 0.10 -0.01 0.01 -0.39 0.35 -0.06 -0.01 -0.30 0.31 -0.02 0.03 -0.38 -0.30 0.22 0.03 -0.08 1.00 0.81 -0.02 -0.09
fr_NH0 0.06 0.05 -0.18 -0.17 -0.42 0.26 -0.32 0.09 -0.21 0.17 0.25 -0.15 -0.34 -0.42 0.20 0.05 -0.03 0.81 1.00 0.21 -0.02
mSaturatedHeterocycles 0.05 0.15 0.07 0.08 -0.08 -0.14 -0.28 -0.03 0.05 -0.33 -0.19 0.12 0.15 -0.04 0.04 -0.11 -0.13 -0.02 0.21 1.00 -0.09
MinEStateIndex 0.04 -0.17 -0.05 -0.02 0.24 -0.18 -0.30 0.11 0.29 0.01 0.21 -0.04 0.33 -0.05 -0.14 0.47 0.29 -0.09 -0.02 -0.09 1.00
CC50, mM
IC50, mM
FpDensityMorgan1
FpDensityMorgan2
HallKierAlpha
BCUT2D_CHGLO
BalabanJ
MinPartialCharge
FractionCSP3
BCUT2D_LOGPLOW
fr_Imine
FpDensityMorgan3
SPS
qed
BCUT2D_MWLOW
VSA_EState5
MinAbsEStateIndex
fr_Ar_N
fr_NH0
NumSaturatedHeterocycles
MinEStateIndex
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Коэффициент корреляции


20 наименьших корреляций с CC50, mM
CC50, mM
MolMR -0.31 1.00 0.99 0.96 0.96 0.96 0.97 0.94 0.97 0.95 0.97 0.98 0.91 0.99 0.75 0.97 0.98 0.87 0.83 0.58 0.85
LabuteASA -0.31 0.99 1.00 0.99 0.99 0.99 0.99 0.98 0.99 0.97 0.99 0.96 0.91 0.98 0.75 0.95 0.97 0.83 0.88 0.61 0.81
MolWt -0.31 0.96 0.99 1.00 1.00 0.99 0.99 1.00 0.99 0.96 0.98 0.93 0.91 0.95 0.75 0.90 0.94 0.78 0.90 0.63 0.74
ExactMolWt -0.31 0.96 0.99 1.00 1.00 0.99 0.99 1.00 0.99 0.96 0.98 0.93 0.91 0.95 0.75 0.90 0.94 0.78 0.90 0.63 0.74
Chi0 -0.31 0.96 0.99 0.99 0.99 1.00 1.00 0.98 0.99 0.98 0.99 0.93 0.92 0.96 0.77 0.93 0.96 0.78 0.88 0.61 0.77
HeavyAtomCount -0.31 0.97 0.99 0.99 0.99 1.00 1.00 0.98 1.00 0.96 0.99 0.93 0.91 0.96 0.75 0.93 0.96 0.78 0.91 0.65 0.77
HeavyAtomMolWt -0.31 0.94 0.98 1.00 1.00 0.98 0.98 1.00 0.98 0.94 0.97 0.90 0.89 0.93 0.74 0.87 0.91 0.74 0.92 0.66 0.69
Chi1 -0.31 0.97 0.99 0.99 0.99 0.99 1.00 0.98 1.00 0.95 0.98 0.93 0.91 0.95 0.75 0.92 0.95 0.77 0.92 0.66 0.75
Kappa1 -0.31 0.95 0.97 0.96 0.96 0.98 0.96 0.94 0.95 1.00 0.99 0.93 0.96 0.96 0.84 0.94 0.97 0.81 0.78 0.50 0.80
NumValenceElectrons -0.30 0.97 0.99 0.98 0.98 0.99 0.99 0.97 0.98 0.99 1.00 0.95 0.93 0.97 0.78 0.95 0.98 0.82 0.85 0.56 0.82
Chi1v -0.30 0.98 0.96 0.93 0.93 0.93 0.93 0.90 0.93 0.93 0.95 1.00 0.86 0.99 0.70 0.98 0.98 0.94 0.76 0.45 0.91
Kappa2 -0.30 0.91 0.91 0.91 0.91 0.92 0.91 0.89 0.91 0.96 0.93 0.86 1.00 0.90 0.94 0.87 0.90 0.68 0.74 0.52 0.67
Chi0v -0.30 0.99 0.98 0.95 0.95 0.96 0.96 0.93 0.95 0.96 0.97 0.99 0.90 1.00 0.74 0.98 0.99 0.91 0.79 0.49 0.89
Kappa3 -0.30 0.75 0.75 0.75 0.75 0.77 0.75 0.74 0.75 0.84 0.78 0.70 0.94 0.74 1.00 0.70 0.74 0.53 0.56 0.42 0.52
Chi1n -0.29 0.97 0.95 0.90 0.90 0.93 0.93 0.87 0.92 0.94 0.95 0.98 0.87 0.98 0.70 1.00 0.99 0.92 0.72 0.43 0.94
Chi0n -0.29 0.98 0.97 0.94 0.94 0.96 0.96 0.91 0.95 0.97 0.98 0.98 0.90 0.99 0.74 0.99 1.00 0.90 0.77 0.48 0.90
Chi2v -0.27 0.87 0.83 0.78 0.78 0.78 0.78 0.74 0.77 0.81 0.82 0.94 0.68 0.91 0.53 0.92 0.90 1.00 0.56 0.24 0.97
BertzCT -0.27 0.83 0.88 0.90 0.90 0.88 0.91 0.92 0.92 0.78 0.85 0.76 0.74 0.79 0.56 0.72 0.77 0.56 1.00 0.82 0.51
NumAromaticCarbocycles -0.26 0.58 0.61 0.63 0.63 0.61 0.65 0.66 0.66 0.50 0.56 0.45 0.52 0.49 0.42 0.43 0.48 0.24 0.82 1.00 0.21
Chi2n -0.26 0.85 0.81 0.74 0.74 0.77 0.77 0.69 0.75 0.80 0.82 0.91 0.67 0.89 0.52 0.94 0.90 0.97 0.51 0.21 1.00
MolMR
LabuteASA
MolWt
ExactMolWt
Chi0
HeavyAtomCount
HeavyAtomMolWt
Chi1
Kappa1
NumValenceElectrons
Chi1v
Kappa2
Chi0v
Kappa3
Chi1n
Chi0n
Chi2v
BertzCT
NumAromaticCarbocycles
Chi2n
№
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Коэффициент корреляции

iii. **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:

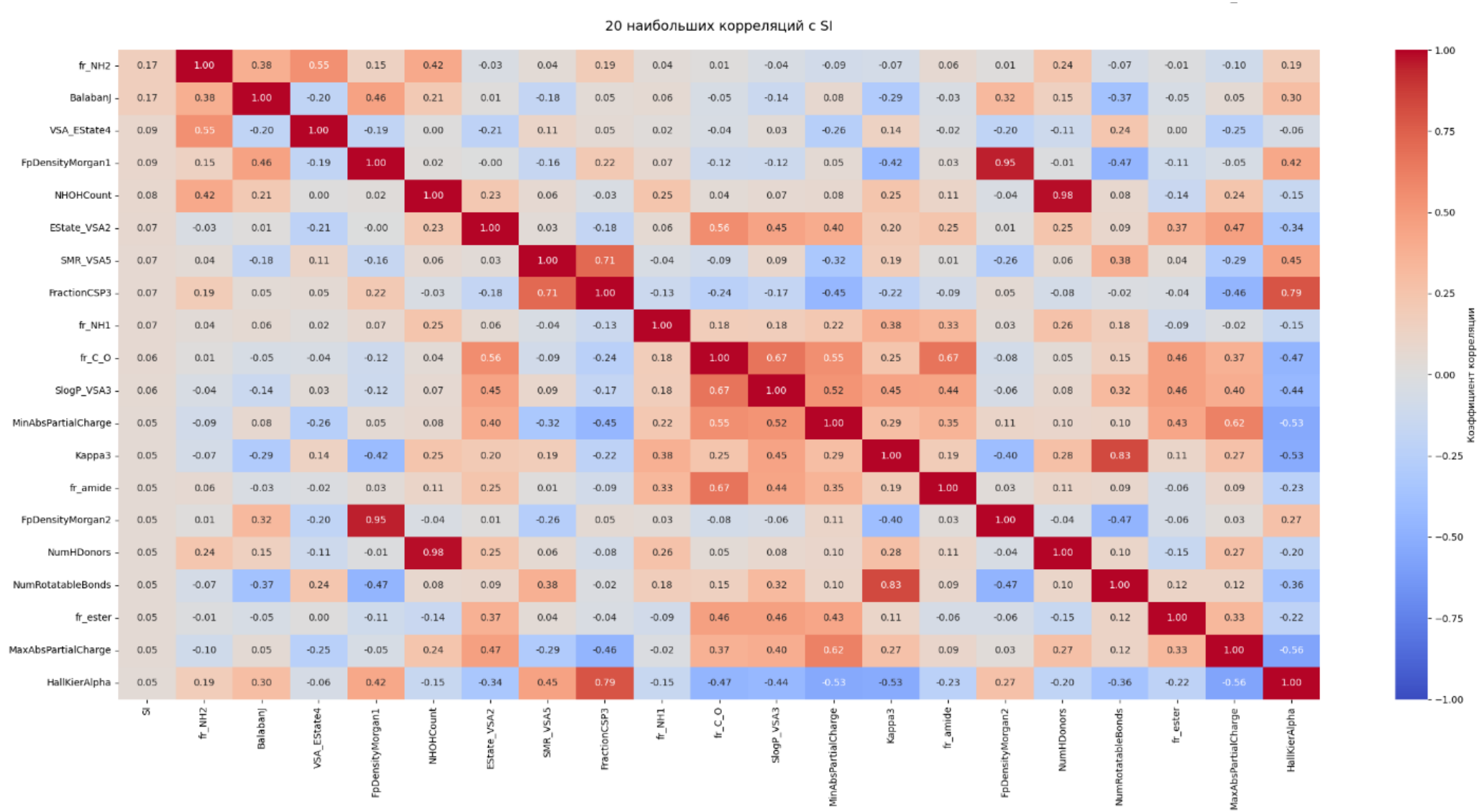
1. RingCount (-0.128**)** – много циклов снижает селективность.
2. NumAromaticRings (-0.090) – ароматические кольца могут ухудшать SI.

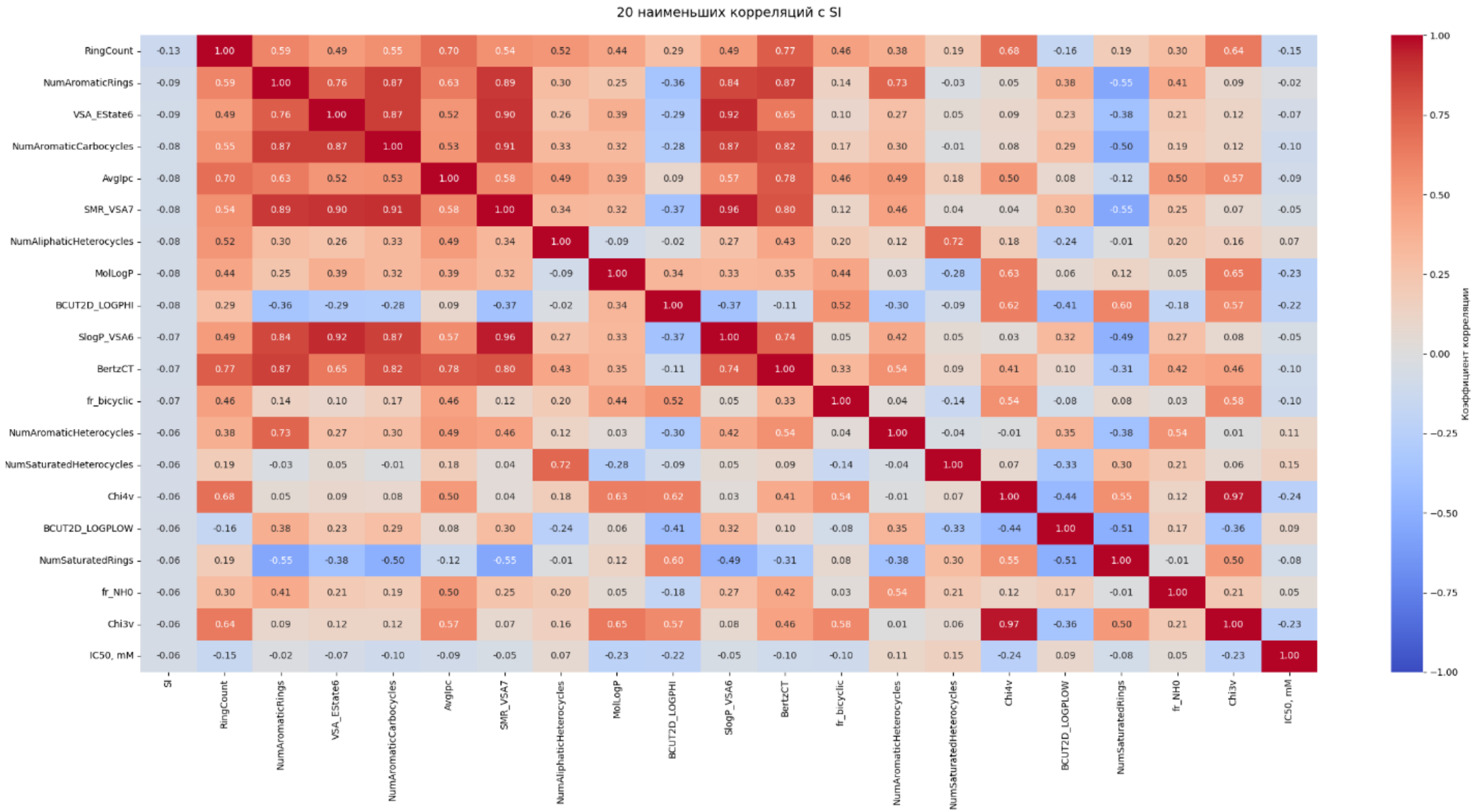
Таким образом при обработке отдельное внимание будет уделено:

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

iv. 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.


20 наибольших корреляций с SI
fr_NH2 0.17 1.00 0.38 0.55 0.15 0.42 -0.03 0.04 0.19 0.04 0.01 -0.04 -0.09 -0.07 0.06 0.01 0.24 -0.07 -0.01 -0.10 0.19
BalabanJ 0.17 0.38 1.00 -0.20 0.46 0.21 0.01 -0.18 0.05 0.06 -0.05 -0.14 0.08 -0.29 -0.03 0.32 0.15 -0.37 -0.05 0.05 0.30
VSA_EState4 0.09 0.55 -0.20 1.00 -0.19 0.00 -0.21 0.11 0.05 0.02 -0.04 0.03 -0.26 0.14 -0.02 -0.20 -0.11 0.24 0.00 -0.25 -0.06
FpDensityMorgan1 0.09 0.15 0.46 -0.19 1.00 0.02 -0.00 -0.16 0.22 0.07 -0.12 -0.12 0.05 -0.42 0.03 0.95 -0.01 -0.47 -0.11 -0.05 0.42
NHOHCount 0.08 0.42 0.21 0.00 0.02 1.00 0.23 0.06 -0.03 0.25 0.04 0.07 0.08 0.25 0.11 -0.04 0.98 0.08 -0.14 0.24 -0.15
EState_VSA2 0.07 -0.03 0.01 -0.21 -0.00 0.23 1.00 0.03 -0.18 0.06 0.56 0.45 0.40 0.20 0.25 0.01 0.25 0.09 0.37 0.47 -0.34
SMR_VSA5 0.07 0.04 -0.18 0.11 -0.16 0.06 0.03 1.00 0.71 -0.04 -0.09 0.09 -0.32 0.19 0.01 -0.26 0.06 0.38 0.04 -0.29 0.45
FractionCSP3 0.07 0.19 0.05 0.05 0.22 -0.03 -0.18 0.71 1.00 -0.13 -0.24 -0.17 -0.45 -0.22 -0.09 0.05 -0.08 -0.02 -0.04 -0.46 0.79
fr_NH1 0.07 0.04 0.06 0.02 0.07 0.25 0.06 -0.04 -0.13 1.00 0.18 0.18 0.22 0.38 0.33 0.03 0.26 0.18 -0.09 -0.02 -0.15
fr_C_O 0.06 0.01 -0.05 -0.04 -0.12 0.04 0.56 -0.09 -0.24 0.18 1.00 0.67 0.55 0.25 0.67 -0.08 0.05 0.15 0.46 0.37 -0.47
SlogP_VSA3 0.06 -0.04 -0.14 0.03 -0.12 0.07 0.45 0.09 -0.17 0.18 0.67 1.00 0.52 0.45 0.44 -0.06 0.08 0.32 0.46 0.40 -0.44
MinAbsPartialCharge 0.05 -0.09 0.08 -0.26 0.05 0.08 0.40 -0.32 -0.45 0.22 0.55 0.52 1.00 0.29 0.35 0.11 0.10 0.10 0.43 0.62 -0.53
Kappa3 0.05 -0.07 -0.29 0.14 -0.42 0.25 0.20 0.19 -0.22 0.38 0.25 0.45 0.29 1.00 0.19 -0.40 0.28 0.83 0.11 0.27 -0.53
fr_amide 0.05 0.06 -0.03 -0.02 0.03 0.11 0.25 0.01 -0.09 0.33 0.67 0.44 0.35 0.19 1.00 0.03 0.11 0.09 -0.06 0.09 -0.23
FpDensityMorgan2 0.05 0.01 0.32 -0.20 0.95 -0.04 0.01 -0.26 0.05 0.03 -0.08 -0.06 0.11 -0.40 0.03 1.00 -0.04 -0.47 -0.06 0.03 0.27
NumHDonors 0.05 0.24 0.15 -0.11 -0.01 0.98 0.25 0.06 -0.08 0.26 0.05 0.08 0.10 0.28 0.11 -0.04 1.00 0.10 -0.15 0.27 -0.20
NumRotatableBonds 0.05 -0.07 -0.37 0.24 -0.47 0.08 0.09 0.38 -0.02 0.18 0.15 0.32 0.10 0.83 0.09 -0.47 0.10 1.00 0.12 0.12 -0.36
fr_ester 0.05 -0.01 -0.05 0.00 -0.11 -0.14 0.37 0.04 -0.04 -0.09 0.46 0.46 0.43 0.11 -0.06 -0.06 -0.15 0.12 1.00 0.33 -0.22
MaxAbsPartialCharge 0.05 -0.10 0.05 -0.25 -0.05 0.24 0.47 -0.29 -0.46 -0.02 0.37 0.40 0.62 0.27 0.09 0.03 0.27 0.12 0.33 1.00 -0.56
HallKierAlpha 0.05 0.19 0.30 -0.06 0.42 -0.15 -0.34 0.45 0.79 -0.15 -0.47 -0.44 -0.53 -0.53 -0.23 0.27 -0.20 -0.36 -0.22 -0.56 1.00
SI
fr_NH2
BalabanJ
VSA_EState4
FpDensityMorgan1
NHOHCount
EState_VSA2
SMR_VSA5
FractionCSP3
fr_NH1
fr_C_O
SlogP_VSA3
MinAbsPartialCharge
Kappa3
fr_amide
FpDensityMorgan2
NumHDonors
NumRotatableBonds
fr_ester
MaxAbsPartialCharge
HallKierAlpha
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Коэффициент корреляции


20 наименьших корреляций с SI
RingCount -0.13 1.00 0.59 0.49 0.55 0.70 0.54 0.52 0.44 0.29 0.49 0.77 0.46 0.38 0.19 0.68 -0.16 0.19 0.30 0.64 -0.15
NumAromaticRings -0.09 0.59 1.00 0.76 0.87 0.63 0.89 0.30 0.25 -0.36 0.84 0.87 0.14 0.73 -0.03 0.05 0.38 -0.55 0.41 0.09 -0.02
VSA_EState6 -0.09 0.49 0.76 1.00 0.87 0.52 0.90 0.26 0.39 -0.29 0.92 0.65 0.10 0.27 0.05 0.09 0.23 -0.38 0.21 0.12 -0.07
NumAromaticCarbocycles -0.08 0.55 0.87 0.87 1.00 0.53 0.91 0.33 0.32 -0.28 0.87 0.82 0.17 0.30 -0.01 0.08 0.29 -0.50 0.19 0.12 -0.10
AvgIpc -0.08 0.70 0.63 0.52 0.53 1.00 0.58 0.49 0.39 0.09 0.57 0.78 0.46 0.49 0.18 0.50 0.08 -0.12 0.50 0.57 -0.09
SMR_VSA7 -0.08 0.54 0.89 0.90 0.91 0.58 1.00 0.34 0.32 -0.37 0.96 0.80 0.12 0.46 0.04 0.04 0.30 -0.55 0.25 0.07 -0.05
NumAliphaticHeterocycles -0.08 0.52 0.30 0.26 0.33 0.49 0.34 1.00 -0.09 -0.02 0.27 0.43 0.20 0.12 0.72 0.18 -0.24 -0.01 0.20 0.16 0.07
MolLogP -0.08 0.44 0.25 0.39 0.32 0.39 0.32 -0.09 1.00 0.34 0.33 0.35 0.44 0.03 -0.28 0.63 0.06 0.12 0.05 0.65 -0.23
BCUT2D_LOGPHI -0.08 0.29 -0.36 -0.29 -0.28 0.09 -0.37 -0.02 0.34 1.00 -0.37 -0.11 0.52 -0.30 -0.09 0.62 -0.41 0.60 -0.18 0.57 -0.22
SlogP_VSA6 -0.07 0.49 0.84 0.92 0.87 0.57 0.96 0.27 0.33 -0.37 1.00 0.74 0.05 0.42 0.05 0.03 0.32 -0.49 0.27 0.08 -0.05
BertzCT -0.07 0.77 0.87 0.65 0.82 0.78 0.80 0.43 0.35 -0.11 0.74 1.00 0.33 0.54 0.09 0.41 0.10 -0.31 0.42 0.46 -0.10
fr_bicyclic -0.07 0.46 0.14 0.10 0.17 0.46 0.12 0.20 0.44 0.52 0.05 0.33 1.00 0.04 -0.14 0.54 -0.08 0.08 0.03 0.58 -0.10
NumAromaticHeterocycles -0.06 0.38 0.73 0.27 0.30 0.49 0.46 0.12 0.03 -0.30 0.42 0.54 0.04 1.00 -0.04 -0.01 0.35 -0.38 0.54 0.01 0.11
NumSaturatedHeterocycles -0.06 0.19 -0.03 0.05 -0.01 0.18 0.04 0.72 -0.28 -0.09 0.05 0.09 -0.14 -0.04 1.00 0.07 -0.33 0.30 0.21 0.06 0.15
Chi4v -0.06 0.68 0.05 0.09 0.08 0.50 0.04 0.18 0.63 0.62 0.03 0.41 0.54 -0.01 0.07 1.00 -0.44 0.55 0.12 0.97 -0.24
BCUT2D_LOGPLOW -0.06 -0.16 0.38 0.23 0.29 0.08 0.30 -0.24 0.06 -0.41 0.32 0.10 -0.08 0.35 -0.33 -0.44 1.00 -0.51 0.17 -0.36 0.09
NumSaturatedRings -0.06 0.19 -0.55 -0.38 -0.50 -0.12 -0.55 -0.01 0.12 0.60 -0.49 -0.31 0.08 -0.38 0.30 0.55 -0.51 1.00 -0.01 0.50 -0.08
fr_NH0 -0.06 0.30 0.41 0.21 0.19 0.50 0.25 0.20 0.05 -0.18 0.27 0.42 0.03 0.54 0.21 0.12 0.17 -0.01 1.00 0.21 0.05
Chi3v -0.06 0.64 0.09 0.12 0.12 0.57 0.07 0.16 0.65 0.57 0.08 0.46 0.58 0.01 0.06 0.97 -0.36 0.50 0.21 1.00 -0.23
IC50, mM -0.06 -0.15 -0.02 -0.07 -0.10 -0.09 -0.05 0.07 -0.23 -0.22 -0.05 -0.10 -0.10 0.11 0.15 -0.24 0.09 -0.08 0.05 -0.23 1.00
SI
RingCount
NumAromaticRings
VSA_EState6
NumAromaticCarbocycles
AvgIpc
SMR_VSA7
NumAliphaticHeterocycles
MolLogP
BCUT2D_LOGPHI
SlogP_VSA6
BertzCT
fr_bicyclic
NumAromaticHeterocycles
NumSaturatedHeterocycles
Chi4v
BCUT2D_LOGPLOW
NumSaturatedRings
fr_NH0
Chi3v
IC50, mM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Коэффициент корреляции

h. Итог по анализу молекулярных данных

i. 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

ii. 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

III. Исследовательский анализ

a. 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.

i. Сравнение эффективности моделей

В ходе исследования были протестированы три модели:

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

Оптимизация гиперпараметров XGBoost позволила улучшить его показатели:

- Снижение RMSE на ~8.4% по сравнению с базовой версией
- Увеличение R^2 на ~27%

Эффективность моделей:

1. Оптимизированный XGBoost показал наилучшую точность (RMSE = 1.3933, R^2 = 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, R^2 = 0.452).
3. Базовый XGBoost без настройки гиперпараметров оказался менее точным, что подтверждает важность оптимизации модели для достижения максимальной предсказательной силы.

ii. Теоретические аспекты и рекомендации

Прогнозирование индекса селективности (IC_{50}) представляет собой сложную задачу регрессии, где ключевую роль играют молекулярные дескрипторы и физико-химические свойства соединений. IC_{50} 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 IC_{50}:

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 IC_{50}.
3. Неоднородность данных
    - Экспериментальные значения IC_{50} могут варьироваться в зависимости от условий проведения тестов (pH, температура, тип клеточной линии), что создает дополнительный шум в данных.

b. 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 (CC_{50}) представляет собой сложную задачу регрессионного анализа, где ключевое значение имеют молекулярные дескрипторы и физико-химические свойства соединений. CC_{50} 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 (CC_{50}):
- Диапазон значений: от 0.1 до 10.0 (логарифмическая шкала)
- Среднее значение: 4.2 ± 1.8

i. Предобработка данных

Для подготовки данных к моделированию были выполнены:

- Нормализация данных:
- Логарифмическое преобразование значений CC_{50}
- Стандартизация числовых признаков

Обработка выбросов:

- Применение метода межквартильного размаха
- Winsorization для экстремальных значений

Отбор признаков:

- Удаление низковариативных дескрипторов
- Корреляционный анализ для устранения мультиколлинеарности

ii. 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 R^2
- Оценочные метрики:
- RMSE (корень из среднеквадратичной ошибки)
- R^2 (коэффициент детерминации)
- Средняя абсолютная ошибка (MAE)

iii. Результаты моделирования

Сравнительные показатели моделей:

Модель	**RMSE**	**R^2**	**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 R^2 на 17%

Эффективность моделей:

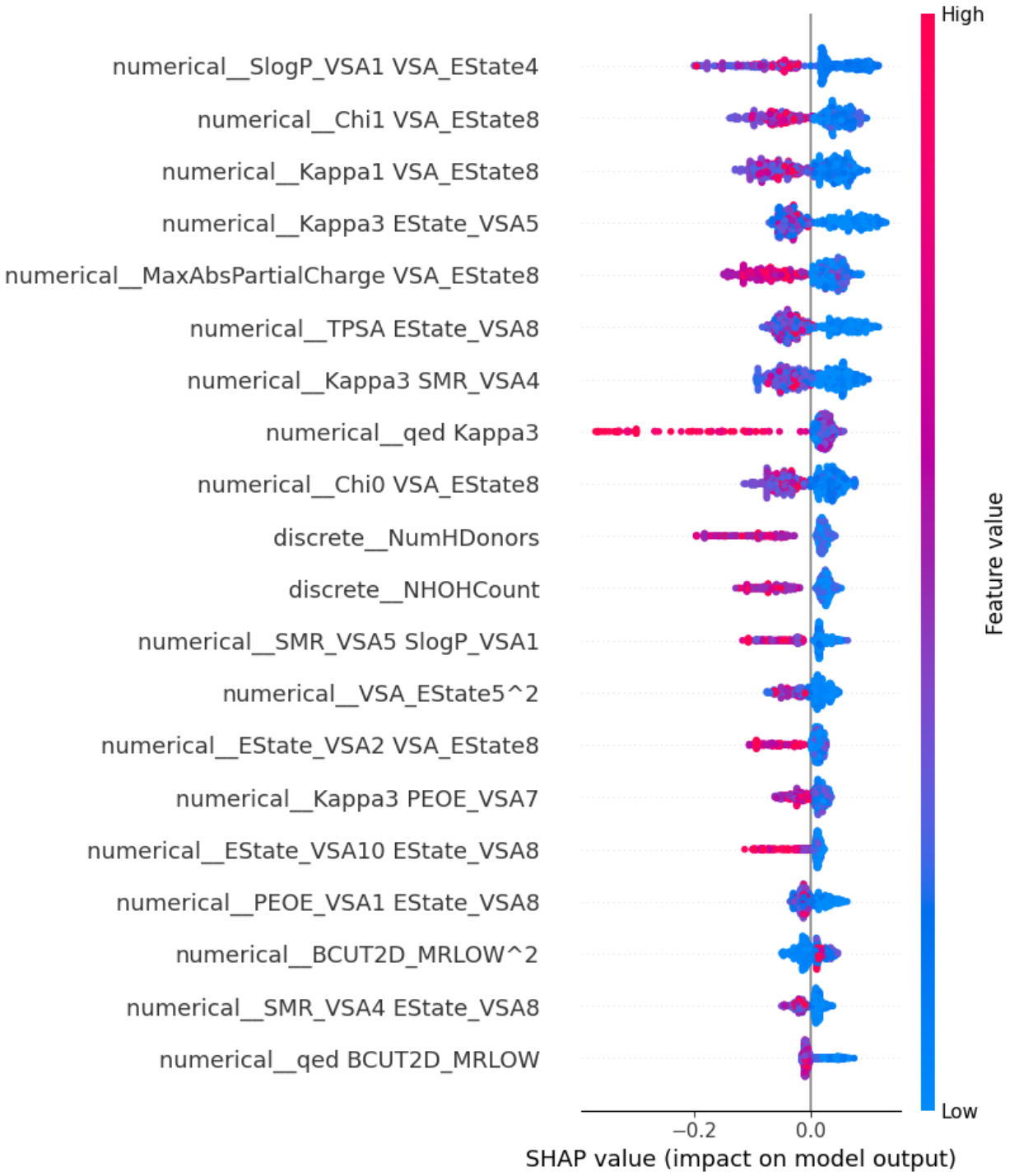
- Random Forest продемонстрировал наилучшую прогностическую способность среди всех тестируемых моделей (RMSE = 1.261, R^2 = 0.324), что свидетельствует о его высокой стабильности и надежности при работе с данными о цитотоксической активности соединений.
- Оптимизированный XGBoost показал значительное улучшение показателей по сравнению с базовой версией (RMSE = 1.281, R^2 = 0.301), подтверждая важность тщательной настройки гиперпараметров для достижения оптимальной производительности модели.
- Базовый XGBoost (RMSE = 1.321, R^2 = 0.257) оказался наименее эффективным среди рассматриваемых подходов, что подчеркивает необходимость обязательной оптимизации алгоритмов бустинга для задач прогнозирования CC_{50}.

Все модели показали умеренную предсказательную способность (R^2 <0.5), что характерно для сложных задач прогнозирования биологической активности, где на конечный результат влияет множество трудноучитываемых факторов.

iv. Анализ важности признаков

Наиболее значимые дескрипторы для прогнозирования CC_{50}:

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



v. Теоретические аспекты и рекомендации

Ключевые аспекты прогнозирования CC_{50}

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

c. Анализ моделей прогнозирования для регрессии индекса селективности (SI)

В данном разделе рассматривалась задача регрессионного анализа для прогнозирования индекса селективности ($SI = CC_{50}/IC_{50}$) на основе молекулярных характеристик соединений. Прогнозирование индекса селективности ($SI = CC_{50}/IC_{50}$**)** 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 (CC_{50}) и противовирусной активностью (IC_{50}). Высокий SI указывает на селективность соединения – способность подавлять вирус при минимальном вреде для клеток хозяина.

i. 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 R^2
- Оценка стабильности модели на тестовой выборке

Оценочные метрики:

- RMSE (корень из среднеквадратичной ошибки)
- R^2 (коэффициент детерминации)
- Сравнение ошибок на тренировочной и тестовой выборках

ii. Результаты моделирования

Сравнительные показатели моделей

Модель	**Train RMSE**	**Test RMSE**	**Train R^2**	**Test R^2**
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 R^2 (0.3002)
2. Random Forest продемонстрировал умеренную предсказательную способность ($R^2 = 0.2106$), но уступает бустинговым методам.
3. XGBoost показал сбалансированные результаты, но немного хуже Gradient Boosting.
4. Decision Tree оказался неэффективен ($R^2 < 0$), что указывает на переобучение.

Таким образом выявлено следующее:

- Оптимизация гиперпараметров улучшила R^2 на 9% для Gradient Boosting и 12% для XGBoost.
- Gradient Boosting – лучшая модель для прогнозирования SI, но общее качество моделей остается умеренным ($R^2 < 0.5$), что характерно для сложных биологических данных.

iii. 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)

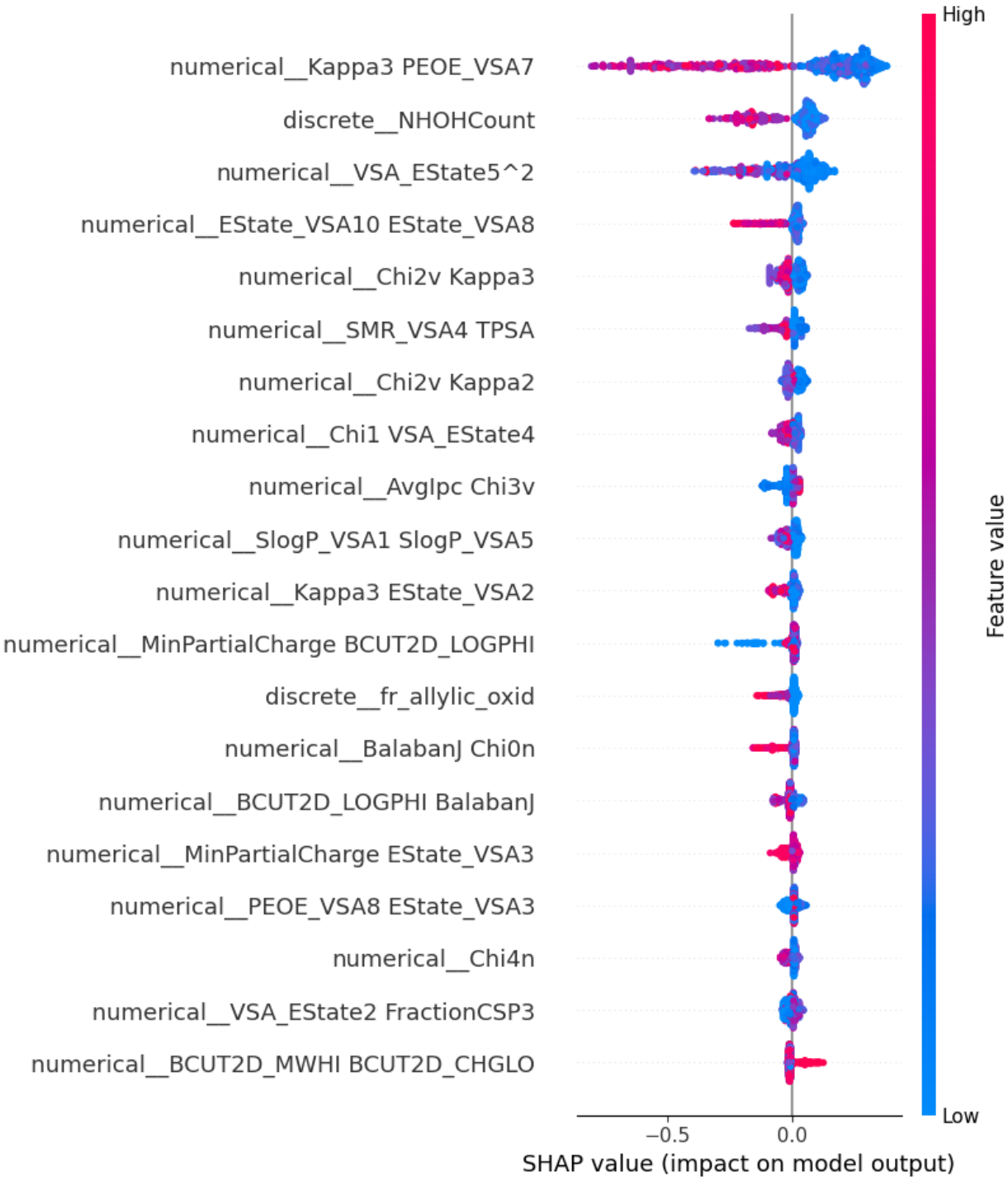
На основе анализа важности признаков и моделей можно сформулировать следующие принципы:

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).



iv. Теоретические аспекты и рекомендации

Аспекты прогнозирования SI:

1. **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.

2. **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.

3. **Особенности биологических данных.**

Прогнозирование SI осложняется рядом факторов:

- Зависимость от типа клеточных линий**:**
  - CC_{50} может значительно варьироваться для разных клеток (например, гепатоциты vs. фибробласты).
- Условия эксперимента:
  - Время инкубации, pH среды, концентрация сыворотки влияют на IC_{50} и CC_{50}.
  - Методологические различия между лабораториями:
    - Разные протоколы измерения IC_{50}/CC_{50} приводят к вариабельности данных.

d. 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 = CC_{50}/IC_{50}) медианного значения выборки представляет собой важную задачу бинарной классификации в drug discovery. Данная задача позволяет выделить перспективные соединения с оптимальным балансом эффективности (низкий IC_{50}) и безопасности (высокий CC_{50}). Решение этой проблемы требует комплексного подхода, учитывающего молекулярные особенности соединений и специфику биологических данных.

i. Этапы работы

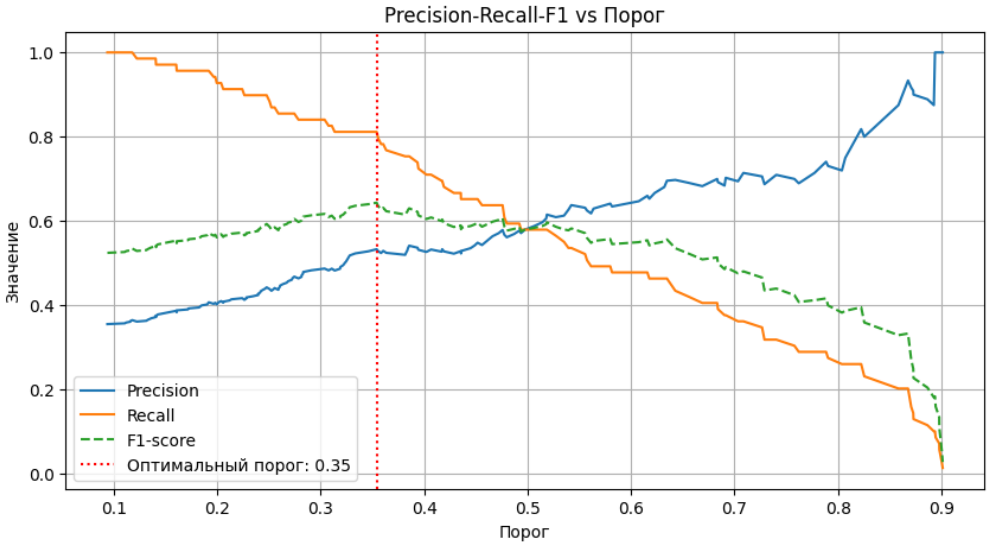
1. Предобработка данных:
   - Удаление константных и квази-константных признаков
   - Проверка на мультиколлинеарность
   - Робастное масштабирование числовых признаков
2. Построение моделей:
   - Random Forest (baseline)
   - XGBoost с параметрами по умолчанию
   - Оптимизированный XGBoost (GridSearchCV)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

ii. Сравнительный анализ моделей

Модель	**RMSE**	**R^2 (тест)**	**R^2 (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**:

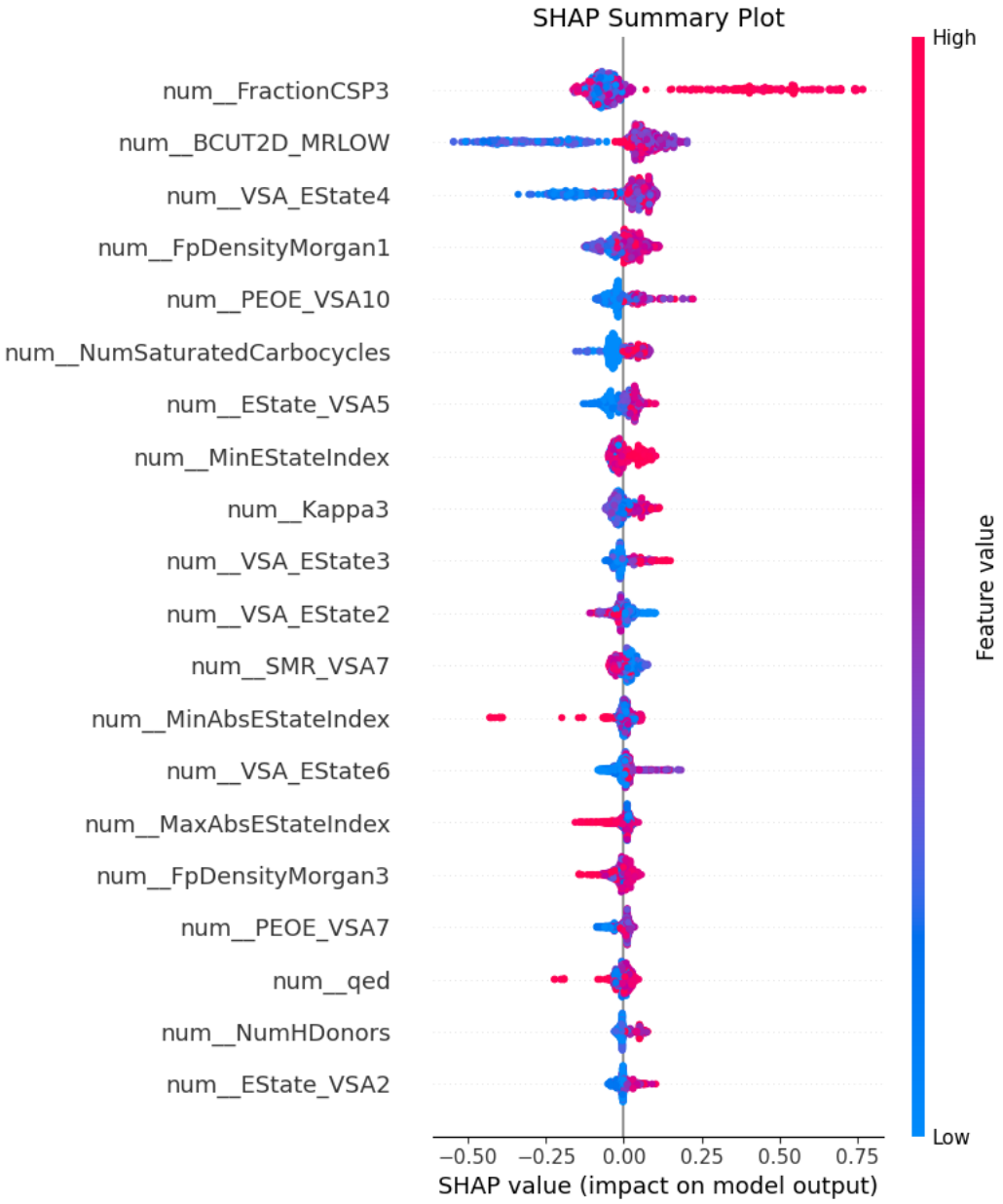
- Бинарная классификация на основе медианного значения SI
- Сильный дисбаланс классов в случае неравномерного распределения SI
- Высокая чувствительность к выбросам из-за широкого диапазона значений SI (0.01-15620.6)
- XGBoost значительно превосходит RandomForest по всем метрикам:
  - Уменьшение RMSE на 38.4%
  - Улучшение R^2 на 51.3%
- Оптимизация гиперпараметров дала незначительное улучшение:
  - R^2 увеличился с 0.828 до 0.828 (практически без изменений)
  - Основной выигрыш - повышение стабильности модели
- Проблемы стабильности**:**
  - Большой разброс R^2 при кросс-валидации (±0.325)
  - Отрицательные значения R^2 для RandomForest указывают на непригодность модели



iii. 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



e. Классификация: превышает ли значение IC50 медианное значение выборки

Прогнозирование превышения IC_{50} медианного значения выборки представляет собой важную задачу бинарной классификации в разработке лекарственных препаратов. IC_{50} (полумаксимальная ингибирующая концентрация) является ключевым параметром, характеризующим эффективность соединения – чем ниже IC_{50}, тем выше противовирусная активность. Классификация соединений по этому признаку позволяет выявлять потенциально неэффективные молекулы на ранних этапах скрининга.

i. 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

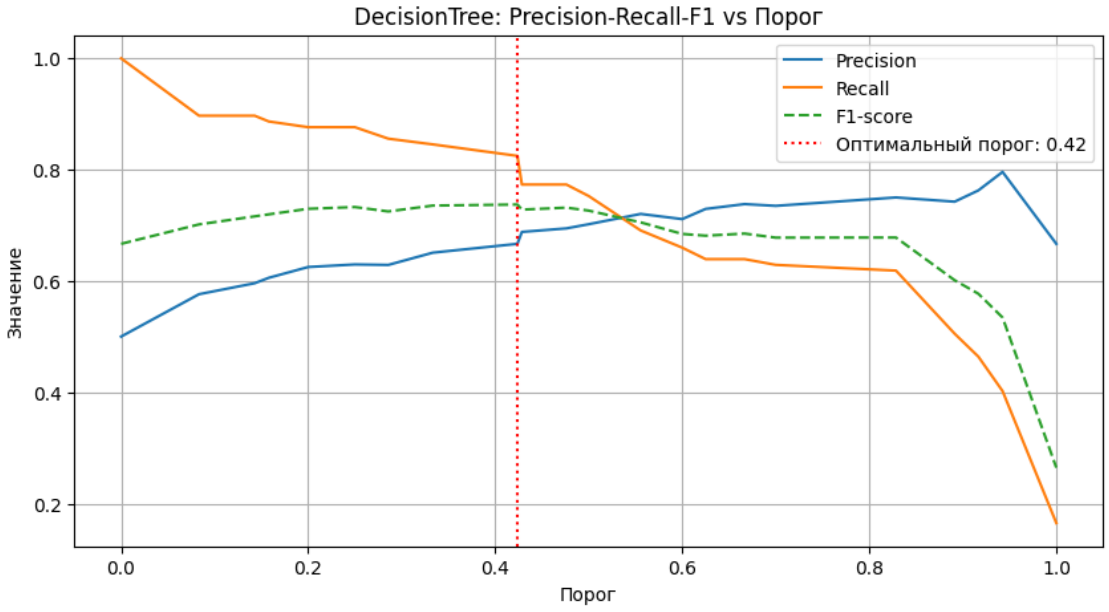
1. Предобработка данных:
    - Удаление константных и коррелирующих признаков
    - Балансировка классов (стратифицированное разбиение)
    - Робастное масштабирование числовых признаков
2. Построение моделей:
    - DecisionTree (baseline)
    - RandomForest
    - GradientBoosting
3. Оценка качества: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

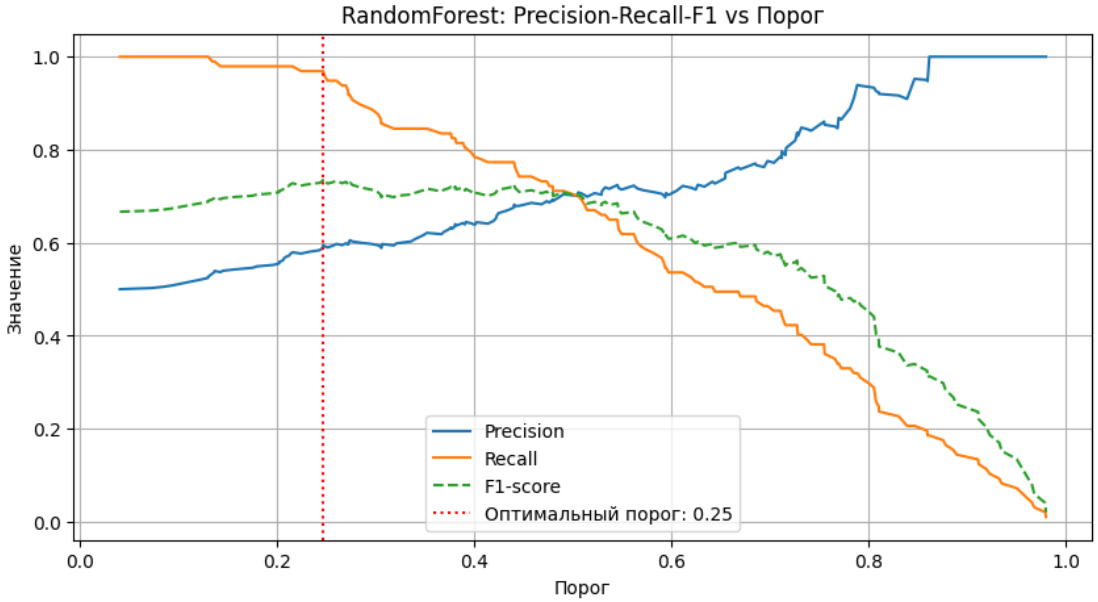
ii. Сравнительный анализ моделей

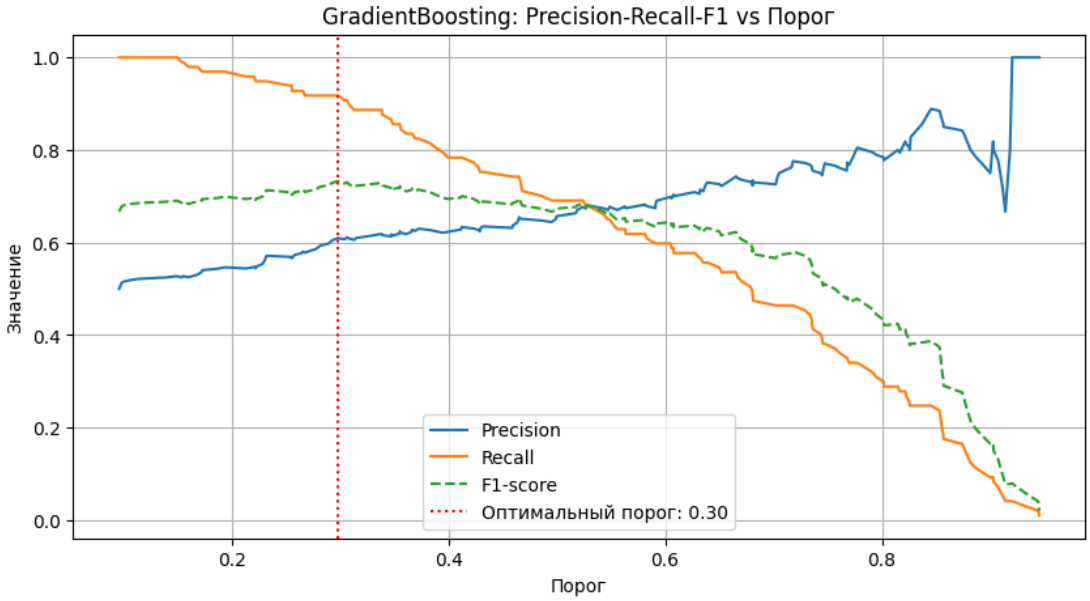
Модель	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.66	0.66	0.66	0.6958
DecisionTree	0.72	0.69	0.71	0.7509
RandomForest	0.70	0.70	0.70	0.7726
GradientBoosting	0.66	0.69	0.68	0.7416

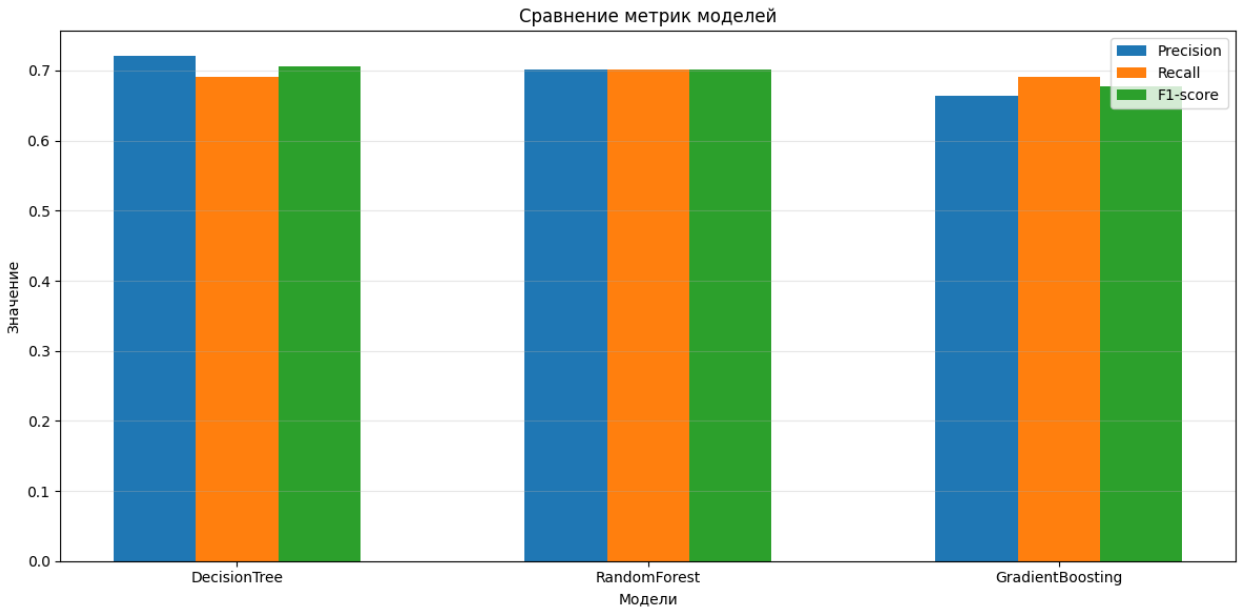
iii. **Ключевые наблюдения**

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.




RandomForest: Precision-Recall-F1 vs Порог
Значение
Порог
Precision
Recall
F1-score
Оптимальный порог: 0.25


GradientBoosting: Precision-Recall-F1 vs Порог
Значение
Порог
Precision
Recall
F1-score
Оптимальный порог: 0.30


Сравнение метрик моделей
Значение
Модели
Precision
Recall
F1-score
DecisionTree
RandomForest
GradientBoosting

iv. Ключевые аспекты прогнозирования

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

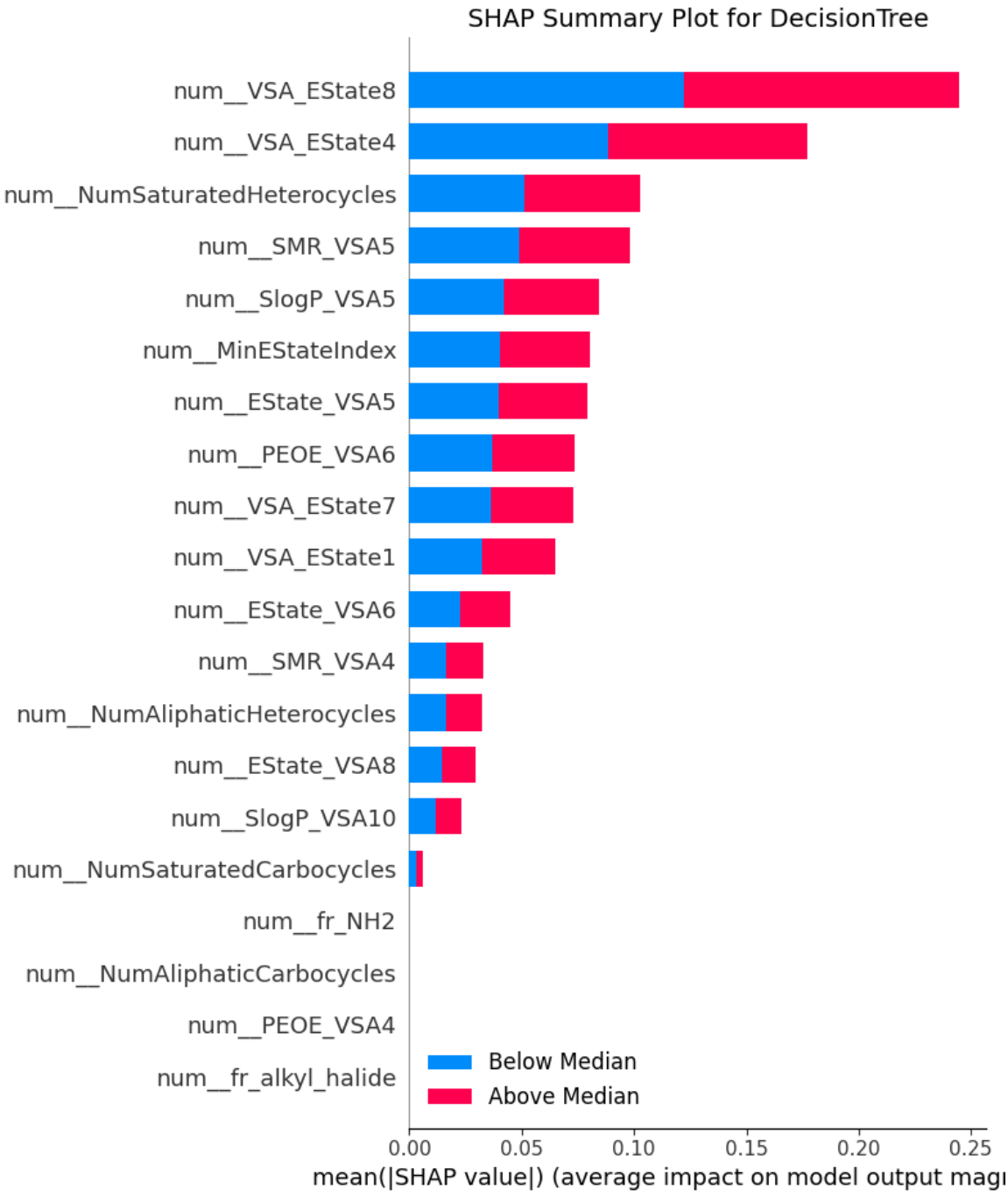
- **Топологические индексы**:
  - *Chi3v* - характеризует молекулярную сложность и разветвленность
  - *Kappa* - описывает форму молекулярного каркаса
- **Электронные параметры**:
  - *EState_VSA2* - электростатический потенциал молекулярной поверхности
  - *PEOE_VSA6* - распределение парциальных зарядов
- **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 IC_{50}
- **Молекулярная масса**:
  - Соединения >500 Da чаще демонстрируют высокие значения IC_{50}
- **Полярность поверхности**:
  - Низкие значения EState_VSA2 (<0.3) указывают на потенциально высокий IC_{50}

Таким образом, необходимо будет обратить внимание на:

- Соединения с высоким IC_{50} (> 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).



f. Классификация: превышает ли значение CC50 медианное значение выборки

Прогнозирование превышения CC_{50} медианного значения выборки представляет собой важную задачу бинарной классификации в токсикологических исследованиях. CC_{50} (полуцитотоксическая концентрация) является ключевым параметром безопасности соединений – чем выше CC_{50}, тем ниже цитотоксичность. Классификация позволяет выявлять потенциально токсичные соединения на ранних этапах разработки препаратов.

1. **Предобработка данных:**

    1. Логарифмическое преобразование CC_{50}
    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)

i. 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:

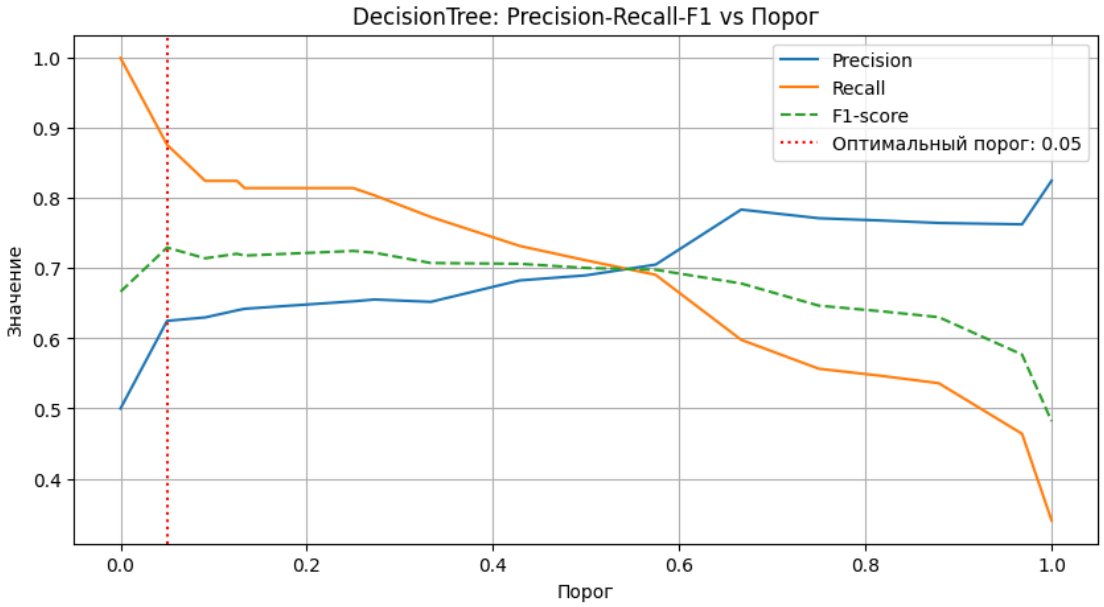
- Основные:
    - Precision, Recall, F1-score
    - ROC-AUC
- Дополнительные:
    - Матрица ошибок
    - Кросс-валидация (5 folds)

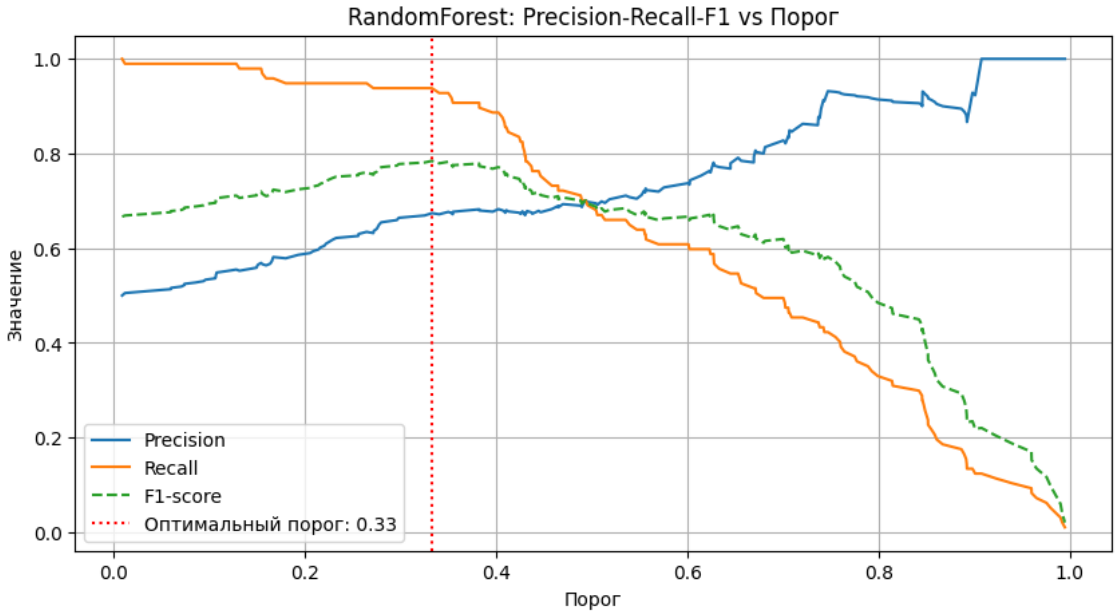
ii. Сравнительный анализ моделей

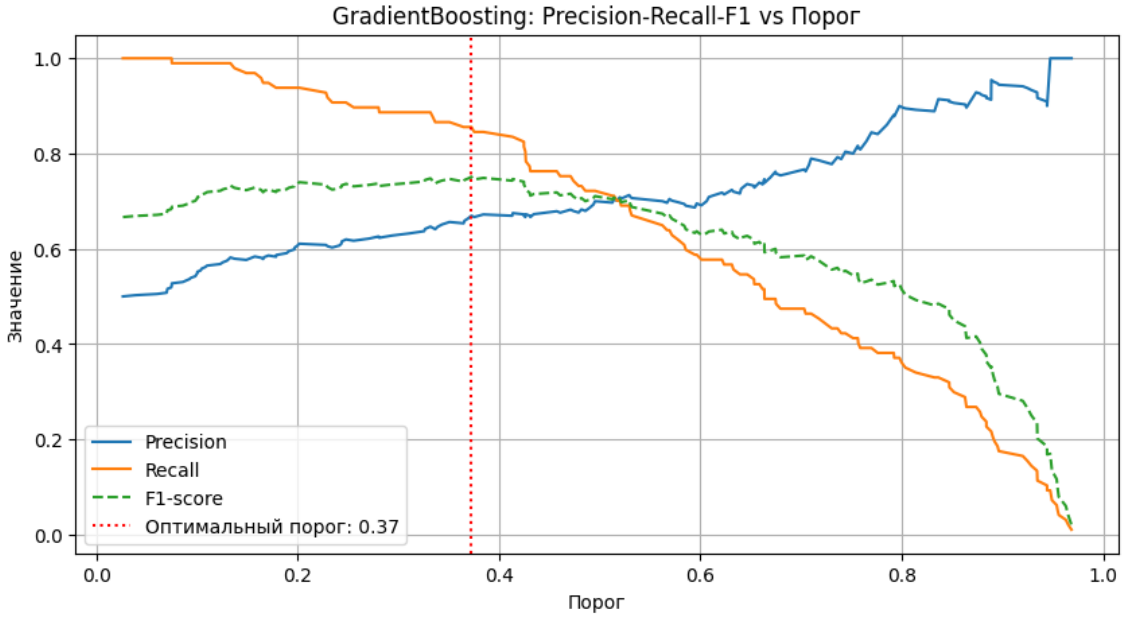
Модель	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Logistic Regression	0.67	0.67	0.67	0.752
Decision Tree	0.70	0.70	0.70	0.761
Random Forest	0.69	0.69	0.69	0.805
Gradient Boosting	0.70	0.70	0.70	0.783

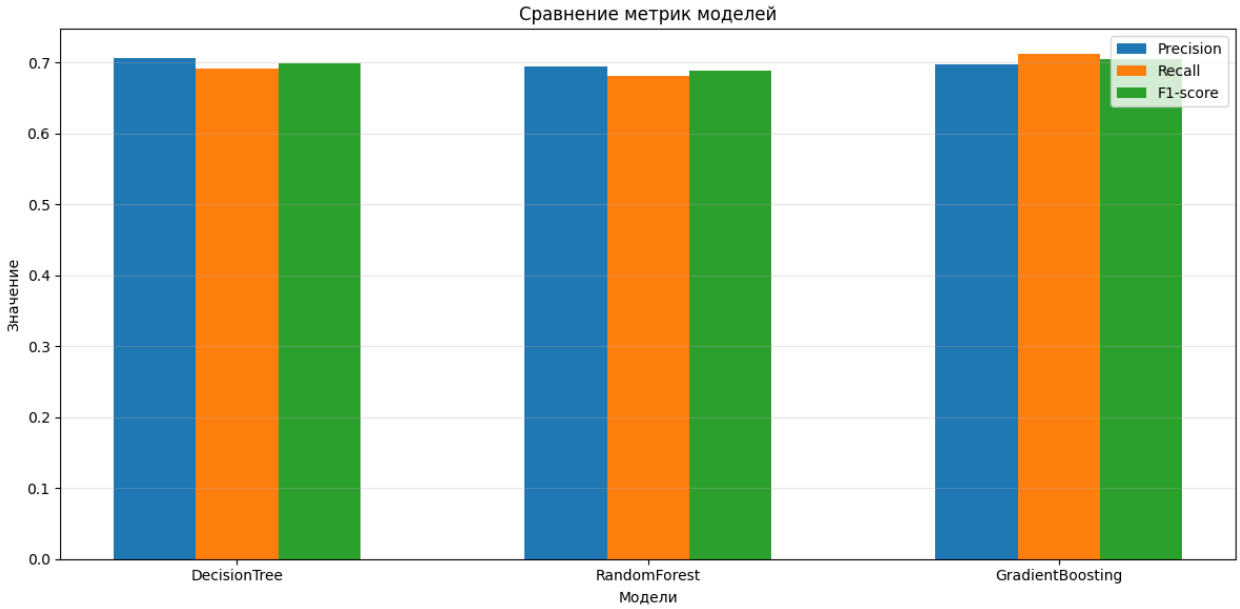
Ключевые выводы:

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









iii. Анализ важности признаков (SHAP)

1. 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

2. Физико-химические свойства

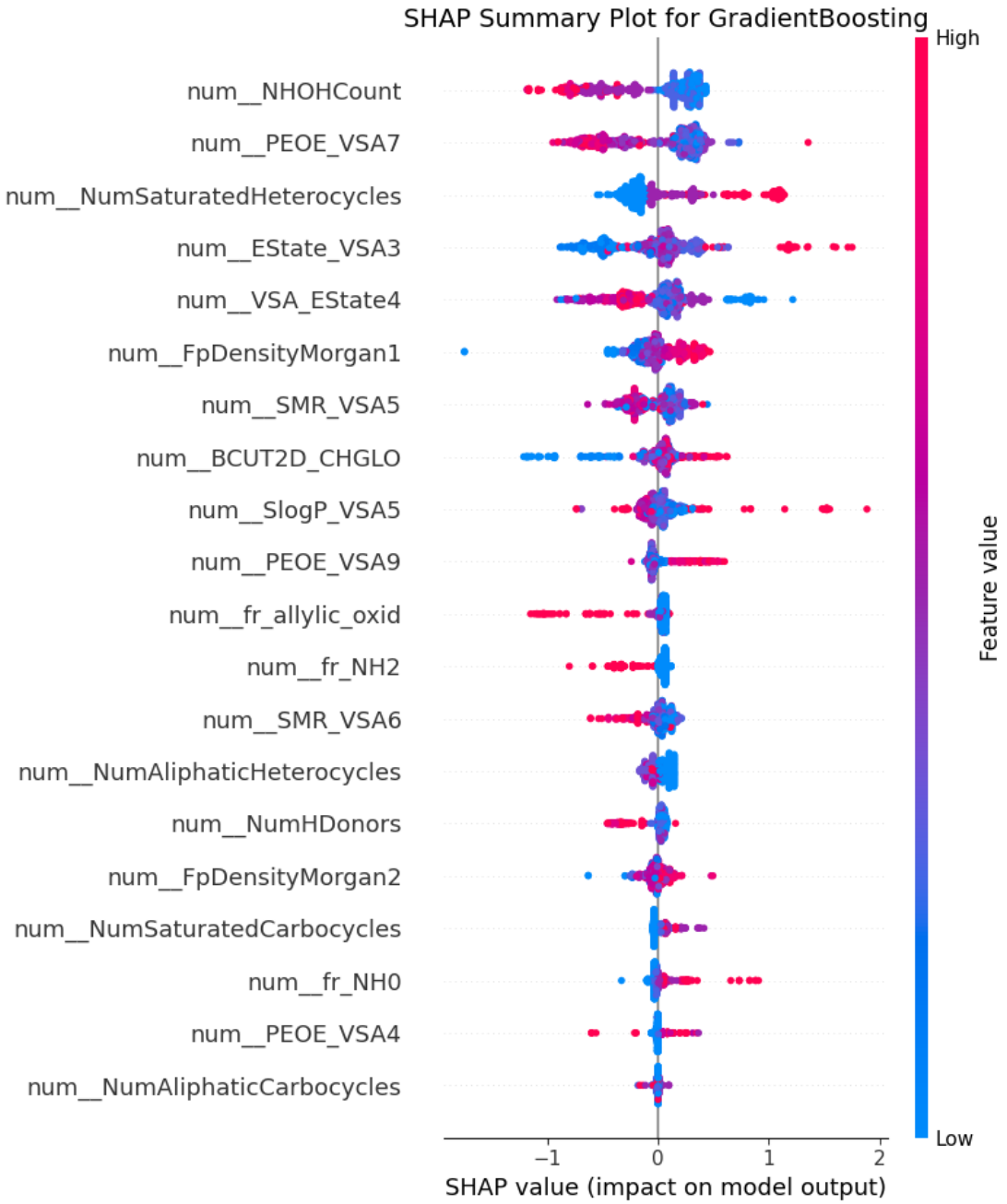
- Липофильность (SlogP_VSA4):
  - Значения >3.0 ассоциированы с высокой цитотоксичностью
- Полярность поверхности:
  - Соединения с EState_VSA2 <0.3 более токсичны
- Молекулярная гибкость:
  - 3-5 вращающихся связей - оптимальный диапазон

3. Функциональные группы

- Аминогруппы (fr_NH2):
  - Наличие 1-2 групп снижает цитотоксичность
- Ароматические системы:
  - 1-2 цикла улучшают профиль безопасности

Практические рекомендации для снижения CC_{50}**:**

- Контролировать электростатические свойства (BCUT2D_CHGLO 0.8-1.2)
- Поддерживать HallKierAlpha в диапазоне 0.1-0.3
- Ограничивать число ароматических циклов (1-2)



g. 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 (IC_{50}) и безопасностью (CC_{50}) потенциального лекарственного соединения. Разработанная модель позволяет эффективно идентифицировать перспективные соединения на ранних этапах скрининга.

i. Исходные данные

2. 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

3. Использованные алгоритмы:

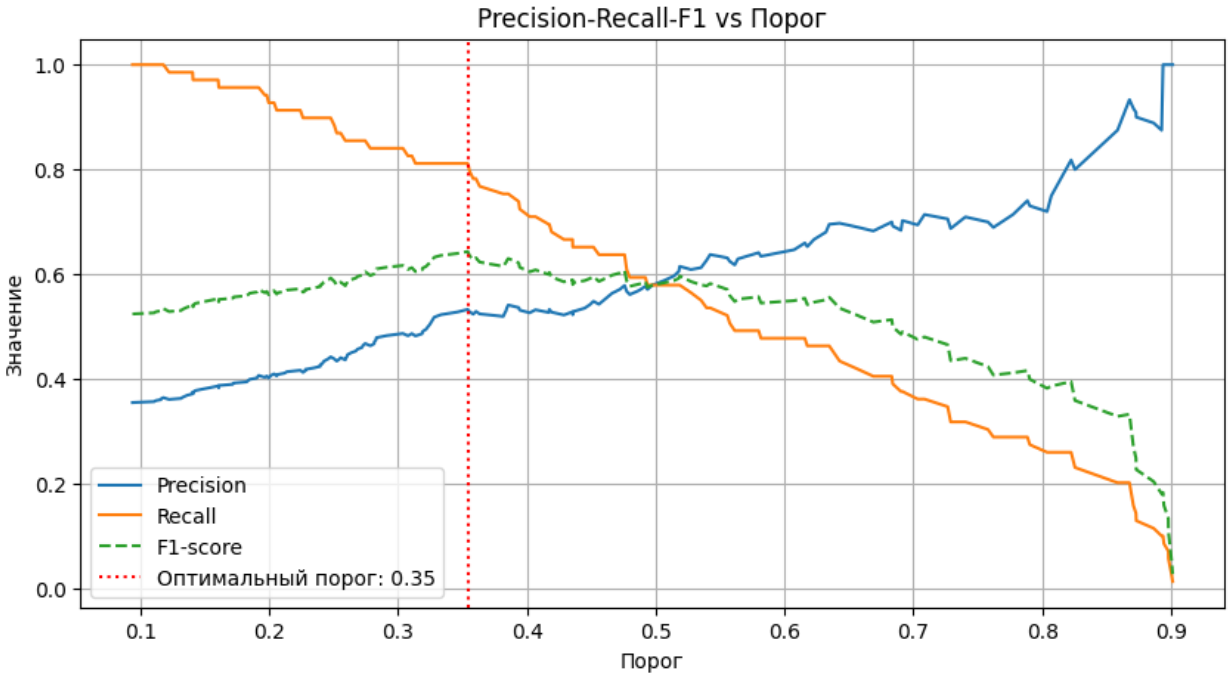
1. Logistic Regression (базовая модель)
2. XGBoost (базовая реализация)
3. XGBoost (с оптимизацией гиперпараметров)

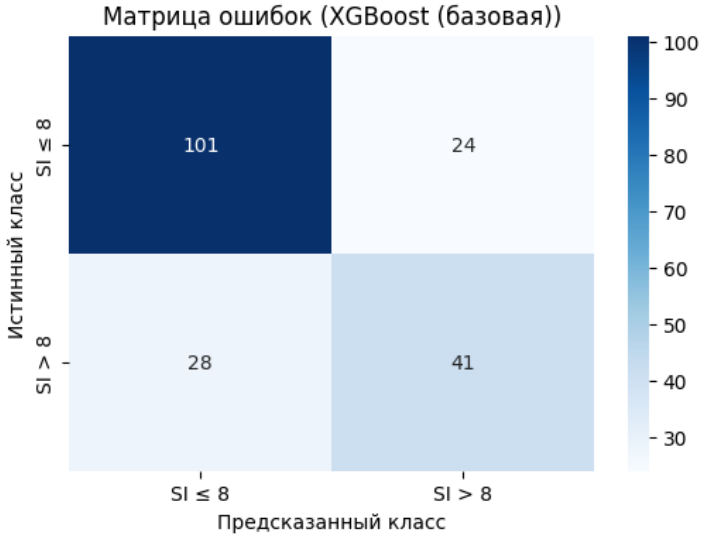
4. Метрики оценки:

- Основные:
  - Accuracy, ROC-AUC
  - F1-score для каждого класса
- Дополнительные:
  - Матрица ошибок
  - Кросс-валидация (5 folds)

iii. 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


Precision-Recall-F1 vs Порог
Значение
Порог
Precision
Recall
F1-score
Оптимальный порог: 0.35


Матрица ошибок (XGBoost (базовая))
Истинный класс
Предсказанный класс
SI ≤ 8
SI > 8
101
24
28
41


Матрица ошибок (XGBoost (оптимизированный))
Истинный класс
Предсказанный класс
SI ≤ 8
SI > 8
97
28
29
40

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

IV. Финальный вывод по проделанной работе

i. 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 (IC_{50}, CC_{50}, 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 IC_{50}: $R^2 = 0.958$, 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 (NH_2, 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.

ii. 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.

V. Источники данных

1. **Датасет молекулярных дескрипторов**: Данные были получены в качестве материалов в рамках курсовой работы
2. **Целевые переменные (IC_{50}, CC_{50}, 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)
3. **Методология расчета дескрипторов**: Для вычисления молекулярных дескрипторов использовались инструменты:
   - RDKit (https://www.rdkit.org/)
   - PaDEL-Descriptor (http://www.yapcwsoft.com/dd/padeldescriptor/)
4. **Дополнительные научные источники**

1. **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.
2. **Engel, T., Gasteiger, J. (Eds.)** *Chemoinformatics: Basic Concepts and Methods*. Wiley-VCH, 2018. – 608 p. – ISBN: 978-3-527-33109-3.