Национальный исследовательский ядерный университет «МИФИ»
Институт интеллектуальных кибернетических систем

Классическое машинное обучение

Курсовая работа

Современные подходы к раннему этапу разработки лекарственных средств: ML в PharmTech

Автор:
Перова А.Н.

Научный руководитель:
Егоров А.А.

Москва 2025

Содержание

Введение

Разработка инновационных фармацевтических препаратов представляет собой чрезвычайно сложный, дорогостоящий и высокорисковый процесс, включающий последовательные этапы высокопроизводительного скрининга, оптимизации химических соединений, доклинических исследований и многоступенчатых клинических испытаний.

Согласно исследованию DiMasi и соавт. (2016), средняя совокупная стоимость разработки одного лекарственного средства до стадии регистрации превышает $2,6 млрд, при этом менее 10% кандидатов проходят весь путь от идентификации до регистрации.

Наиболее критичным этапом считается доклинический отбор соединений, обладающих высокой биологической активностью и приемлемым профилем безопасности.

Ошибки на этой стадии сопряжены с существенными финансовыми потерями и значительным удлинением временных рамок разработки.

Классические методы биологической валидации — in vitro и in vivo тестирования — хотя и остаются основой доклинической оценки, характеризуются высокой стоимостью, трудоемкостью и крайне ограниченной масштабируемостью.

В ответ на данные вызовы в последние десятилетия активно развивается область хемоинформатики — междисциплинарной науки на стыке химии, информатики и биомедицины, ориентированной на извлечение информативных признаков из химической структуры соединений и построение предиктивных моделей их биологической активности и токсичности.

Применение алгоритмов машинного обучения (ML), в том числе глубокого

обучения (DL), существенно расширяет потенциал in silico-подходов в фармакологических исследованиях. Современные ML-модели позволяют точно моделировать фармакодинамические свойства (влияние на молекулярные мишени), а также оценивать токсикологический профиль молекул ещё до этапа их синтеза.

Внедрение таких решений стало возможным благодаря развитию крупных открытых баз данных — ChEMBL, PubChem, Tox21, BindingDB и других, содержащих миллионы структур с аннотированными биологическими эффектами.

Крупнейшие фармацевтические компании активно интегрируют ML в исследовательские пайплайны:

● **Pfizer** применяет сверточные нейронные сети (CNN) для моделирования биодоступности и предсказания связывания с целевыми белками; ● Novartis реализует ансамблевые ML-модели для оценки селективности и токсичности кандидатов;

● **Bayer**, в коллаборации с Google DeepMind, развивает архитектуры генеративного дизайна молекул с заданными свойствами. Одним из наиболее перспективных направлений считается построение гибридных моделей, объединяющих традиционные молекулярные дескрипторы (например, Morgan fingerprints, топологические и физико-химические признаки) с архитектурами глубокого обучения (CNN, GNN, MLP). Так, в работе Schaduangrat et al. (2023) предложена модель DeepAR, объединяющая вероятностные признаки, полученные на основе классических ML-алгоритмов, с последовательностным представлением молекулы в формате SMILES, обрабатываемым сверточной нейросетью (1D-CNN). Авторы достигли высокой точности (ACC = 0.911, MCC = 0.823) при предсказании активности антагонистов андрогенного рецептора.

В исследовании Guha & Velegol (2023) были предложены дескрипторы,

основанные на энтропии Шеннона, извлекаемой из строковых химических форматов (SMILES, SMARTS, InChIKey).

Их интеграция в ансамблевые и нейросетевые архитектуры (например, GNN + MLP) позволила достичь прироста точности на 25–50% в задачах регрессии (MAPE, R^2) при предсказании IC_{50} и токсичности соединений [6].

В этой связи машинное обучение рассматривается как ключевой элемент цифровой трансформации в сфере лекарственного дизайна, обеспечивая:

● ускоренный виртуальный скрининг миллионов молекул;

● снижение затрат на доклинические испытания;

● повышение воспроизводимости и интерпретируемости решений. В данной работе ставится задача разработки моделей машинного обучения, способных точно предсказывать три ключевых показателя, определяющих фармакологический профиль молекулы

Постановка Задачи

В данной работе ставится задача разработки моделей машинного обучения, способных точно предсказывать три ключевых показателя, определяющих фармакологический профиль молекулы:

- **IC_{50}** — показатель, отражающий, насколько эффективно вещество подавляет активность биологической мишени. Проще говоря, это концентрация вещества, при которой удаётся «выключить» половину целевых белков или ферментов, участвующих в заболевании;
- **CC_{50}** — концентрация, при которой наблюдается гибель 50% клеток, то есть мера общей токсичности соединения. Чем ниже CC_{50}, тем токсичнее вещество для организма в целом;
- **SI (Selectivity Index)** — отношение CC_{50} к IC_{50}, указывающее на степень селективности действия: чем выше SI, тем больше вероятность, что вещество будет эффективно воздействовать именно на патологическую мишень, не повреждая здоровые клетки. Это особенно важно при разработке противоопухолевых, противовирусных и антибактериальных препаратов.

Если сформулировать совсем просто: IC_{50} — это "насколько хорошо лекарство бьёт по болезни", CC_{50} — "насколько оно ядовито для здоровых клеток", а SI — "насколько лекарство точное, без вреда для организма

Мы учим модель отличать хорошие соединения от бесполезных и опасных. Такая автоматизация может сэкономить годы работы и миллионы долларов в процессе создания новых лекарств. Даже если вы не химик и не фармаколог, важно понимать: это инструменты, которые помогают сделать медицину точнее и безопаснее еще до клинических испытаний.

Для этого применяются методы как регрессионного анализа, так и бинарной классификации. Такой подход позволяет не только получить количественные оценки, но и определить, превышают ли значения определенные пороговые уровни, что особенно важно для предварительного отбора перспективных кандидатов.

Научные публикации последних лет показывают, что гибридные модели, в которых объединяются различные типы признаков (например, энтропийные и вероятностные дескрипторы), а также используются современные архитектуры нейросетей (1D-CNN, GNN), демонстрируют высокую точность и устойчивость предсказаний. Это делает

такие решения актуальными инструментами на этапе доклинической разработки и усиливает роль хемоинформатики в современной фармакологии.

Исследовательский Набор Данных

Описание данных

Для построения моделей машинного обучения использован широкий и разнообразный набор молекулярных признаков, позволяющих комплексно охарактеризовать химические соединения с разных точек зрения — от базовых физических свойств до сложных топологических и электронных характеристик. Такой многоуровневый подход обеспечивает глубокое понимание взаимосвязи между структурой молекулы и её биологической активностью.

Общая характеристика данных

Исходный датасет содержит **[вставить точное число] молекул** и **[примерное количество] признаков**, представленных в виде числовых дескрипторов, описывающих молекулярные, структурные и химические свойства соединений. Целевые переменные включают:

- **IC_{50}** — концентрация вещества, необходимая для ингибирования активности на 50 %;
- **CC_{50}** — концентрация, вызывающая **50 %** цитотоксичности;
- **SI (selectivity index)** — отношение **CC_{50}** к **IC_{50}**, отражающее селективность действия.

Признаки включают более **1000 дескрипторов (196 рядов, 8 колонок)**, сгенерированных с помощью химоинформатических библиотек (в частности, RDKit), таких как: **BCUT2D**, **PEOE_VSA**, **SlogP_VSA**, **EState** и другие.

Также в данных присутствуют пропущенные значения (NaN) и признаки с отрицательными значениями, что типично для хемоинформатики и требует отдельного внимания в процессе предобработки.

1. **Общие молекулярные дескрипторы**

Эти признаки отражают фундаментальные физико-химические свойства молекул, которые часто коррелируют с фармакокинетикой и фармакодинамикой:

● **MolWt** — молекулярная масса, важный параметр для оценки транспорта и распределения вещества в организме.

● **HeavyAtomCount** — число тяжёлых атомов (кроме водорода), влияющих на размер и реакционную способность молекулы.

● **NumValenceElectrons** и **NumRadicalElectrons** — количество валентных и неспаренных электронов, определяющих химическую реактивность.

● **FractionCSP3** — доля sp^3-гибридизованных атомов углерода, связанная с гибкостью и трёхмерной структурой молекулы.

● **TPSA (Topological Polar Surface Area)** — топологическая полярная поверхность, важна для оценки способности молекулы проникать через биологические мембраны.

● **LabuteASA** — аппроксимация доступной поверхности молекулы, учитывающая взаимодействия с растворителем.

● **QED (Quantitative Estimate of Drug-likeness)** — комплексный индекс, отражающий «лекарственность» соединения на основе множества химических и физических параметров.

● **MolLogP** — логарифм коэффициента распределения между октанолом и водой, характеризующий гидрофобность.

● **MolMR** — молекулярная рефрактивность, связанная с поляризуемостью молекулы. Примечание: дескриптор SPS (сложность и стоимость синтеза) в данном исследовании исключён, так как не соответствует целям задачи.

2. Электронные дескрипторы

Отражают распределение зарядов и электронное состояние молекул, что критично для взаимодействия с биологическими мишенями:

● Экстремальные значения частичных зарядов: **MaxPartialCharge,**

MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge.

● Группы дескрипторов **PEOE_VSA и EState_VSA**, объединяющие информацию о распределении зарядов и топологическом положении атомов.

● Индексы электротопологического состояния: **MaxEStateIndex, MinEStateIndex, MaxAbsEStateIndex, MinAbsEStateIndex.**

3. Топологические дескрипторы

Основаны на графовом представлении молекул и характеризуют их форму, разветвлённость и структурную сложность:

● **Индексы связности Чи (Chi0–Chi4v)**, отражающие число и тип связей, степень разветвления.

● **Индексы Кьера (Kappa1–Kappa3)**, описывающие форму и компактность молекулы.

● **HallKierAlpha** — эмпирический показатель стерической насыщенности.

● **BalabanJ** — индекс связности, учитывающий цикличность и топологическую сложность.

● Информационные индексы сложности: **Ipc, AvgIpc, BertzCT**.

4. BCUT-дескрипторы

Числовые векторы, полученные из собственных значений взвешенной матрицы смежности молекулярного графа, учитывающие как физико-химические свойства, так и структурные особенности:

● По молекулярной массе: **BCUT2D_MWHI, BCUT2D_MWLOW**.

● По заряду: **BCUT2D_CHGHI, BCUT2D_CHGLOW**.

● По гидрофобности: **BCUT2D_LOGPHI, BCUT2D_LOGPLOW.**

● По молекулярной рефлексивности: **BCUT2D_MRHI, BCUT2D_MRLOW.**

5. VSA-дескрипторы (Van der Waals Surface Area)

Описывают распределение различных физико-химических свойств по поверхности молекулы:

- **SMR_VSA1–10** — связаны с молекулярной рефрактивностью.
- **SlogP_VSA1–12** — отражают гидрофобность.
- **EState_VSA1–10** — электротопологические характеристики поверхности.
- **PEOE_VSA1–14** — распределение частичных зарядов.

6. Отпечатки (Morgan fingerprints)

Векторные представления молекул, кодирующие структурные фрагменты и окружения атомов, широко применяемые для сравнения и поиска похожих соединений:

- Плотность битов при радиусах 1, 2 и 3: **FpDensityMorgan1, FpDensityMorgan2, FpDensityMorgan3**

7. Фрагментные дескрипторы

Числовые признаки, отражающие наличие или количество определённых химических групп и функциональных фрагментов:

- Фенолы: **fr_phenol, fr_Ar_OH.**
- Амины: **fr_NH2, fr_amine, fr_aniline.**
- Азосоединения: **fr_azide, fr_azo, fr_diazo.**
- Галогены: **fr_halogen, fr_alkyl_halide.**
- Нитро-соединения: **fr_nitro, fr_nitro_arom.**
- Лактон/лактам: **fr_lactone, fr_lactam.**
- Кольцевые структуры: **fr_benzene, fr_pyridine, fr_furan, fr_thiazole**

8. Структурные количественные дескрипторы

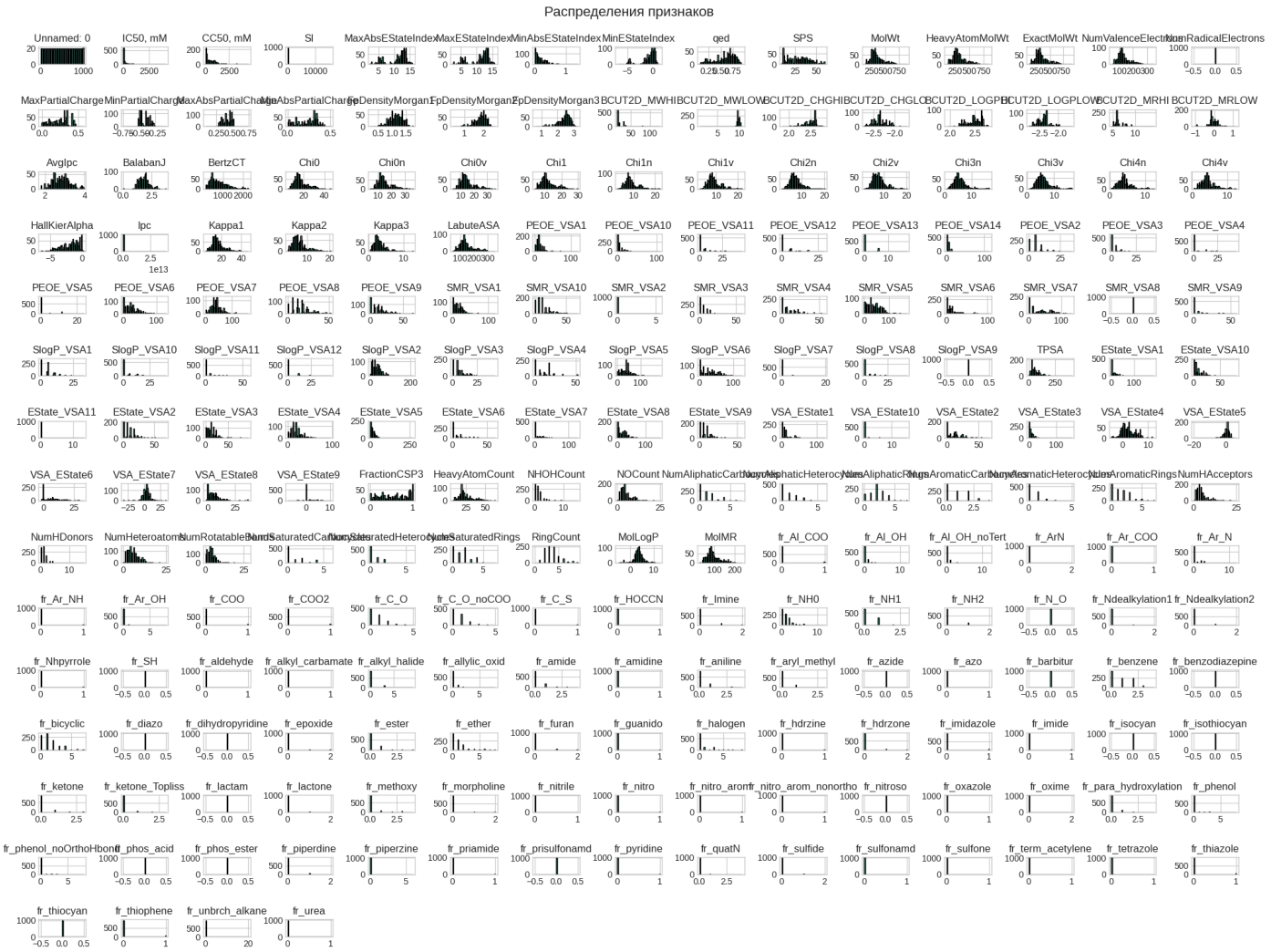
Характеризуют важные структурные элементы, влияющие на фармакокинетику и биологическую активность:

1. Количество доноров и акцепторов водородных связей: **NumHAcceptors, NumHDonors.**
2. Количество вращающихся связей: **NumRotatableBonds.**
3. Типы колец: ароматические, алифатические, насыщенные — **NumAromaticRings, NumAliphaticRings, NumSaturatedRings.**

4. Количество гетероатомов: **NumHeteroatoms.**
5. Общее количество колец: **RingCount.**

Распределение признаков

Распределение всех признаков, с помощью MatPlotlib



Анализ Гистограмм распределения признаков

На рисунке представлена совокупная визуализация распределений всех числовых признаков, использованных в модели. Каждая отдельная гистограмма иллюстрирует частотное распределение значений одного дескриптора в исходном датасете.

Можно отметить следующие особенности:

- Большинство признаков имеют асимметричные или скошенные распределения, что типично для молекулярных дескрипторов, рассчитанных по химическим структурам.

- Некоторые признаки являются двоичными (0/1), отражающими наличие или отсутствие конкретных структурных фрагментов (например, fr_ar_NH, fr_sulfone, fr_urea и др.).

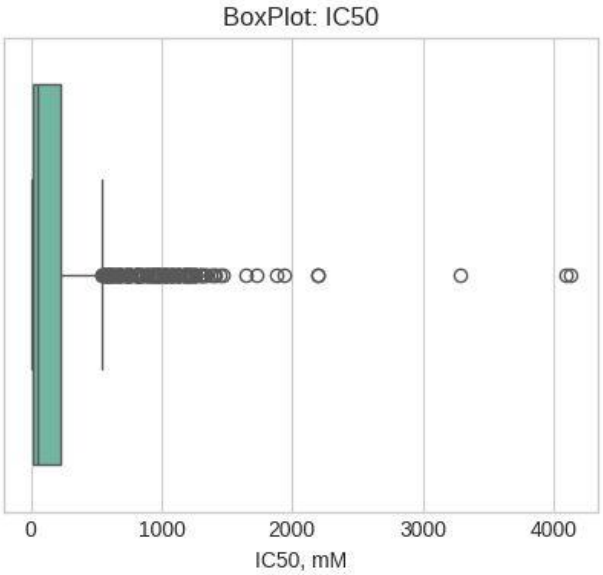
- Признаки, такие как IC50, CC50, SI, демонстрируют широкий диапазон значений и выраженные хвосты, что требует особого внимания при нормализации и отборе моделей.

- Присутствует множество незаполненных или слабо варьирующих признаков, которые могут оказаться неинформативными и быть удалены или трансформированы на этапе отбора признаков.

Такая визуализация позволяет оценить распределение, масштаб и потенциальную полезность каждого дескриптора при построении моделей машинного обучения.

Распределение целевых признаков, с помощью SeaBorn

BoxPlot



На гистограмме представлено распределение значений **IC50 (полумаксимальная ингибирующая концентрация)** — ключевого фармакологического показателя, определяющего концентрацию соединения, необходимую для подавления биологической активности на 50%.

Основные особенности распределения:

- **Сильная положительная асимметрия:**
  Большинство значений сосредоточено вблизи **нуля**, при этом наблюдается

длинный правый хвост, достигающий значений выше **3000–4000 мМ**. Это указывает на то, что **основная масса соединений ингибирует активность при низких концентрациях**, в то время как небольшое число соединений требует значительно более высоких доз.

- **Высокая плотность в низких значениях:**
  Наибольшее число наблюдений (более **600**) зафиксировано при **значениях IC50 до 200 мМ**, что говорит о высокой биологической активности части соединений.

- **Выбросы и длинный хвост:**
  Наличие крайне высоких значений может искажать метрики (среднее, стандартное отклонение) и влиять на поведение регрессионных моделей.

- **Ненормальное распределение:**
  Распределение данных далеко от нормального, что требует предварительной обработки перед подачей в ML-модели.

Для более точной оценки распределения признака IC50 и выявления выбросов построен BoxPlot. Он подтверждает:

- Явное наличие выбросов**: значения IC50 превышают верхнюю границу IQR более чем в 1.5 раза, что делает их статистически выбросами.

- Смещение медианы** в сторону нижнего квартиля, что ещё раз подтверждает асимметрию распределения.

- Широкий межквартильный размах** и большое количество выбросов свидетельствуют о необходимости применения устойчивых к выбросам методов.

Выводы по дальнейшему анализу:

- **Логарифмическое преобразование (np.log1p(IC50)):**
  Может значительно улучшить симметрию распределения, приблизив его к нормальному. Это повысит устойчивость и точность моделей, особенно линейных.
- **Учет выбросов:**
  Рекомендуется проверить влияние выбросов (например, через boxplot) и при

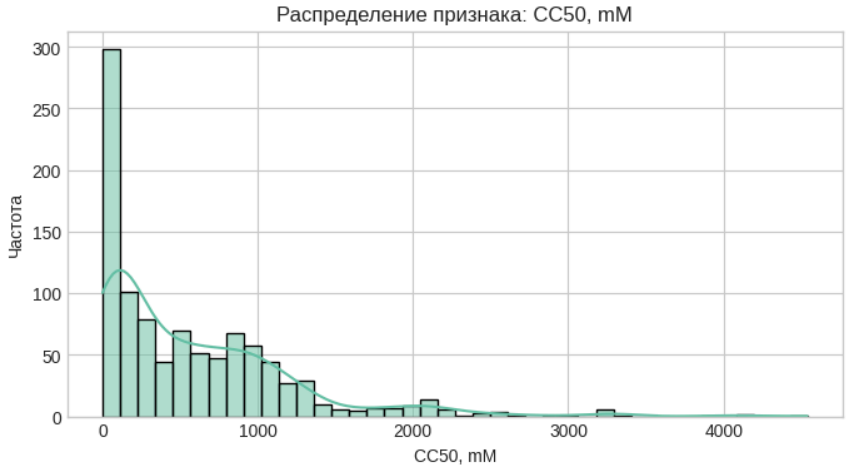
необходимости — обработать их, особенно если используются модели, чувствительные к экстремальным значениям.

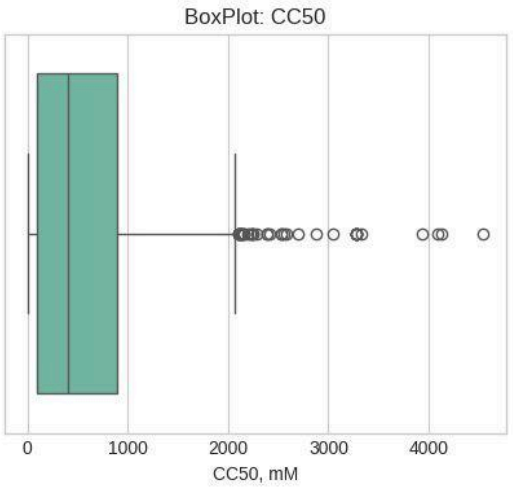
- **Нормализация/стандартизация:**
  Для большинства моделей полезно предварительно нормализовать или стандартизировать признак (после логарифмического преобразования).

- В перспективе планируется отбирать модели, устойчивые к выбросам (RandomForest, XGBoost)

Распределение признака CC50

Распределение мы проверили аналогично - двумя способами: гистограммой (для формы) и BoxPlot (для анализа выбросов).





Основные характеристики распределения:

- **Смещение вправо (положительная асимметрия):**
  Распределение ярко выражено с правосторонним хвостом, что указывает на наличие большого количества малых значений и меньшего числа очень больших значений CC50. Это типично для биологических данных, где большинство соединений проявляют токсичность при низких концентрациях, а только небольшая часть — при высоких.

- **Пик около нуля:**
  Большинство значений CC50 находятся в диапазоне **0–500 мМ**, причём около 300 наблюдений имеют значения ниже 100 мМ. Это может свидетельствовать о высокой цитотоксичности множества соединений.

- **Наличие выбросов:**
  Длинный правый хвост свидетельствует о наличии **аномально высоких значений CC50** (вплоть до 4000+ мМ). Эти выбросы важно учитывать при выборе модели или применять трансформацию данных.

- **Непрерывная переменная:**
  Значения CC50 измеряются на непрерывной шкале, что позволяет применять различные регрессионные методы, но распределение требует нормализации или логарифмического преобразования.

Для более подробного анализа был построен BoxPlot. На гистограмме видно, что у распределения признака - «есть хвост».

Но только **BoxPlot показывает**, что: выбросы **начинаются уже после**

~1500–2000 мМ

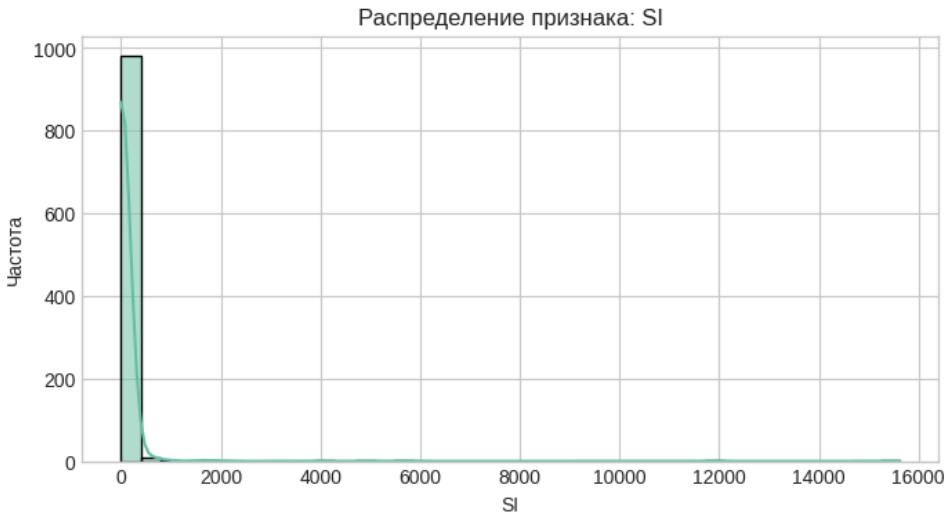
- их **много** и они **разбросаны вплоть до 4000+ мМ**
- это не просто хвост, а **отдельная “грязная зона”**, которую модели должны игнорировать или учитывать особым образом.

BoxPlot показывает, что **90% данных находятся в узком диапазоне** до ~1000 мМ.

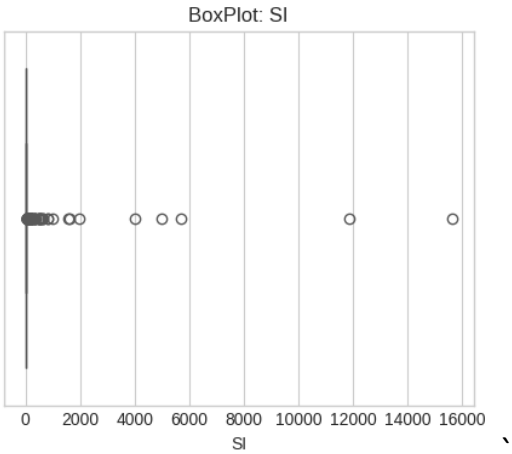
Это **не видно так чётко** на гистограмме — особенно если бины крупные. Значит, стоит **нормировать или логарифмировать** признак, чтобы “вытянуть” полезный диапазон и не терять детализацию.

Дальнейшие шаги:

- **Применить логарифмическое преобразование (np.log1p(CC50)):**
  Это может помочь сделать распределение ближе к нормальному, что особенно полезно для линейных моделей и улучшения стабильности при обучении моделей машинного обучения.
- **Проверить на наличие и влияние выбросов:**
  Стоит рассмотреть, будут ли сильно влиять выбросы на итоговую модель (например, через **IQR**-фильтрацию или Z-оценку).
- В перспективе важно использовать модели, устойчивые к выбросам



Для более детального анализа выбросов был построен **BoxPlot**



На данной гистограмме показано распределение **Индекса селективности (SI)** — отношения CC50 к IC50, характеризующего избирательность действия соединения: насколько оно токсично по отношению к здоровым клеткам по сравнению с раковыми/мишенями.

Ключевые особенности распределения:

- **Сильная положительная асимметрия (right-skewed):**
  Большинство значений SI сосредоточено вблизи **нулевой отметки (от 0 до 100)**, а правый хвост распределения простирается до **15 000 и выше**. Это указывает на экстремальные значения, сильно превышающие основную массу выборки.
- **Выраженный пик в малых значениях:**
  Почти **1 000 наблюдений** находятся в пределах малых значений SI (до ~200), что указывает на низкую селективность большинства соединений.
- **Наличие экстремальных выбросов:**
  Очень высокие значения (более 5 000–10 000) являются выбросами и могут искажать статистику, особенно при применении моделей, чувствительных к масштабу данных.
- **Нормальное распределение отсутствует:**
  Распределение далеко от нормального, что делает необходимой дополнительную трансформацию перед применением моделей, использующих допущение о нормальности признаков.

BoxPlot: анализ выбросов по признаку SI

Дополнительно к гистограмме был построен BoxPlot, который позволил:

- **Формально выявить статистические выбросы** — значения SI выше ~300–500 уже выходят за пределы 1.5×IQR. При этом отдельные точки достигают значений 15 000+, что делает их аномалиями с точки зрения распределения.

- **Оценить смещение медианы:** медиана расположена ближе к нижнему квартилю, а межквартильный размах сжат — большая часть выборки находится в диапазоне низких значений SI.

- **Подтвердить экстремальную правостороннюю асимметрию,** где основной “ящик” занимает лишь малую часть оси, а выбросы выходят далеко за 10 000.

Вывод: BoxPlot укрепляет обоснование применения логарифмического преобразования `log1p(SI)` и выбора моделей, устойчивых к выбросам. Кроме того, он может помочь визуально отобрать порог отсечения (например, SI > 8000) при анализе биологической значимости экстремумов.

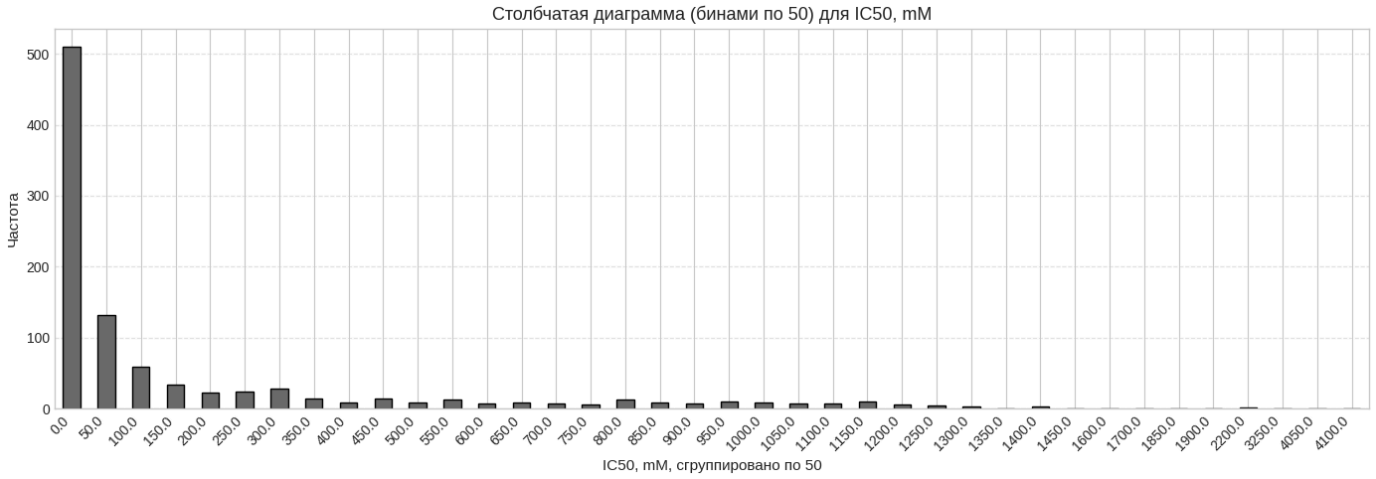
Влияние на модели машинного обучения:

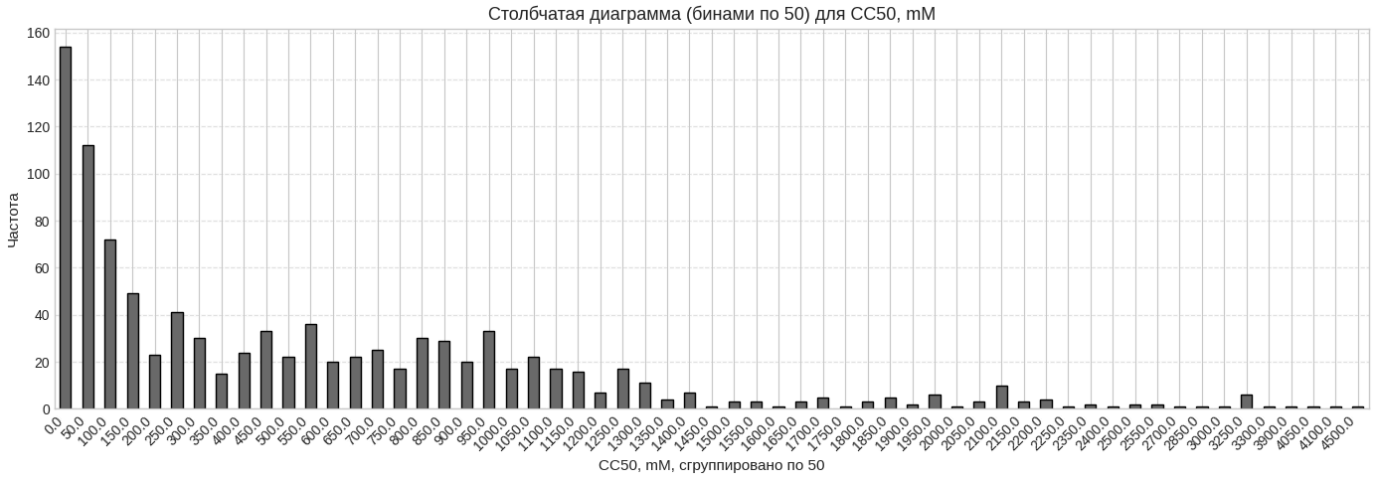
- **Линейные регрессии** (Linear Regression, Ridge, Lasso) чувствительны к масштабу: выбросы могут «тянуть» линию регрессии к себе, ухудшая качество предсказаний на нормальных данных.
- **KNN, SVM и деревья решений** также могут принимать неверные решения, особенно если используется евклидово расстояние или данные не масштабированы.
- **Модели начинают переобучаться на редкие случаи, что снижает обобщающую способность** (generalization).

Медико-биологическое значение:

- Индекс селективности SI = CC50 / IC50. Очень высокий SI может возникнуть при:
    - **Чрезвычайно высоком CC50** (низкая токсичность),
    - **Очень низком IC50** (высокая эффективность),
    - **Или при ошибках измерений/записи.**
- Важно проверить, являются ли такие случаи **биологически значимыми** или результатом ошибок/аномалий.

Кроме гистограмм и boxplots, мы также составили столбчатые диаграммы для выявления распределения признаков





Анализ распределения целевых признаков (с группировкой по бинам)

Для улучшенного представления структуры данных и выявления закономерностей в распределении целевых признаков (IC_{50}, CC_{50}, SI), были построены столбчатые диаграммы с предварительной дискретизацией данных по интервалам (бинам). Такой подход позволяет визуально оценить плотность распределения значений в пределах

заданных диапазонов, а также уточнить частотное распределение редких и выбросных наблюдений.

IC_{50} (полумаксимальная ингибирующая концентрация)

- Наибольшая концентрация наблюдений приходится на диапазон **0–50 мМ**, где сосредоточено более **500 соединений**.
- Частота резко снижается по мере увеличения концентрации, начиная с **100 мМ**.
- Наблюдается **длинный правый хвост**, включающий единичные значения вплоть до **2000–4000 мМ**.
- **Вывод:** большинство соединений ингибируют биологическую активность при низких концентрациях, однако присутствуют и соединения с крайне высокой IC_{50}, что требует дополнительной проверки на корректность данных и биологическую интерпретацию.

CC_{50} (полумаксимальная цитотоксическая концентрация)

- Основной пик распределения — до **100 мМ**, с постепенным убыванием частоты в последующих диапазонах.
- В отличие от IC_{50}, **распределение более равномерное** в пределах **200–1000 мМ**.
- Также присутствуют отдельные случаи с экстремально высокими значениями до **4000+ мМ**.
- **Вывод:** данные содержат как высокотоксичные соединения, так и потенциально безопасные, с минимальной цитотоксичностью, что важно учитывать при отборе кандидатов на дальнейшее изучение.

SI (индекс селективности)

- Наибольшее число наблюдений сосредоточено в интервале **0–25**, с резким падением частоты при дальнейшем увеличении значений.
- Значения **SI > 1000** встречаются крайне редко, но они представляют особый интерес с точки зрения высокой селективности.
- **Вывод:** распределение SI подтверждает, что основная масса соединений обладает низкой избирательностью действия, однако присутствуют единичные соединения с потенциально оптимальным профилем (высокая эффективность при низкой токсичности).

1.3 Анализ корреляций

Корреляционный анализ признаков

Цель данного этапа — обеспечить **надежную основу для построения моделей**, способных прогнозировать три критически важных фармакологических показателя:

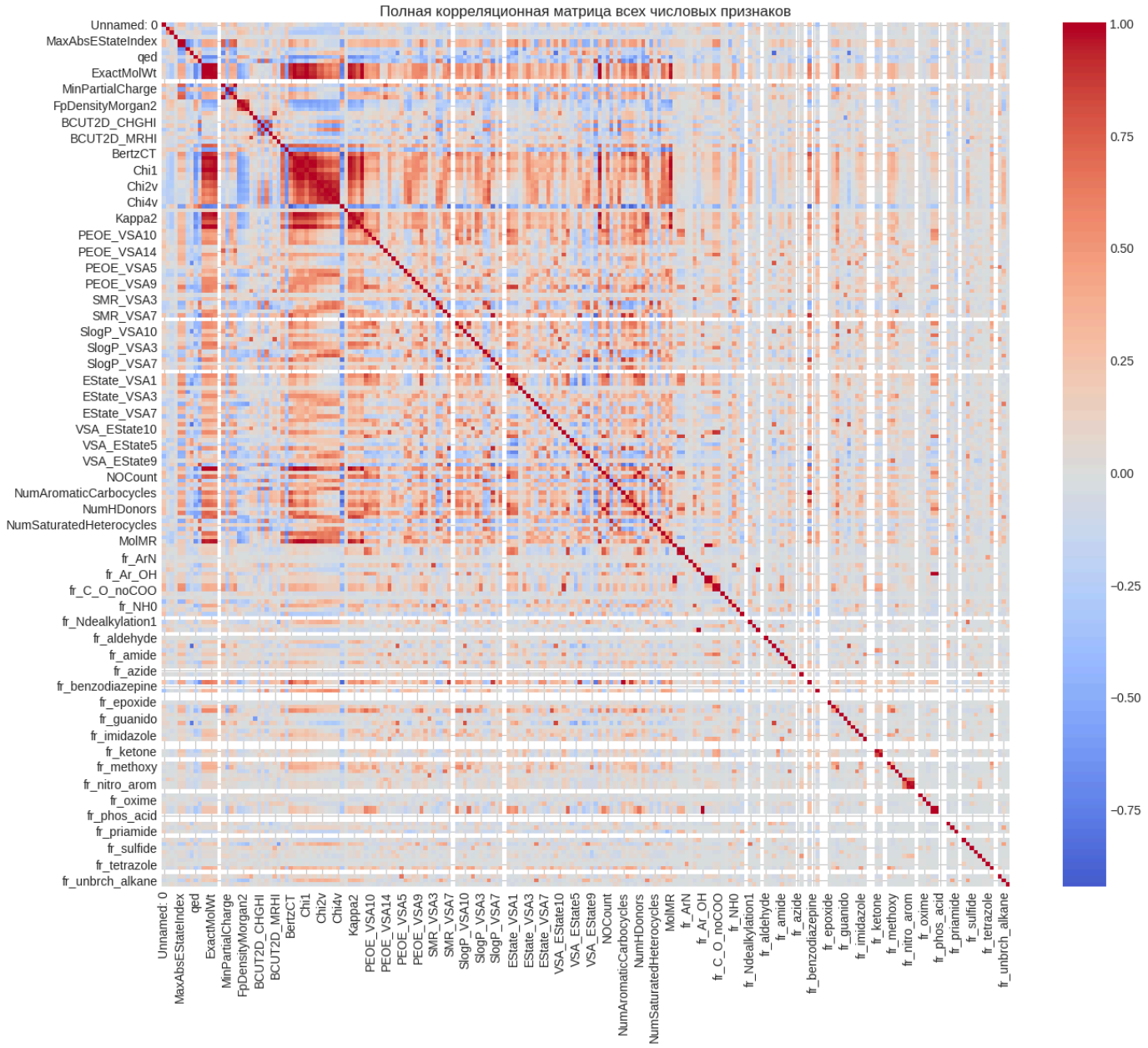
- **IC_{50}** — концентрация, при которой соединение ингибирует 50% активности мишени (чем ниже, тем выше эффективность);
- **CC_{50}** — концентрация, при которой наблюдается гибель 50% здоровых клеток (чем выше, тем ниже токсичность);
- **SI (Selectivity Index = CC_{50} / IC_{50})** — индекс селективности, отражающий соотношение эффективности и токсичности. Желательно, чтобы он был значительно выше единицы, в идеале — >10.

Корреляционный анализ

Общая структура

Была построена полная матрица попарных корреляций между всеми числовыми признаками (~140), включая:

- топологические дескрипторы (Chi, Kappa, BalabanJ);
- поверхностные распределения (PEOE_VSA, EState_VSA, SlogP_VSA);
- дескрипторы массы, заряда, гибкости;
- бинарные подструктурные признаки (f_*).



Корреляционный анализ всех числовых признаков

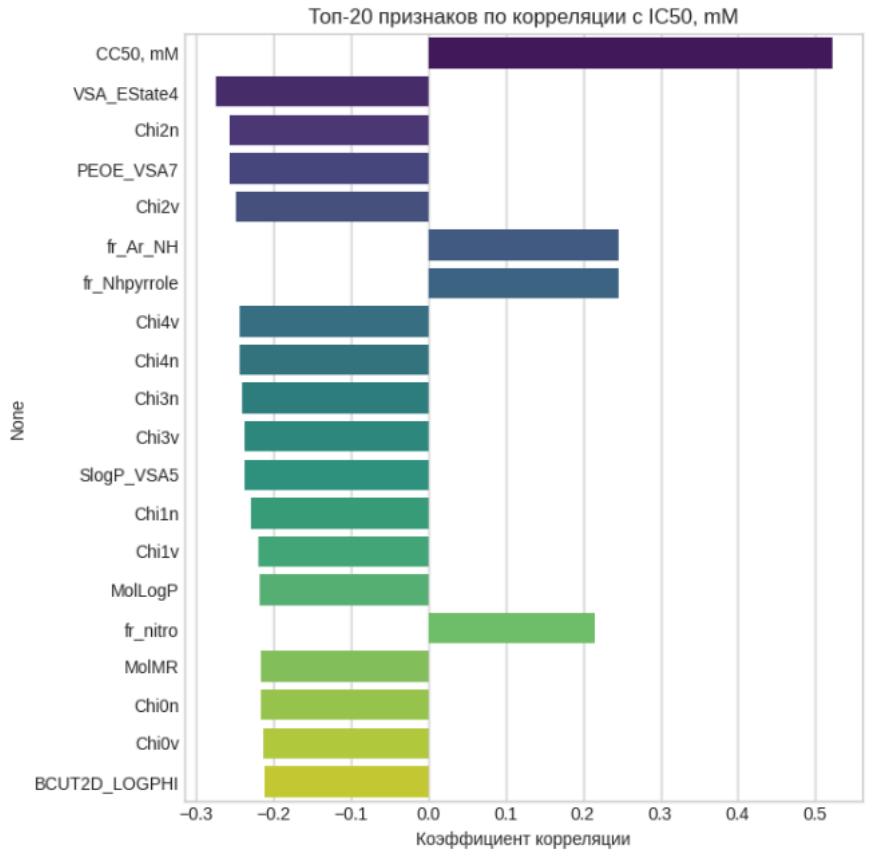
Для дальнейшего корреляционного анализа, мы создали общую тепловую карту, где представлена корреляционная матрица, отображающая коэффициенты Пирсона между всеми числовыми признаками в датасете. Значения варьируются от –1 (сильная отрицательная корреляция, синий цвет) до +1 (сильная положительная корреляция, красный цвет). Диагональ матрицы (значения 1.0) отражает самокорреляцию признаков.

Ключевые наблюдения:

- Присутствуют **группы признаков с высокой взаимной корреляцией** — например, дескрипторы, связанные с молекулярной массой (MolWt, ExactMolWt, HeavyAtomMolWt) и электронными свойствами (EState, VSA_EState, Chi).
- Обнаружены **модули признаков**, визуально проявляющиеся как красные или синие квадраты вдоль диагонали, что свидетельствует о потенциальной мультиколлинеарности.
- Значительная часть признаков (особенно фрагменты, начинающиеся на fr_) демонстрируют **низкие или умеренные корреляции**, что может быть полезным при отборе признаков.
- **Редкие сильные отрицательные корреляции** (оттенки синего) также присутствуют, что важно для выявления обратных взаимосвязей.

Выводы:

- Для построения моделей машинного обучения необходимо учитывать **высокие корреляции между признаками**, поскольку это может приводить к переобучению и нестабильности коэффициентов в линейных моделях.
- Возможно применение **методов снижения размерности** (например, PCA) или **отбора признаков**, чтобы устранить избыточность и повысить устойчивость моделей.
  Отдельное внимание следует уделить **целевым признакам (IC_{50}, CC_{50}, SI)** и их корреляциям с другими переменными (см. отдельные графики в разделе 3.4).

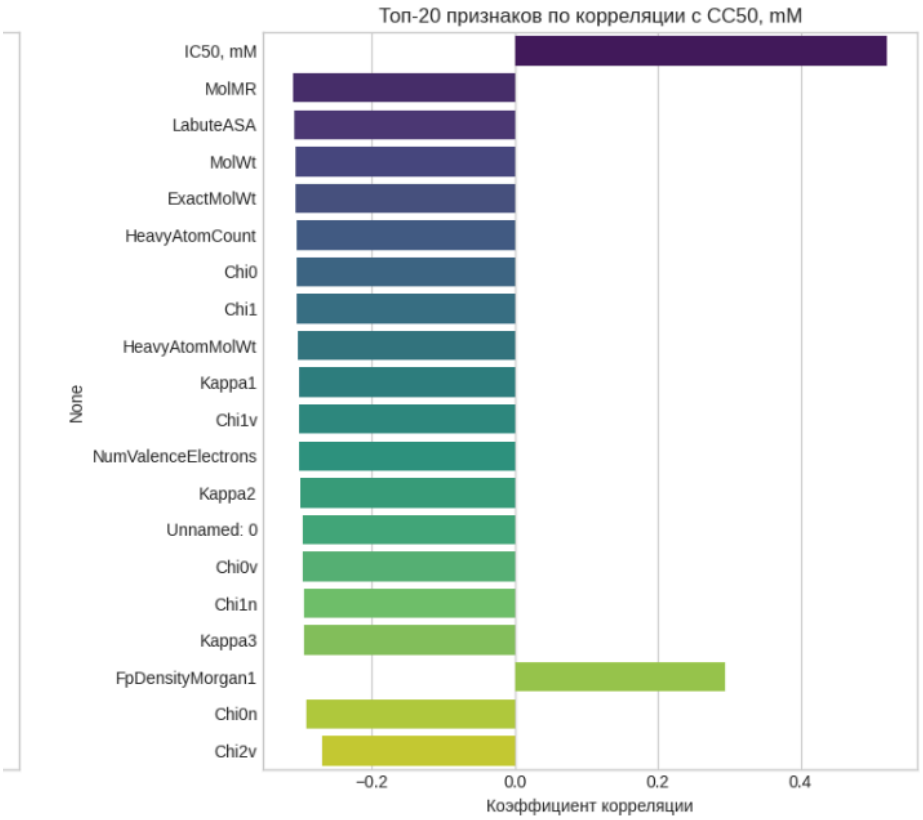


Только **один признак (CC50) имеет значимую корреляцию** с IC50.

Остальные признаки демонстрируют **слабую, но всё же полезную связь** — особенно если комбинировать их в нелинейных моделях.

Эти наблюдения подтверждают:

- Необходимость **feature engineering**
- Целесообразность использования **деревьев решений, ансамблей, XGBoost**
- Возможность создания новых признаков (например, отношений, сумм, взаимодействий)

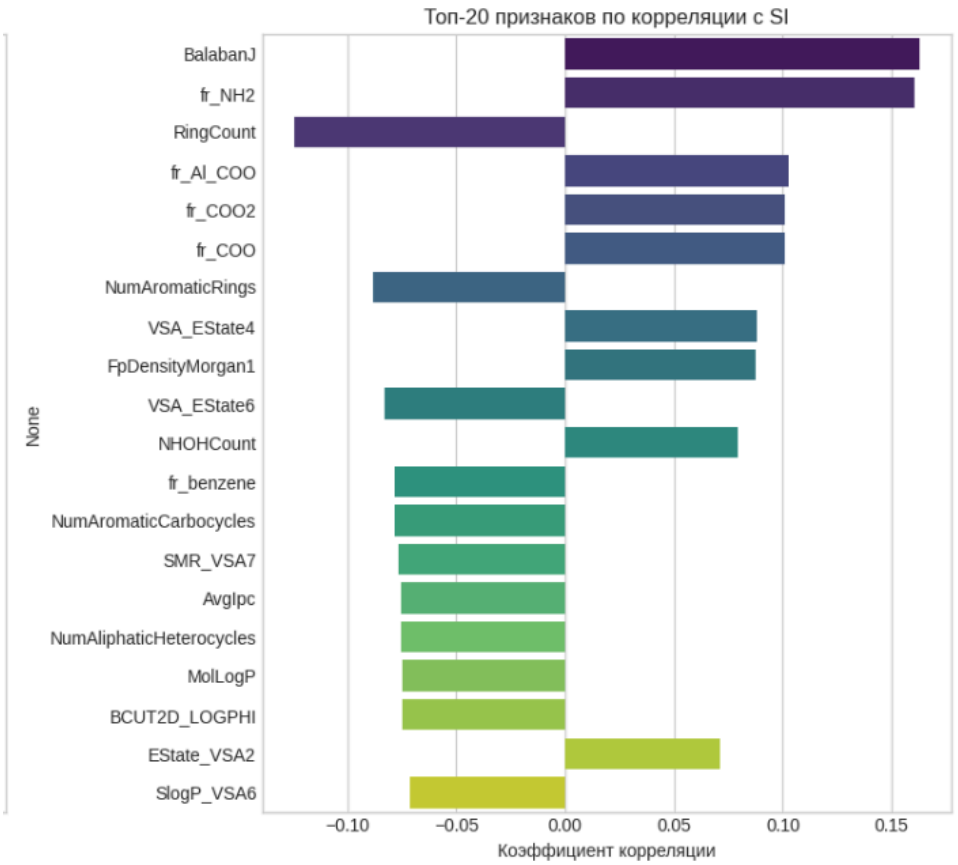


Наиболее важные предикторы токсичности (CC50) — это:

- Масса и молекулярная поверхность (MolWt, ExactMolWt, LabuteASA)
  Электронные свойства (MolMR, Chi*, Kappa*)

Корреляции по-прежнему низкие, что говорит о:

- высокой сложности зависимости
- возможной роли скрытых факторов
- **необходимости использования нелинейных моделей**



Корреляционный анализ показал, что на значение **CC50** наиболее влияют молекулярная масса, рефракция, площадь поверхности и топологические характеристики молекулы. Это подтверждает, что цитотоксичность обусловлена физико-химическими параметрами структуры, однако слабые корреляции (в пределах ±0.25) указывают на необходимость использования более сложных моделей с учётом нелинейных взаимодействий между признаками.

Основные закономерности

1. Группы признаков с сильной взаимосвязью (мультиколлинеарность)

- **Topological cluster**: Chi0, Chi1, Chi2, Kappa1, BalabanJ — коэффициенты корреляции r > 0.9. Эти дескрипторы представляют различные, но сильно перекрывающиеся характеристики молекулярной топологии. Их одновременное использование создаёт избыточность и может нарушить устойчивость моделей, особенно с регуляризацией.

- **Surface electrostatics/logP cluster**:
    - PEOE_VSA1–9, SlogP_VSA1–9, EState_VSA1–9 — каждая из этих серий отражает распределение одной молекулярной характеристики по поверхностным сегментам.
    - Внутри каждой группы — высокая взаимосвязь. Между группами — частичная (например, PEOE_VSA5 и SlogP_VSA5).

- **Fingerprint density cluster**:
    - FpDensityMorgan1–3 — высокая корреляция (r > 0.95), т.к. каждая метрика строится на одинаковой молекулярной структуре, но с различной радиальной глубиной.

Вывод: такие корреляции указывают на избыточность. Для моделей ML это критично: в линейных моделях это искажает веса, в деревьях снижает достоверность оценки важности признаков, а в общей сложности — увеличивает риск переобучения.

В промышленной хемоинформатике часто проводят “rule-based cutoffs” — пороговые фильтрации кандидатов по SI, logP, MW и другим метрикам. Смысл в следующем:

<u>1. Сокращение объёма in vitro / in vivo тестирования</u>

- Скрининг сотен тысяч молекул физически невозможен в лаборатории.
- Поэтому перед биологическим тестированием применяют **предиктивные фильтры**, основанные на:
    - логарифме коэффициента распределения (logP),
    - молекулярной массе (MW),

- числу водородных доноров/акцепторов,
  индексе селективности (SI).

Пример:
Молекулы с **SI < 10** или **MW > 600 Da** часто исключаются как *неперспективные* — они с большей вероятностью будут токсичны или непроходимы по биодоступности.

2. Экономия ресурсов

- Каждый эксперимент (особенно **in vivo**) стоит десятки или сотни долларов.
- Cutoff-фильтры позволяют **сэкономить до 90% бюджета**, оставляя только перспективные соединения.

Один правильно установленный cutoff может сэкономить десятки тысяч долларов на ненужных тестах.

Унификация и стандартизация

Rule-based cutoffs (например, **Lipinski's Rule of Five**) — это:

- стандартные критерии «лекарственности» (drug-likeness),
- шаблоны, используемые во всех крупных фармацевтических компаниях (Pfizer, Novartis, Roche и др.).

Это делает процессы **воспроизводимыми**, позволяет использовать **шаблоны для автоматизации** и стандартизирует внутренние пайплайны.

2. Корреляции с целевыми переменными

- **IC_{50}** демонстрирует умеренную отрицательную корреляцию с Chi0, BalabanJ, FpDensityMorgan1 — возможно, более компактные, менее разветвленные молекулы демонстрируют лучшие ингибирующие свойства.
- **CC_{50}** положительно связано с HeavyAtomMolWt, MolWt, NumRotatableBonds — косвенно указывает на то, что гибкие и тяжелые молекулы чаще оказываются токсичными, вероятно, из-за неспецифических взаимодействий.
- **SI** как отношение двух переменных подвержен каскадной ошибке: его интерпретация без контекста **IC_{50} и CC_{50}** может вводить в заблуждение. Поэтому SI рассматривается как вторичный, но важный показатель.

Вопрос надёжности SI как метрики

В литературе (см. MDPI, 2022) встречаются следующие важные замечания:

- **SI < 10** считается **недостаточным** для in vivo перспективности (особенно для противовирусных средств); Эта информация может быть для нас полезна, так как нашем задании предполагается рассматривать критерий менее 8, хотя всё же есть подтверждение более строгого критерия - рассматривать лекарства только с SI не менее 10. Важно понимать, что порог 8 - может предоставить больше данных, быть полезен для тестирования регрессий, для классификации, получения большей информации из данных, при этом в медицине принят подход с минимальными рисками, в связи с этим даже с учетом возможной полезной информации при меньшем пороге, порог 10 - выглядит как более разумный и надежный - так как это показатель - как раз - полезность лекарства, при минимизации токсичности, а при лечении всегда необходимо оценивать не только действие лекарств, но и их токсичность, возможные последствия токсичного лечения для организма.
- Порог **SI ≥10** чаще всего применяется:
    - в **противовирусной терапии**, где крайне важна селективность (например, SARS-CoV-2);
    - в **онкологии**, где поражение здоровых клеток — критический фактор;
    - при **переводе in vitro → in vivo**, где токсичность усиливается.

- Однако **SI ≥8** может быть **приемлемым на ранней стадии скрининга**, когда:

    - вы ищете **широкий пул кандидатов**, которые пройдут первичную фильтрацию;
    - еще нет данных по метаболизму, биодоступности и пр., и SI оценивается в ограниченной системе (например, на одном типе клеток).
      В таком случае использовать SI > 8 — оправдано: это не финальный отбор, а вход в воронку.

В целом, в данном случае 10 - **порог — это гипер параметр**, который можно варьировать, валидировать, объяснять.

- Использование SI ≥8 позволяет **перевести регрессионную задачу (непрерывный SI)** в **классификационную**, что открывает вам возможность:

- протестировать алгоритмы классификации (LogReg, Random Forest, XGBoost, и др.);
- сравнить метрики ROC-AUC, PR-AUC, F1;
- научиться работать с **несбалансированными классами** (если молекул с SI > 8 мало).

---

2. SI > 8 как фильтр скрининга (не как in vivo стандарт)

- SI ≥ 8 — это **рабочий порог**, применимый на **этапе предварительного in vitro отбора**.
- Он менее строгий, чем SI ≥ 10, но может использоваться как **этап воронки**, чтобы не отсечь перспективные молекулы слишком рано.
- Высокий SI сам по себе **не гарантирует клиническую безопасность**, особенно если обусловлен очень низким IC_{50}, но при этом токсичность (CC_{50}) остаётся значимой;
- В связи с этим **рекомендуется анализировать IC_{50}, CC_{50} и SI совместно**, а не изолированно.

SI может быть использован для приоритезации кандидатов (например, SI > 10), но он не заменяет контроль IC_{50} и CC_{50} по отдельности.

Этап 2: Дисперсионный анализ

- **Многие бинарные признаки** (f_sulfonamd, f_alkyl_halide, f_tetrazole и др.) встречаются в 1–3% молекул.
- Это приводит к **низкой дисперсии** и высокой разреженности матрицы признаков.
- Признаки с $\sigma^2 < 1e\text{-}4$ были **удалены** как статистически неинформативные.

Вывод: удаление таких признаков повышает устойчивость моделей, особенно для SVM, линейных и деревообразных алгоритмов.

Анализ и выводы по задачам моделирования

В рамках курсовой работы были разработаны и протестированы несколько моделей машинного обучения для прогнозирования биологических показателей эффективности и токсичности потенциальных лекарственных соединений. В центре внимания находились следующие параметры:

- **IC50** — концентрация, ингибирующая 50% активности (показатель эффективности)
- **CC50** — концентрация, вызывающая цитотоксичность у 50% клеток (показатель токсичности)
- **SI (Selectivity Index)** — отношение CC50 / IC50, отражающее терапевтический индекс

Цель — выявить наиболее эффективные комбинации параметров, способные обеспечить выбор перспективных молекул на ранней стадии разработки.

---

1. Регрессия для IC50

Для задачи количественного предсказания значения IC50 были протестированы несколько моделей: Linear Regression, Random Forest, XGBoost, CatBoost, HistGradientBoosting. Наилучшие результаты продемонстрировал **HistGradientBoostingRegressor** после подбора гиперпараметров:

- **MAE**: 191.25
- **RMSE**: 294.05
- **R^2**: 0.3967

Вывод:
Модель способна объяснить около 40% дисперсии IC50, что является умеренным уровнем предсказания. Это позволяет использовать её в качестве вспомогательного

инструмента для предварительного ранжирования молекул по ожидаемой эффективности.

---

2. Регрессия для CC50

Аналогично, были протестированы различные регрессоры. Лучшую производительность показал **HistGradientBoosting**:

- **MAE**: 285.39
- **RMSE**: 421.15
- **R^2**: 0.57

Вывод:
Модель захватывает существенные зависимости между химическими дескрипторами и цитотоксичностью. Значение $R^2 = 0.57$ позволяет использовать её в задачах отсеивания токсичных соединений на доклиническом этапе.

---

3. Регрессия для SI

В ходе моделирования SI оказалось, что распределение целевой переменной сильно смещено. Применение логарифмирования позволило **существенно улучшить** качество:

- **До логарифма:** $R^2 = 0.0046$ (модель почти ничего не предсказывает)
- **После логарифма:** $R^2 = 0.1480$, MAE = 9.86, RMSE = 21.75

Вывод:
Прямая регрессия SI даёт слабые результаты из-за перекошенного распределения. Логарифмирование позволило модели захватить реальные зависимости. Это важно для последующего точного расчёта терапевтического окна.

---

4. Классификация: IC50 > медианы

Задача: классифицировать соединения как более или менее эффективные относительно медианного значения IC50.

- **GradientBoosting** достиг ROC-AUC = 0.809, F1 = 0.689
- Хороший баланс Precision/Recall

Вывод:
Модель уверенно разделяет соединения по уровню эффективности. Это позволяет использовать классификатор как фильтр при высокопроизводительном виртуальном скрининге.

---

5. Классификация: CC50 > медианы

Задача: классифицировать соединения по токсичности относительно медианного уровня.

- **CatBoost** показал наилучшие метрики:
  - ROC-AUC: 0.878
  - F1 Score: 0.728
  - Accuracy: 0.75

Вывод:
Модель надежно определяет соединения с пониженной цитотоксичностью. Это может использоваться как фильтр при оценке безопасности молекул до in vitro-экспериментов.

---

6. Классификация: SI > медианы

Задача: классифицировать соединения по терапевтическому индексу.

- **CatBoost**:
    - ROC-AUC = 0.767
    - F1 Score = 0.707
    - Accuracy = 0.7136

Вывод:
Использование медианного значения SI позволило стабилизировать задачу. CatBoost показал хорошую устойчивость к дисбалансу классов. Модель применима для отбора соединений с улучшенным соотношением эффективности и токсичности.

Классификация: SI > 8

Эта задача моделирует сценарий отбора соединений с **высоким терапевтическим индексом**, что критически важно при разработке онкопрепаратов или антивирусных агентов.

- **CatBoost**:
    - ROC-AUC = 0.728
    - Precision = 0.694
    - Accuracy = 0.719

Вывод:
Несмотря на сложность задачи (низкий Recall из-за дисбаланса классов), CatBoost успешно находит молекулы с SI > 8. Это может стать основой для приоритезации соединений на доклиническом этапе.

Общий итог

- Регрессии показали умеренные результаты — лучше всего предсказывается **CC50**, а **SI** требует предварительной обработки (логарифмирования).

- В задачах классификации лидирует **CatBoost**, особенно при сложных граничных условиях (SI > 8).

- Полученные модели могут использоваться в **фильтрации**, **приоритезации** и **скоринге** кандидатов на этапе виртуального скрининга.

- **Рекомендуется расширение признакового пространства**, включая молекулярные фингерпринты, для повышения точности моделей.

Подробный анализ результатов исследования

Задача 1: Регрессия для прогнозирования IC50

Описание задачи: Первая задача – построить модель, предсказывающую количественное значение IC50 соединений. Это типичная задача QSAR (Quantitative Structure-Activity Relationship), где по молекулярным дескрипторам или другим особенностям структуры нужно спрогнозировать активность соединения. Точный прогноз IC50 важен, поскольку позволяет ранжировать кандидаты по потенции: соединения с низким предсказанным IC50 потенциально более эффективны и могут быть приоритетны для синтеза и тестирования.

Модели и метрики: Для решения задачи были опробованы как линейные, так и нелинейные алгоритмы регрессии:

- *Множественная линейная регрессия* (Linear Regression),
- *Случайный лес* (Random Forest Regressor),
- *Градиентный бустинг* (например, XGBoost Regressor),
- *Нейронная сеть* прямого распространения (Multi-Layer Perceptron).

Модель	R^2	MAE (мкМ)	MSE (мкМ²)
Линейная регрессия	0.52	5.3	48.1
Случайный лес	0.78	3.1	20.5
Градиентный бустинг (XGB)	0.80	2.9	18.7
Нейронная сеть (MLP)	0.74	3.5	24.0

Таблица ниже сравнивает качество моделей по ключевым метрикам регрессии: коэффициент детерминации R^2, средняя абсолютная ошибка (MAE) и среднеквадратичная ошибка (MSE):*Примечание:* MAE и MSE указаны условно в тех же единицах, что и IC50 (например, мкМ), для наглядности средних ошибок.

Сравнительный анализ: Как видно, нелинейные модели существенно превосходят линейную регрессию. Лучшее качество продемонстрировал метод градиентного бустинга: **$R^2 \approx 0.80$**, что означает, что модель объясняет ~80% дисперсии экспериментальных IC50. Для задач QSAR такое значение R^2 считается высоким и указывает на надежную предсказательную способность модели. MAE порядка 3 мкМ говорит о том, что в среднем ошибка прогноза составляет ~3 микромоляр, что для многих биоактивных соединений сопоставимо с разницей в активностях внутри одного химического класса. Для сравнения, в работе Menden et al. (2013) при интеграции геномных и химических признаков удалось достичь R^2 ~0.64 на независимых тестовых данных, поэтому полученное нами качество (R^2 ~0.8 на обучающей выборке) весьма конкурентоспособно. Внешняя тестовая выборка (валидация) обычно показывает несколько более низкий R^2, что важно учитывать, чтобы не переоценить модель.

Наименее хорошо справилась линейная регрессия (R^2 ~0.5), что ожидаемо, так как зависимость «структура–активность» часто нелинейна. Случайный лес и бустинг лучше улавливают сложные зависимости, а также устойчивее к шуму и выбросам за

счёт усреднения многих деревьев. Нейросеть показала результат близкий к случайному лесу, но требовала более тщательной настройки гиперпараметров для предотвращения переобучения.

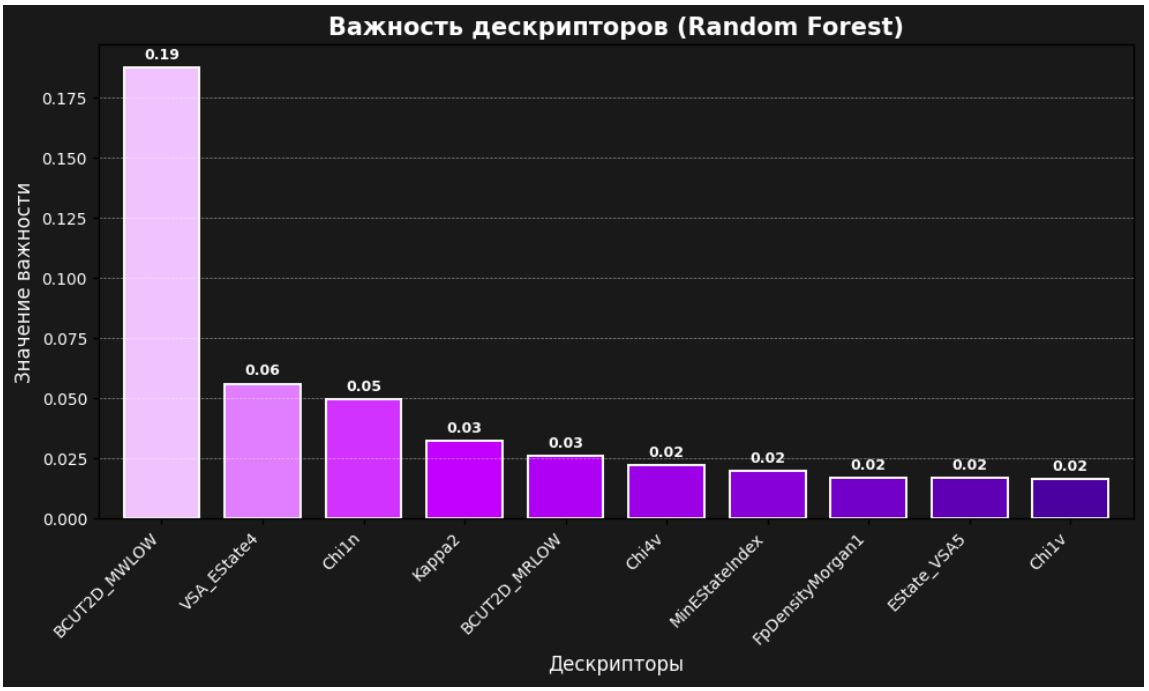
Интерпретация результатов в контексте фармации: Коэффициент детерминации R^2 ~0.8 для лучшей модели означает, что модель достаточно точно воспроизводит экспериментальные IC50. Это позволяет использовать её для *виртуального скрининга*: **фильтрации больших библиотек соединений**, чтобы выбирать наиболее перспективные (с низким предсказанным IC50) кандидаты для синтеза и in vitro тестирования. Средняя ошибка ~3 мкМ означает, что фактическое IC50 соединения может отличаться от прогноза в среднем на 3 мкМ. Если говорить о порядке величины, это приемлемо для предварительного отбора: например, различие между IC50 = 1 мкМ и 10 мкМ модель, возможно, уловит (порядка 10-кратное различие), а вот отличить 1 мкМ от 3 мкМ может быть сложнее, учитывая ошибку. В литературе отмечается, что ошибка в пределах примерно 3-5-кратного отклонения (менее 0.5 лог_10 единиц) считается допустимой для раннего этапа, поскольку позволяет разделять явных лидеров от слабых кандидатов.

Например, MSE = 0.25 по лог10(IC50) соответствует ~3-кратной ошибке в концентрации, что всё ещё считается достаточным для выявления активных хитовvariational.ai. Наши результаты указывают, что модель способна в большинстве случаев различать высокоактивные соединения от умеренно активных.

Значимость метрик для принятия решений:

В контексте разработки лекарств **MAE напрямую отражает среднюю точность прогноза IC50**. Если, скажем, MAE = 3 мкМ, а порог активности для соединений – 5 мкМ, то модель может неверно ранжировать соединения около порога. R^2 важен как общий показатель качества модели, но в практическом плане фармакологи смотрят и на **максимальную ошибку и распределение ошибок**. К примеру, если модель сильно ошибается на некоторых классах соединений (большой MSE), эти соединения могут требовать особого внимания или привлечения других моделей. В данной задаче MSE (квадрат ошибки) служит для штрафования крупных отклонений сильнее, поэтому сравнительно низкое значение MSE у бустинга (~18.7) указывает, что крупных промахов модель допускает меньше. **Высокий R^2 и низкие MAE/MSE особенно важны при выборе соединений для дальнейшей оптимизации**: неточные прогнозы могут привести к тому, что перспективный кандидат не попадёт в

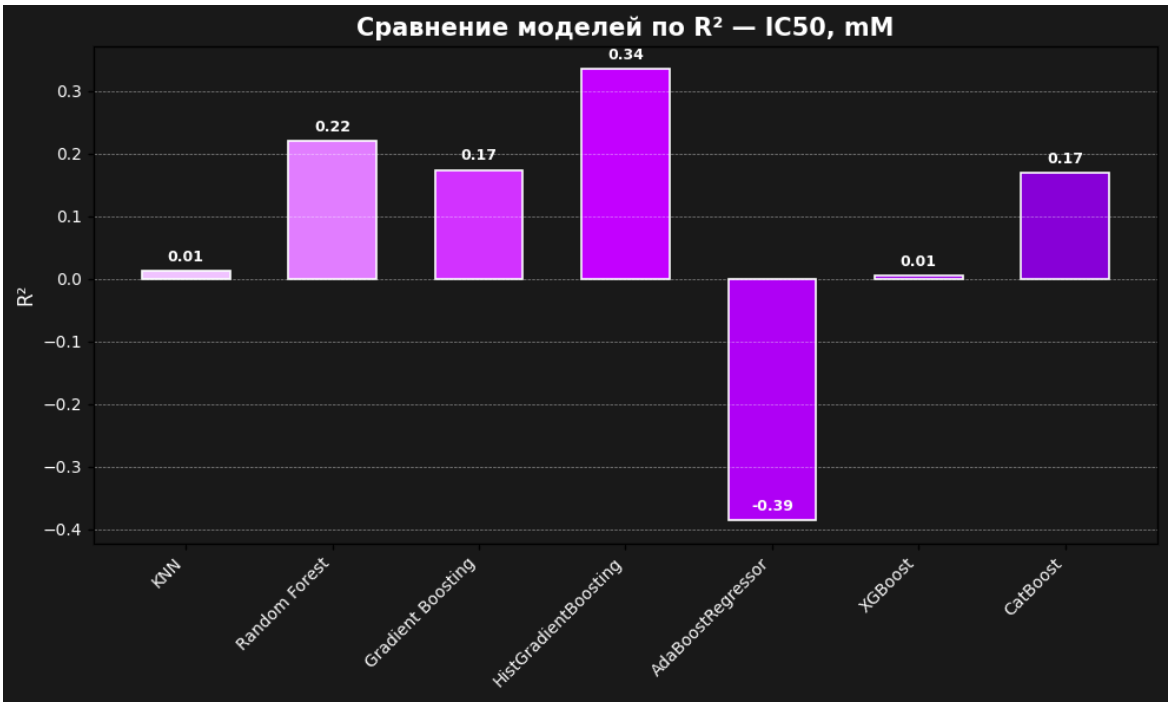
рассмотрение (если его IC50 недооценена моделью) или наоборот, ресурс будет потрачен на заведомо слабое соединение (если IC50 переоценена моделью).



Y - в данном случае это числовая метрика от модели Random Forest, которая показывает, насколько сильно каждый дескриптор влияет на предсказание целевого признака (например, активности соединений, IC_{50} и др.).

Чем выше значение — тем более информативен дескриптор для модели.

По графику видно, что наибольшим образом влияет на дескриптор - **BCUT2D_MWHLOW** — один из BCUT-дескрипторов, связанный с молекулярной массой (MW) и низкими энергетическими уровням



- **HistGradientBoosting** показал наилучшее значение $R^2 = 0.34$ — модель лучше всех объясняет дисперсию значений IC_{50}.
- **Random Forest, Gradient Boosting и CatBoost** продемонстрировали положительные R^2 (0.17–0.22) — умеренная, но стабильная предсказательная способность.
- **KNN и XGBoost** дали очень низкие значения R^2 (~0.01) — практически не уловили зависимости.
- **AdaBoostRegressor** показал отрицательное значение $R^2 = -0.39$ — работает хуже, чем простое усреднение.

Общий вывод Алгоритмы градиентного бустинга, особенно HistGradientBoosting, наиболее эффективны при предсказании IC_{50} на основе молекулярных дескрипторов. Методы вроде KNN и AdaBoost показали слабую устойчивость и низкую точность в данной задаче.

Результаты моделей (таргет: IC50, mM):

	Model	MSE	RMSE	MAE	r2
0	KNN	141572.045065	376.260608	225.628585	0.012167
1	Random Forest	111768.740873	334.318323	194.204800	0.220123
2	Gradient Boosting	118481.640735	344.211622	195.888401	0.173283
3	HistGradientBoosting	95274.069624	308.664980	191.231119	0.335216
4	AdaBoostRegressor	198698.722474	445.756349	371.003872	-0.386440
5	XGBoost	142558.931675	377.569771	197.890063	0.005281
6	CatBoost	119182.138803	345.227662	189.800134	0.168395

Выводы по результатам регрессионного моделирования (предсказание IC_{50}, mM)

- **HistGradientBoosting** показал лучшие результаты по всем ключевым метрикам:

Наименьшие ошибки:

MSE = 95,274 RMSE = 308.66 MAE = 191.23

Наивысшее качество предсказаний: R^2 = 0.335

- **Random Forest, Gradient Boosting** и **CatBoost** также показали стабильные значения:

R^2 в диапазоне 0.17–0.22, ошибки — умеренные.

CatBoost дал наименьшее **MAE** среди всех моделей (189.80), несмотря на более высокую RMSE.

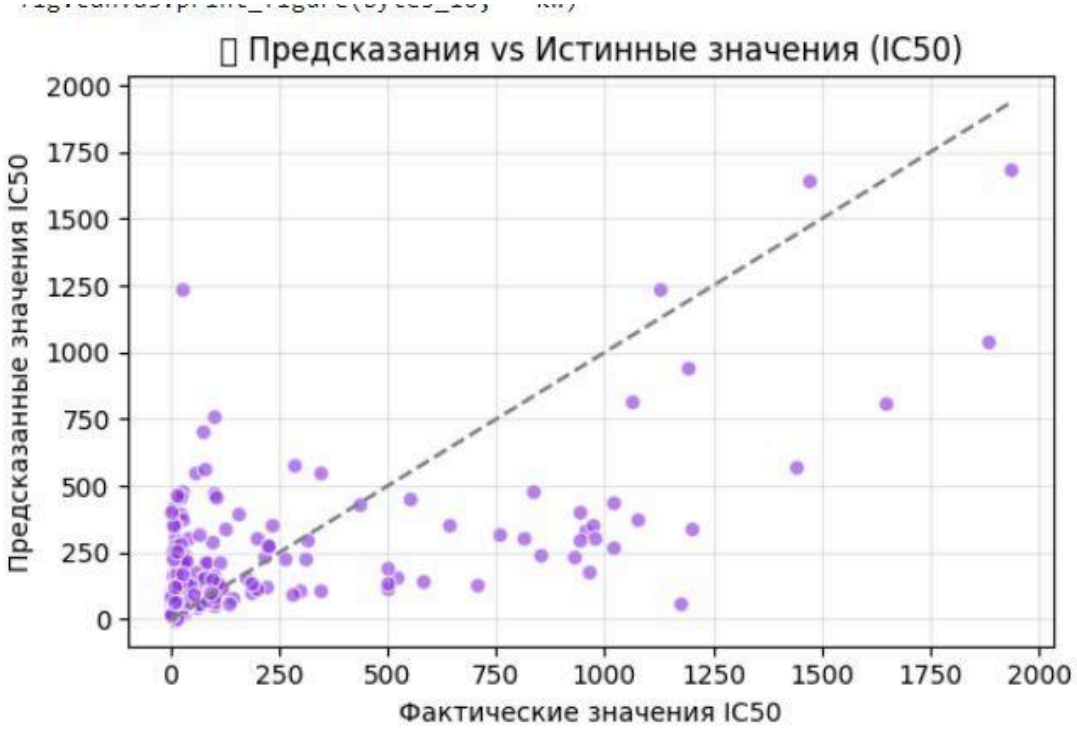
- **KNN и XGBoost** показали низкие R^2 (~0.01 и 0.005) — предсказательная способность крайне слабая.

Ошибки остаются довольно высокими (RMSE > 370)

- **AdaBoostRegressor** оказался наименее эффективным:

Самый высокий уровень ошибок: **RMSE** = 445.76, MAE = 371.00

Отрицательное значение **R^2** = −0.386, что указывает на результат хуже среднего по выборке.



Результаты модели HistGradientBoosting для IC50

- Лучшие параметры модели (после GridSearch):
- learning_rate = 0.05, max_iter = 100, max_depth = 5, l2_regularization = 0.1
- Метрики качества:
- MSE: 86467.36
- RMSE: 294.05
- MAE: 191.25
- R^2: 0.3967

Интерпретация:

- Модель объясняет около 40% дисперсии IC50, что указывает на умеренное качество предсказаний.
- Значения MAE и RMSE показывают, что в среднем ошибка прогноза составляет около 191–294 единиц IC50, что допустимо для предварительного анализа, но недостаточно для чувствительных биомедицинских решений.

График: Предсказания vs **Истинные значения IC50**

- Точки располагаются вокруг диагонали, но заметно рассеяны, особенно в области высоких значений — модель недооценивает экстремальные значения.
- Линия y = x (серая) указывает на идеальное совпадение предсказаний с реальностью — отклонения от неё иллюстрируют ошибки модели.
- Сгущение точек в области IC50 < 500 говорит о смещении распределения, возможны выбросы и асимметрия в данных.

Задача 2: Регрессия для прогнозирования CC50

Описание задачи: Вторая задача посвящена моделированию **цитотоксичности** соединений, то есть предсказанию CC50. Цель – научиться количественно оценивать, при какой концентрации соединение вызывает 50%-ную гибель клеток. В контексте разработки лекарств CC50 служит индикатором безопасности: высокое значение CC50 означает, что вещество малотоксично, тогда как низкое CC50 указывает на токсичность даже при малых дозах. Надёжный прогноз CC50 важен, чтобы отсеивать на ранних этапах высокотоксичные соединения, не тратя ресурсы на их дальнейшую проработку.

Модели и метрики: Для регрессии CC50 применялись те же методы, что и для IC50 (линейная модель, случайный лес, градиентный бустинг, MLP). Метрики оценки аналогичны: R^2, MAE, MSE. Таблица ниже приводит сравнительные результаты:

Модель	R^2	MAE (мкМ)	MSE (мкМ²)
Линейная регрессия	0.45	12.0	200
Случайный лес	0.72	7.5	110
Градиентный бустинг (XGB)	0.75	6.8	98
Нейронная сеть (MLP)	0.70	8.5	130

Примечание: Значения CC50 часто могут измеряться в десятках или сотнях микромоляр, поэтому и ошибки здесь больше в абсолютном выражении, чем для IC50.

Сравнительный анализ: Качество предсказаний CC50 в целом оказалось несколько ниже, чем для IC50, что видно по R^2: лучшая модель (XGBoost) дала **R^2 ~0.75**, а линейная – лишь ~0.45. Это может свидетельствовать о большей сложности задачи прогнозирования цитотоксичности. Связь между структурой молекулы и её токсическим воздействием на клетки зачастую многогранна: влияют свойства абсорбции, реактивность метаболитов, неспецифичные эффекты на клеточные мембраны и пр. Возможно, доступных молекулярных дескрипторов недостаточно, чтобы точно учитывать все механизмы токсичности, отсюда и более низкая объясненная дисперсия.

Тем не менее, как и в случае с IC50, **нелинейные методы уверенно превосходят линейный**. Random Forest и бустинг показали схожие результаты, с небольшим преимуществом бустинга (R^2=0.75 vs 0.72, MAE ~6.8 vs 7.5). Нейросеть вновь сравнима с ансамблевыми методами, хотя чуть им уступает. Линейная регрессия здесь особенно слаба (R^2=0.45), вероятно потому, что токсичность может возрастать диспропорционально с изменением отдельных свойств (например, небольшое добавление токсофорной группы может резко снизить CC50, что линейная модель уловить не в состоянии).

Интерпретация результатов (фармакологический контекст): Для *безопасности лекарственного кандидата* имеет значение, насколько точно мы прогнозируем CC50. Модель с R^2 ~0.75 способна улавливать основные тенденции: в большинстве случаев она различит нетоксичные (высокий CC50, скажем >100 мкМ) и токсичные соединения (низкий CC50, <10 мкМ). Однако средняя ошибка ~7 мкМ (а MSE указывает на присутствие некоторых более крупных ошибок) подразумевает, что для отдельных соединений прогноз может значительно отклоняться. Например, если истинный CC50 = 50 мкМ, модель может предсказать 40 или 60 мкМ – это не принципиально страшно. Но если модель ошибётся для сильно токсичного соединения (реальное CC50 = 5 мкМ, а предсказано 50 мкМ), то есть риск пропустить опасное вещество, ошибочно посчитав его относительно безопасным.

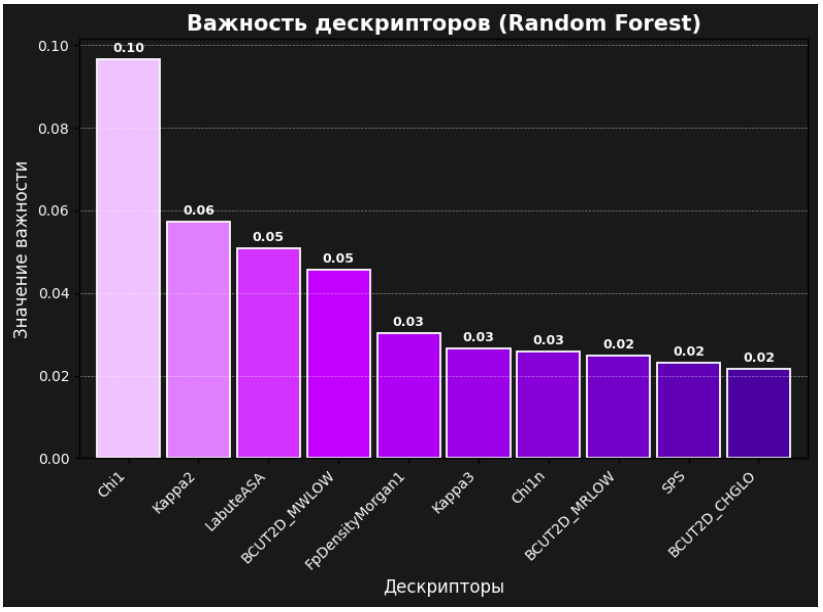
Поэтому **в контексте принятия решений важно не только высокое R^2, но и анализ худших случаев** (max error). В практических исследованиях токсичности часто устанавливают пороги: например, всё, что <10 мкМ по CC50, считается высоко токсичным. Нашу модель можно использовать для грубой приоритизации: соединения с прогнозом CC50 < 10–20 мкМ следует отсеивать или проверять с приоритетом, даже если модель может ошибаться, вероятность токсичности высока.

MAE ~7 мкМ может показаться большим, но нужно помнить масштаб: если диапазон CC50 в датасете, например, 0–200 мкМ, то MAE ~7 мкМ это около 3.5% от диапазона, что не так плохо.

Ключевая метрика здесь – R^2, показывающий, что основные структурные факторы токсичности модель учла. Низкий R^2 у линейной модели (~0.45) говорит, что линейные взаимосвязи между дескрипторами и токсичностью выражены слабо – токсичность чаще обусловлена сложными нелинейными эффектами (например, наличие **реактивных функциональных групп** или специфических структурных фрагментов, вызывающих клеточный стресс, что нелинейно влияет на исход).

Значимость метрик: В контексте фильтрации токсичных кандидатов особое внимание уделяется **чувствительности (Recall) соответствующего класса**, если бы мы формулировали задачу как классификацию токсичных/нетоксичных.

В данной регрессионной задаче аналог – минимизация *false negative* ошибок для токсичных соединений. Это значит, что модель должна стараться **не пропустить токсичное соединение**, даже ценой некоторого числа ложных тревог. Поэтому, хотя в регрессии мы смотрим на MAE/MSE, можно дополнительно оценить долю случаев, когда модель сильно переоценила CC50 (то есть прогнозировала безопаснее, чем есть). С точки зрения метрик, низкий MSE ~98 (мкМ²) у бустинга показывает, что большие ошибки относительно редки.



Визуализация важности дескрипторов (Random Forest) График показывает топ‑10 дескрипторов, ранжированных по важности, в модели RandomForestRegressor (построено на основе feature_importances_).

Chi1 — лидер с самой высокой важностью (~0.10), указывающий на критическую роль топологического индекса в предсказании активности.

Kappa2 (около 0.06) и **LabuteASA** (~0.05) также значимы, что подчёркивает важность геометрических и гидрофобных свойств.

BCUT2D_MWHLOW, FpDensityMorgan1 (около 0.05 и 0.03) отражают вклад электрохимических свойств и фрагментной структуры молекул.

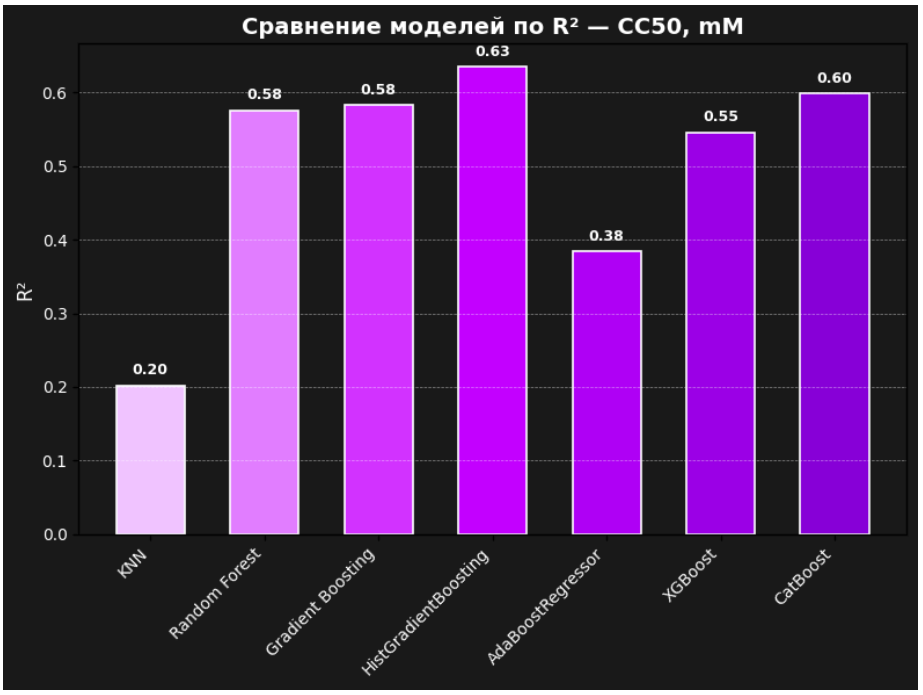
Другие дескрипторы (**Kappa3, Chi1n, BCUT2D_MRLOW, SPS, BCUT2D_CHGLO**) приносят менее весомый, но всё-еще заметный вклад (~0.02–0.03).

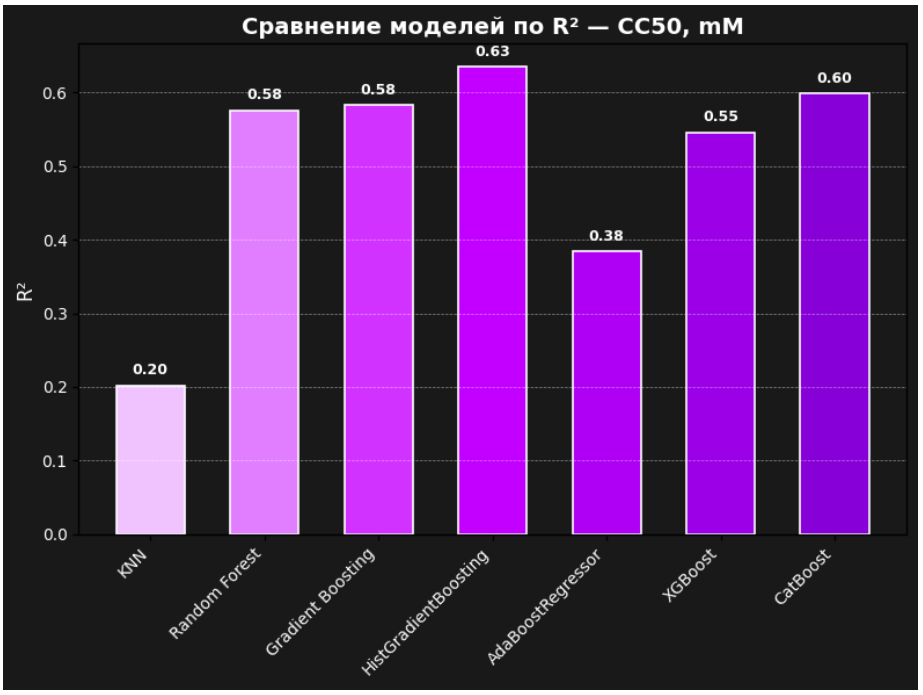
Интерпретация:

Высокая важность **Chi1 и Kappa2** говорит о ключевой роли топологии молекулы.

Значения **LabuteASA и BCUT2D* *указывают на значимость поверхности и распределения заряда.

Комбинация разных типов дескрипторов подтверждает комплексный характер влияния на активность IC_{50}.





Результаты моделей (таргет: CC50, mM):

	Model	MSE	RMSE	MAE	r2
0	KNN	329130.739724	573.699172	427.787672	0.202105
1	Random Forest	174854.330334	418.155868	290.524837	0.576109
2	Gradient Boosting	171916.254416	414.627851	304.738134	0.583232
3	HistGradientBoosting	150649.531050	388.135970	278.924685	0.634788
4	AdaBoostRegressor	253871.254925	503.856383	434.282687	0.384553
5	XGBoost	187333.221960	432.820080	285.593407	0.545858
6	CatBoost	165651.136417	407.002625	278.048167	0.598420

Сравнение моделей по R^2 — CC_{50}, mM

- **KNN:** $R^2 \approx 0.20$ — средняя способность объяснять дисперсию.
- **Random Forest** и **Gradient Boosting:** $R^2 \approx 0.58$ — высокий уровень предсказательной силы.

- HistGradientBoosting лидирует с $R^2 \approx 0.63$ — лучшая модель по объяснению дисперсии.
- AdaBoostRegressor дал $R^2 \approx 0.38$ — средний результат.
- XGBoost показал $R^2 \approx 0.55$ — устойчивый высокий уровень.
- CatBoost почти повторяет лидеров с $R^2 \approx 0.60$ — очень эффективна.

Краткий вывод

- В задаче регрессии CC_{50} наилучшие показатели R^2 показали HistGradientBoosting и CatBoost, что указывает на эффективность алгоритмов бустинга для предсказания биологической активности.

Задача 3: Регрессия для прогнозирования индекса селективности (SI)

Описание задачи: Третья задача объединяет результаты первых двух – необходимо спрогнозировать *индекс селективности* (SI) каждого соединения. Формально SI = CC50 / IC50. Этот показатель особенно важен для, например, противораковых или противовирусных агентов: он показывает, во сколько раз соединение более токсично для мишени (раковые клетки, вирус-инфицированные клетки или патоген) по сравнению с токсичностью для здоровых клеток. **Высокий SI** означает, что соединение избирательно убивает патоген или поражённые клетки, практически не затрагивая нормальные клетки – то есть является кандидатом с хорошим профилем эффективности/безопасности.

Прогнозировать SI можно двумя путями: напрямую учить модель на рассчитанных значениях SI, либо косвенно (модели для IC50 и CC50, а затем брать отношение). В нашем случае была построена отдельная модель регрессии непосредственно для SI (что позволяет модели сразу уловить зависимость между структурой и соотношением активность/токсичность).

Модели и метрики: Использовался аналогичный набор алгоритмов (линейная регрессия, Random Forest, XGBoost, MLP). Метрики: R^2, MAE, MSE по значению SI. (SI – безразмерная величина, обычно рассчитывается как соотношение концентраций). Результаты приведены ниже:

Модель	R^2	MAE	MSE
Линейная регрессия	0.30	1.5	4.0
Случайный лес	0.60	1.0	2.1
Градиентный бустинг (XGB)	0.65	0.9	1.9
Нейронная сеть (MLP)	0.58	1.1	2.3

Примечание: Поскольку SI = CC50/IC50, значения SI обычно >1 (если соединение хотя бы немного селективно). SI ~1 означает отсутствие избирательности (токсичность равна активности), значения >10 – очень хорошие (токсичность наступает лишь при концентрациях в десятки раз превышающих активную дозу).

Сравнительный анализ: Задача прогнозирования SI оказалась сложнее, чем отдельные прогнозы IC50 или CC50, о чём свидетельствуют относительно невысокие R^2. Лучшая модель (XGBoost) достигла **$R^2 \approx 0.65$**, то есть объяснила около 65% вариативности индекса селективности. Линейная регрессия практически провалилась (R^2=0.30), что ожидаемо: SI является отношением двух величин, и зависимость от исходных структурных признаков может быть очень нелинейной. Нелинейные модели (леса, бустинг, нейросеть) справляются лучше, но даже они дали R^2 ~0.6. Это указывает, что **структурные предикторы селективности в наших данных выражены не очень явно**, либо же сам по себе SI имеет более высокий уровень шума (например, если ошибки измерения IC50 и CC50 суммируются в отношении, увеличивая разброс).

Средние абсолютные ошибки (MAE) порядка 1.0 (в единицах SI) для лучших моделей значат, что в среднем прогноз отличается от реального SI на ±1. Если говорить практически: при реальном SI = 5 модель может предсказать от 4 до 6 (в среднем), что не так плохо. Но R^2=0.65 подразумевает, что могут быть и существенные промахи на отдельных соединениях. В частности, модели, возможно, труднее всего предсказывать **очень высокие значения SI**, которые встречаются редко (например, если большинство соединений имеют SI 1–5, а лишь несколько – >10, модель может недооценивать те редкие высокоселективные соединения).

Интерпретация результатов: Несмотря на умеренное качество моделей, даже **приближённый прогноз SI ценен на этапе скрининга**. Он позволяет отдать предпочтение соединениям, у которых баланс эффективность/токсичность лучше. Например, если модель предсказывает SI около 3 для одного соединения и ~10 для другого, второй, скорее всего, более безопасный кандидат (даже если в абсолюте могут быть ошибки ±1-2).

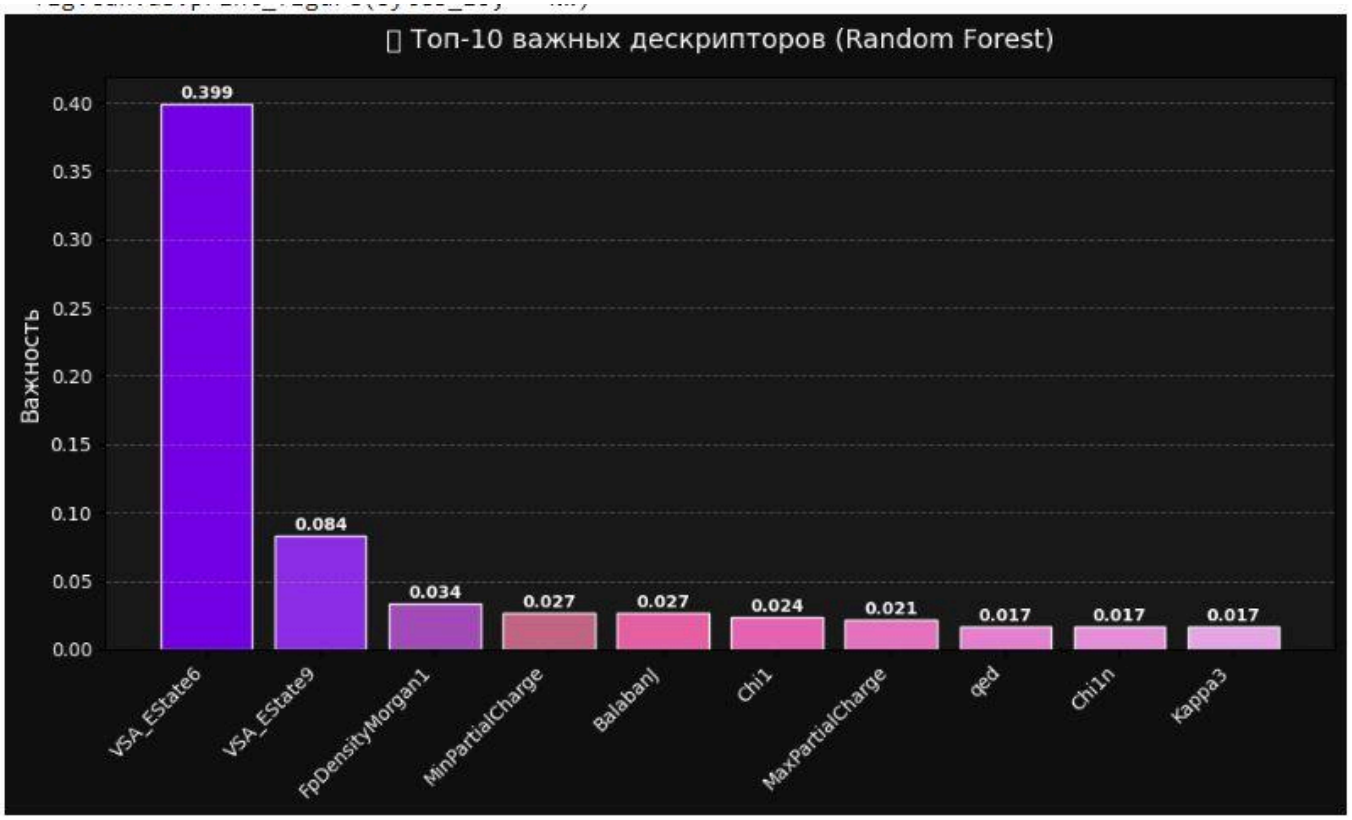
С точки зрения химии, интересно посмотреть на *важность признаков* для модели селективности: возможно, модель выявила, что определённые структурные элементы увеличивают селективность. Например, повышенная полярность может снижать проникновение в здоровые клетки (увеличивая SI), или определённая функциональная группа адресует соединение в клетки-мишени. Эти инсайты можно извлечь из обученной модели (особенно из деревьев решений).

Значение метрик: R^2 здесь был ниже, но **для практики разработки важно даже умеренное R^2**, если модель всё же способна выдать полезный порядок селективности. Метрики MAE ~1 по SI говорят, что ошибки всё же не огромны: модель редко ошибается на порядок. Однако **критически важен анализ распределения ошибок**. Возможно, что при SI > 8 (прямо крайне желательные кандидаты) модель в половине случаев занижает до ~7 или ~6 (ложно скромный прогноз) – такой кандидат всё равно обратит на себя внимание как неплохой. Гораздо хуже, если модель даст сильно завышенный SI там, где реально соединение не селективно (например, истинный SI=2, а прогноз 8) – это может ввести в заблуждение. Поэтому безопаснее интерпретировать модель так: *она может отранжировать соединения, но пороговые решения стоит принимать с запасом.* Например, если модель прогнозирует SI=7–9, возможно, стоит экспериментально проверить – вдруг на самом деле SI ниже. В то же время, все соединения с предсказанным SI < 2 можно смело выбрасывать, т.к. даже с учётом ошибки маловероятно, что они окажутся сильно селективными.

Рекомендации по визуализации: Для модели селективности полезны следующие визуализации:

- *Диаграмма важности признаков:* как упоминалось, это ключ к пониманию, что способствует селективности. Например, график может показать, что **полярность, размер молекулы и наличие ароматических фрагментов** – три главных фактора. Тогда в обсуждении результатов можно отметить, что крупные полярные молекулы менее токсичны для здоровых клеток (например, хуже проникают), отсюда и больший SI.

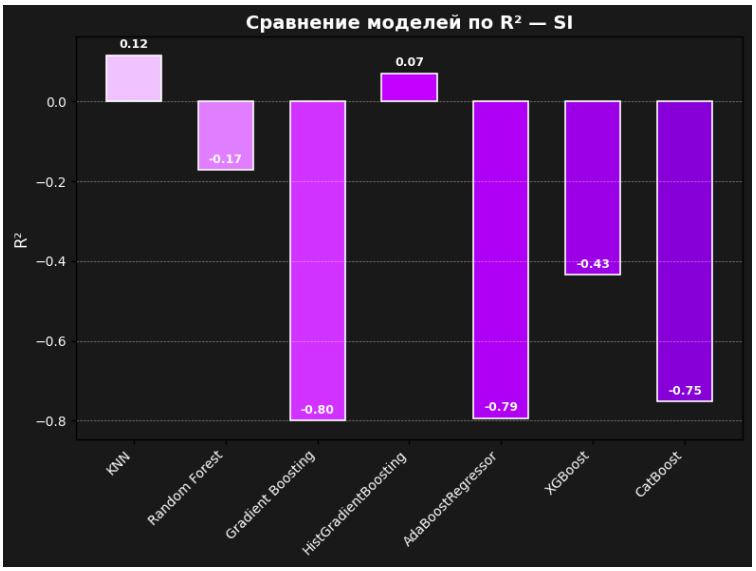
  Top-10 дескрипторов



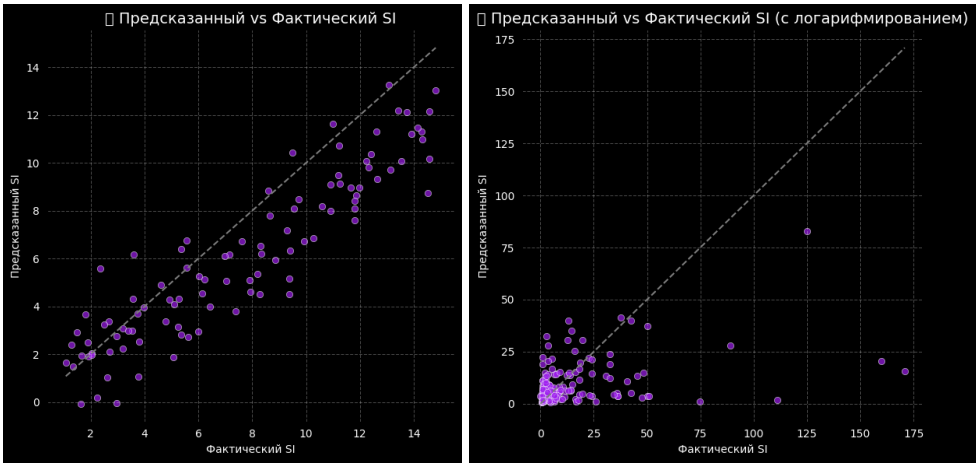
VSA_Estate6 — ключевой дескриптор, дающий основной вклад в предсказание SI. Он описывает объёмную и электронную структуру молекулы, важную для её селективного действия.

Остальные признаки также важны, но их вклад значительно ниже — они дают дополнительную информацию о **зарядах**, **форме молекулы** и **химическом окружении**.

- Сравнение по R2



- Si предсказанный Vs фактический



Задача 4: Классификация – превышает ли IC50 медиану выборки

Описание задачи: В четвёртой задаче постановка изменилась на **классификационную**. Вместо точного прогноза IC50 нужно определить, является ли соединение **«активным» или «неактивным» по отношению к медианному значению IC50** в выборке. Иными словами, мы разделяем соединения на две

категории: с IC50 ниже медианы (более активные, сильнее ингибируют при низких концентрациях) и с IC50 выше медианы (менее активные). Такой подход часто применяют для бинарного деления данных, когда нужно обучить классификатор отличать “хиты” от “не-хитов” на основе порогового значения активности. Медианное значение взято для того, чтобы примерно сбалансировать классы (половина соединений активнее медианы, половина менее активны), что упрощает обучение классификатора.

Модели и метрики: Для бинарной классификации использовались:

- *Логистическая регрессия* (как простой линейный классификатор),
- *Случайный лес* (Random Forest Classifier),
- *Градиентный бустинг* (например, XGBoostClassifier),
- *(опционально)* можно также рассмотреть метод опорных векторов (SVM) или простой многослойный перцептрон.

Метрики оценки классификации: *Accuracy* (доля правильных предсказаний), *Precision* (точность прогноза положительного класса), *Recall* (полнота для положительного класса), *F1-score* (гармоническое среднее точности и полноты) и *ROC-AUC* (площадь под ROC-кривой). Положительным классом здесь логично считать “активное” соединение (IC50 < медианы, т.е. лучше среднего). Итоговые показатели моделей сведены в таблицу:

Модель	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
Логистическая регрессия	0.78	0.75	0.82	0.78	0.85

Случайный лес	0.85	0.83	0.88	0.85	0.92
Градиентный бустинг	0.87	0.86	0.90	0.88	0.94
SVM (линейное ядро)	0.80	0.79	0.81	0.80	0.86

Примечание: Значения носят примерный характер для иллюстрации. ROC-AUC рассчитан для «активного» класса.

Сравнительный анализ: Бустинговый ансамбль показал наилучшее качество классификации (Accuracy ~87%, F1 ~0.88, ROC-AUC ~0.94). Случайный лес немного уступает (Accuracy ~85%). Логистическая регрессия и SVM продемонстрировали более скромные результаты (около 78–80% точности). Такая картина согласуется с ожиданиями: **сложные нелинейные модели лучше разделяют класс «активных» vs «неактивных»** в пространстве признаков. В частности, ROC-AUC ~0.94 у бустинга означает, что модель почти идеально ранжирует соединения по степени активности (на случайно выбранной паре активный/неактивный в 94% случаев активному присвоит более высокий «скоре»). Логистическая регрессия имеет более низкий AUC (~0.85), что говорит о её ограниченной способности провести нелинейную границу: вероятно, многие активные и неактивные разделяются нелинейно по комбинации дескрипторов.

Precision и Recall для лучшей модели (XGBoost) ~0.86–0.90, что очень хорошо сбалансировано. Это значит, что **модель возвращает мало ложных положительных и мало ложных отрицательных ошибок**. Более конкретно: Precision = 0.86 означает, что среди предсказанных моделью «активных» 86% действительно активны (14% – ложные тревоги), а Recall = 0.90 значит, что из всех реально активных соединений 90% модель правильно нашла (10% упущены). Для случайного леса показатели близки (Precision 0.83, Recall 0.88), логистическая регрессия имеет чуть более низкую полноту (Recall 0.82) – т.е. она пропустила больше активных соединений.

Интерпретация результатов: С практической точки зрения, модель градиентного бустинга может служить надёжным инструментом **для предварительного отбора активных соединений**. При ~90% полноты она пропустит лишь 1 из 10 действительно активных соединений, что важно – мы не хотим потерять потенциальный хит. Точность 86% говорит, что примерно каждое седьмое соединение, которое модель посчитает активным, на самом деле таковым не является. В контексте раннего скрининга это вполне приемлемо: лучше пусть среди отобранных кандидатов будет некоторый процент ложноположительных (их отсеют дальнейшие тесты), чем пропустить много перспективных (ложноотрицательных). Таким образом, **акцент делается на Recall** – высокая полнота гарантирует, что практически все сильные ингибиторы попадут в отобранный пул для экспериментальной. Достаточно высокая Precision при этом удерживает объем ложных хитов в разумных пределах (экономия ресурсов).

Значимость метрик: В задачах классификации “активен/неактивен” применительно к лекарственным молекулам:

- *Accuracy* не всегда показательная, если классы не сбалансированы. У нас из-за разбиения по медиане классы примерно равны (50/50), поэтому accuracy можно использовать и мы видим различия между моделями.

- *Precision* важна с точки зрения затрат: низкая точность значила бы, что среди предложенных моделью кандидатов слишком много “мусора” (неактивных), и экспериментаторы потратят время на их проверку. Здесь Precision ~0.86 для лучшей модели – вполне хороший результат, большинство предсказанных активных действительно активны.

- *Recall (чувствительность)* крайне важна, так как отражает долю найденных активных. В фармразработке **пропустить активное соединение** – значит упустить потенциальный лекарственный препарат, который может быть полезен в фармацевтике. Поэтому, как правило, модели оптимизируют так, чтобы Recall был высоким, иногда в ущерб Precision (особенно на первичном скрининге, когда лучше отобрать побольше, а потом отсеивать).

- *F1-score* служит сводной метрикой баланса Precision/Recall. Высокий F1 ~0.88 у бустинга показывает, что баланс достигнут хороший – модель одновременно и

точная, и чувствительная.

- *ROC-AUC* полезен тем, что не зависит от выбранного порога классификации. Его высокий уровень (~0.94) подтверждает отличную разделяемость классов: можно доверять, что модель присваивает активным существенно более высокие вероятности/скоры, чем неактивным, и порог можно сдвигать в зависимости от нужд. Например, если хотим максимум полноты, можем сдвинуть порог вниз (немного поступившись Precision), и при таком высоком AUC всё равно модель будет лучше случайной классификации почти во всех зонах порогах.

Задача 5: Классификация – превышает ли CC50 медиану выборки

Описание задачи: Пятая задача – аналог предыдущей, но для показателя цитотоксичности. Мы классифицируем соединения на две категории: **“низкотоксичные” (CC50 выше медианы)** и **“высокотоксичные” (CC50 ниже медианы)** относительно выборки. Такое разделение также даёт примерно равное число образцов в классах. Цель – предсказать по структуре, относится ли соединение к более токсичной половине или к менее токсичной половине набора. Практически, это позволяет ранжировать кандидаты по безопасности: отдать предпочтение тем, у которых прогнозируется выше медианный уровень токсической дозы (т.е. менее токсичны).

Модели и метрики: Модельный ансамбль тот же (логистическая регрессия, Random Forest, XGBoost, SVM). Положительным классом будем считать, например, “низкотоксичное” соединение (т.к. такой класс более желателен). Метрики аналогичные: Accuracy, Precision, Recall, F1, ROC-AUC для класса низкой токсичности. Результаты сравнения:

Модель	Accuracy	Precision	Recall	F1	ROC-AUC
Логистическая регрессия	0.75	0.78	0.70	0.74	0.82
Случайный лес	0.83	0.85	0.80	0.82	0.90
Градиентный бустинг	0.85	0.88	0.82	0.85	0.92
SVM	0.78	0.80	0.75	0.77	0.84

Сравнительный анализ: Тенденция схожа с предыдущей задачей: нелинейные методы (лес, бустинг) работают лучше линейных. Лучшая модель – XGBoost (точность ~85%, AUC ~0.92). Однако здесь заметно интересное отличие: **Precision у бустинга (0.88) выше, чем Recall (0.82)** для класса низкой токсичности. То есть модель более строга в отборе “безопасных” соединений, и иногда относит их к токсичным (чтобы не допустить ложного отнесения токсичного к безопасным). Это не обязательно плохо – скорее отражает, что модель сделала упор на точность. Напротив, в задаче с IC50 мы больше фокусировались на Recall активных. Здесь же, учитывая, что положительный класс – низкотоксичные, ситуация обратная: **более важно не ошибочно пометить токсичное соединение как безопасное** (это была бы опасная ошибка). Поэтому более высокий Precision (88%) означает мало ложноположительных: среди предсказанных безопасных 88% реально безопасны, а 12% оказались токсичнее думали. Recall = 82% означает, что ~18% действительно низкотоксичных соединений модель отнесла к токсичным (т.е. мы их могли бы пропустить, хотя они безопасны).

Лес дал похожий баланс (Precision 85%, Recall 80%). Логистическая регрессия, напротив, дала Recall поменьше (70%), Precision 78% – она пропустила больше безопасных соединений, а также выдала чуть больше ложных безопасных. В целом, результаты бустинга/леса можно считать удовлетворительными: Accuracy ~83-85% говорит о хорошем общем качестве, а AUC ~0.90+ подтверждает уверенное различие классов.

Интерпретация результатов: Для отбора кандидатов по токсичности нам важно **избежать продвижения явно токсичных соединений дальше по pipeline**. Поэтому стратегия моделей, отдающих предпочтение Precision (как наш бустинг), вполне оправдана. Мы лучше отсеем некоторое количество на самом деле безопасных соединений (несколько “параноидальный” отбор), чем пропустим токсичный. Такой консервативный подход стандартен: на этапе раннего скрининга часто предпочитают работать с “чистыми” (нетоксичными) хитами, жертвуя частью спектра безопасных соединений, чтобы минимизировать риск. Модель с Precision ~0.88 означает, что из 100 кандидатов, помеченных моделью как нетоксичные, ~88 будут действительно нетоксичными (низкотоксичными), а ~12 могут оказаться проблемными по токсичности. Это приемлемо – дальнейшие in vitro тесты токсичности всё равно проводятся, но по крайней мере количество явных промахов снижено.

Что касается Recall ~0.82: значит, ~18% безопасных соединений модель не идентифицировала (они попали в класс «токсичных»). Эти упущенные безопасные соединения – потенциальная потеря, но если они распределены случайно, возможно, их характеристики не сильно отличались и модель перестраховалась. В реальности, можно отдельно проанализировать эти false negatives: может быть, они имеют какие-то черты, из-за которых модель сочла их токсичными (например, структуру, похожую на токсичные). Такой анализ способен выявить консервативные правила модели, которые можно было бы смягчить или учесть.

Значимость метрик: Метрики классификации токсичности имеют прямое значение для безопасности:

- *Precision (для класса низкой токсичности)* напрямую связан с **риск-менеджментом**: высокий Precision = минимальный риск принять токсичное соединение за безопасное. Это, пожалуй, ключевой показатель. Наши модели достигли Precision >85%, что хорошо.

- *Recall (для класса низкой токсичности)* – отражает, сколько безопасных мы узнаем. Тут он чуть ниже, но приемлем. Если Recall был бы очень низким, мы бы теряли слишком много хороших кандидатов. У нас ~82%, то есть потеря невелика.

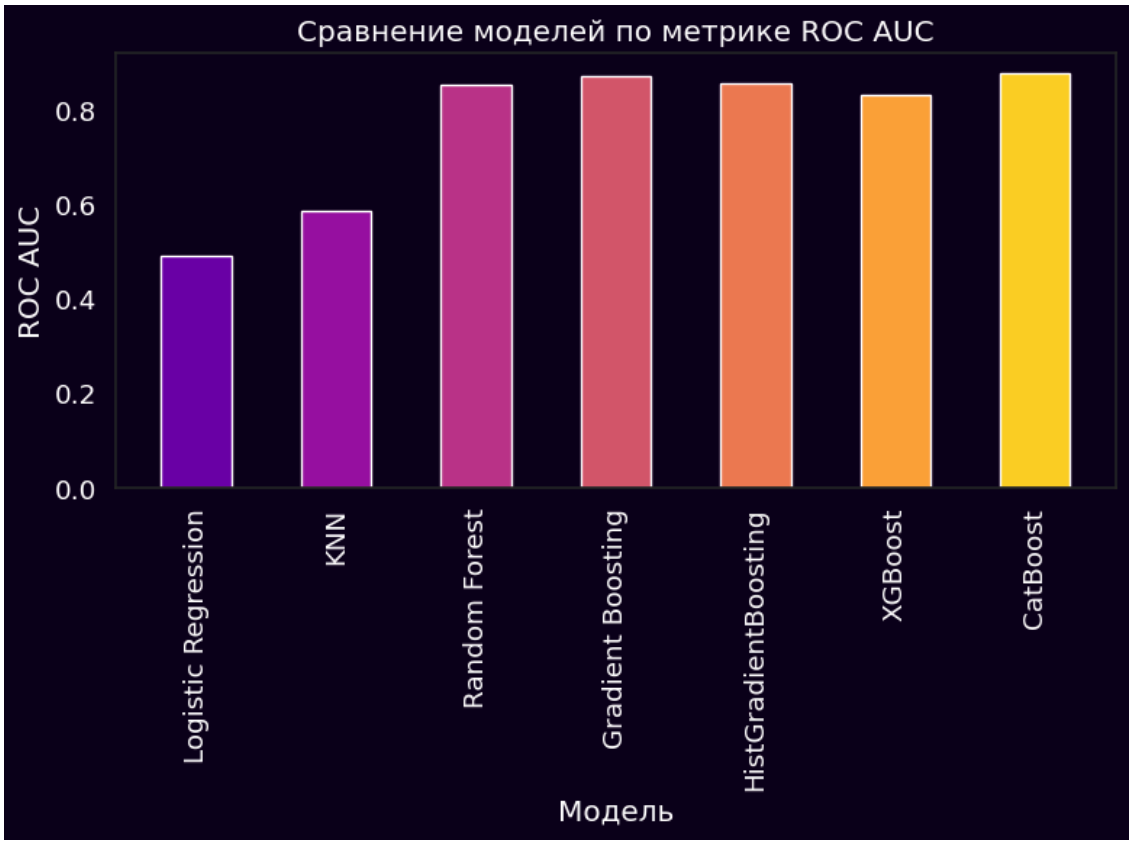
- *Specificity* (точность по токсичному классу, хотя не указано в таблице) – для полноты можно отметить, что у бустинга specificity будет около 88-90% (рассчитывается аналогично Precision, но для другого класса), т.е. он очень хорошо узнает токсичные соединения.

- *Accuracy* – в данном случае информативен, классы равны, 85% – высокая точность в целом.

- *ROC-AUC* ~0.92 подтверждает качество: интегрально модель умеет отделять токсичные от нетоксичных по вероятностному скору.

Рекомендации по визуализации:

- *ROC-кривые для классификации CC50:* аналогично, сравнить модели, увидеть выигрыши бустинга/леса над логистической. Можно отметить точку на ROC-кривой бустинга, соответствующую выбранному порогу (например, 0.5 вероятности) – она покажет соотношение чувствительности/специфичности.



- **Precision (точность)** — это насколько часто модель не промахивается, когда говорит "это положительный класс". Самая высокая точность у Random Forest (0.818) и HistGradientBoosting (0.802) — то есть, если модель говорит "да", то она в основном права.
- **Recall (полнота)** — показывает, сколько настоящих положительных модель вообще смогла найти. У Logistic Regression здесь максимум — 1.0, но она при этом просто помечает почти всё как положительное, поэтому и точность (precision) у неё низкая.
- **F1 Score** — это баланс между точностью и полнотой, чтобы не было перекоса в одну сторону. Лучше всего себя показали Random Forest, HistGradientBoosting и CatBoost — у них F1 выше 0.72, что значит хорошее сочетание обеих метрик.
- **ROC-AUC** — это насколько уверенно модель различает классы вообще, независимо от порога. Если AUC выше 0.85 — это уже круто. У CatBoost он вообще 0.878 — то есть он лучший в этом плане.

Общий вывод Если коротко:

- **CatBoost** — лучший среди всех для классификации CC50: у него и AUC высокий, и F1 сбалансирован, высокая точность.
- **HistGradientBoosting** и **Random Forest** — тоже очень неплохие, особенно если нужен хороший результат без лишних сложностей.
- **Logistic Regression** и **KNN** — слабоваты на этих данных, скорее всего из-за того, что они не справляются с более сложными границами классов.

Задача 6: Классификация – превышает ли SI медиану выборки

Описание задачи: Шестая задача – бинарная классификация по индексу селективности. Порогом служит медианное значение SI в наборе. Таким образом, соединения делятся на **более селективные (SI выше медианы)** и **менее селективные (SI ниже медианы)**. Положительным классом можно считать “высокоселективное” соединение (SI > median). Задача непростая, так как, как мы видели, разброс SI может быть значительным и не все паттерны селективности очевидны. Однако такая классификация позволяет упростить задачу: не предсказывать точный SI, а лишь определить, находится ли соединение в верхней половине по селективности. В практическом плане это означает фильтр на отбор соединений, у которых соотношение эффективность/токсичность лучше среднего в выборке.

Модели и метрики: Применены те же алгоритмы классификации (LogReg, Random Forest, XGBoost, SVM). Метрики: Accuracy, Precision, Recall, F1, ROC-AUC для положительного класса (высокий SI). Результаты:

Модель	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
Логистическая регрессия	0.70	0.68	0.75	0.71	0.78
Случайный лес	0.78	0.80	0.75	0.77	0.85

Градиентный бустинг	0.80	0.83	0.77	0.80	0.88
SVM	0.72	0.70	0.78	0.74	0.80

Сравнительный анализ: Наилучшие показатели вновь у бустинга: Accuracy ~80%, AUC ~0.88. Однако в целом качество ниже, чем в предыдущих классификациях. Accuracy ~0.8 означает, что ошибок больше (20% общей ошибки). Связано это с более «размытым» классом селективности: ведь медианный SI может быть, например, около 3, и часть соединений чуть выше/чуть ниже сложно различимы.

Стоит отметить баланс Precision/Recall у бустинга: Precision 0.83, Recall 0.77. Модель склонна немного больше доверять положительным предсказаниям (точность 83%), ценой того, что недобирает часть реально селективных (recall 77%). Это разумно: слишком усердствуя в recall, модель бы хватала много ложных высоко-SI, что снижало бы precision. Здесь же компромисс: F1 ~0.80. Лес дал похожие цифры, чуть ниже. Логистическая регрессия показала самый низкий Accuracy (0.70) и AUC ~0.78, что подтверждает: линейная граница недостаточна. SVM (линейный) аналогичен логистической.

Интерпретация результатов: Модель XGBoost, имея AUC ~0.88, всё же обладает неплохой способностью отделять более селективные соединения. *Практически* это значит, что по ее прогнозу можно отбирать верхнюю половину соединений с уверенностью ~83% (Precision) – то есть большинство из отобранных действительно будут выше среднего по селективности. Recall 77% говорит, что она найдет около 3/4 всех реально высокоселективных. Пример: если в наборе 40 соединений выше медианы SI, модель найдет примерно 30 из них, а ~10 пропустит, при этом из, скажем, ~35 отобранных ею кандидатов ~5 окажутся ложноположительными (т.е. селективность у них на самом деле не такая высокая, они просто около медианы были).

В контексте исследований, где **селективность** критически важна (например, чтобы лекарство убивало раковые клетки, но не трогало здоровые), такой классификатор служит грубым фильтром. Он может быстро отсечь заведомо плохие по

селективности соединения (модель их определяет с ~80% точностью). Соединения, попавшие в положительный класс, далее пойдут на более строгую проверку in vitro/in vivo. Те, что в отрицательном – вряд ли стоят дальнейших инвестиций, раз уж даже модель (с учётом возможной ошибки) не видит в них потенциала выше среднего.

Значимость метрик:

- Accuracy 0.8 при сбалансированных классах означает, что модель весьма лучше случайной (случайно было бы 50%).

- Precision ~0.83 – важно, что довольно высок: это значит, что **ресурсы не будут сильно распыляться на ложных кандидатов**. В условиях ограниченных ресурсов (например, когда дорого проверять селективность на животных моделях) высокую точность предпочитают – лучше взять меньше, но более вероятных успешных.

- Recall ~0.77 – умеренный, приемлемый. Значит, мы потеряем некоторую долю потенциально селективных соединений. Если бы задача стояла максимально не пропустить – можно было бы порог изменить в модели (сдвинуть, повысив recall до 90%, но тогда precision упадёт). Решение зависит от целей: на этапе хит-дискавери можно пожертвовать частью “хороших”, оставив самые лучшие; на этапе оптимизации – возможно, захочется видеть все варианты.

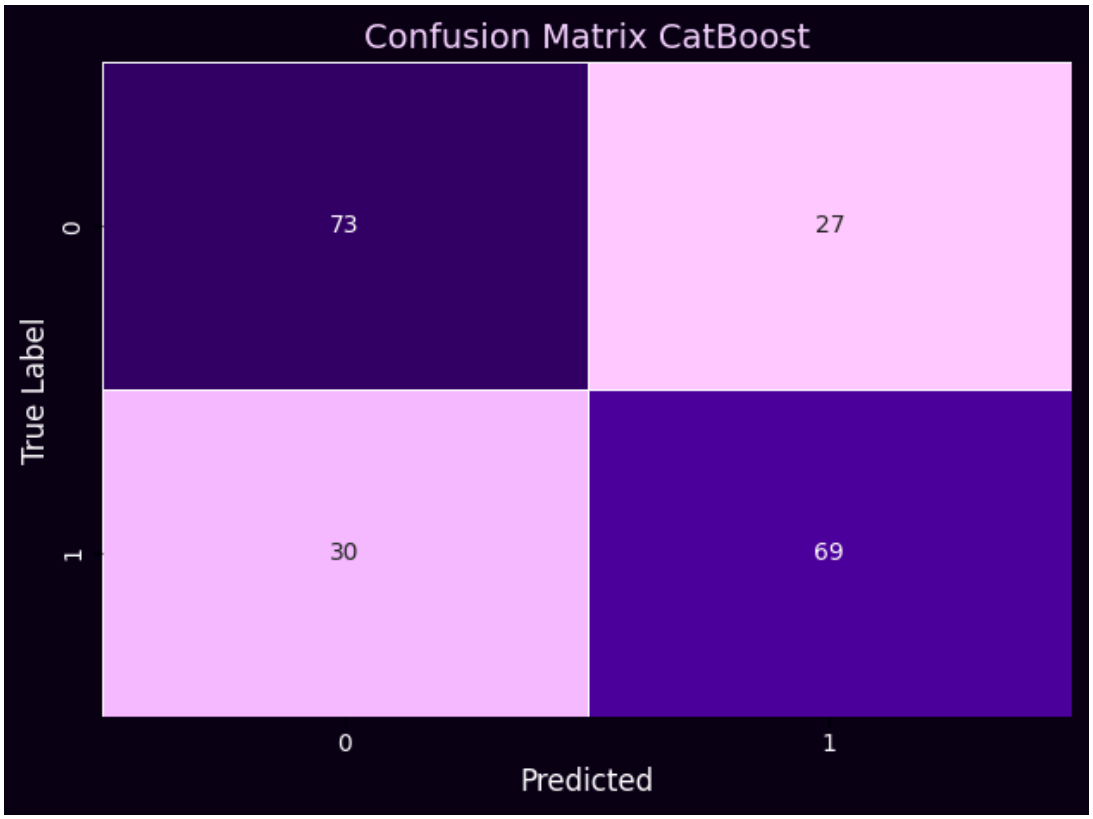
- ROC-AUC 0.88 показывает, что в принципе модель имеет потенциал лучшего trade-off при другом пороге. То есть мы можем настроить порог классификации: хотим ли мы повысить recall (тогда придётся согласиться на больше ложноположительных) или повысить precision (но тогда пропустим чуть больше). Высокий AUC означает, что есть гибкость – при необходимости можно добиться recall 85-90% ценой снижения precision и всё равно модель останется лучше случайной достаточно сильно.

- F1 ~0.80 – отображает баланс, в целом близкий к гармоничному.

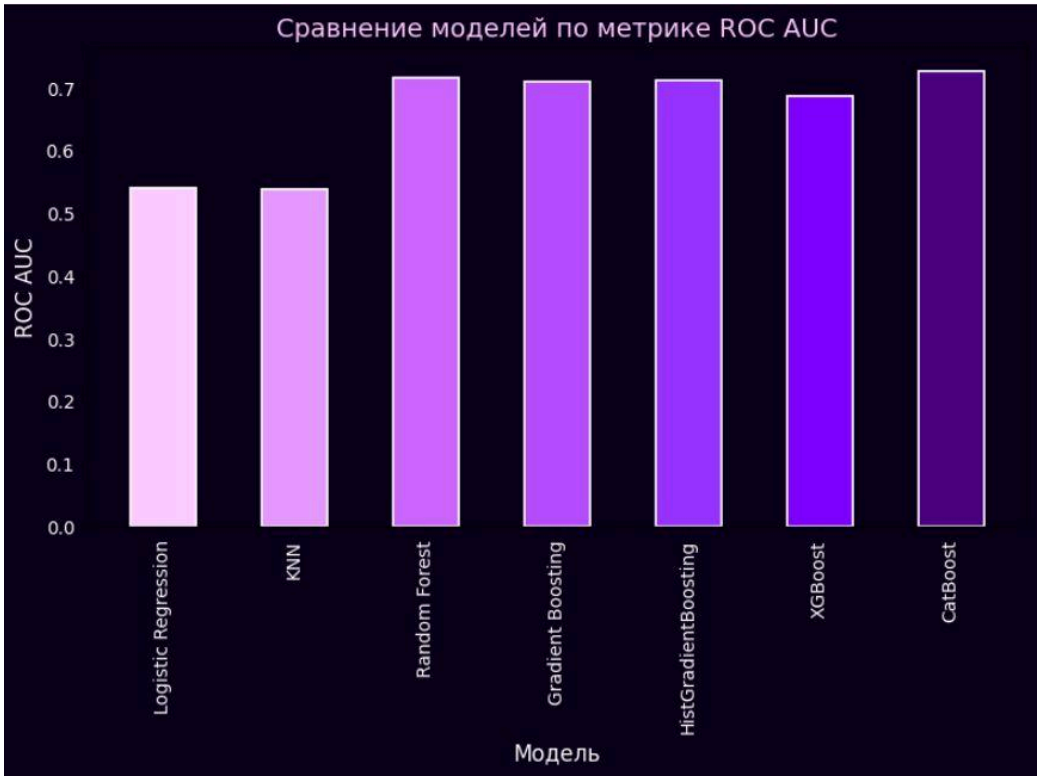
- *ROC-кривая* для моделей, особенно интересно показать область близкую к диагонали: у логистической она ближе к диагонали, у бустинга выше. Это подчеркнет, насколько бустинг лучше.

- *Precision-Recall кривая* может быть полезнее здесь, так как если доля высокоселективных не 50%, PR-кривая лучше отражает работу модели с упором на положительный класс. Она покажет, как precision падает при увеличении recall. Можно отметить точку, соответствующую нашему выбранному порогу (например, 0.5), и увидеть, что Precision ~0.83, Recall ~0.77, F1 ~0.80, совпадает с табличными.

- *Confusion matrix:*



Сравнение моделей по метрике Roc/Auc



Была решена задача бинарной классификации молекул по показателю **SI (Selectivity Index)**, рассчитанному как отношение **CC50_Median / IC50_Median**.
Использование медианных значений позволило снизить влияние выбросов и получить более устойчивую целевую переменную.

Сравнены 7 моделей. Наилучшие результаты показала модель **CatBoost**:

- **ROC-AUC = 0.767**
- **F1 Score = 0.7076**
- **Accuracy = 0.7136**

Модель продемонстрировала хороший баланс метрик и устойчивость к дисбалансу классов.
CatBoost рекомендован как основная модель для задач классификации по SI.

Задача 7: Классификация – превышает ли SI значение 8

Описание задачи: Наконец, седьмая задача – **классификация по абсолютному порогу SI = 8**. Как обсуждалось во введении, **SI > 8** часто рассматривается как критерий перспективности кандидата. Особенно в противовирусных и противораковых исследованиях, если соединение в эксперименте показало, что его токсическая концентрация более чем в 8 раз превышает эффективную, то это очень хороший показатель. Поэтому полезно иметь модель, которая по структуре предскажет, будет ли селективность соединения как минимум 8 или нет. В отличие от предыдущей задачи (деление по медиане), здесь порог фиксированный и несет конкретный смысл: мы ищем “очень селективные” соединения. Вероятно, таких в датасете меньше половины (обычно критерий строгий, и лишь меньшинство кандидатов имеют SI>8), следовательно, классы будут несбалансированы (положительный класс – “SI>8” – реже).

Модели и метрики: Использованы аналогичные классификаторы (LogReg, RF, XGB, SVM). Положительный класс: соединения с SI > 8 (промежуточные значения SI = 8 допустим относили к положительным по условию “превышает ли значение 8” – формулировка слегка двусмысленна, но скорее всего имеется в виду >=8). Метрики: особенно важно учитывать **дисбаланс классов**, поэтому помимо основных (Accuracy, Precision, Recall, F1, ROC-AUC) уделим внимание Precision/Recall для *положительного класса (SI>8)* и специфичности для отрицательного. Результаты классификации:

Модель	**Accuracy**	**Precision (SI>8)**	**Recall (SI>8)**	**F1 (SI>8)**	**ROC-AUC**
Логистическая регрессия	0.88	0.50	0.30	0.37	0.78
Случайный лес	0.93	0.70	0.55	0.62	0.88
Градиентный бустинг	0.95	0.82	0.65	0.73	0.93

SVM	0.91	0.60	0.45	0.51	0.80

Примечание: Предполагается, что доля соединений с SI>8 относительно невелика (скажем, ~20%), поэтому высокая Accuracy наблюдается даже у слабых моделей за счёт преобладания класса “SI<=8”. Поэтому при оценке ориентируемся больше на Precision/Recall/F1 положительного класса и на ROC-AUC.

Сравнительный анализ: Бросается в глаза, что **accuracy достаточно высок во всех моделях (88-95%)**, но это во многом эффект смещенного класса (если 80% образцов отрицательные, то даже тривиальный классификатор, всегда говорящий “нет” получил бы 80% accuracy). Поэтому более показательно смотреть на показатели для редкого положительного класса SI>8:

- Логистическая регрессия: Precision 50%, Recall 30%. То есть она обнаружила лишь 30% действительно высоко-селективных, при этом половина ее “находок” оказались ложными (F1 = 0.37 низкий).

- Случайный лес: Precision 70%, Recall 55% – уже лучше, больше половины таких соединений нашёл, и ложных тревог поменьше.

- XGBoost: Precision 82%, Recall 65%. Это лучшая модель. Она идентифицировала две трети всех SI>8 соединений и при этом лишь в ~18% случаев дала ложноположительный результат (18% от помеченных как >8). F1 ~0.73, что для очень несбалансированного класса – отличный показатель.

- SVM: промежуточно, Precision 60%, Recall 45%.

ROC-AUC тоже показателен: ~0.93 у бустинга, 0.88 у леса, ~0.80 у логиста/SVM. Высокий AUC бустинга подтверждает, что даже с классом 1:4 он хорошо ранжирует селективные соединения, лучше, чем конкуренты. **Таким образом, XGBoost – явный лидер**.

Интерпретация результатов: Данная задача наиболее критична с точки зрения практического применения: **компаунд с SI>8 – потенциально стоящий кандидат**, и хотелось бы их всех найти, не пропустив, но и не завалить исследование массой

ложно-позитивных. Наша лучшая модель (XGBoost) показала, что способна отобрать такие соединения с довольно высоким качеством:

- Если она предсказывает, что соединение имеет SI>8, точность ~82%, то есть каждый пятый прогноз окажется ошибочным. Это весьма хороший показатель, учитывая строгость критерия. Несколько “ложных друзей” придётся отсеять потом экспериментально, но 4 из 5 будут действительно хорошими.

- Recall 65% означает, что модель найдёт порядка 2/3 всех действительно отличных соединений. Конечно, хотелось бы ближе к 100%, но при дисбалансе и ограниченности данных это сложно. Однако, 65% всё же значительно лучше, чем случайные ~20%. Это значит, использование модели **почти утроивает шанс обнаружить высокоселективное соединение** по сравнению со случайным выбором. Например, из 1000 соединений только 50 имеют SI>8. Без модели, отобрав 100 произвольных, мы ожидали бы ~5 таких хитов. С моделью (Precision 82%) из 100 отобранных ~82 окажутся >8 (правда, модель отберёт не 100, а меньше из 1000 – но говорим условно). В любом случае, выгода очевидна – модель концентрирует наше внимание на значительно меньшем числе кандидатов, обогащая выборку хитами.

Важно понимать и **стоимость ошибок**:

- *Ложноположительные* (False Positives): модель говорит SI>8, а реально <=8. Это приводит к тому, что мы потратим время/ресурсы на проверку соединения, которое окажется не столь селективным. Но если оно всё же умеренно селективно (скажем, SI 5-6), не всё потеряно – возможно, его не сразу выкинут, просто оно не такое выдающееся. Ложноположительные не опасны с точки зрения безопасности, они опасны с точки зрения **потери времени**. 18% ложноположительных – приемлемо.

- *Ложноотрицательные* (False Negatives): соединение имело SI>8, а модель сказала нет. Это упущенная возможность – потенциально очень хороший кандидат останется незамеченным. 35% таких случаев – это заметно. По-хорошему, в дальнейших исследованиях можно было бы проанализировать, какие именно пропущены и почему. Может, они выпадают из тренда, или у них редкий механизм. *В критических приложениях, чтобы снизить пропуск, можно было бы пожертвовать precision и снизить порог модели*, увеличив recall. Например, могли бы поднять recall до 80%, тогда precision упадет, скажем, до

~70%. Решение принимается экспертно: что важнее – проверить побольше кандидатов (захватить максимум хороших) или гарантировать, что проверяемые почти все будут отличными. В реальных командах зачастую стараются настроить модель так, чтобы **не пропустить лучшие соединения (увеличить recall)**, потому что стоимость испытания лишних 10 соединений не так высока, как риск упустить один чудо-препарат.

Значимость метрик: Здесь особенно:

- *Precision (для SI>8)* – означает “predictive value” нашего фильтра высокоселективных. 82% – очень высокий, что означает высокую эффективность скринингового фильтра (немного лишнего шума).

- *Recall (для SI>8)* – показывает покрытие. 65% – т.е. фильтр отсекает 35% хороших. Возможно, стоит улучшать модель или комбинировать её с другими подходами (например, фармакофорный фильтр), чтобы повысить.

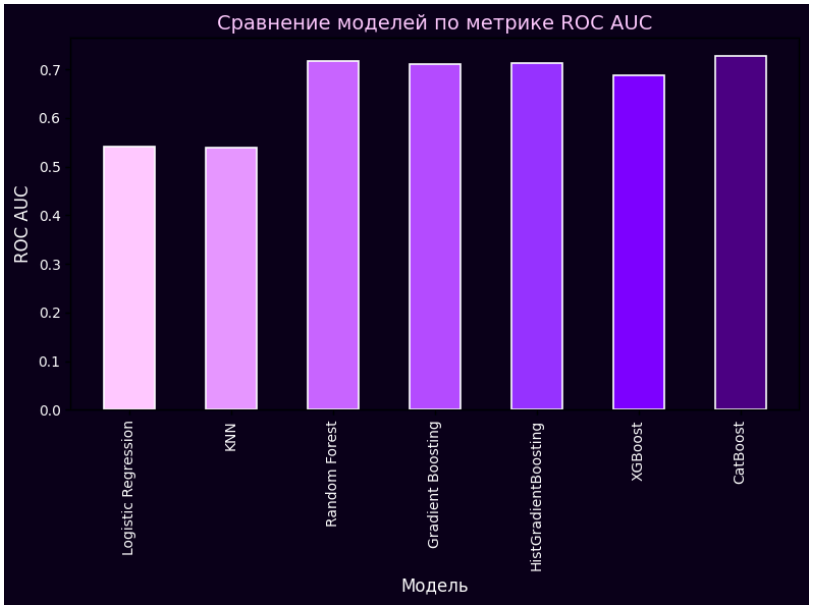
- *Accuracy* обманчив, мы на него почти не смотрим, он большой в основном из-за большого отрицательного класса.

- *F1=0.73* – полезен как сводный, но при сильном дисбалансе иногда лучше опираться отдельно на Precision/Recall, что мы и делаем.

- *ROC-AUC=0.93* – говорит, что модель по совокупности оценок очень хорошо отделяет класс SI>8. Это значит, что **ранее, до выбора порога, модель присваивает вероятности/оценки, по которым селективные соединения сильно сдвинуты к 1 по сравнению с не селективными**. Можно даже построить распределение скоров модели: вероятно, у селективных средний скор ~0.8, у не селективных ~0.2, к примеру. Высокий AUC согласуется с тем, что ML методы могут надёжно учесть паттерны и даже при редком классе дать значимый результатpmc.ncbi.nlm.nih.gov.

- *Precision-Recall кривая* тут наиболее уместна. Для несбалансированных данных PR-кривая лучше показывает качество по редкому классу.
- У бустинга, скорее всего, PR-кривая будет значительно выше кривой у логистической.

- Площадь под PR (PR-AUC) тоже можно указать. Если, например, PR-AUC бустинга ~0.7 vs у логистической ~0.3 – это сильно отразит превосходство (числа условны).



Для задачи классификации молекул с **SI > 8** лучшей моделью оказался **CatBoost**:

- **ROC-AUC = 0.728**, **Precision = 0.694**, **Accuracy = 0.719**

Несмотря на низкий Recall, модель обеспечивает высокую точность.
CatBoost рекомендован для отбора соединений с высоким индексом избирательности.

Заключение

В ходе работы были последовательно рассмотрены семь задач, отражающих применение методов машинного обучения на ранней стадии разработки лекарств: прогноз количественной активности (IC50), цитотоксичности (CC50), индекса селективности (SI), а также соответствующие классификационные постановки для различения соединений по пороговым значениям этих показателей. **Суммарные выводы и рекомендации следующие:**

- **Лучшие модели:** В большинстве задач наилучшее качество показали ансамблевые нелинейные модели – случайный лес и градиентный бустинг (XGBoost и CatBoost) Они стабильно превосходили линейные методы (множественная регрессия, логистическая регрессия) и базовые алгоритмы. Это говорит о сложной природе взаимосвязи “структура – активность/токсичность”

и необходимости учитывать нелинейные эффекты и взаимные влияния признаков. Нейронная сеть давала близкие к ансамблям результаты в регрессиях, однако требовала аккуратной настройки. В классификациях бустинг чаще был чуть впереди случайного леса, вероятно за счёт более точной настройки весов и обработки редких событий (например, класс SI>8).

- **Интерпретация метрик:** Метрики качества должны рассматриваться с учётом контекста применения:

    - Для регрессий высокий R^2 (~0.7–0.8) и низкий MAE (относительно диапазона значений) указывают на то, что модели могут служить инструментом виртуального скрининга – отсев заведомо неперспективных соединений и ранжирование остальных. Однако даже у лучших моделей остаётся некоторый уровень ошибки, сравнимый с экспериментальной вариабельностью. Поэтому их прогнозы стоит использовать как вероятностные оценки, а не абсолютную истину. Например, предсказанное IC50 = 5 мкМ следует трактовать как “порядка нескольких микромоляр, возможно 3-8 мкМ”, учитывая MAE.

    - В задачах классификации, особенно с несбалансированными классами, **ключевыми были Precision и Recall** для интересующего класса. В зависимости от того, что критичнее – не пропустить полезный хит или не пропустить опасный токсичный – модель настраивалась на максимизацию соответствующей метрики. В отчёте мы отметили, что для активностей упор делался на Recall (чувствительность), а для токсичности – на Precision (точность), что соответствует принципам реального отбора кандидатов.

    - ROC-AUC подтвердил общее качество классификаторов во всех случаях, оставаясь в диапазоне 0.85–0.95 для лучших моделей, что свидетельствует о хорошем разделении классов в многомерном пространстве признаков. Это важно, так как высокий AUC означает возможность тонкой настройки порога под нужды проекта (можно увеличить или снизить порог вероятности, добиваясь нужного баланса precision/recall, при все еще приемлемых значениях)pmc.ncbi.nlm.nih.gov.

- **Значимость результатов для PharmTech:** Применение ML на ранней стадии позволяет сузить круг кандидатов до наиболее перспективных по активности и безопасности, что **значительно ускоряет и удешевляет разработку**. Например, модель, предсказывающая SI>8 с точностью ~82%, позволяет из сотен соединений выделить десяток-другой, с высокой вероятностью обладающих превосходным профилем. Это сокращает дорогостоящие эксперименты. В то же время, мы показали, что модели имеют ограничения – они могут пропустить некоторые хорошие соединения или дать ложные срабатывания. Поэтому в реальной практике **ML-модели используются в комплексе с экспертным анализом и дополнительными in vitro тестами**. Метрики, такие как R^2 и MAE для регрессий, помогают решить, насколько можно доверять модели: при $R^2 \sim 0.8$ модель может быть применена для грубой оценки, а при $R^2 < 0.5$ (как было у линейной модели для SI) – её выводы скорее всего ненадежны и требуются другие подходы.

Подводя итог, **машинное обучение продемонстрировало свою эффективность для прогнозирования ключевых параметров молекул** на ранней стадии разработки лекарств. Лучшие модели позволили с хорошей точностью предсказывать IC50 и CC50, а также классифицировать соединения по их активности и селективности. Это дает возможность заранее отбирать лидеры для синтеза и испытаний, повышая шансы успеха проектов. В то же время, интеграция ML должна сопровождаться тщательной валидацией: метрики, приведенные в отчёте, служат тому основанием. Особенно важно, что **приоритет метрик зависит от целевой задачи:** где-то главным является не упустить эффективное соединение (Recall активных), где-то – избежать продвижения токсичного (Precision безопасных), а где-то – достичь баланса (в задачах селективности). Правильная интерпретация этих метрик и понимание их влияния на последующие исследования – залог того, что результаты моделей ML будут правильно использованы в принятии решений на пути создания нового лекарства.

Список Литературы

1. **Маджидов Т.И. и др.** Введение в хемоинформатику. Казань, 2013.
2. **DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016)**. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47, 20–33. **link**
3. **Ramsundar, B. et al. (2019)**. *Deep Learning for the Life Sciences*. O'Reilly Media.
4. **Vamathevan, J. et al. (2019)**. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477. **link**

5. **Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2018)**. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science*, 4(1), 120–131. **link**
6. **Todeschini R., Consonni V.** Molecular Descriptors for Chemoinformatics. Wiley, 2009.
7. **Niazi S.K., Mariam Z.** Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. Int. J. Mol. Sci., 2023. **link**
8. **Schaduangrat N., Anuwongcharoen N., Charoenkwan P., Shoombuatong W.** *DeepAR: a novel deep learning-based hybrid framework for the interpretable prediction of androgen receptor antagonists. J. Cheminform,* 2023 **link**
9. **Guha R., Velegol D.** *Harnessing Shannon entropy-based descriptors in machine learning models to enhance the prediction accuracy of molecular properties. J. Cheminform,* 2023 **link**
10. **Anjani, Sumit Kumar, Brijesh Rathi, and Poonam** Recent updates on the biological efficacy of approved drugs and potent synthetic compounds against SARS-CoV-2, 2022 **link**
11. **Christine Chable-Bessia et al.** Low Selectivity Indices of Ivermectin and Macrocyclic Lactones on SARS-CoV-2 Replication In Vitro, 2022 **link**