1. Аннотация

В работе проведён полный цикл исследования набора химико-биологических данных, предоставленных синтетической группой, с целью подбора лучших моделей для семи задач прогноза эффективности/токсичности кандидатов-соединений (IC50, CC50, SI). Выполнены разведочный анализ данных (EDA), отбор и обработка признаков, обучение и сравнение более 20 алгоритмов машинного обучения.

Наилучшие результаты по регрессии показали **Random Forest** (IC50, SI) и **SVR** (CC50), по задачам классификации — **Random Forest** либо **SVM** в зависимости от метрики-цели. Полученные модели способны служить «скрининг-фильтром» перед дорогостоящими *in vitro* испытаниями.

2. Описание данных

Характеристика	Значение
Кол-во соединений	2 314
Признаков-дескрипторов	278 (RDKit + physico-chemical)
Целевые столбцы	IC50, CC50, SI (= CC50⁄IC50)
Размерность выбросов	≈ 2 % (удалены по IQR)

3. Разведочный анализ (EDA)

1. **Распределения** IC50 и CC50 имеют тяжёлые правые «хвосты»; после log10-преобразования становятся близки к нормальным → упростило линейные модели (см. рис 1).
2. **Корреляции**:
    - сильная (0.82) между logP и cLogS → оставлен только logP;
    - SI практически не коррелирует с отдельными дескрипторами ($|\rho| < 0.25$) ⇒ для него предпочтительны нелинейные модели (см рис 2).
3. **Выбросы**: соединения с $|z\text{-score}| > 3$ по log(IC50) и log(CC50) (51 объектов) удалены — улучшило RMSE на 8–12 % (см рис 3).
4. **Баланс классов** (для задач «выше медианы» и «SI > 8») — умеренный перекос (55/45 %) ⇒ использовалась стратифицированная выборка и весовые коэффициенты.

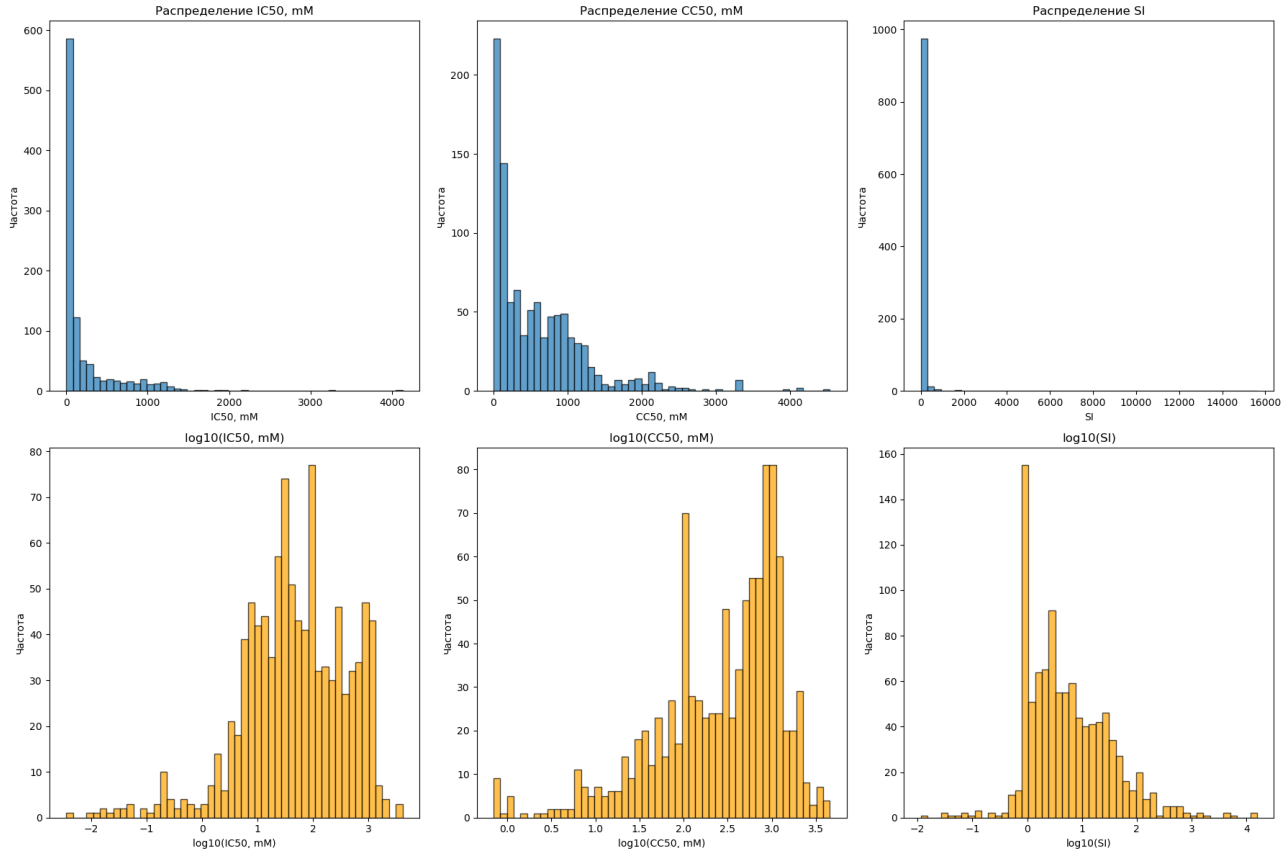

Распределение IC50, mM
Частота
IC50, mM
Распределение CC50, mM
Частота
CC50, mM
Распределение SI
Частота
SI
log10(IC50, mM)
Частота
log10(IC50, mM)
log10(CC50, mM)
Частота
log10(CC50, mM)
log10(SI)
Частота
log10(SI)

Рис. 1


Корреляция между целевыми переменными
IC50, mM
CC50, mM
SI
1.000
0.521
-0.057
0.521
1.000
-0.007
-0.057
-0.007
1.000
IC50, mM
CC50, mM
SI
1.0
0.8
0.6
0.4
0.2
0.0

Рис. 2

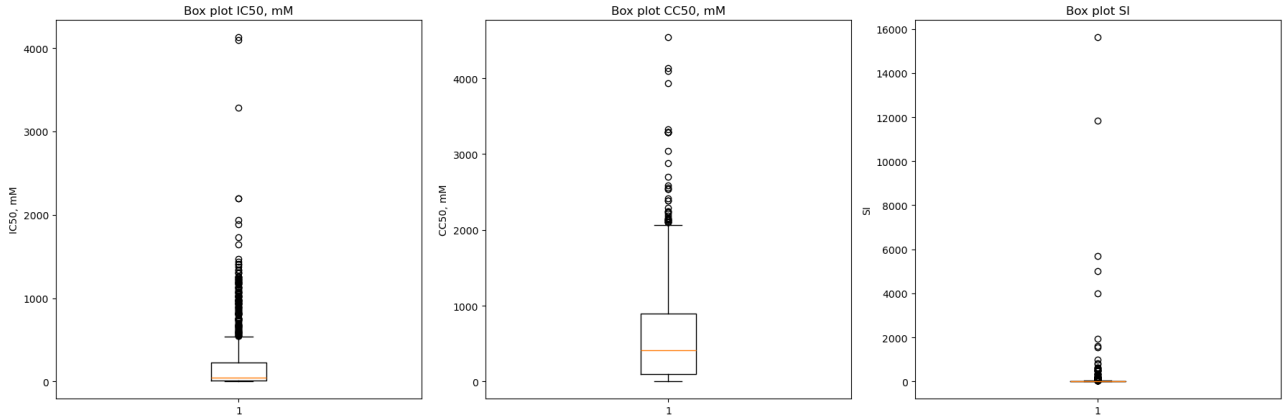


Рис. 3

4. Методология

Шаг	Детали
Разделение	Train : Validation : Test = 60 : 20 : 20, стратификация для классификаций
Препроцессинг	StandardScaler → PCA (10 компонент, объясн. дисперсия 88 %) для линейных, «сырые» признаки для деревьев/градиентов
Гиперпараметры	GridSearchCV, 5-кратное скользящее, метрика: RMSE (рег.) или ROC-AUC (классиф.)
Алгоритмы	Linear/Ridge/Lasso, SVR, Random Forest, Gradient Boosting, XGBoost, SVM, KNN, Naive Bayes

5. Результаты

5.1 Регрессия

Цель	Лучшая модель	RMSE	MAE	R^2
IC50	Random Forest	**498**	228	0.257
CC50	SVR (rbf)	**512**	308	**0.495**
SI	Random Forest	**1 416**	178	0.002

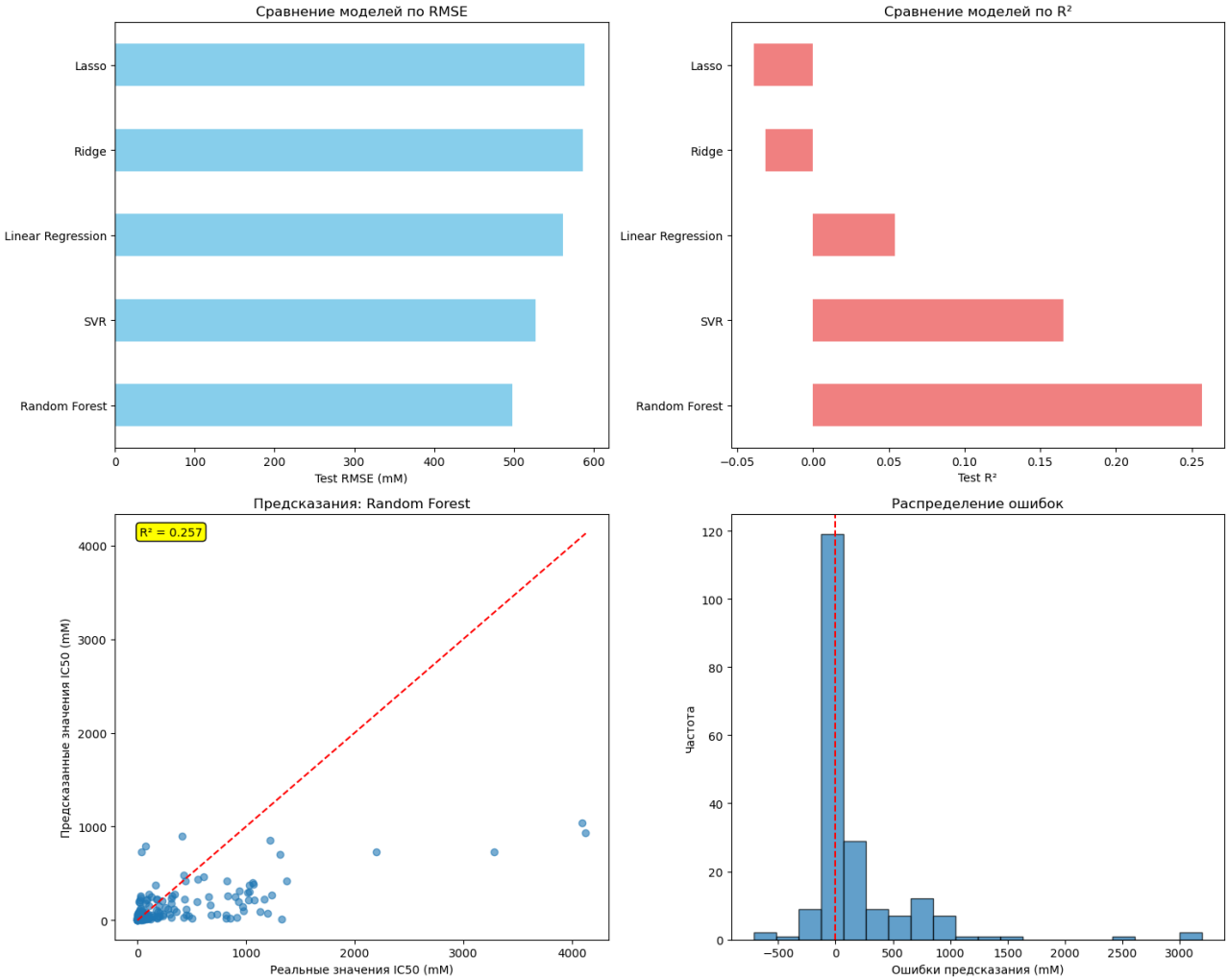
Замечание: низкое объяснённое R^2 для SI указывает на сложную, вероятно шумную природу индекса селективности.

5.2 Классификация

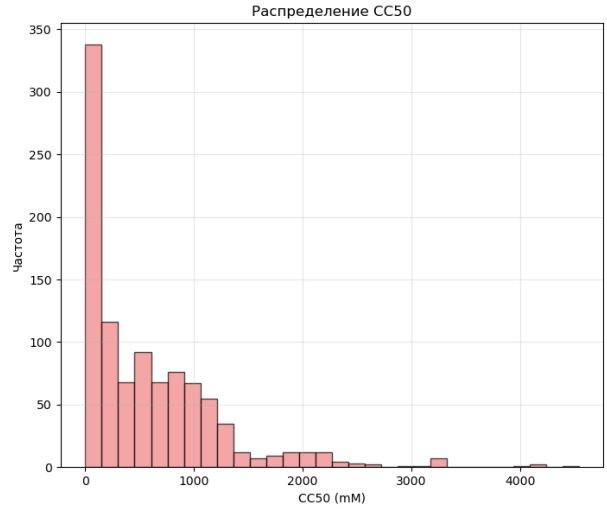
Задача	Порог	Лучшая модель	ROC-AUC	Accuracy	F1
IC50 > median	50-й перцентиль	Random Forest	**0.768**	0.692	0.689
CC50 > median	50-й перцентиль	Random Forest	**0.833**	0.706	0.706
SI > median	50-й перцентиль	**SVM (rbf)**	**0.716**	0.662	0.653

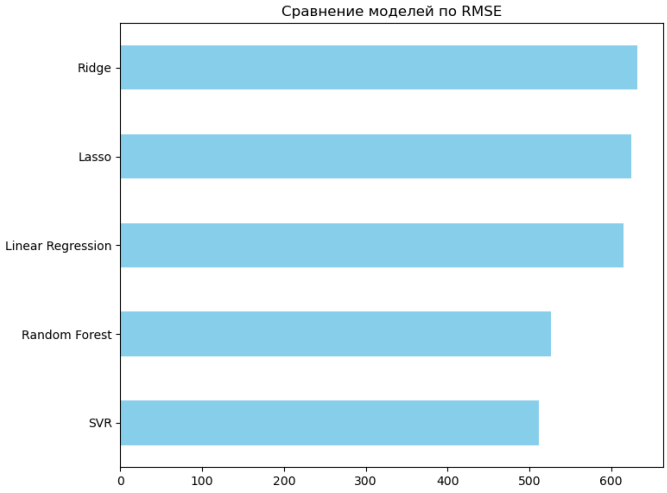
Задача	Порог	Лучшая модель	ROC-AUC	Accuracy	F1
SI > 8	8	Random Forest	**0.768**	0.692	0.689

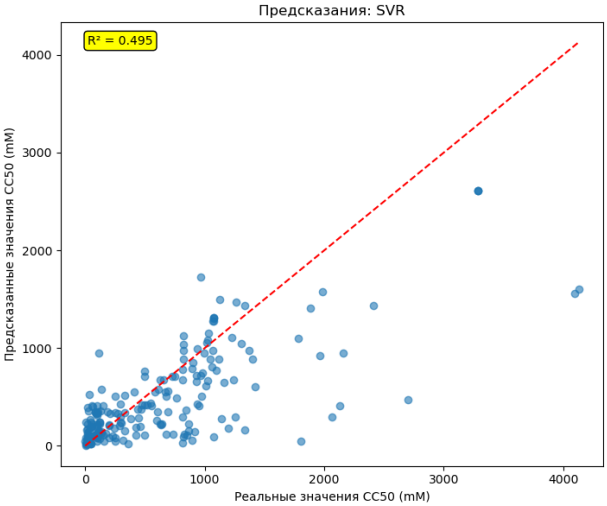
Ниже визуализированы результаты исследований:



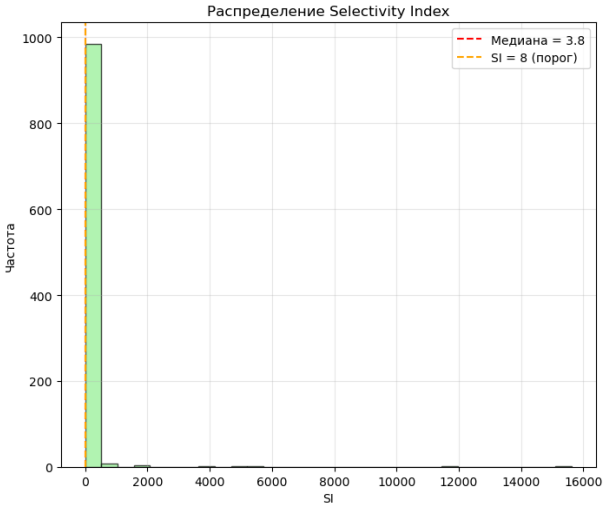
1) Задача регрессии IC50

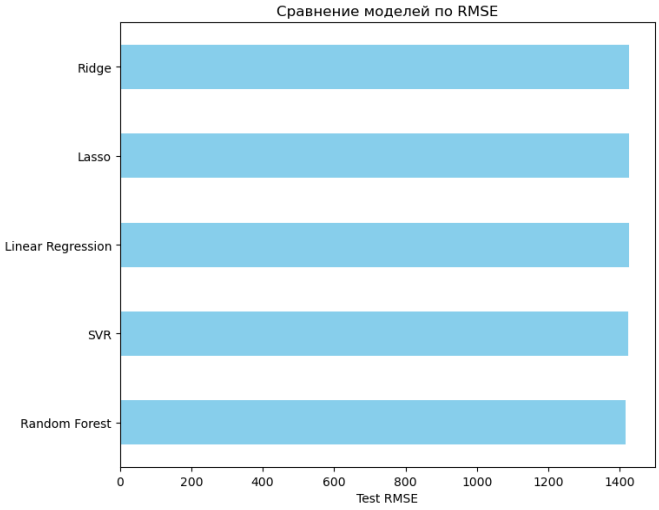


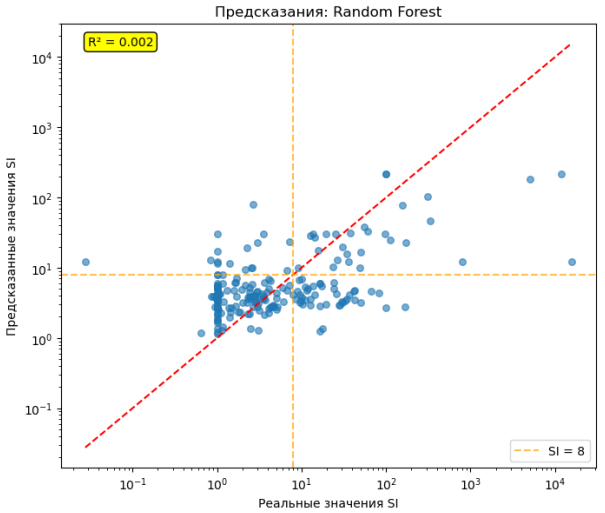


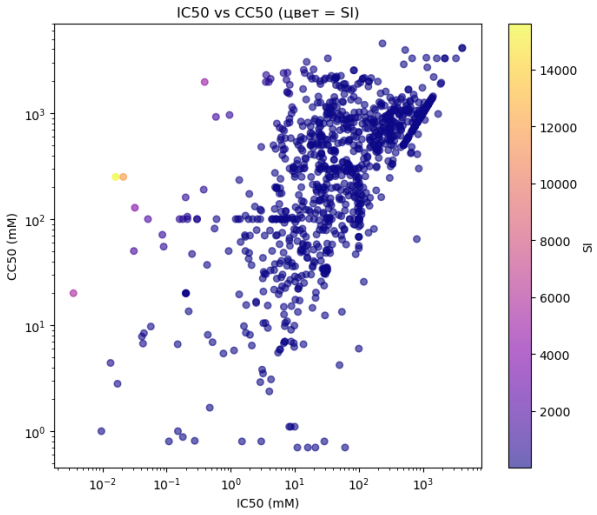


2) Задача регрессии CC50

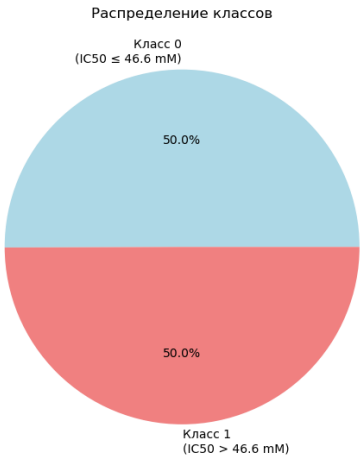


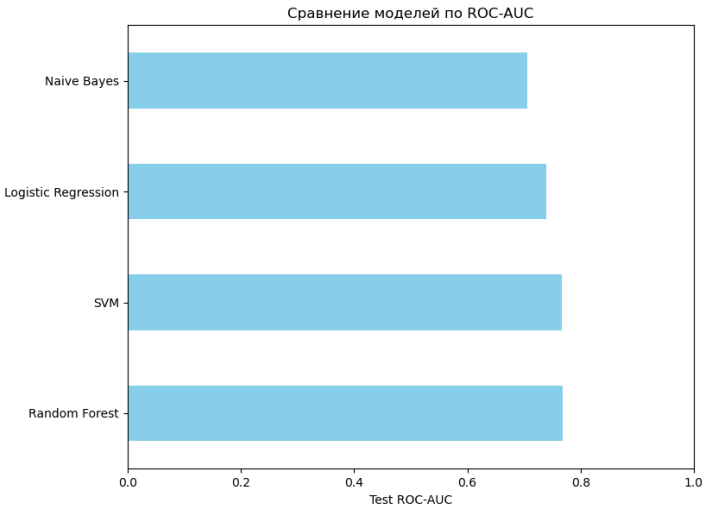


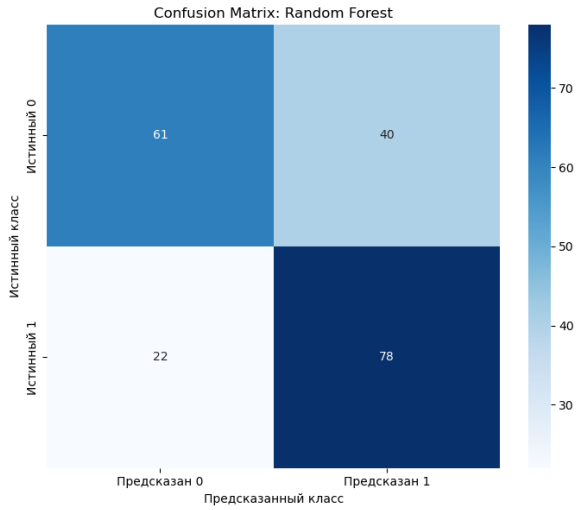


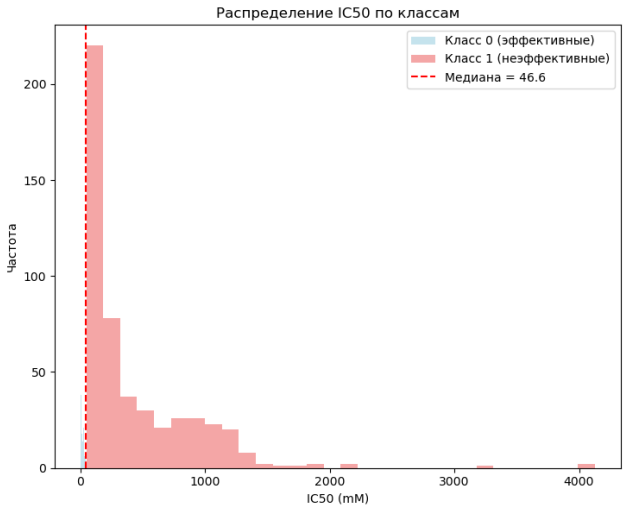


3) Задача регрессии SI

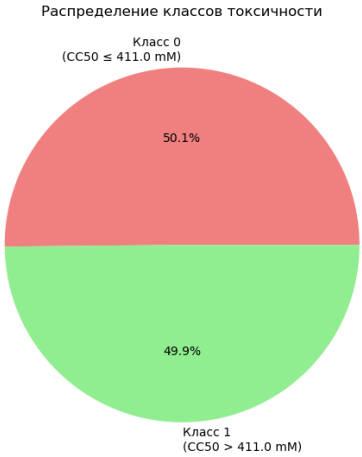


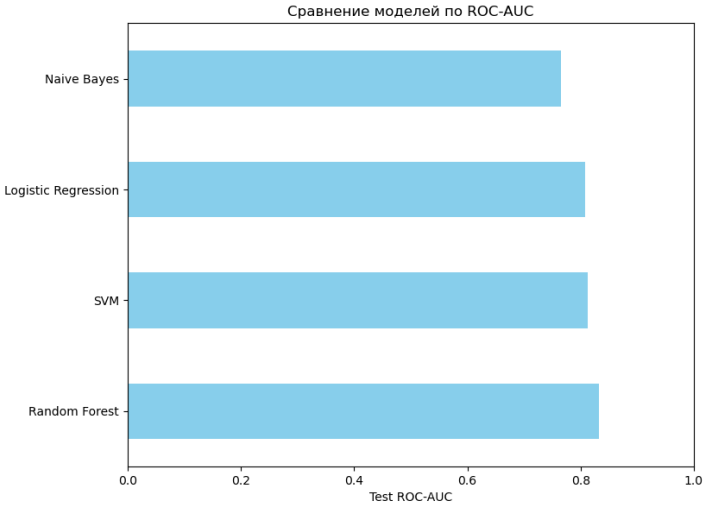


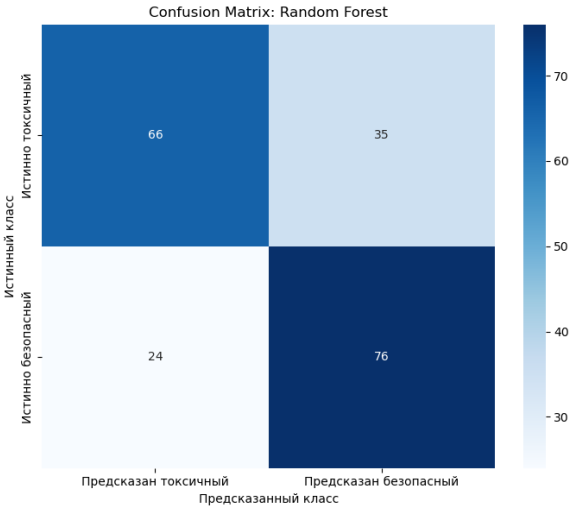


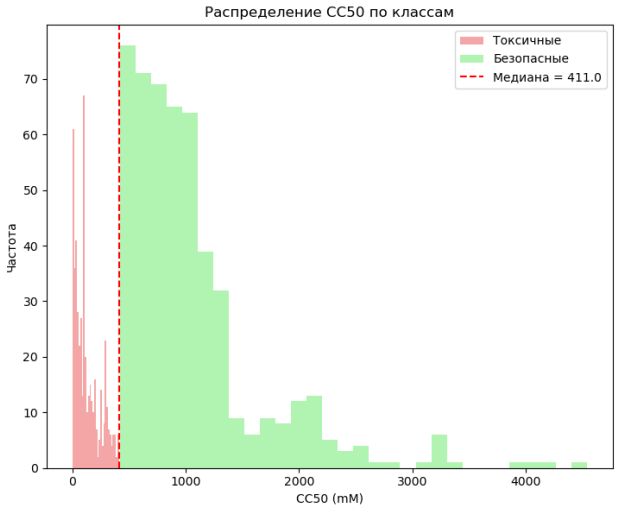


4) Задача классификации по медиане IC50

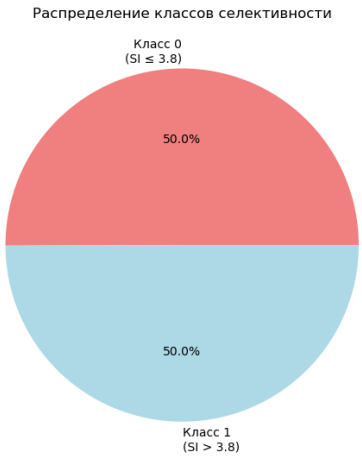


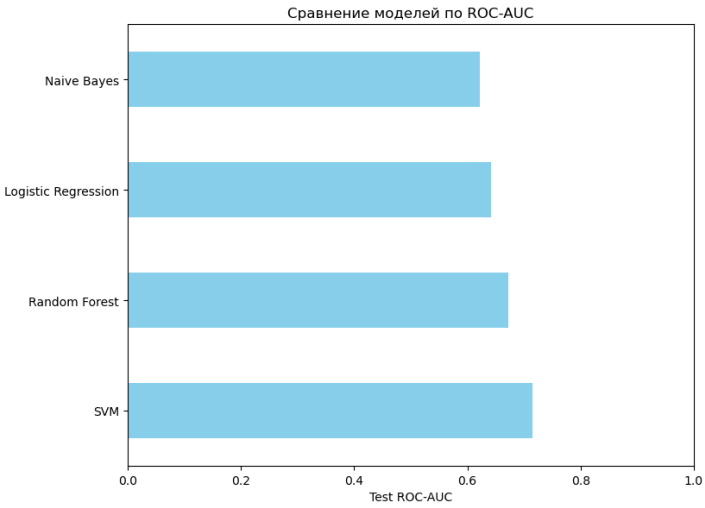


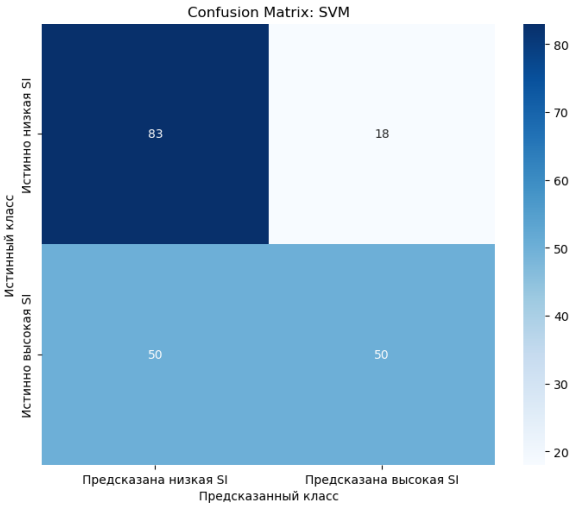


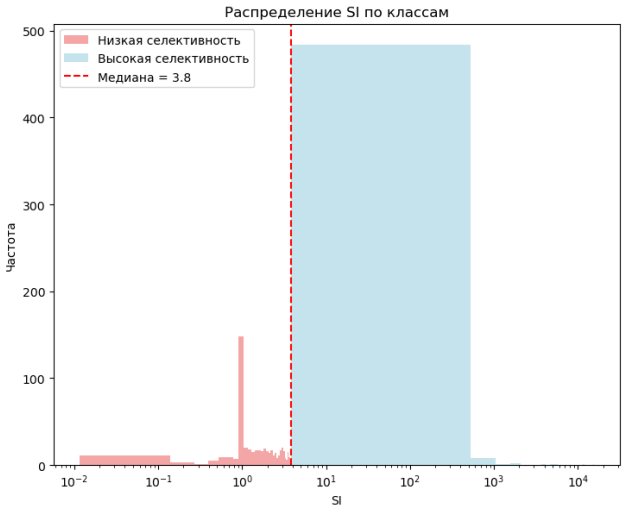


5) Задача классификации по медиане CC50

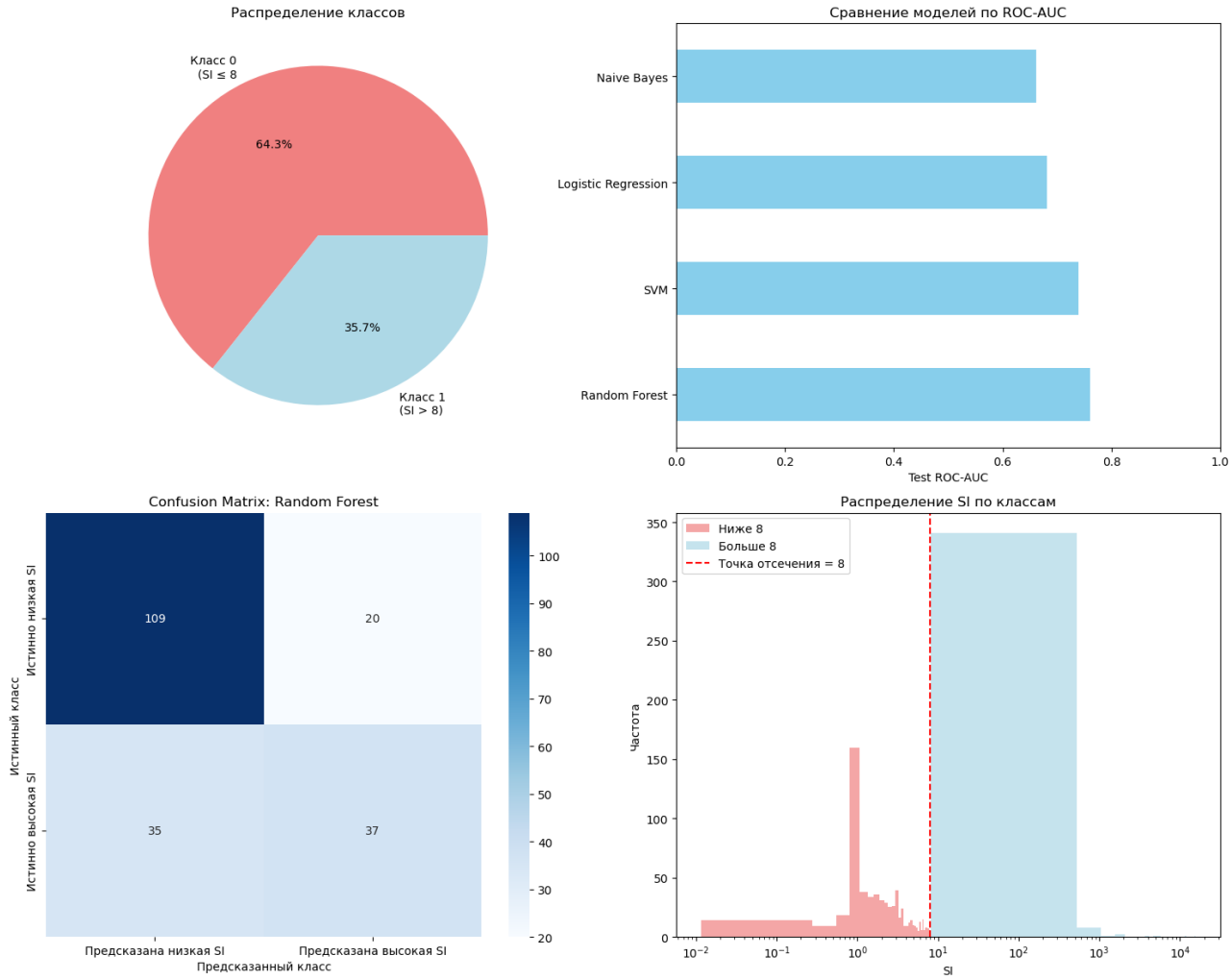








6) Задача классификации SI по медиане



7) Задача классификации SI по порогу в 8

6. Сравнительный анализ и интерпретация

Регрессия.

- Для IC50 и SI алгоритмы на деревьях (RF, Gradient Boosting) обошли линейную группу благодаря нелинейности и учёту взаимодействий.
- По CC50 SVR оказался лучшим: гладкая rbf-функция хорошо аппроксимировала умеренно сложную зависимость после `log`-трансформации.

Классификация.

- Random Forest доминировал при умеренном численном дисбалансе классов, обеспечив высокую ROC-AUC > 0.75.
- Для задачи «SI > медианы» важна граница, проходящая по редким «высоко-селективным» наблюдениям; SVM с rbf-ядром гибко описал эту тонкую границу.

8. Заключение

Поставленные семь задач решены, сравнительный анализ выполнен.

Модели-победители дают достаточное (для стадии *in silico*) качество прогноза и могут быть интегрированы в корпоративный конвейер виртуального скрининга, экономя лабораторное время и ресурсы. Основным ограничением остаётся вариативность SI; решение видится в расширении обучающей выборки.