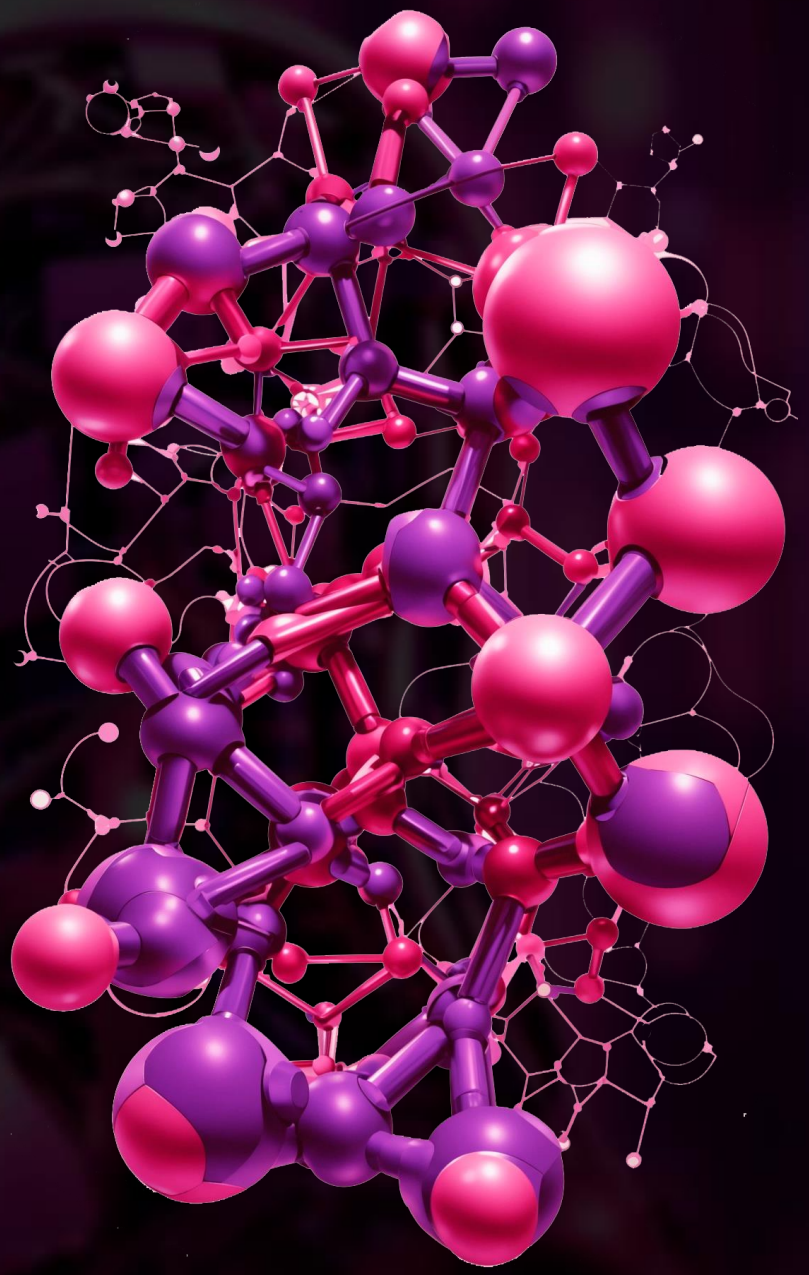
ATOMIC HACK

МИФИсты
ft.МИСИСт

Вступление

На основе данных о структурах химических соединений необходимо предсказать показатели CC50, IC50 и SI.

Это поможет определить является ли химическое соединение лекарством или нет, что позволит уменьшить время на биологические эксперименты.

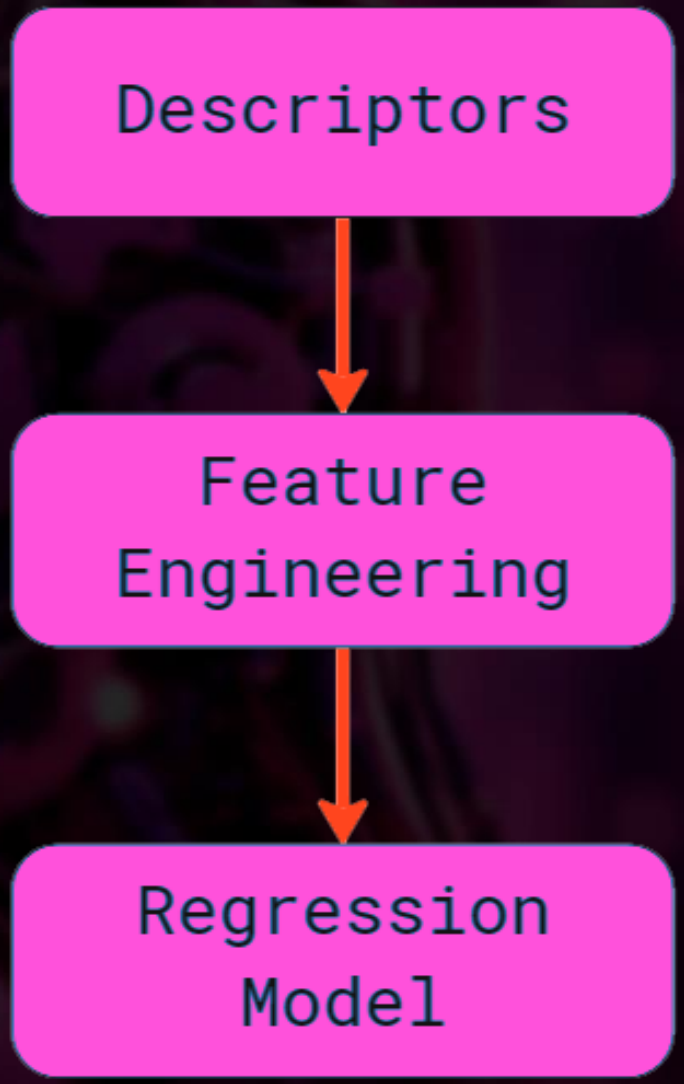


Вступление

Descriptors:
1. RDkit descriptors
2. Векторайзер
3. FingerPrints

Feature engineering:
1. Присвоить класс молекуле с помощью кластеризации
2. Попарно перемножить признаки
3. Отбросить малозначимые признаки

Regression model:
1. CatBoost
2. LGBM



Анализ датасета - удаление дубликатов

1. 1400 записей
   - CC50
   - IC50
   - SI
2. 35 тыс. записей
   - IC50

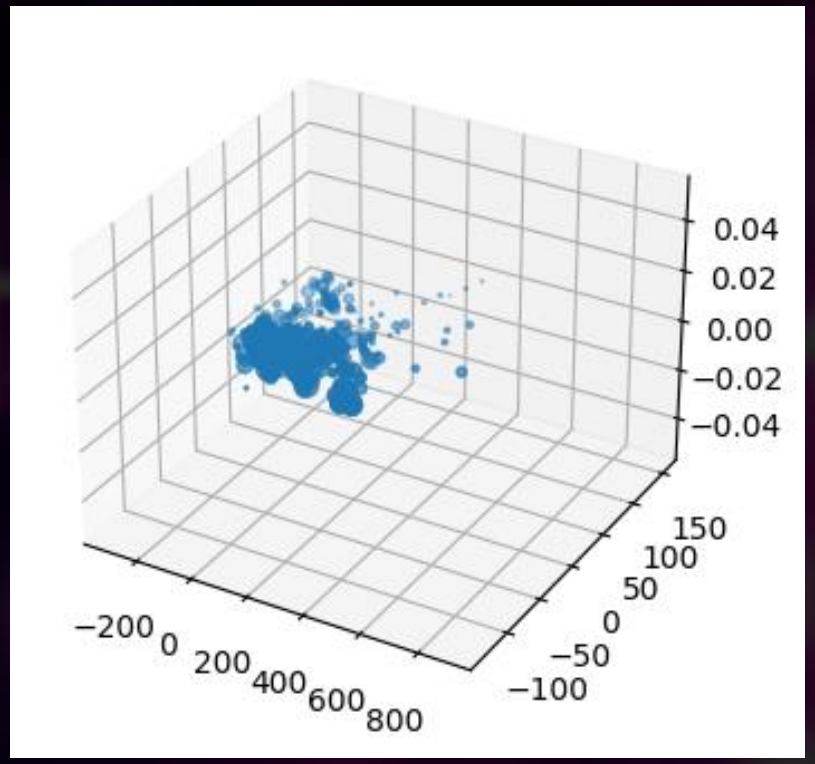
Проблемы

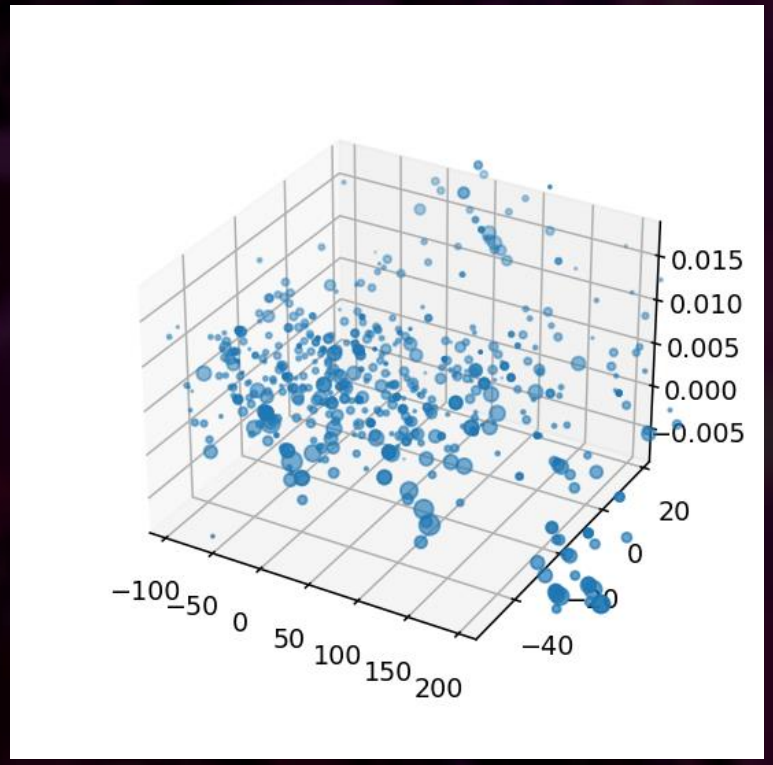
- Разнородные датасеты
- Наличие выбросов и дубликатов

Решение

- Работали по отдельности с датасетами
- Обработали дубликаты
- Удалили выбросы

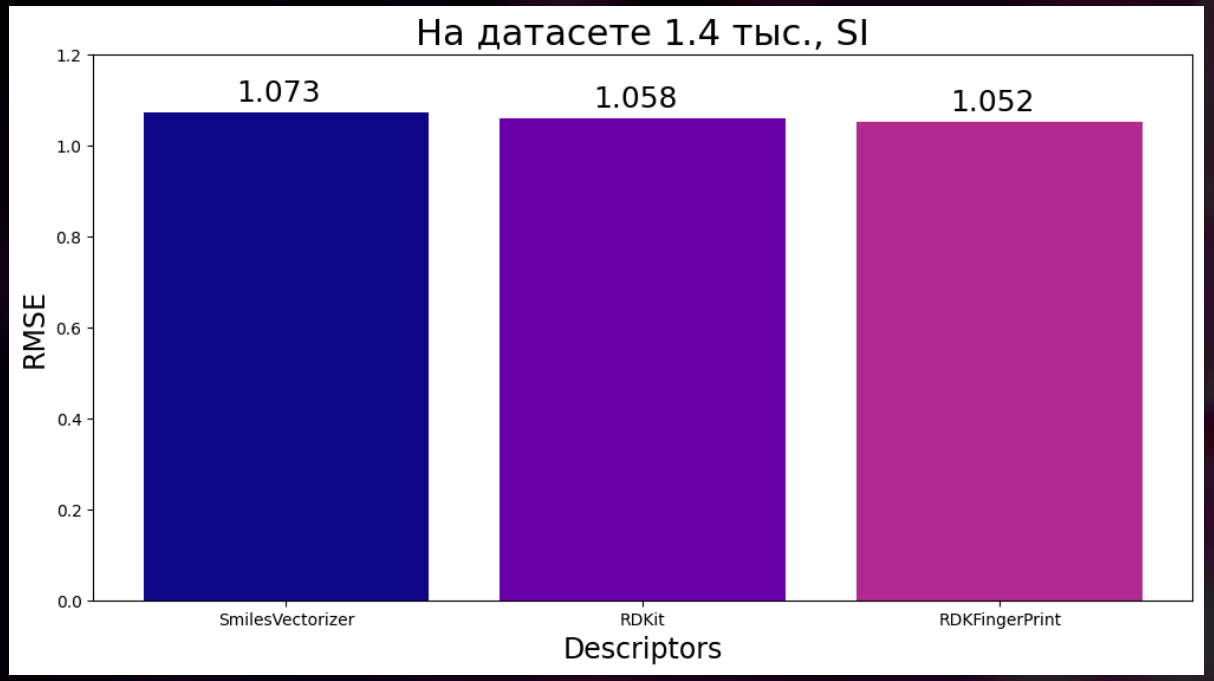
Кластеризация



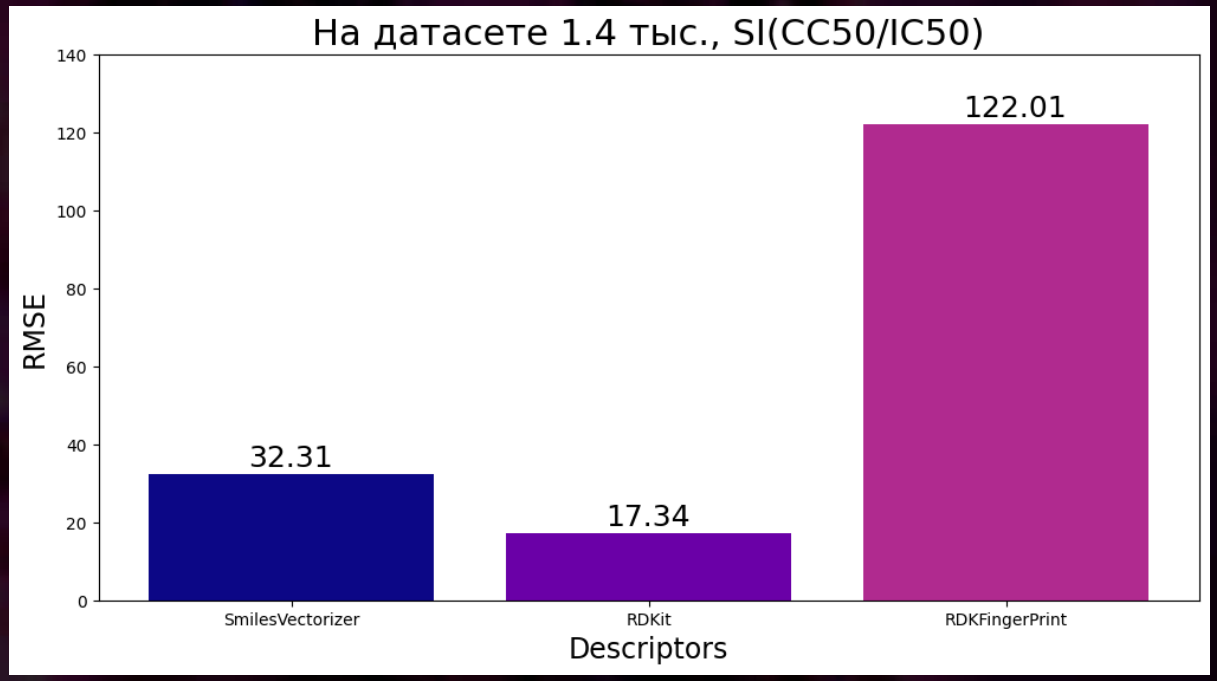


Итог: высокая плотность – не можем кластеризовать

Датасет 1400(SI)



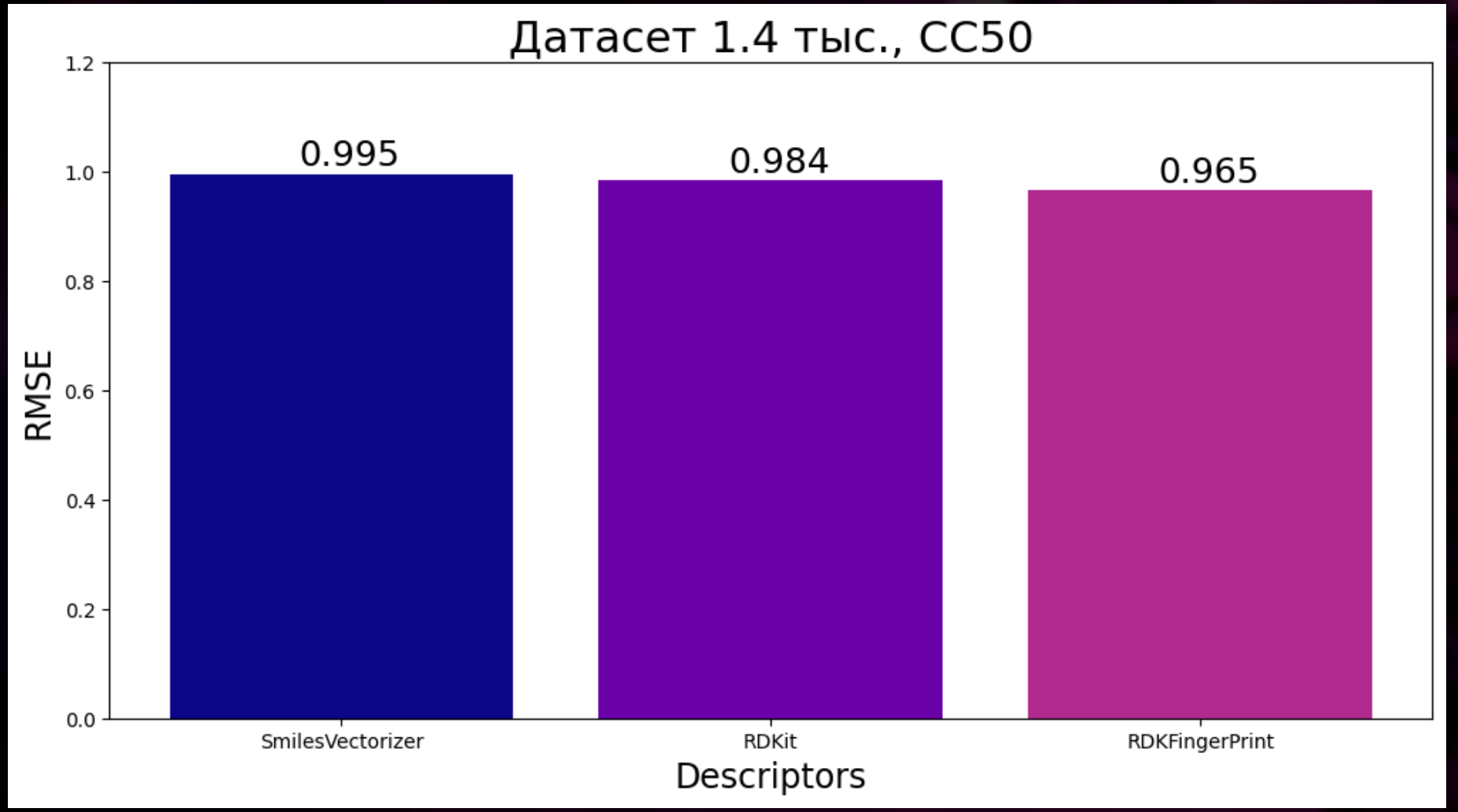
RMSE получен на кросс-валидации



RMSE получен на тестовой выборке

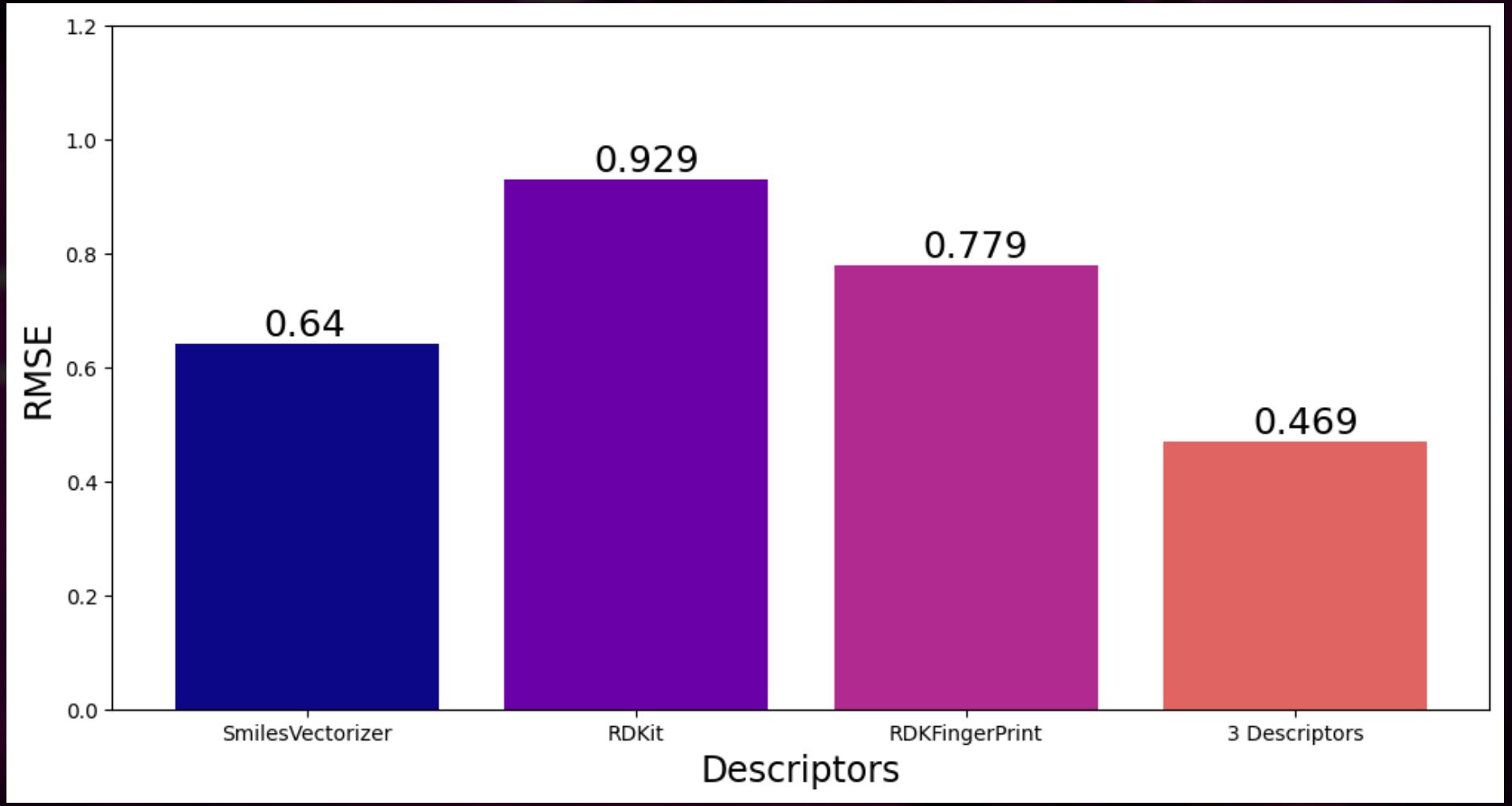
Regression Model: **Catboost**

Датасет 1400(CC50)



Regression Model: **Catboost**
RMSE получен на кросс-валидации

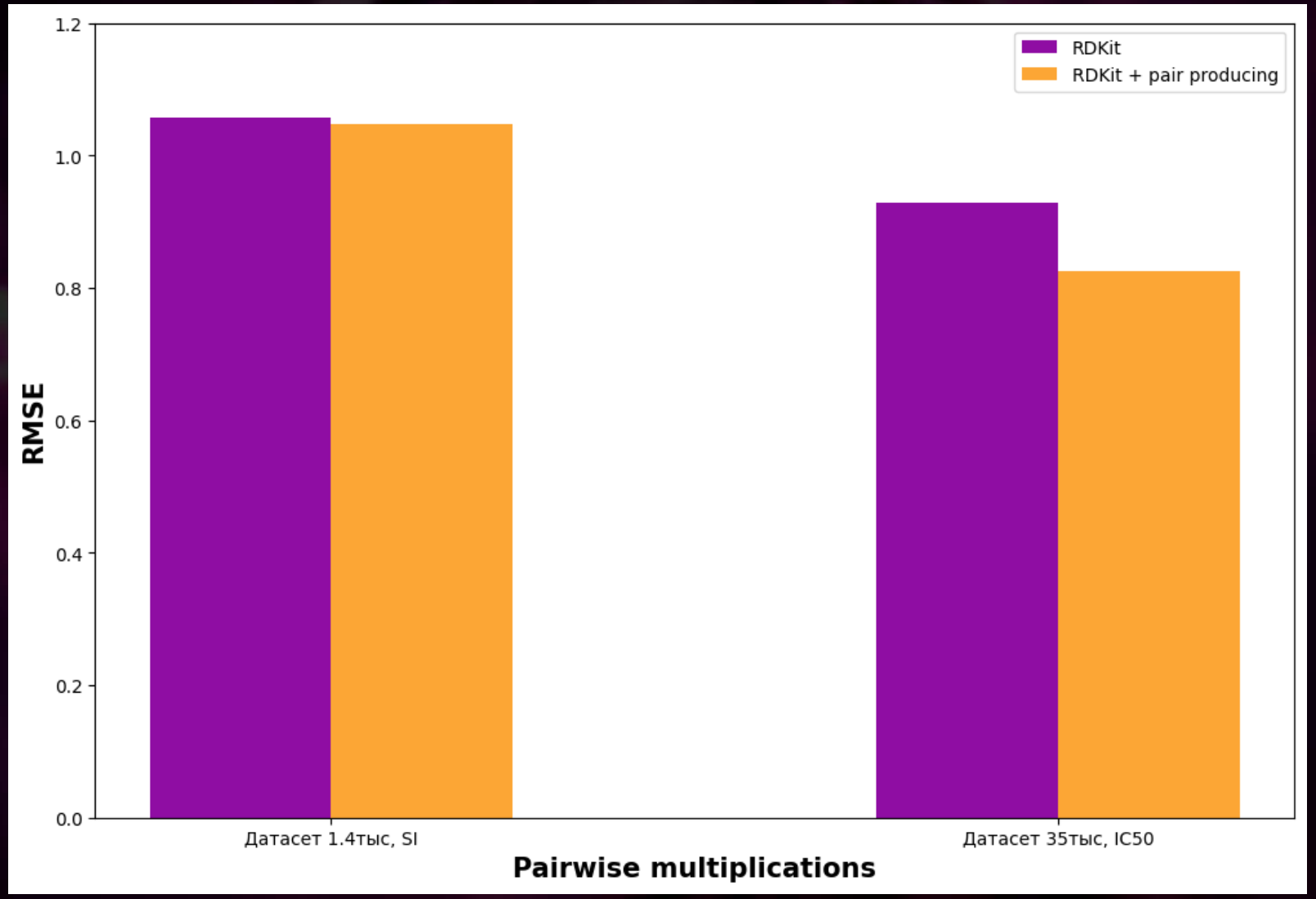
Датасет 35000(IC50)



Regression Model:
Catboost

RMSE получен на кросс-валидации

Pairwise multiplications



Regression Model: **Catboost**

RMSE получен на кросс-валидации

Выводы

Final RMSE:

- CC50 – 103.11
- IC50 – 84.57
- SI – 16.82

Final RMSE(standartized):

- CC50 – 0.72,
- IC50 – 0.876
- SI – 0.93

- Попарное перемножение улучшает показатели
- Использование совокупности дескрипторов даёт прирост в качестве

Дальнейшие планы

- GNN в качестве дескриптора
- Усреднение IC50 на двух датасетов
- Подбор гиперпараметров для Regression Model
- Создание ансамбля моделей