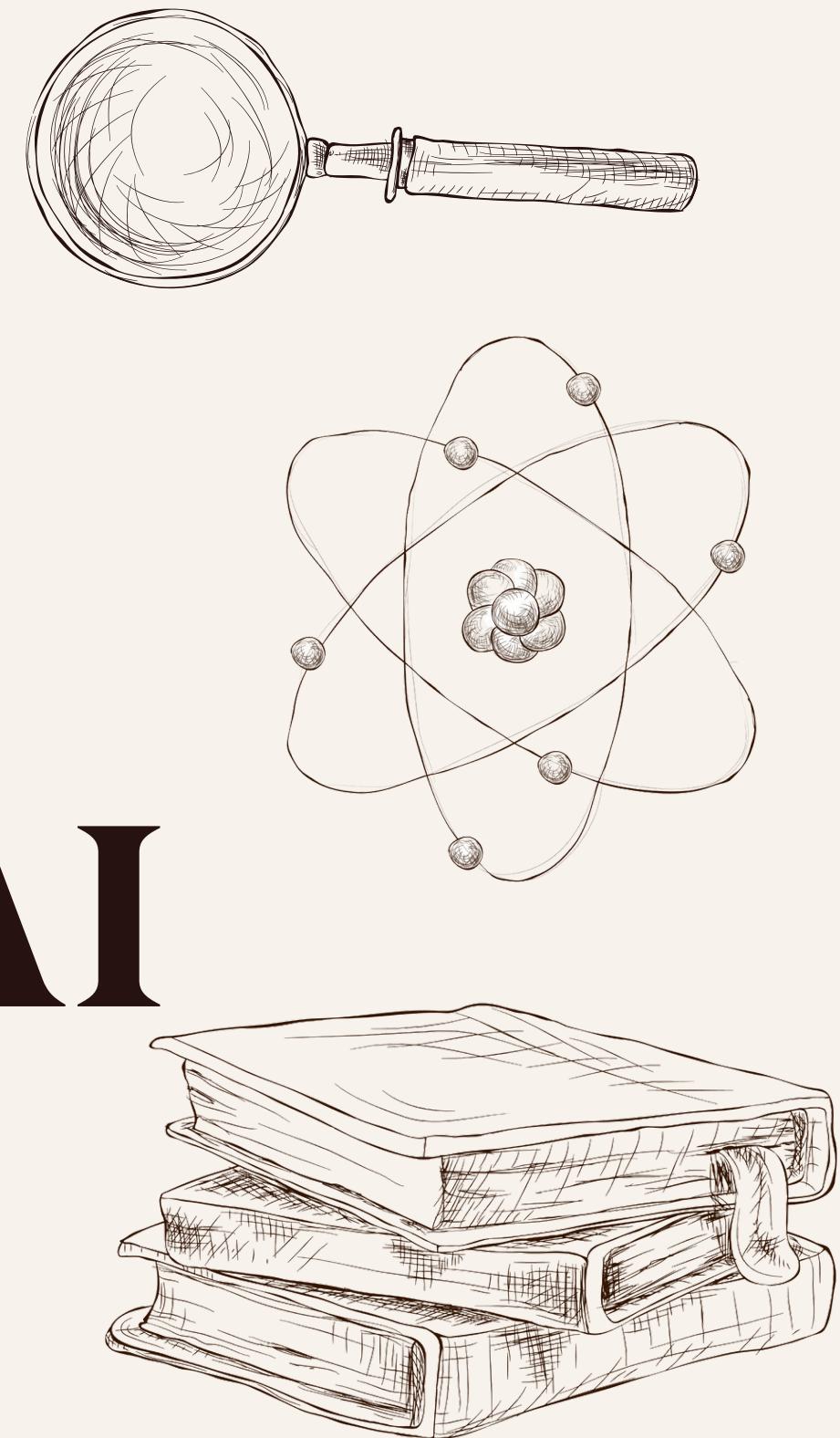
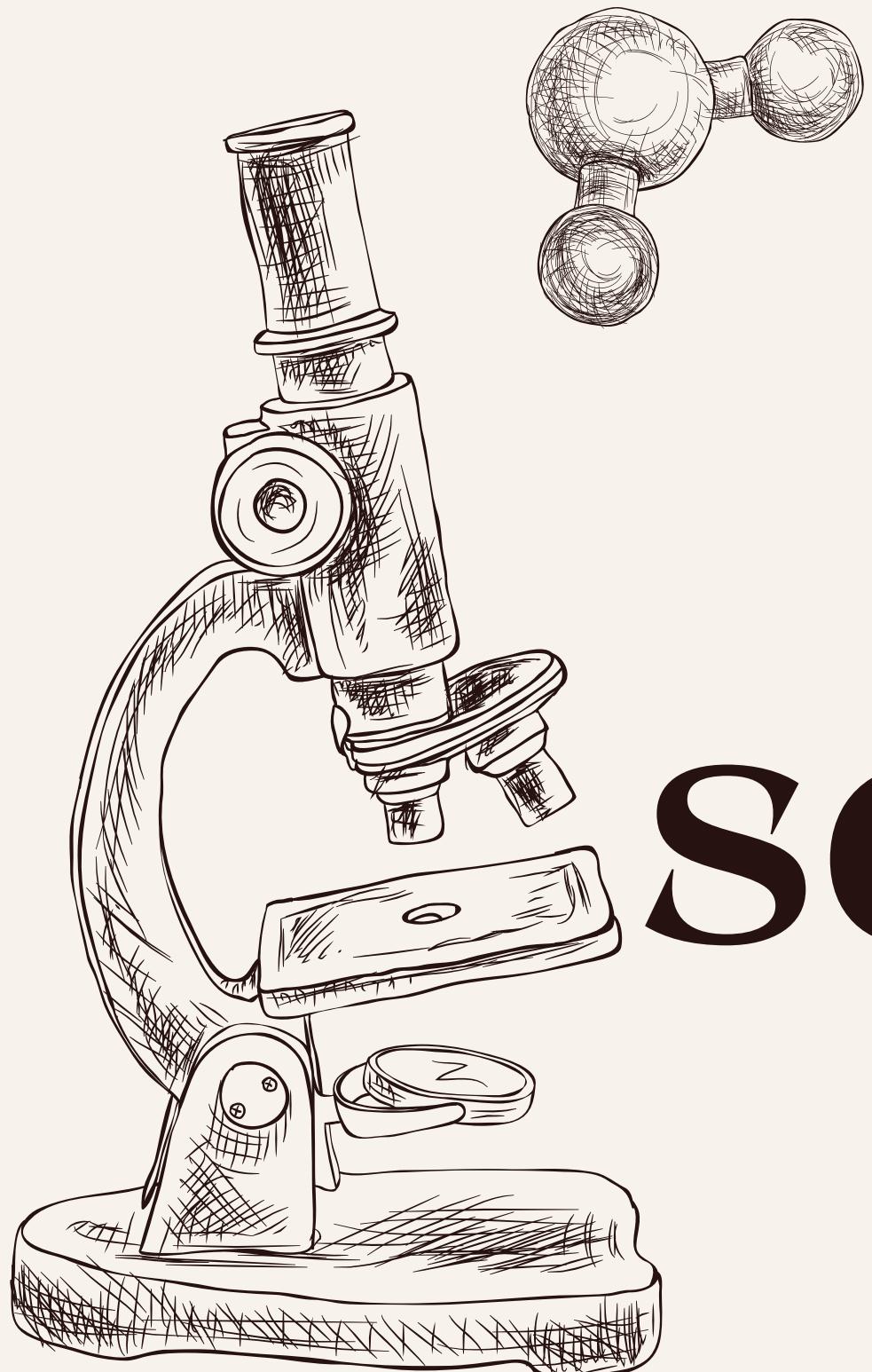


VIRUSES SCANNER AI

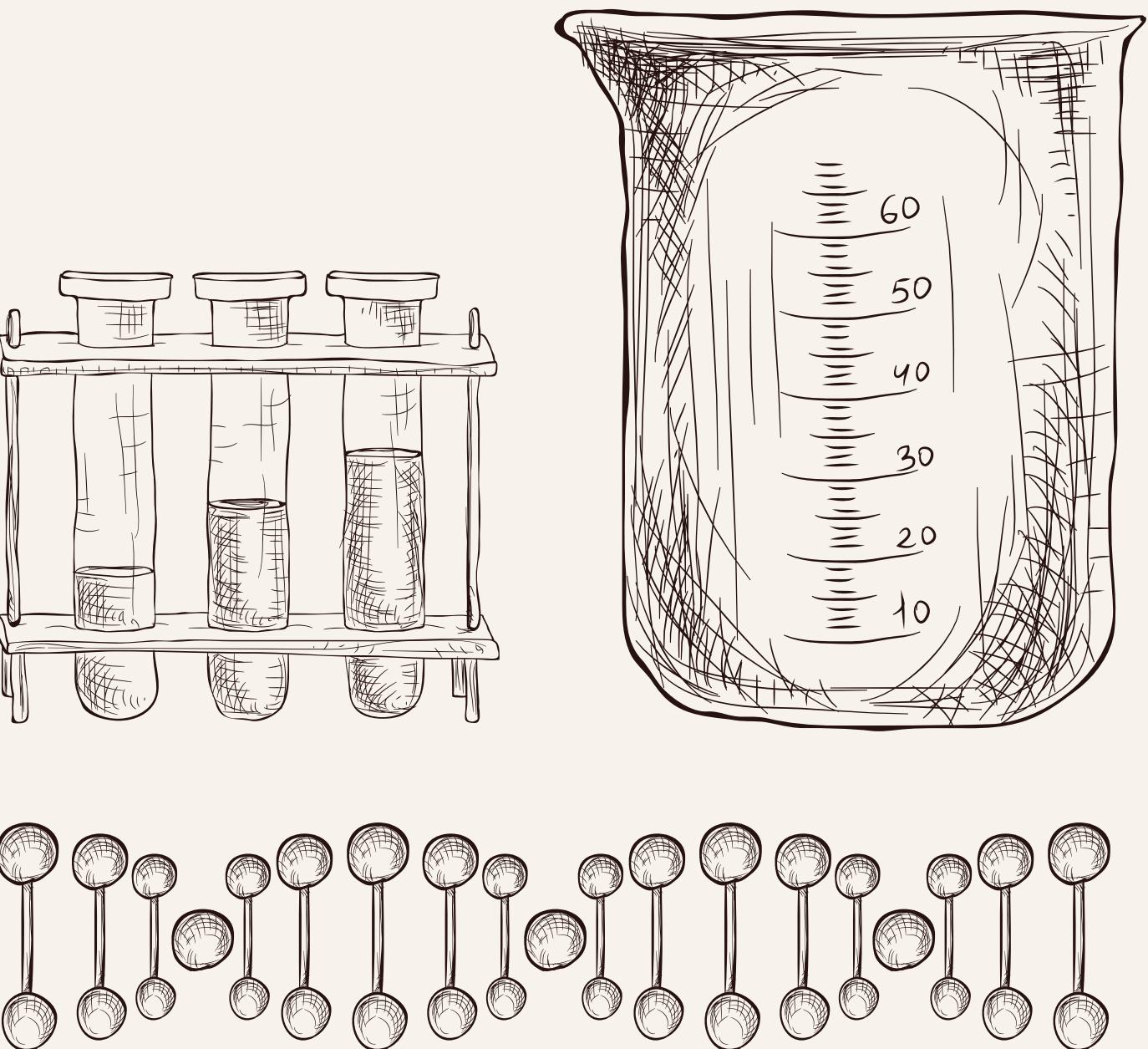


Подготовлено:
Блувштейн Елизавета
Сафонов Илья
Кулаков Дмитрий

ЦЕЛЬ ПРОЕКТА

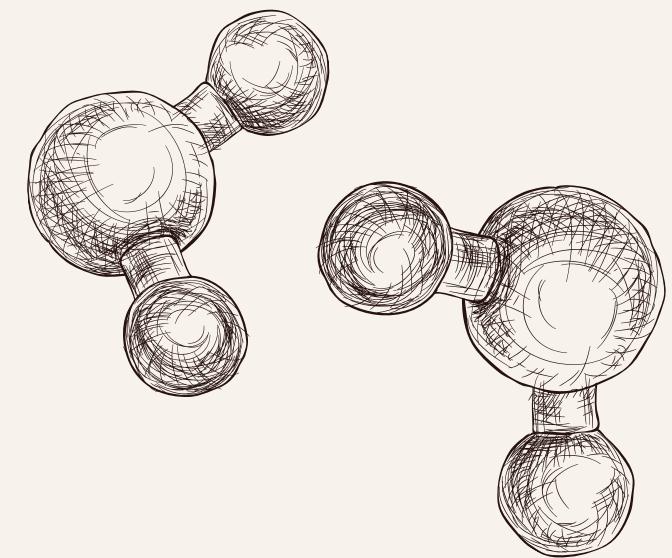
Целью проекта является помочь людям посредством создания модели машинного обучения, способной:

- Предсказывать противовирусную активность молекул вириуса SARS-CoV-2 .
- Снизить затраты и время, необходимые для проведения экспериментов.
- Повысить эффективность поиска новых противовирусных препаратов.





Этапы работы



01

02

03

04

05

Сбор и подготовка данных о химических соединениях и их противовирусной активности

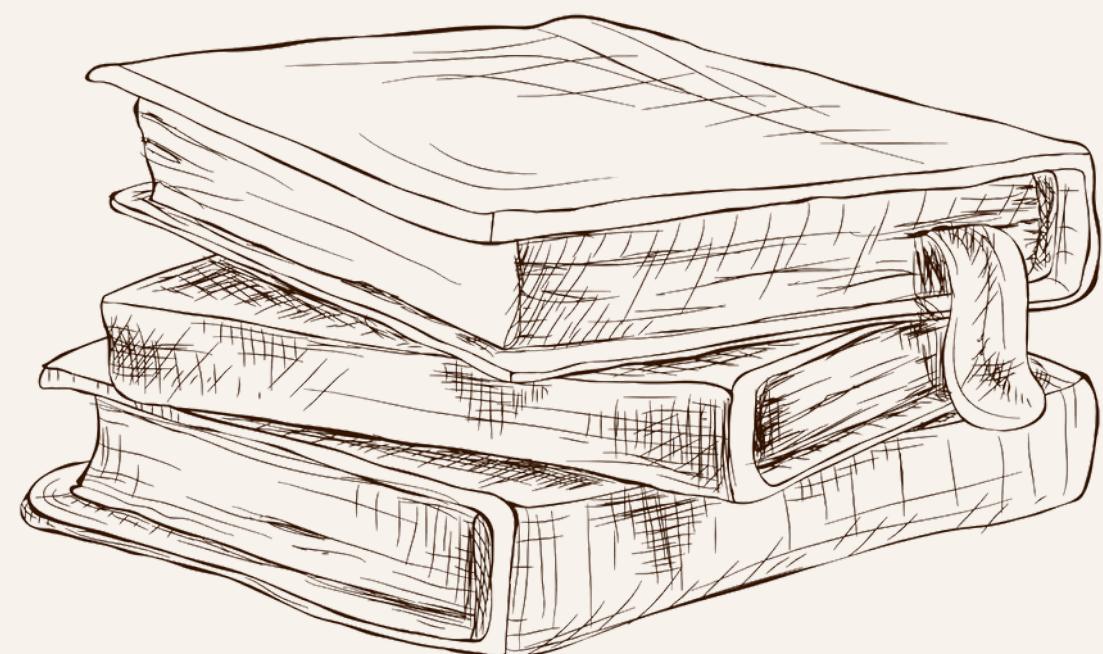
Преобразование данных в формат, подходящий для модели машинного обучения

Обучение разных моделей машинного обучения на собранных данных

Оценка точности разных моделей, выбор наилучшей

Разработка интерфейса для взаимодействия с моделью

ПРОБЛЕМАТИКА



1. Нет существующих решений
2. Катастрофически мало данных с селективным индексом
3. Необходимы глубокие знания в области биохимии

Используемые Технологии

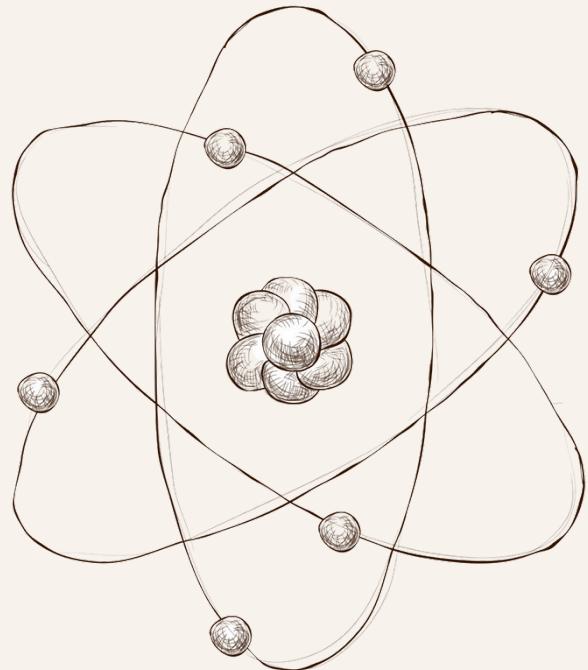
Машинное обучение:

- Модель XGBoost
- Библиотеки: sklearn, tensorflow, rdkit

Обработка данных:

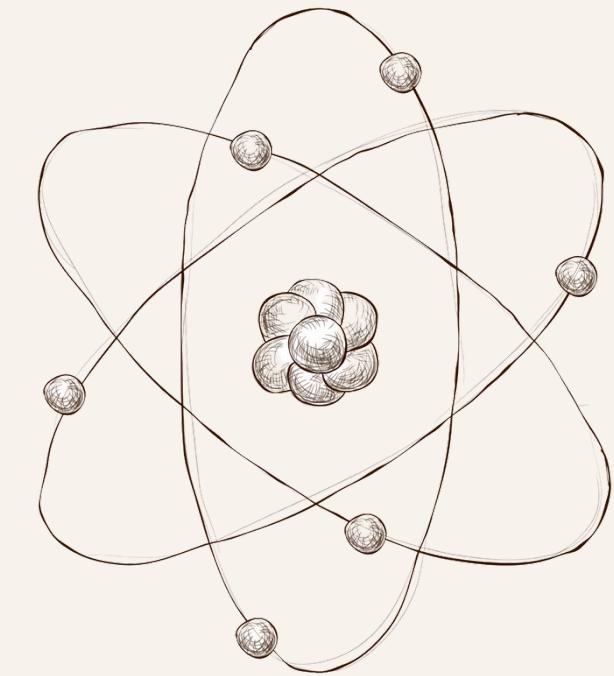
- Данные о молекулярных соединениях и их противовирусной активности были собраны из базы данных ChEMBL

Сбор и обработка данных:



1. Извлечение признаков из SMILES-строк происходило через генерацию Morgan Fingerprints и набора двухмерных дескрипторов, описывающих различные физико-химические свойства молекулы. Это дало около 2 тысяч признаков

2. С помощью масштабирования, удаления признаков с низкой дисперсией и сильно коррелирующих признаков и модели РСА было снижено их количество
3. Для решения проблемы широкого разброса целевой переменной была применена логарифмическая функция



Обучение и оценка модели

01

Было решено создать 2 модели которые будут предсказывать параметр CC50 и IC50 информации о которых было получено из исследований на клетках Vero E6

02

Для предсказаний параметра CC50 была выбрана XGBoost с точностью:

MSE = 2.9341

MAE = 1.1357

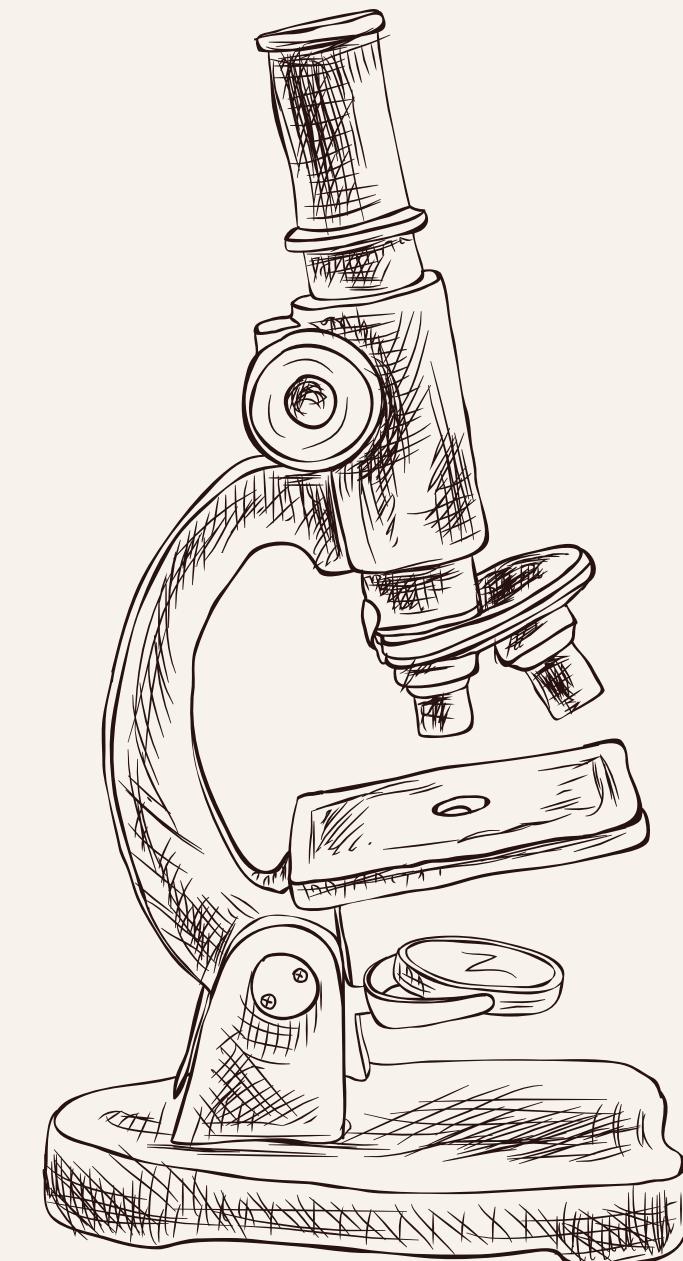
для логарифмированной целевой переменной

Для предсказаний параметра IC50 была выбрана XGBoost с точностью:

MSE = 2.1597

MAE = 0.749

для логарифмированной целевой переменной



SI = CC50/IC50

Интерфейс

Простой интерфейс, позволяющий легко проводить исследования

VirusesScannerAI

SMILE:

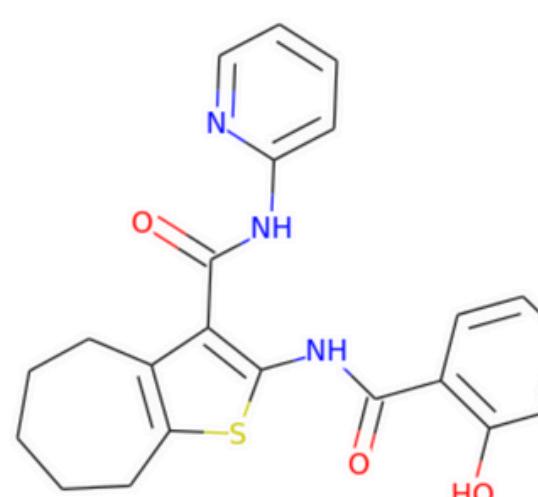
Ctrl+V take picture

load from file

C1[C@@H](O)CC(C)C[C@@]1(C)CNc(nc2)nc(c23)cccc3
C1[C@@H](O)CC(C)C[C@@]1(C)CNc(n2)sc(c23)cccc3
C1[C@@H](O)CC(C)C[C@@]1(C)CNc(n2)sc(c23)cc(F)cc3
C1[C@@H](O)CC(C)C[C@@]1(C)CNc(n2)sc(c23)cc(Cl)cc3
C1[C@@H](O)CC(C)C[C@@]1(C)CNc2ccccn2
FC(F)F)c1ccc(nc1)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
FC(F)F)c1cccc(n1)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cccc2
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2c(Cl)cccc2
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(Cl)cc2
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2ccc(Cl)cc2
c1cccc(c1C(F)F)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
FC(F)F)c1cc(ccc1)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
FC(F)F)c1cc(cc(Cl)c1)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(F)cc(Cl)c2
FC(F)F)c1c(Cl)ccc(c1)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(ccc2)Oc3cccc3
NC(=O)c1cc(ccc1)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(S(=O)(=O)C)cc2
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(S(=O)(=O)N)cc2
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(F)cc(c2)S(=O)(=O)N
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc2cc(S(=O)(=O)N)cc(Cl)c2
NS(=O)(=O)c1cc(cc1C(F)F)NC[C@]2(C)CC(C)(C)C[C@H](C2)O
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc(cc2)cc(c2)S(=O)(=O)N
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc(cc2)cc(S(=O)(=O)N)c2OC
C1[C@@H](O)CC(C)(C)C[C@@]1(C)CNc(c2)ccc(Cl)c2S(=O)(=O)N
C1[C@@H](O)CC(C)(C)C[C@H](CC2)[C@@]1(C)C[C@H](CC3)[C@]2(C)C[C@](C[C@H](O)CC(C)(C)C[C@H](CC2)[C@@]1(C)C[C@H](CC3)[C@]2(C)C[C@](n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(F)cccc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4cc(Cl)cc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(OC)cccc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(O)cccc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(O)cccc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(O)cccc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4cc(O)ccc4
n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4ccc(O)cc4
c1cc(Cl)ccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(OC)cccc4
c1cc(Cl)ccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(O)cccc4
c1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(O)cccc4

Make prediction All Download

n1cccc1NC(=O)c2c(sc(c23)CCCCC3)NC(=O)c4c(O)cccc4



Будущее проекта

- Расширить набор данных, включив информацию о новых соединениях
- Перенести сервис VirusesScannerAI в облако на отечественные сервера
- Исследовать применение технологии для других научных областей



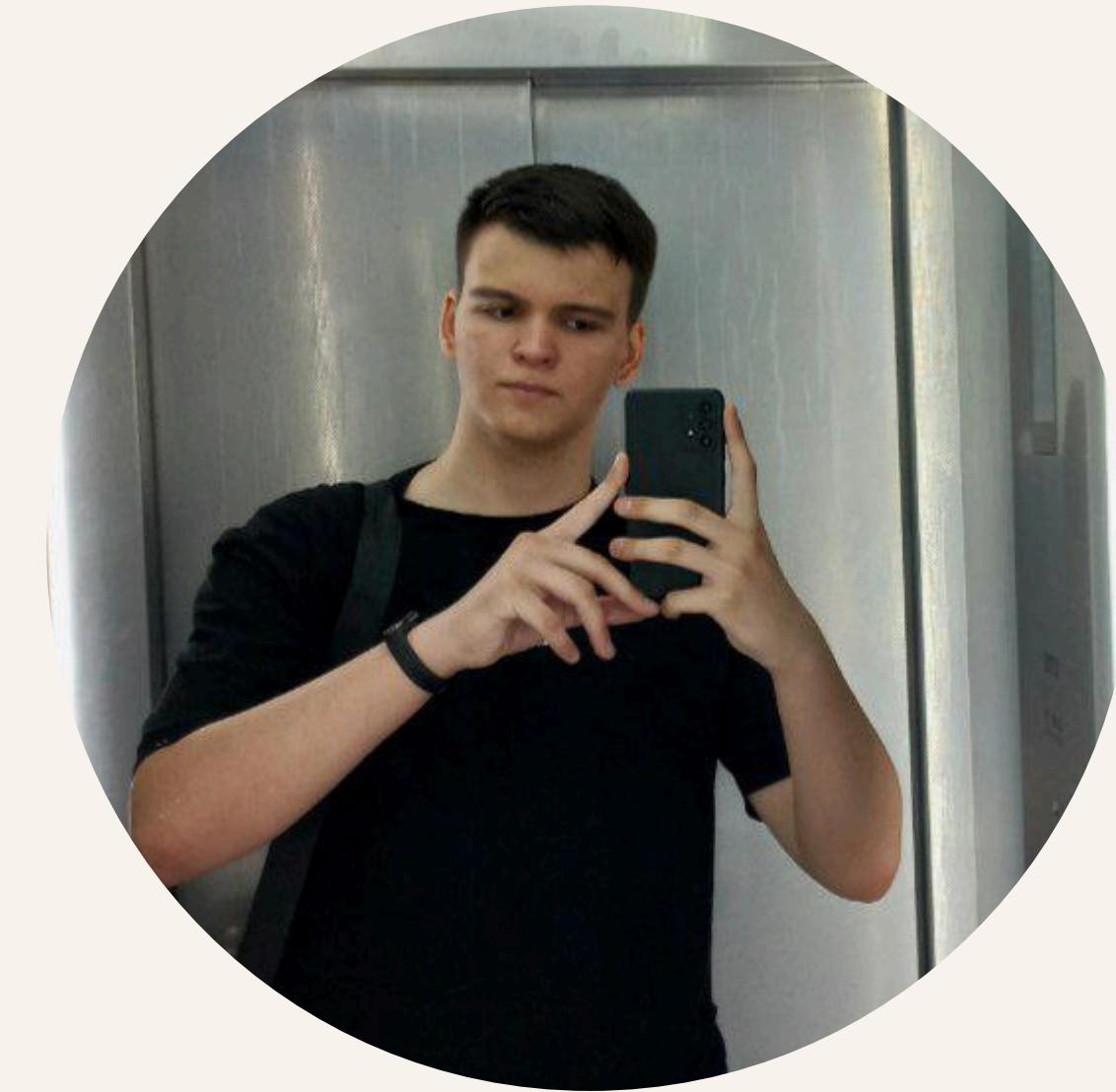
Команда



Кулаков Дмитрий
[Анализ данных]
@Dmiiy



Блувштейн Елизавета
[Машинное обучение]
@izzkabl



Сафонов Илья
[Разработка окружения]
@Resk_QuiT

Готовы ответить на ваши вопросы