

# Инструкция пользователя: Интерфейс предсказаний для молекулярных данных

## 1. Общая информация

Приложение предназначено для предобработки данных о молекулах и создания предсказаний с использованием машинного обучения. Пользователь может загружать собственные датасеты, выбирать модели, предобрабатывать данные и визуализировать результаты.

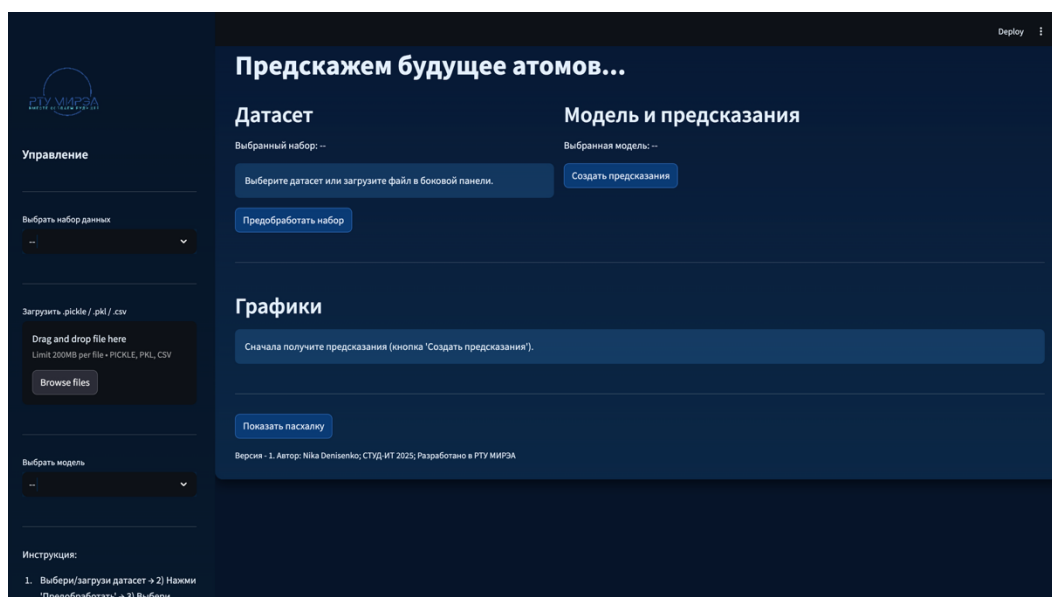
Версия приложения: **1.0**

Автор: **Вероника Денисенко, СТУД-ИТ 2025**

Разработка: **PTU MIRZA**

## 2. Запуск и интерфейс

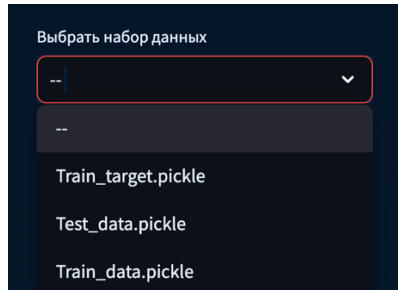
- Запустите приложение Streamlit. `streamlit run app streamlit.py`
- Интерфейс разделён на две основные колонки:
  - Левая колонка:** Работа с датасетом
  - Правая колонка:** Выбор модели и создание предсказаний



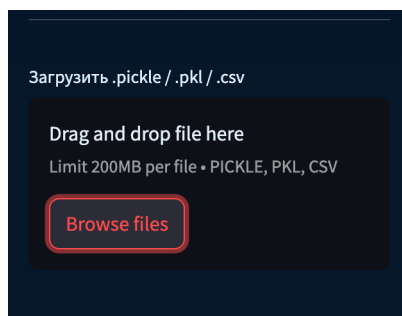
## 3. Работа с датасетом

### 3.1 Выбор или загрузка данных

- В боковой панели выберите один из доступных локальных файлов (.csv, .pickle, .pkl) в списке **"Выбрать набор данных"**.



- Для загрузки собственного файла используйте **"Загрузить .pickle / .pkl / .csv"**.



- После загрузки файл автоматически добавляется в сессию.

### 3.2 Просмотр данных

- После выбора датасета в левой колонке отображается предварительный просмотр первых строк.
- Если в колонках содержатся сложные структуры (списки, массивы), для них показываются укороченные примеры.
- Разверните вкладки с вложенными колонками для просмотра детализированных данных.

### Датасет

Выбранный набор: Train\_data.pickle

id	positions
289,436	"[[2.7177422 0.18937249 1.18467343]\n [1.03808224 -0.4019455 1.065485
172,291	"[[-2.80976963 0.42855179 -0.96908551]\n [-2.6878438 0.6895228 -2.49647
16,145	"[[-0.68493009 1.16981983 0.74941361]\n [-1.05636311 -0.23196419 0.4479
12,345	"[[4.15824986 0.05714491 -1.70589423]\n [4.95895386 0.6903069 -2.43322
45,824	"[[-0.47792548 0.71881342 -1.29738855]\n [0.43063852 -0.3190276 -0.96598

Вложенные колонки (примеры)

positions — показать примеры

elements — показать примеры

0: "[L'i 'F' 'P' 'F' 'F' 'F' 'F']"

1: "[C' 'C' 'O' 'C' 'O' 'C' 'O' 'H' 'H' 'H' 'H' 'H' 'H' 'C' 'C' 'O' 'C' 'O' 'C' 'O' 'H' 'H' 'H' 'H' 'H']"

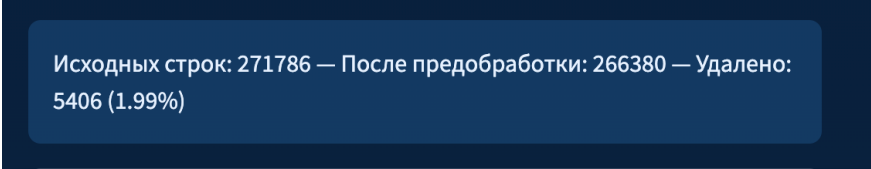
2: "[C' 'O' 'C' 'O' 'C' 'H' 'H' 'O' 'H' 'H']"

3: "[C' 'O' 'O' 'O' 'C' 'C' 'H' 'H' 'H' 'H' 'H' 'H' 'C' 'O' 'O' 'O' 'C' 'C' 'C' 'H' 'H' 'H' 'H' 'H']"

4: "[C' 'O' 'C' 'O' 'C' 'H' 'H' 'O' 'H' 'H']"

### 3.3 Предобработка данных

1. Нажмите кнопку **"Предобработать набор"**.
2. Приложение выполнит:
  - Очистку данных (удаление строк без координат, с выбросами энергии или заряда).
  - Генерацию новых признаков (количество атомов, связи, расстояния, градиенты и дипольные моменты).
  - Удаление исходных колонок с массивами.
3. После завершения предобработки появится информация:
  - Количество исходных строк
  - Количество удалённых строк
  - Список признаков после предобработки
4. Если строки были удалены, их можно просмотреть в отдельной вкладке **"Показать примеры удалённых строк"**.



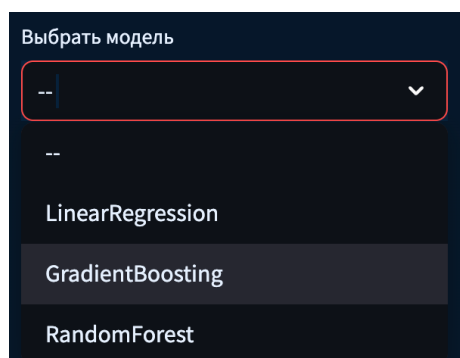
Исходных строк: 271786 — После предобработки: 266380 — Удалено: 5406 (1.99%)

---

## 4. Работа с моделями

### 4.1 Выбор модели

- В боковой панели выберите модель из списка:
  - Загруженные модели с диска (.pkl)
  - Встроенные модели: LinearRegression, RandomForest, GradientBoosting



### 4.2 Создание предсказаний

1. Убедитесь, что выбран предобработанный датасет и модель.
2. Нажмите **"Создать предсказания"**.
3. Результаты сохраняются в сессии и отображаются в правой колонке:
  - Таблица предсказаний
  - Возможность скачать CSV-файл с результатами

- Просмотр конкретного предсказания по выбранному индексу с указанием примерной неопределённости

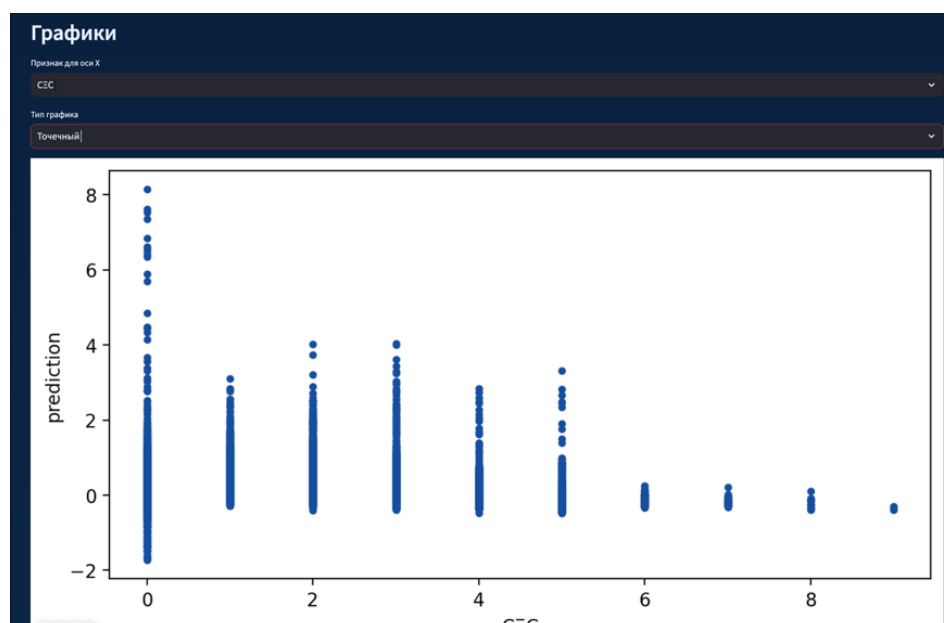
**Таблица предсказаний**

id	prediction
289,436	-0.0033
172,291	-0.1041
16,145	-0.014
12,345	-0.0112
45,824	-0.0183
91,156	-0.0209
159,539	-0.1296
56,089	-0.0356
145,464	-0.0512
313,427	0.5745

Скачать предсказания (CSV)

## 5. Визуализация предсказаний

1. В разделе "Графики" выберите:
  - Признак для оси X (index или любой числовой признак после предобработки)
  - Тип графика: **Линейный**, **Точечный**, **Гистограмма**
2. Приложение автоматически фильтрует некорректные значения (NaN, Inf).
3. Графики отображаются непосредственно в интерфейсе.



## 6. Дополнительно

- **Информация об авторе:** кнопка "Об авторе" в боковой панели
  - **Пасхалка:** кнопка "Показать пасхалку" активирует визуальные эффекты
- 

## 7. Рекомендации

- Для больших датасетов рекомендуется использовать предобработку и визуализацию частями (head/фильтры), чтобы ускорить работу.
  - Перед созданием предсказаний убедитесь, что все данные корректно предобработаны.
  - Загружайте модели только из доверенных источников.
-