Предсказание молекулярных свойств

При помощи глубокого обучения

Больщиков Костя

Алиса Аленичева

Мама-Мама, это ты?

Морозов Антон

Пальченков Иван

Беляева Ульяна

Roadmap

Мама, я в Питере

Было больно и ничего непонятно, но потом всё стало +- Ок

Мама, я глупый

Ну, мы потыкали палкой MNIST - узнали, что не тонет

Мама, я умный

Стандардизировали SMILES/в столовой не оч вкусно

Roadmap

4 → 5 → 6

Мама, не работает!

В 13.37 LSTM выдал:
Training loss = 2.28

Мама, красиво!

Transformer и графики - ну, прикольно, че - порисовали

Мама, конец...

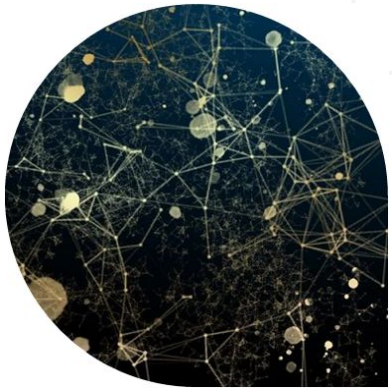
Ну че, погнали на защиту, получается

ЦЕЛЬ

Предсказать растворимость в воде и жирах

Проблемы

- Много молекул (в млн)
- Исследования каждой - очень дорого
- Мало баз данных с исследованиями



Применение

- Лекарства
- Воздействие загрязнителей на организмы

Краткое описание проекта

- **Baseline**
- **NLP**
- **Визуализация**

Данные

01

ESOL

Данные о растворимости в воде органических молекул

1k молекул

02

Lipophilicity

Данные о растворимости в жирах молекул

4k молекул

03

Chembl

Различные свойства молекул с низкой точностью

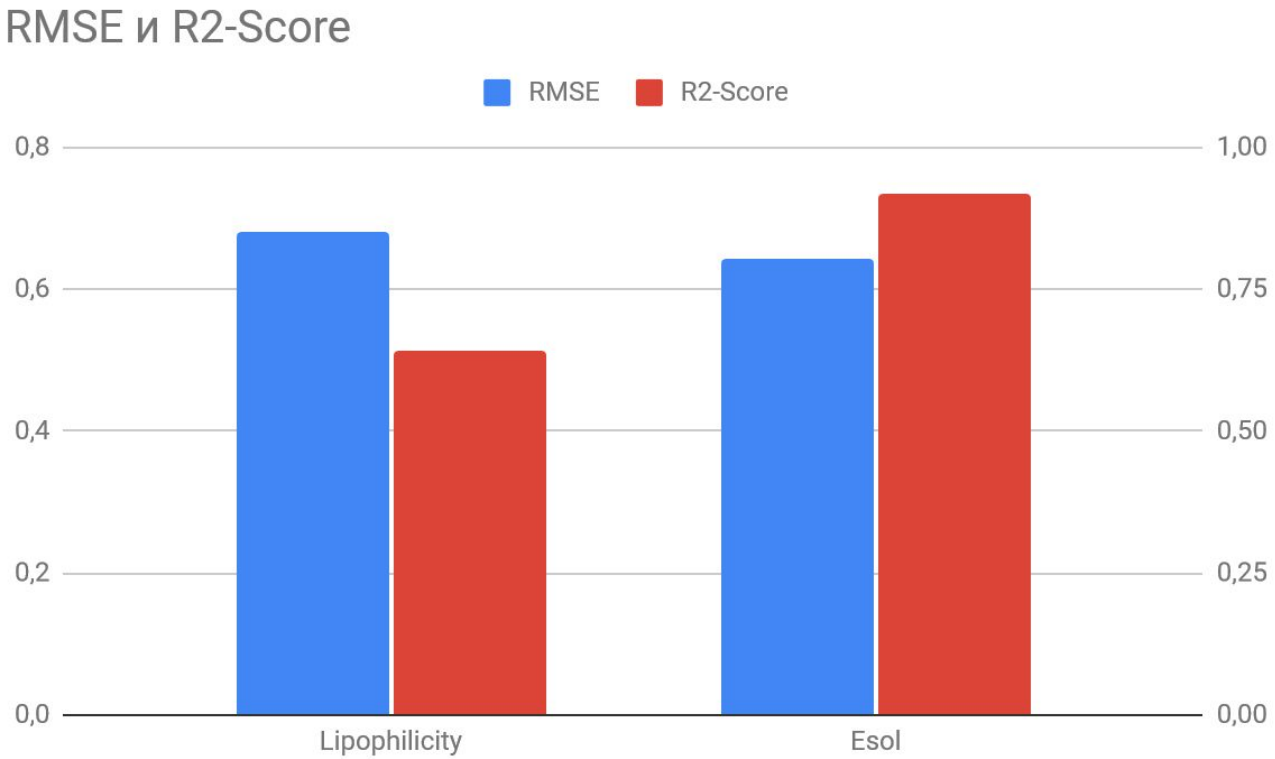
2M молекул

Модели

BaseLine

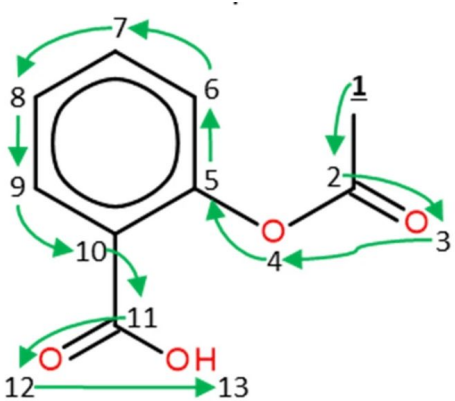
- Feature-engineering: считаем 200 свойств молекул
- Модель - XGBRegressor
- Поиск оптимальных гиперпараметров: GridSearchCV

BaseLine: результаты



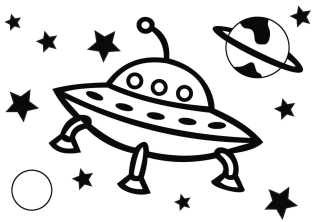
NLP модели

Smiles



CC(=O)Oc1ccccc1C(=O)O

SMILES - строчное представление молекул - работаем с NLP-моделью (НЛО?)



Предобработка данных

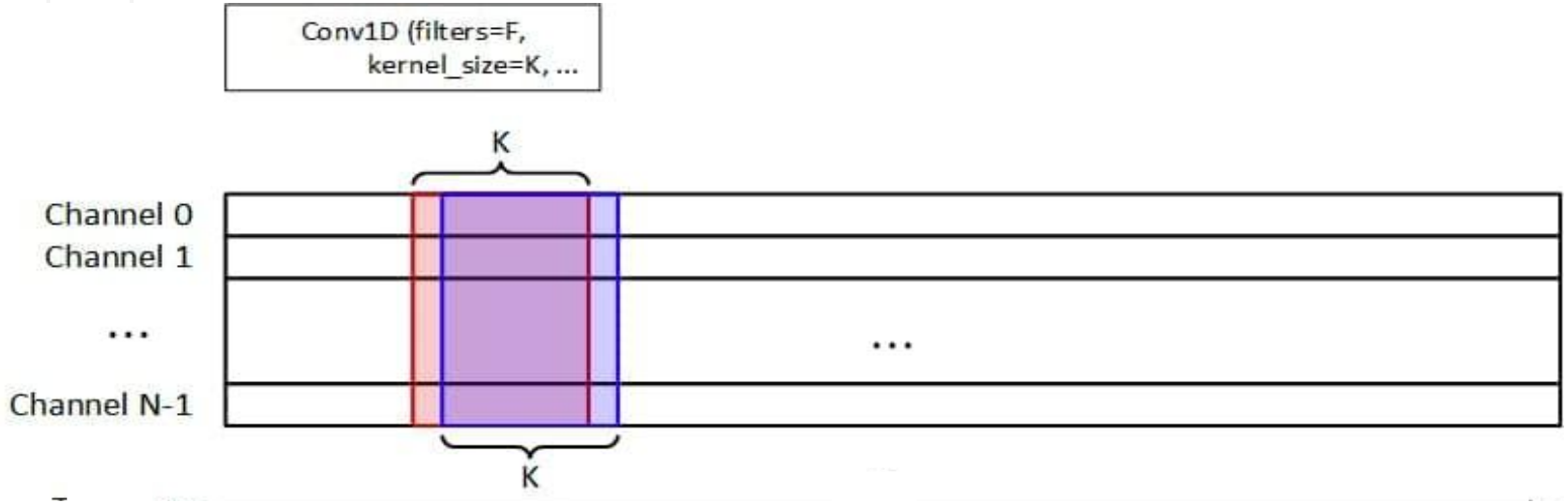
CClBr#I → C Cl Br # I

Токенизация

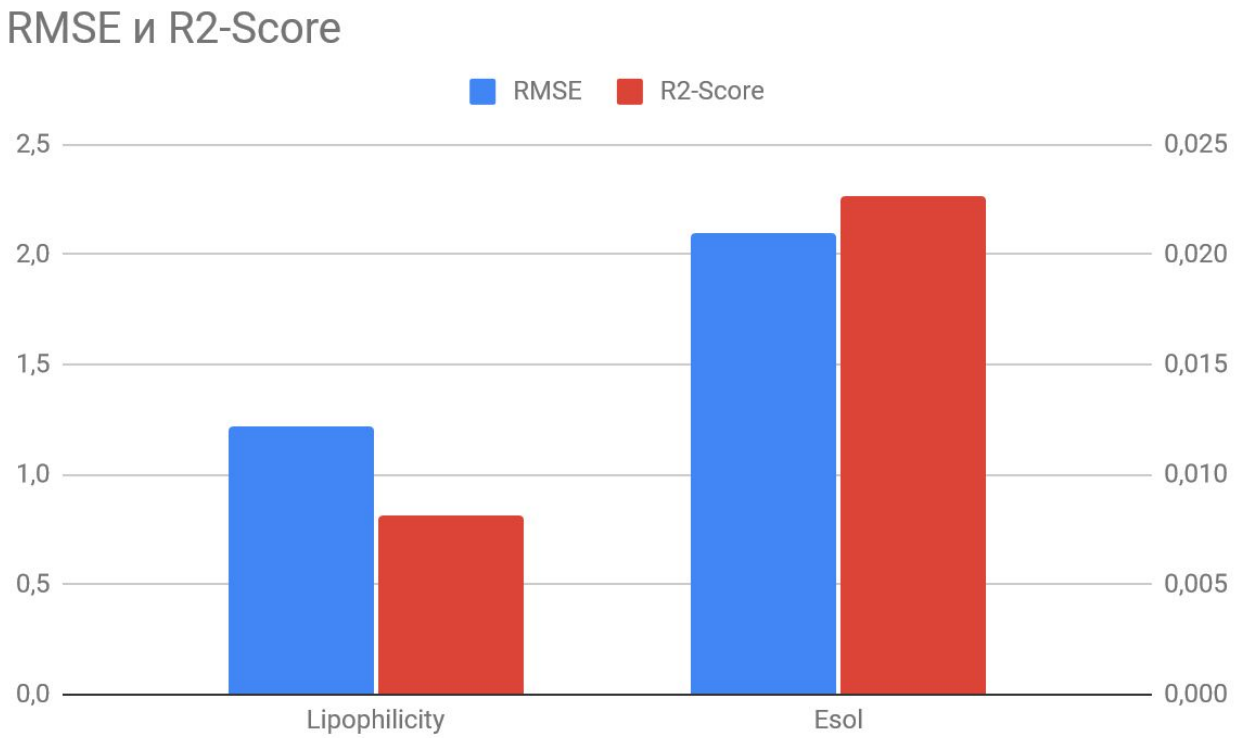
- Части молекулы -> чиселки
- Не разбиваем посимвольно!!!
- Ценим личное пространство каждого химического элемента

Convolution

Conv - довольно стремная штука



Convolution: результаты

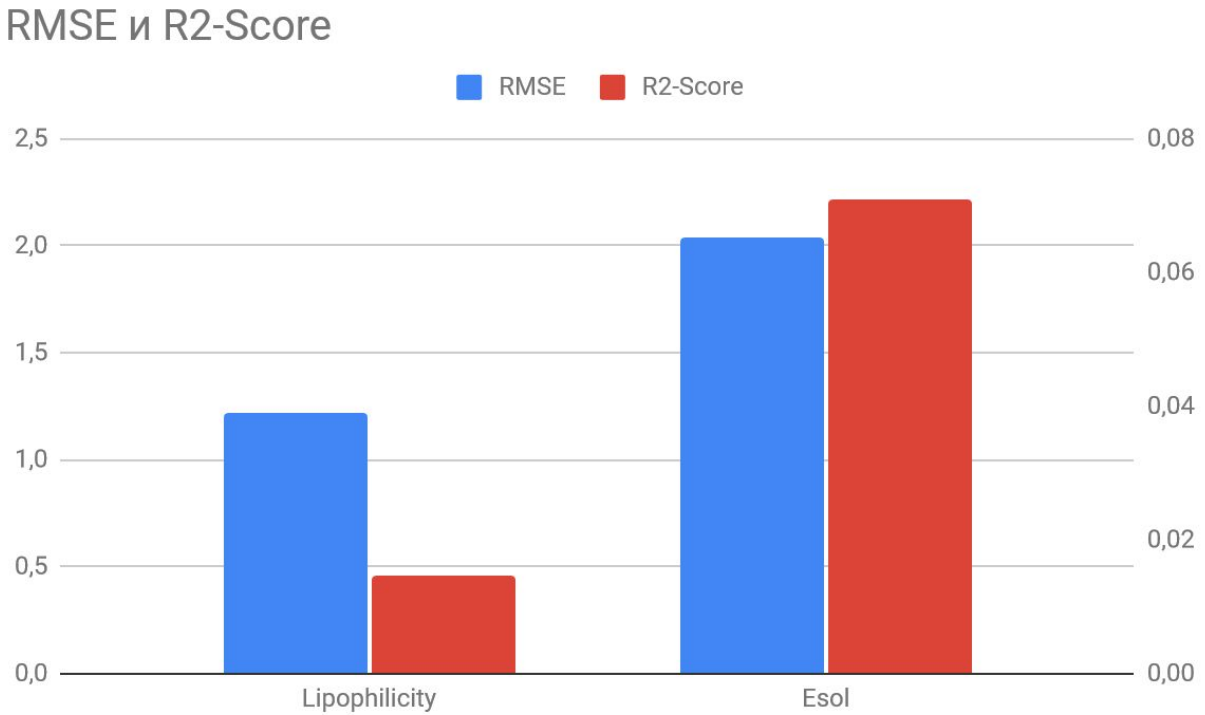


LSTM

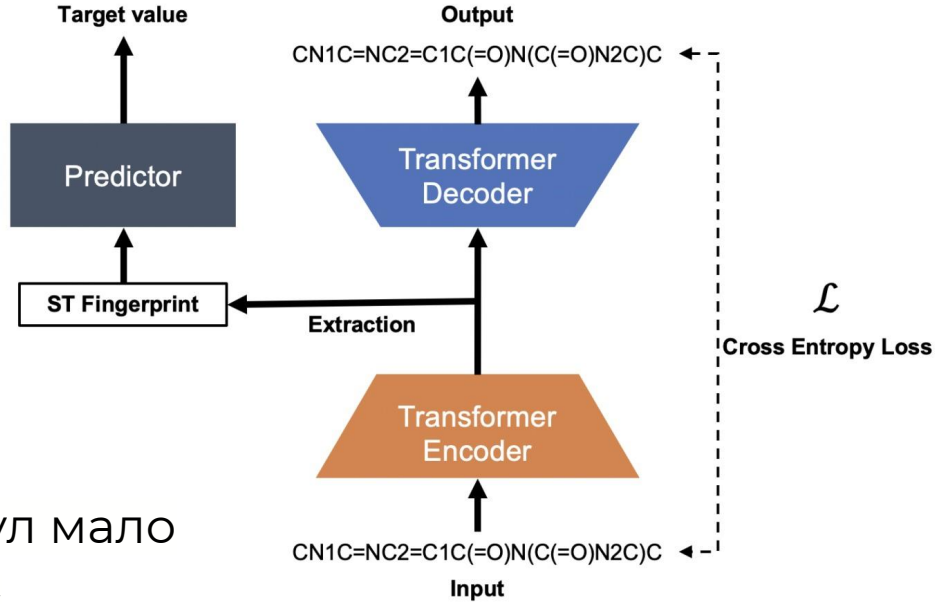
LSTM - архитектура, придуманная специально для последовательных данных

Сохраняет информацию от далеких элементов последовательности

LSTM: результаты

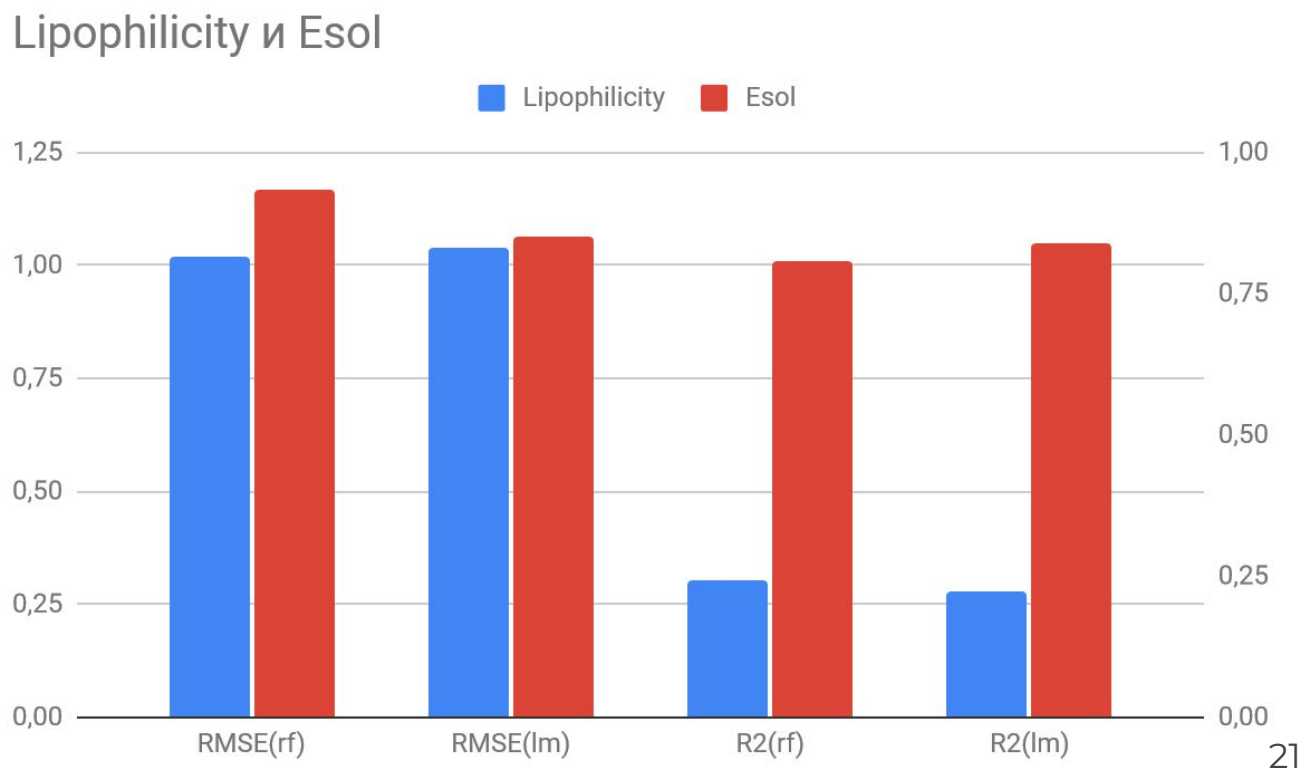


Pretrain Transformer



- Pretrain: Для свойств молекул мало экспериментальных данных
- Pretrain - autoencoder - transformer
- Но в Chembl - 2M молекул, можно взять модель обученную только на них

Pretrain результаты Transformer

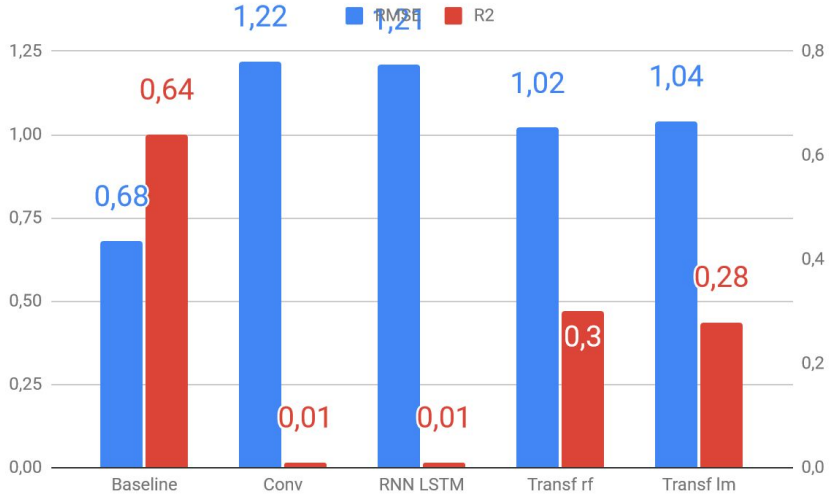


Результаты

Результаты

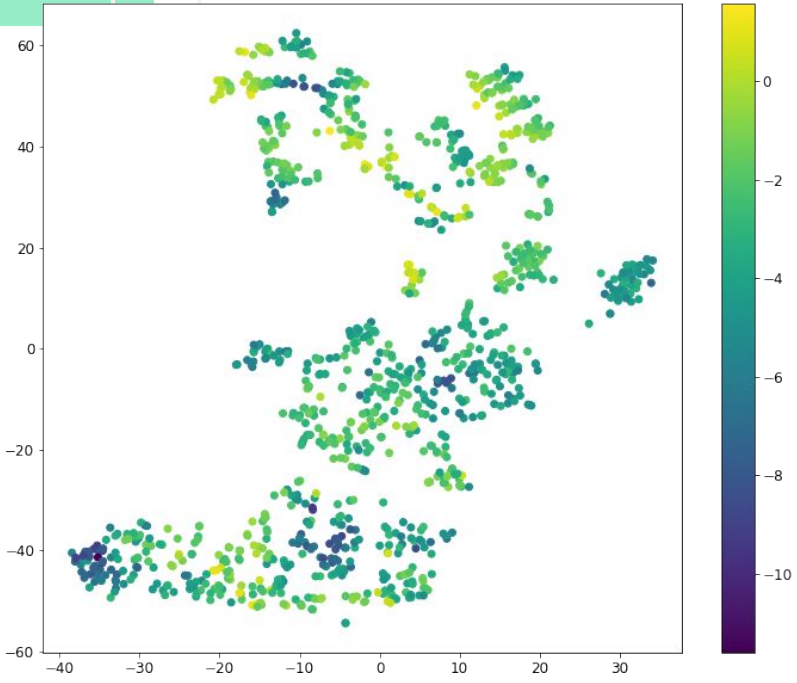
Esol

Lipophilicity

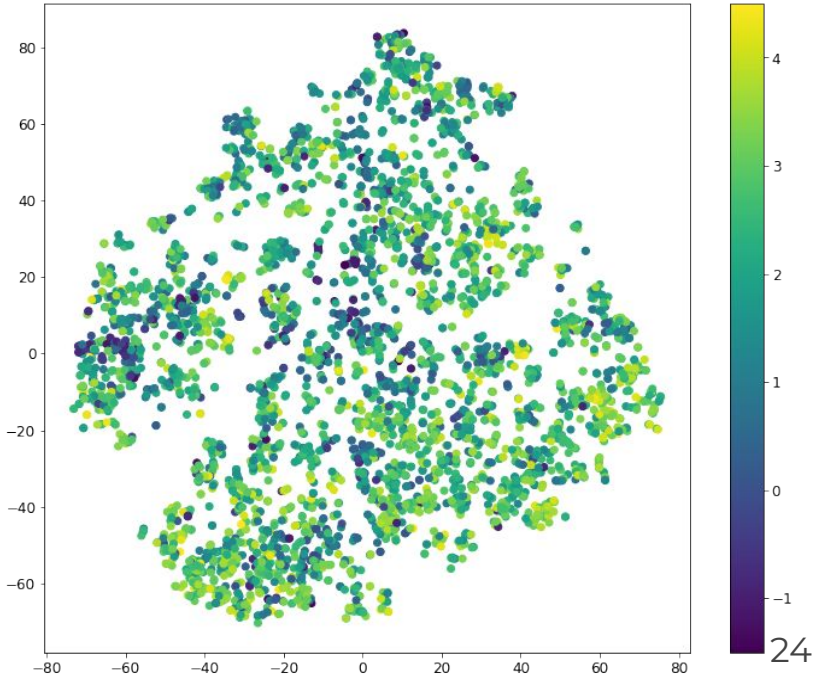


Сравнительная характеристика датасетов + визуализация

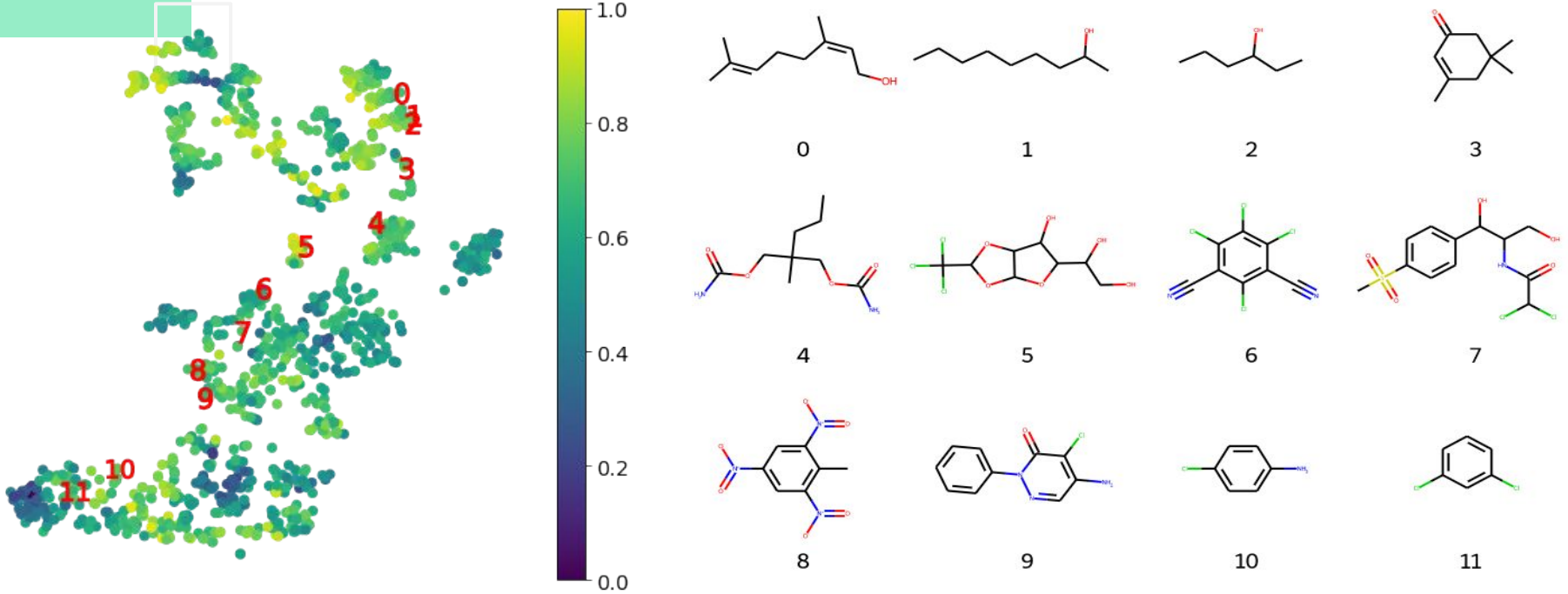
Esol



Lipophilicity

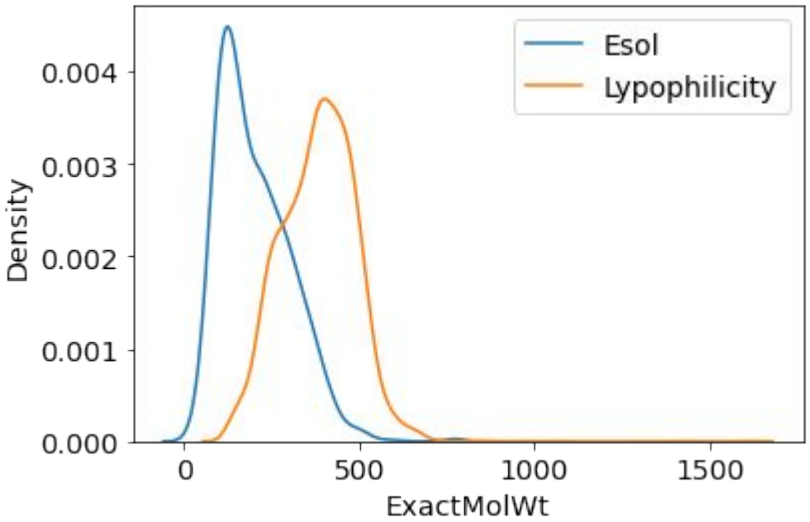


Сравнительная характеристика датасетов + визуализация

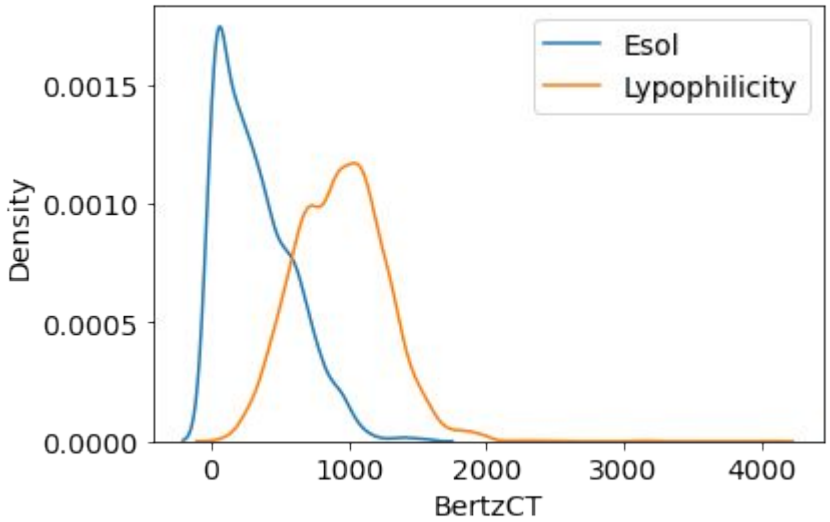


Сравнительная характеристика датасетов + визуализация

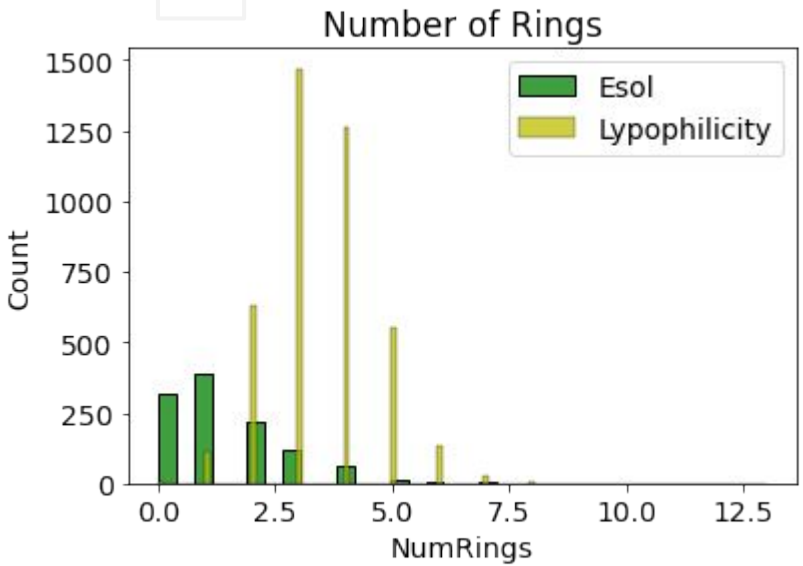
Молярная масса молекулы

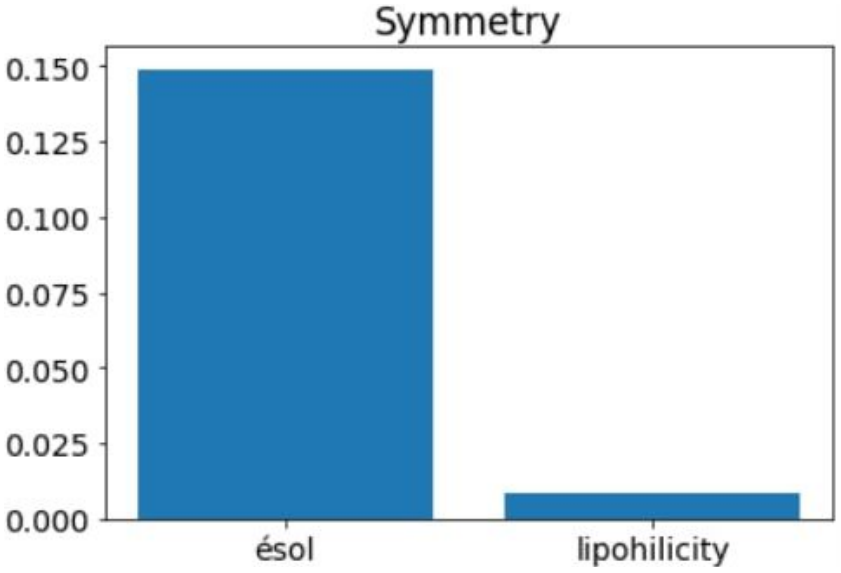


Топологическая ‘сложность’ молекулы



Сравнительная характеристика датасетов + визуализация





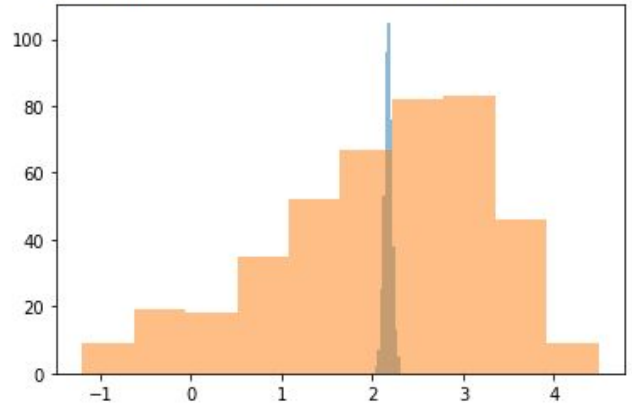
Выводы

- **Baseline**
- **NLP**
- **Визуализация**

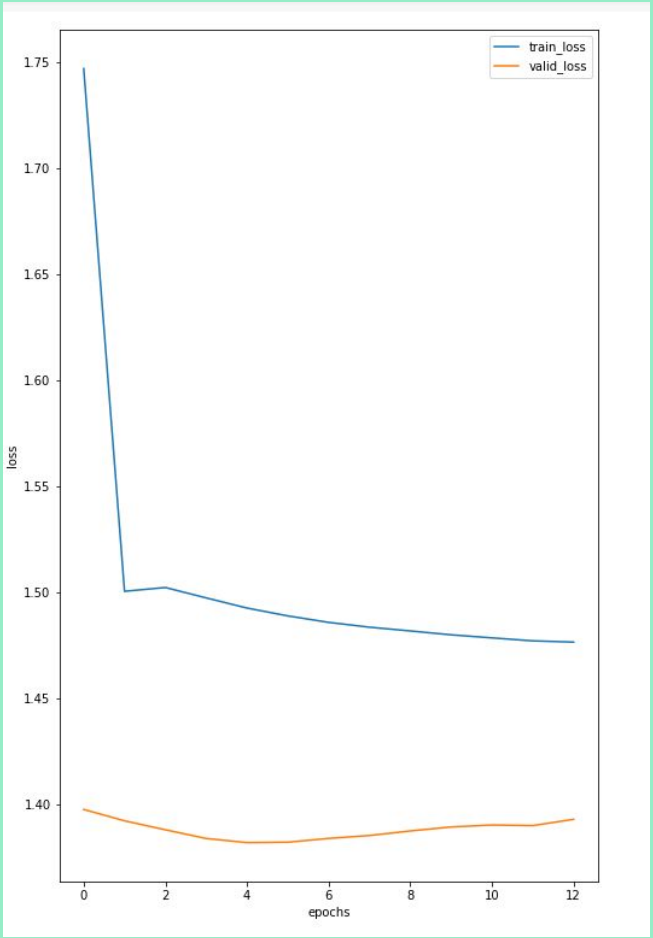
Бонус

Lipophilicity

```
MannwhitneyuResult(statistic=81626.0, pvalue=0.03077180313039602)
(1.3816576858116951, 0.012027260240867088)
```



Обучение



Вопросы?

GitHub