DataCon 2024

Команда 5 | Падение Сов

Заходят как-то в бар…

Заходят как-то в бар…

Химик

Артемий Лобач

Заходят как-то в бар…

Биоинформатик

Максим Щепетов

Химик

Артемий Лобач

Заходят как-то в бар…

Биоинформатик

Максим Щепетов

Химик

Артемий Лобач

ML-щик

Герман Кавкаев

Заходят как-то в бар…

Химик

Артемий Лобач

Биоинформатик

Максим Щепетов

ML-щик

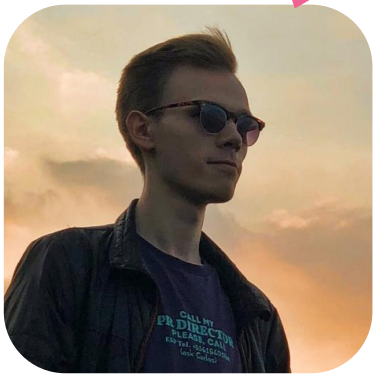
Герман Кавкаев

Какой-то клоун

Тимофей Рыко

А бармен им говорит:

Разработаете модель для предсказания CPP?

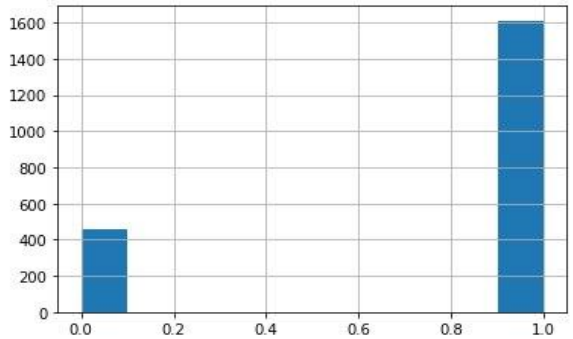


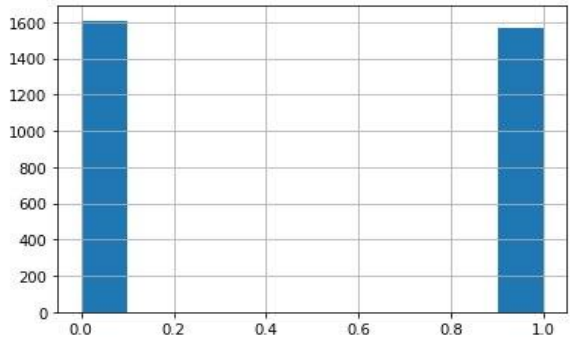
Мы попробовали, и кое-что даже получилось!

Предобработка данных

Бинарный датасет:

- Обогатили последовательностями CPP из базы CPPsite 2.0 (удостоверившись, что дубликатов нет)
- Побороли Imbalanced Data, обогатив пептидами из протеома человека





Датасет POSEIDON для регрессии

- Разобрались с единицами измерения, по-возможности конвертировали
- Почистили Time и Temp., заполнили недостающие значения медианой
- Удалили «плохую» колонку Conc.
- И, главное, почистили **Uptake** — целевую переменную

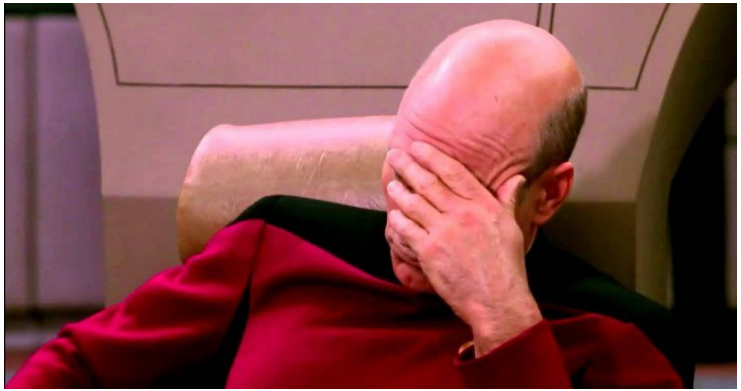
Чем же так плоха Conc.?

[31]

```
poseidon_df['Conc.'].unique()
```

Table Raw Visualize Statistics

```
array(['12.5 uM', '1.8 uM', '44 uM', '40 uM', '2.5 uM', '200 uM',
   '1600 uM', '0.05 umol', '50 uM', '10 uM', '5 uM', '5 umol/L',
   'Charge ratio = 5:1 (MPG/DNA)', 'Charge ratio = 10:1', '300 uM',
   '5 umol/l', '6 uM', '15 uM', '4 uM', '2 uM', nan,
   '100:1 molar excess of siRNAs', 'N/P ratio 20', '100 nM siRNA',
   '100 nM', '5 mM', 'N/P ratio 10',
   'Equal to a DOX dose of 10 ug/mL', '3 uM', '1 uM', '10 umol/L',
   '100 uM siRNA', 'Charge ratio = 0.5 / Charge ratio 5',
   '10^4 particles', '2 ug/ml', '25 uM', '57 umol/L', '5 ug/ml',
   '1 nM', '200 ug/ml', '8 mM Trehalose', '13 uM', '8 uM', '16 uM',
   '8 uM MCoTI-II / 16 uM SFTI-1', '100 nmol/kg', '10 uM BSH-11R',
   '10 uM BSH-11R (Boron concentration)', '1.5 ug/mL lipossomes',
   '40 ug/ml', '30 ug/ml', '25 µg/ml', '4 uM CPP', '150 uM',
   'Final Concentration between 1/5 uM', '10000 nM', '0.15 uM',
   '100 uM', '500 nM', 'Charge ratio = 10', '10 nM', '2.1 uM',
   '10 ug/ml', 'charge ratio = 2', '3 nmol',
   'Molar ratio 50 (v/w) or (w/w) eGFP', '1.0 mg/ml', '1 umol',
   'N/P ratio = 3', '100 ug/ml', '0.5 uM', 'N/P ratio 6',
   'plasmid/PDL ratio 1:4', '1.5 nM', '5 ug',
   'pDNA/peptide ratio 1:2500', '10 ug/mL',
   '2 uM dox equivalent concentration', '50 ug/ml', '0.5uM', '1uM',
   '2uM', '5uM', '10uM', '15uM', '20uM', '30uM', '40uM', '50uM',
   '1.9 uM', '5.6 uM', '16.7 uM', '6.3 uM', '25uM', '37 kBq', '20 uM',
   '12 uM', '0.2uM', '2.5uM', '7.5uM', '6uM', '100 ug/mL', '4uM',
   '100nM', '40 ug /mL', '0.1uM', '30 uM', '1.25uM', '9.0 uM', '25nM',
   '50nM', '200nM', '0.002ug/mL', '0.25uM', '25 ug/mL', '20 ug/ml',
   '100ug/ml', '0.1ug/ml', '100 mg/L', '20ug', '2ug/mL', '7uM', '3uM',
   '2.5', '3.9mg/ml', '16uM', '6ug/mL', '12ug/mL', '25ug/mL', '100uM',
   '0.01uM', '3000 pmol', '25ug/ml', '2mg/ml', '10ug', '600 ug/ml',
   '1 mg/ml', '2ug/ml', '5ug/ml', '100ug'], dtype=object)
```

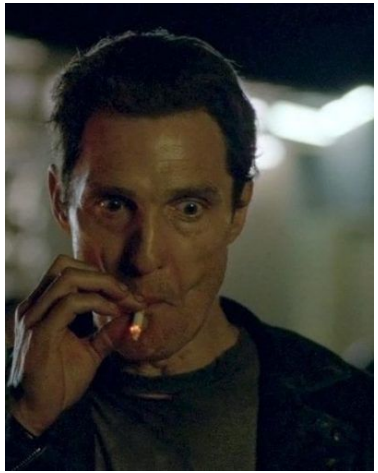


И так везде…

25 %

Последовательностей в датасете POSEIDON — странные

Вот лишь некоторые из последовательностей…



Mpa(luc)-KTRVLKRWKL-NH2

rXrrXrrXrrXr

Nbtg-Nspe-Nspe-Nbtg-Nspe-Nspe-Nbtg-Nspe-Nspe-Nbtg-Nspe-Nspe

(Acp)-KKKKKRFSFKKSFKLSGFSFKKNK

RRWWWWRR-E12

C37H66N7O17P3S-rrrrrrrrrrr

c[DKP-RGD]-PEG4-GLRKRLRK(CF)FRNKIKEK-CONH2

CH3(CH2)16-CONH-GGGGLRKRLRKFRNKIKEK-NH2

Мы чистили, чистили, и наконец почистили до…

21 %

Остальное — удалили, потому что почти никакие тулы с такими последовательностями не работают


Это немного, но это честная работа

3121

Строк в бинарном датасете

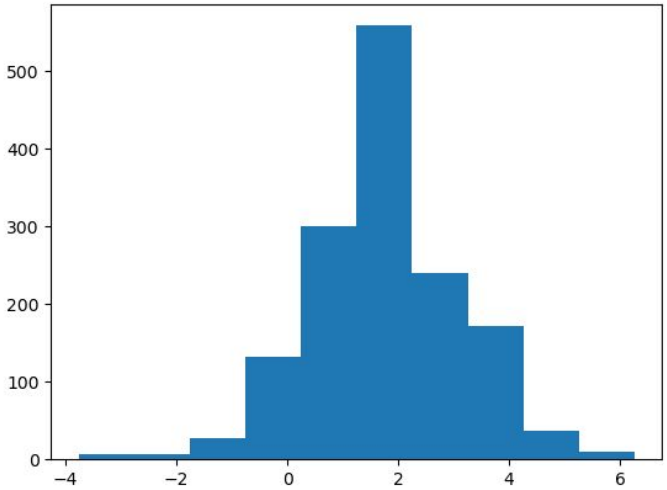
1524

Строки в датасете POSEIDON для регрессии

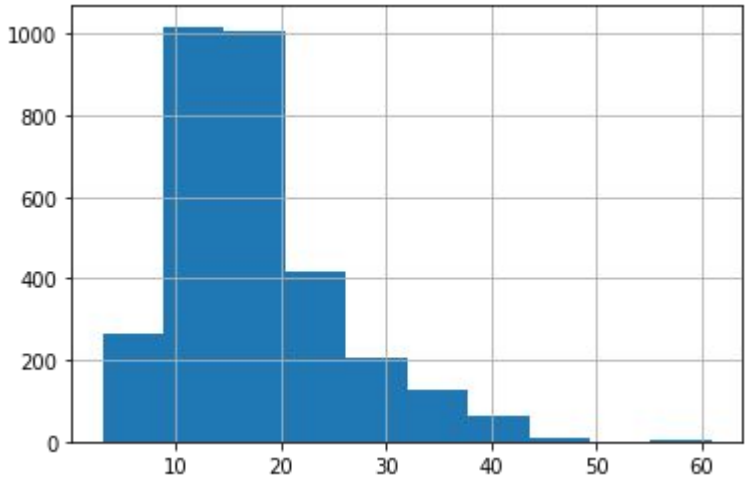
EDA & Feature engineering

Распределение целевой переменной Uptake

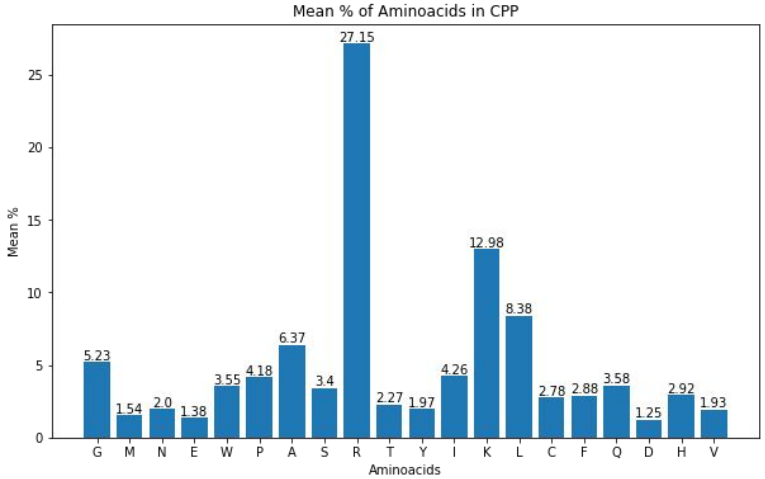
- Изначально распределение было ужасным
- Поэтому мы его логафримировали (на сайте POSEIDON тоже это делают)

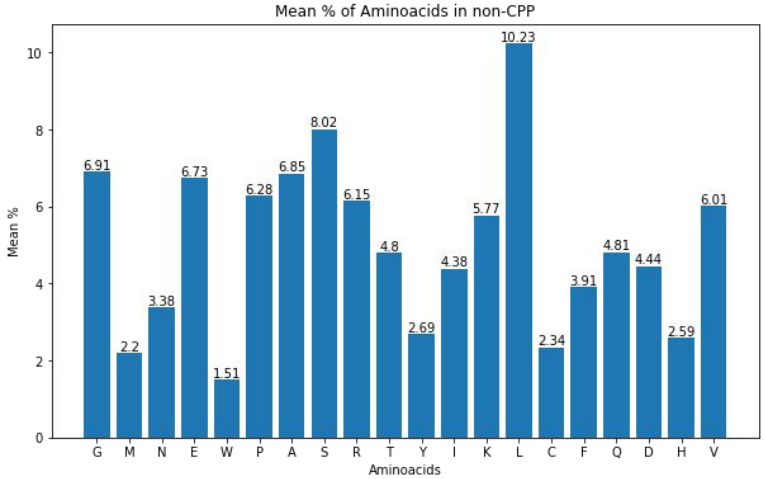


Длина последовательности



Распределение по аминокислотам





Feature engineering

- Рассчитали дескрипторы PyBioMed и Biopython
- В некоторых подходах добавили доли важных АК

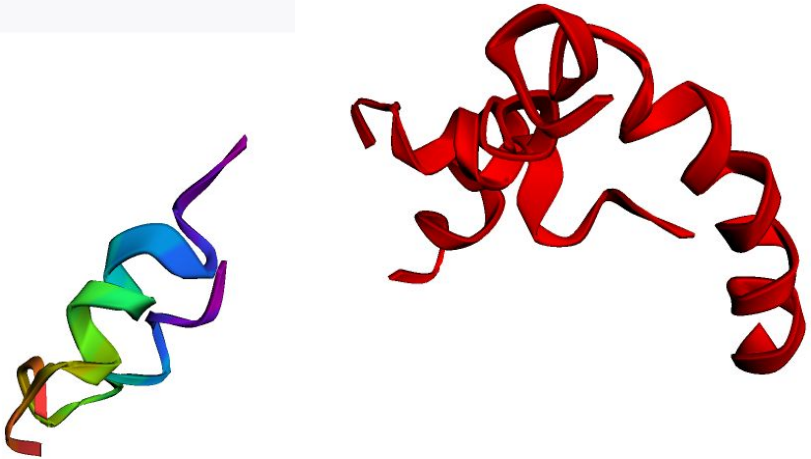
Анализ пространственных паттернов

- Многие CPPs имеют альфа-спирали в своей вторичной структуре, которые расположены в середине последовательности или на ее концах
- CPP, не имеющие регулярной вторичной структуры также встречаются
- В перспективе можно будет добавить предикторы:
  - Длина предсказанной (например при помощи alphaphold) самой длинной альфа-спирали,
  - Количество предсказанных альфа-спиралей и процент последовательности, имеющий регулярную вторичную структуру
  - В целом, эта информация есть в дескрипторах

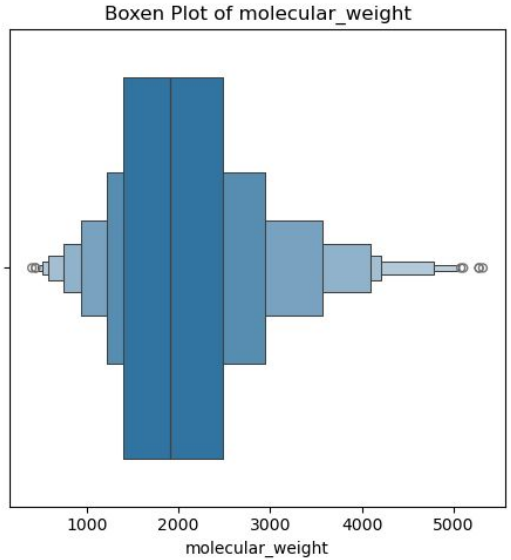
Красивое

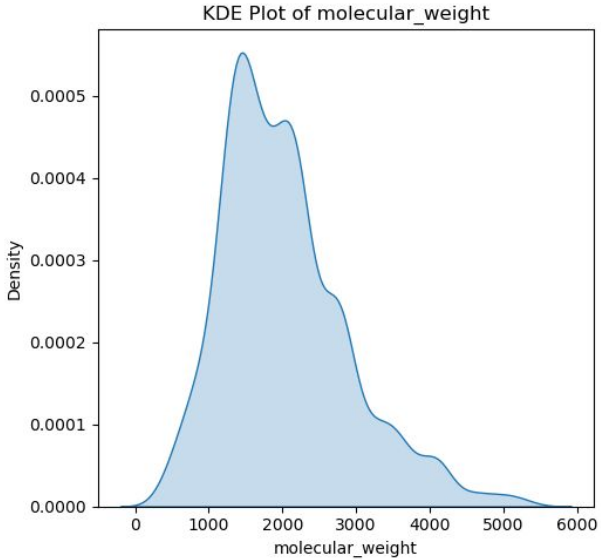
[176]

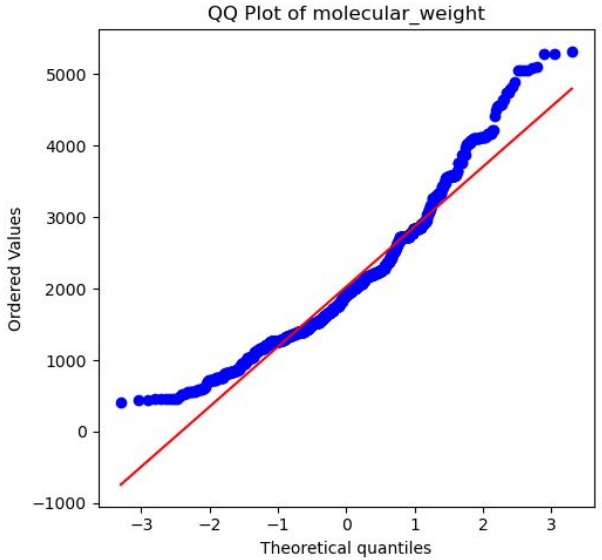
```
view = py3Dmol.view(width=800, height=400)
view.addModel(pdb_file4, 'pdb')
view.addModel(pdb_file5, 'pdb')
view.setStyle({'cartoon': {'color':'spectrum'}})
view.show()
```



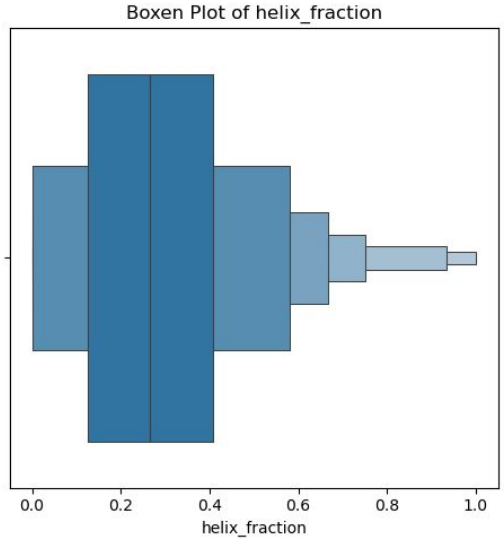
Молекулярный вес

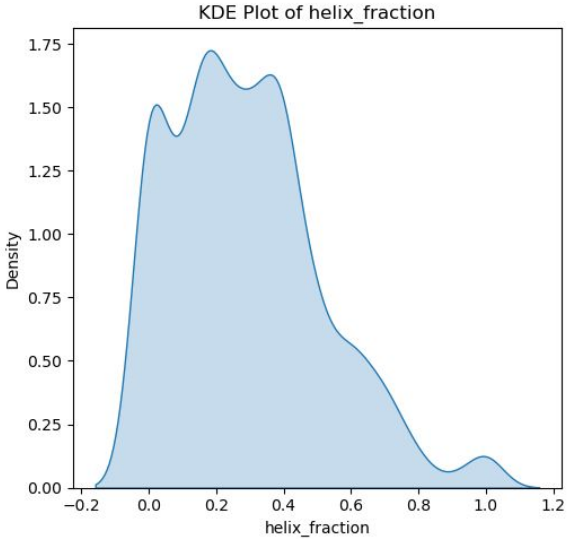


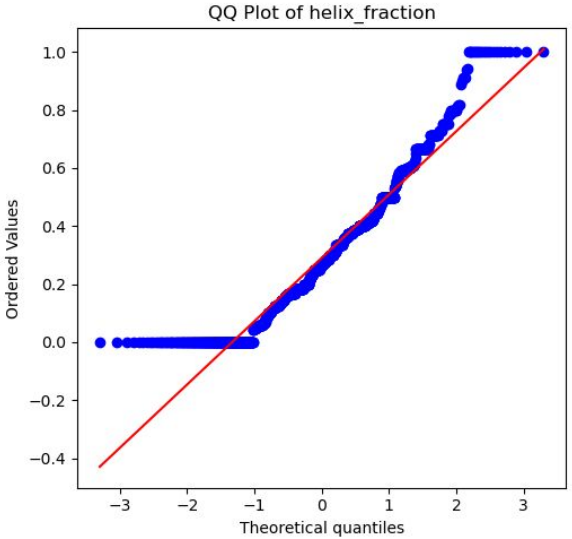




Доля альфа-спиралей







Как мы работали с ML?

- Тестировали разные модели
- Все перспективные модели улучшали
  - Feature selection with correlation analysis
  - Hyperparameter Optimization with optuna (cross-validation)
- Для классификации смотрели на такие метрики:
  - F1 score
  - AUC
  - Precision (так как ложноположительные результаты стоят дороже в нашем случае)
- Для регрессии:
  - MSE
  - MAE
  - R^2 (идеально для сравнения самых разных подходов)

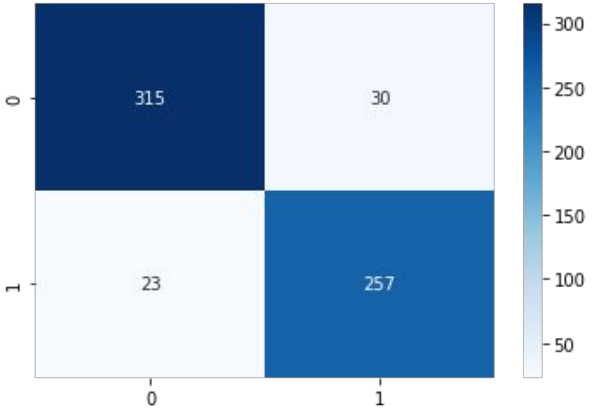
Классификация

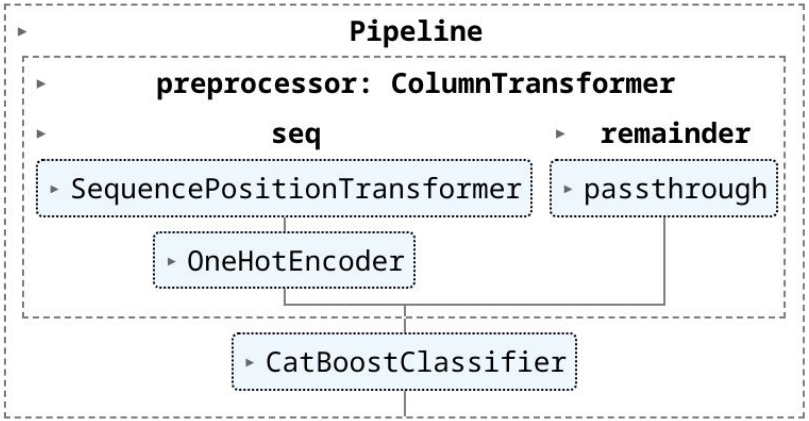
Лучший — CatBoostClassifier

- Простой one-hot энкодинг аминокислот
- Предикторы BioPython

Final F1 Score: 0.906525573192239 9

Final AUC: 0.961023029503329 7





Регрессия

Попробовали разные подходы

- Попробовали разные подходы:
  - Удалить или оставить аутлаеры
  - Выбрать самые важные фичи или работать со всеми
- По итогу оказалось эффективнее использовать все данные и все фичи
- Протестированные модели: модели классификации и регрессии

Лучшая модель — XGboost Regression

- Полные данные, без обрезки выбросов в параметрах
- Логарифмирование таргетного параметра
- Обрезка датасета по 95 перцентилю таргета
- Использование всех параметров в датасете
- XGBoost архитектура

Лучшие метрики

```
XGBoost Regression Model:
  Mean Squared Error (MSE): 1.5975
  Mean Absolute Error (MAE): 0.8404
  Root Mean Squared Error (RMSE): 1.2639
  R-squared (R^2): 0.7945
```

Наиболее важные фичи

- **SolventAccessibility** — этот параметр описывает, насколько определенный участок пептида будет взаимодействовать с растворителем
- **Polarizability** — описывает поляризуемость пептида
- **SecondaryStr** — описывает наличие альфа-спиралей (бета-листов) пептида
- **Charge, isoelectric_point, charge_at_ph** — Заряд пептида может существенно влиять на его взаимодействие с клеточной мембраной
- **Polarity** — описывает полярность пептидов
- **NormalizedVDWVT** — описывает Ван-дер-Ваальсов объем
- **Hydrophobicity**, **gravy** — являются мерой гидрофобности пептида
- **Turn_fraction**, **sheet_fraction** — описывают гибкость и конформацию пептида
- **Instability_index** — является мерой стабильности пептида

Эмбеддинги пептидных последовательностей

- Попробовали многое (альфафолд, ProtBert…)
- Не успели имплементировать (беды с зависимостями, слишком долго, не хватает вычислительных мощностей)
- Реализовали **blomap** — эмбеддинги, основанные на BLOSUM
- Модель, казалось бы, получилась хорошая, но…

За час до дедлайна мы выяснили, что...

Судя по всему, произошла утечка из TRAIN в TEST

	TRAIN	TEST
MSE	0.00034	0.00303
MAE	0.00325	0.010895
R^2	0.99976	0.99793

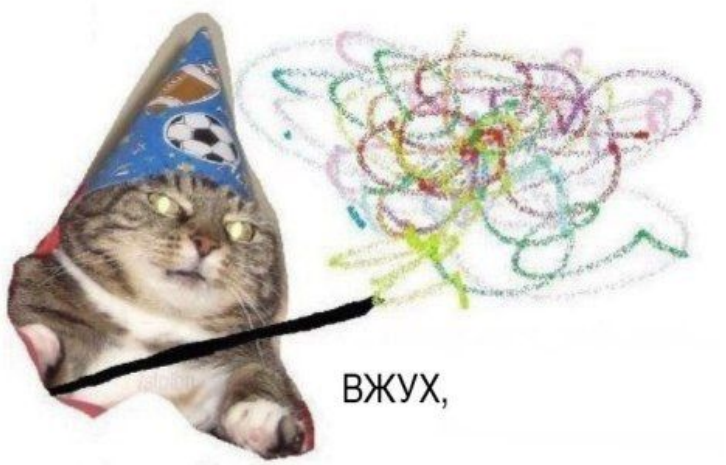
- Модель видит в TEST такие же или очень похожие последовательности
- И предсказывает то, что видела в TRAIN
- Информация — в BLOSUM эмбеддингах

Но модели без эмбеддингов тоже работают неплохо!

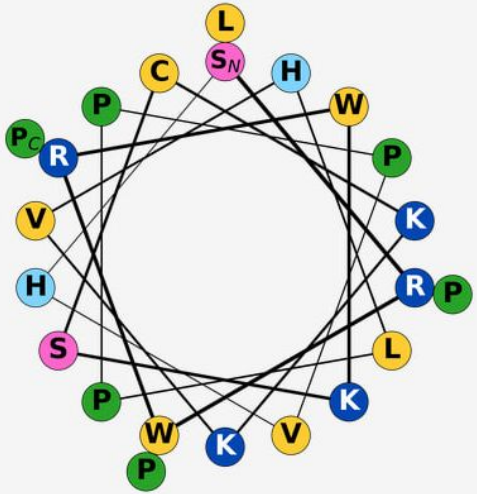
Генерация

Как оно работает?

- Реализовали прототип генетического алгоритма
- В качестве функции приспособленности использовали нашу регрессию



Genetics go brrrrr



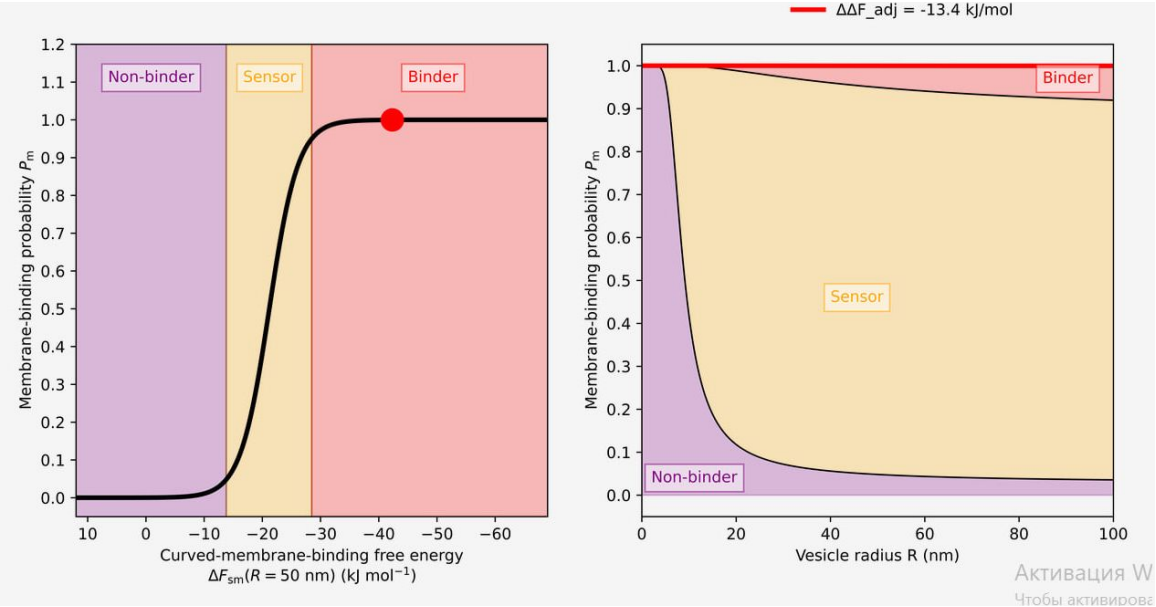
SRWRWKSCKKVHLPPPVHLPPP

Predicted ΔΔF: -7.909 kJ/mol
Predicted ΔΔF_L24: -8.75 kJ/mol
Predicted ΔΔF_adj: -13.4 kJ/mol

Negatively charged membrane
Calculated from ΔΔF_adj:
Predicted ΔF_sm(R=50): -42.333 kJ/mol

Sequence length: 22
Sequence charge: 5
Hydrophobicity: 0.518
Hydrophobic moment: 0.119

Умеренное, но уверенное связывание



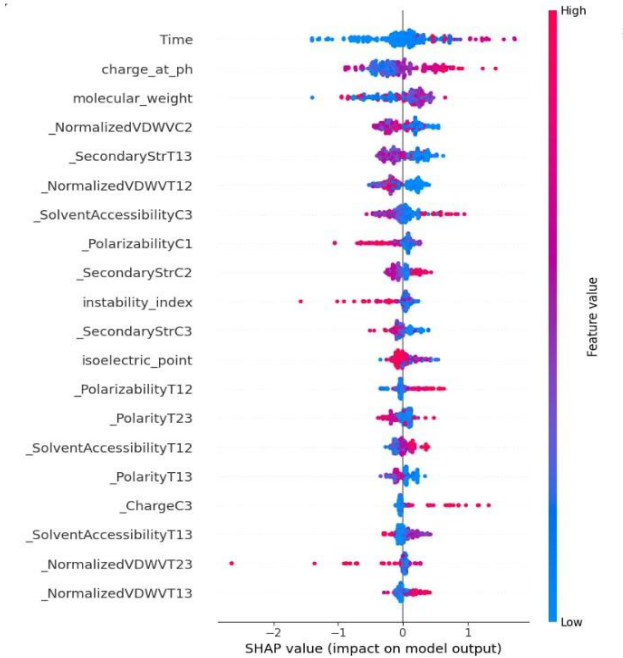
Non-redundant

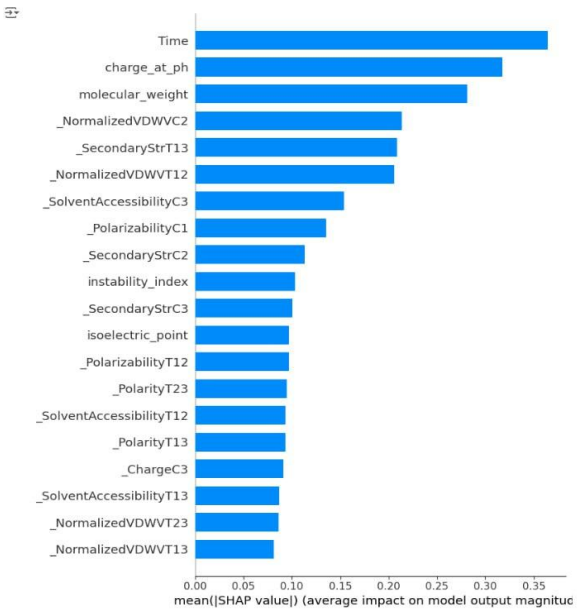
- Длина пептида 22
- Найденный идентичный участок в CPPsite 2.0 длины 13

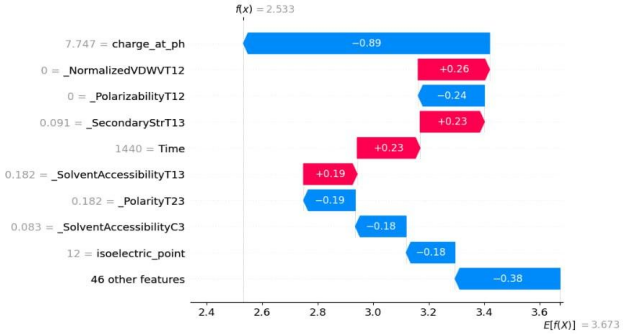
```
>1193
          Length = 13

 Score = 23.1 bits (47), Expect = 0.030,   Method: Compositional matrix adjust.
 Identities = 10/10 (100%), Positives = 10/10 (100%)

Query: 1  SRWRWKSCKK 10
          SRWRWKSCKK
Sbjct: 4  SRWRWKSCKK 13
```

Спасибо за внимание! Давайте поговорим 🙂

SHAP Feature Importances (Regression)







Беды со стандартизацией: Method имеет значение