

АНАЛИЗ СВЯЗЫВАНИЯ ИОНОВ МАГНИЯ С РНК

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

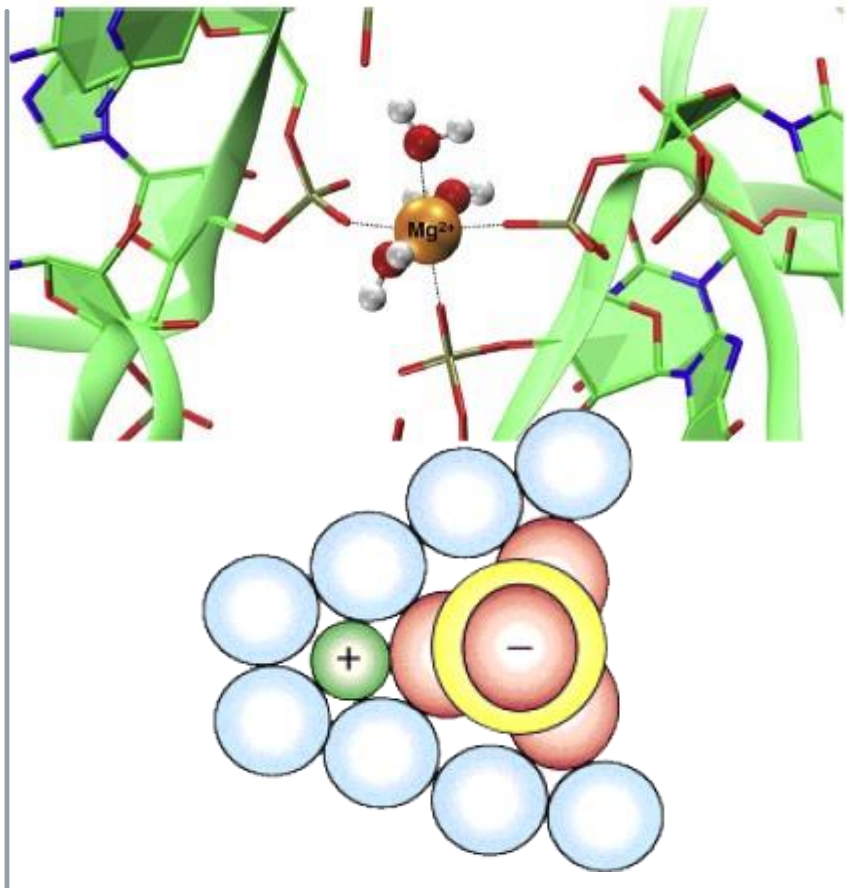
Студент:
Тихонова Полина

Руководитель:
М.А. Ройтберг
И.И. Цитович

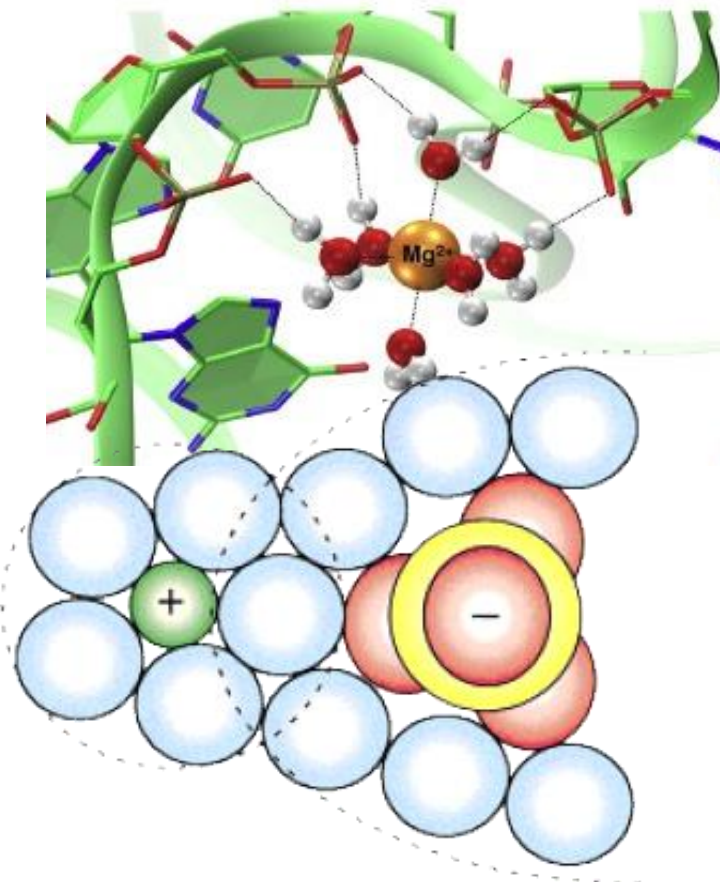
АКТУАЛЬНОСТЬ РАБОТЫ



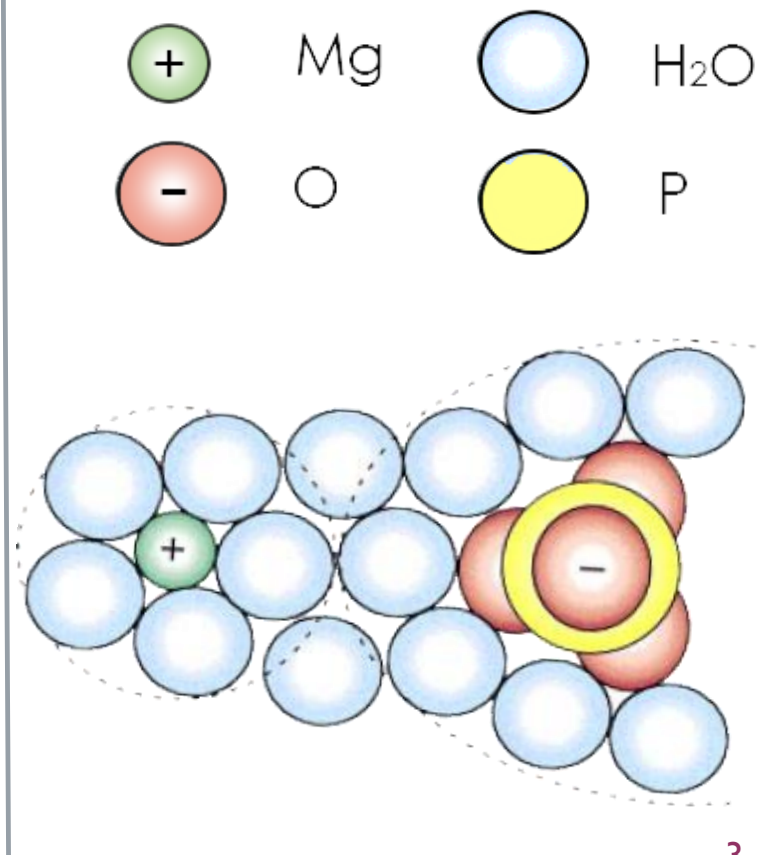
ТИПЫ СВЯЗЫВАНИЙ ИОНОВ МАГНИЯ



1. Сайт-специфическое связывание



2. Специфическое связывание
через воду

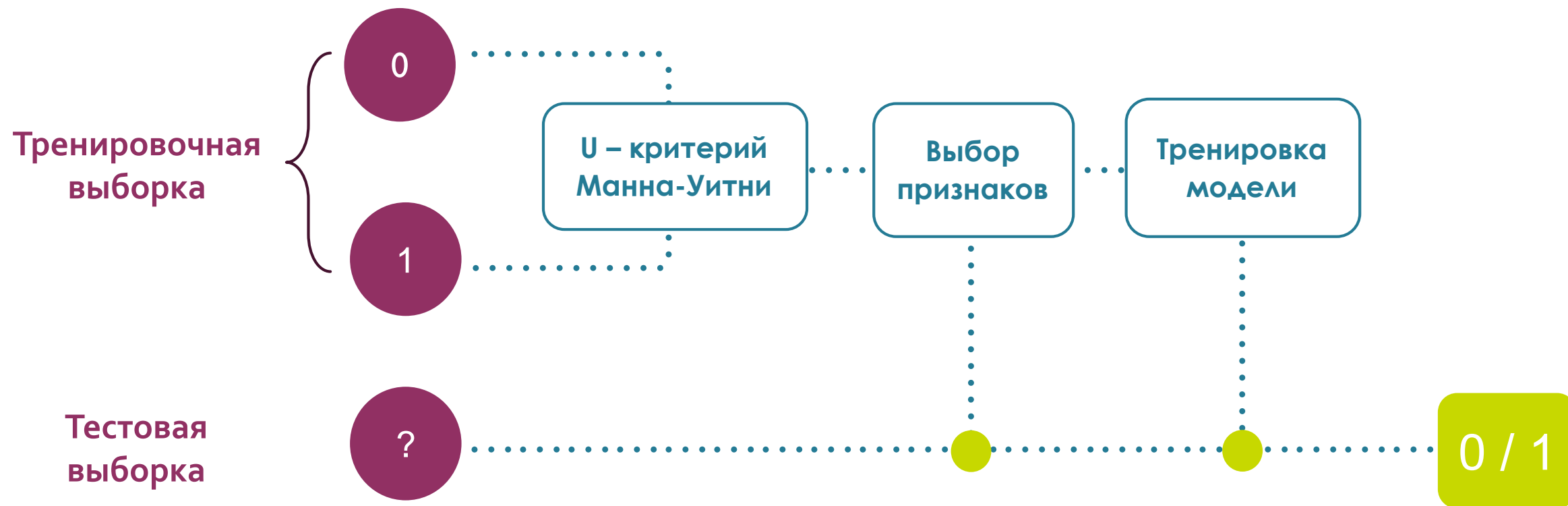


3. Диффузионное

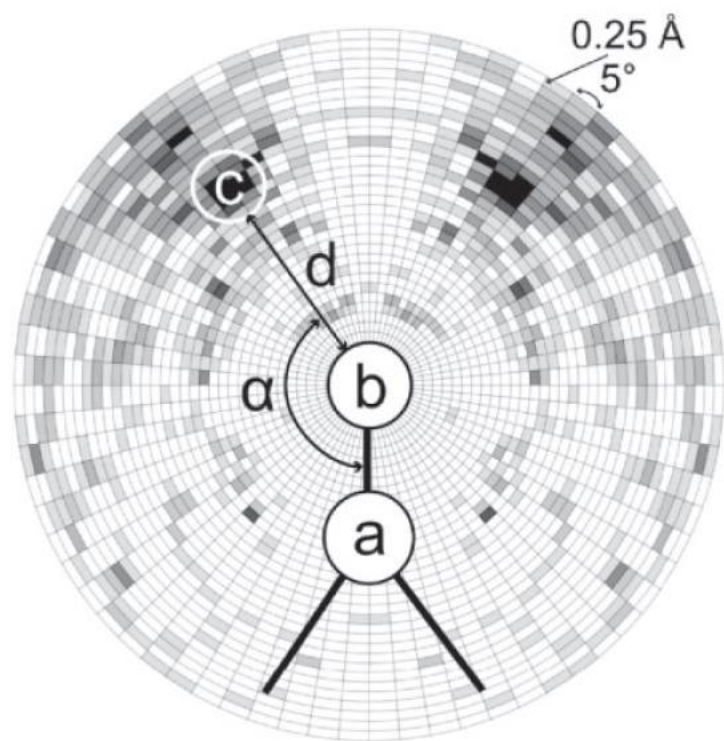
ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

WEBFEATURE, METALIONRNA

WEBFEATURE



METALIONRNA

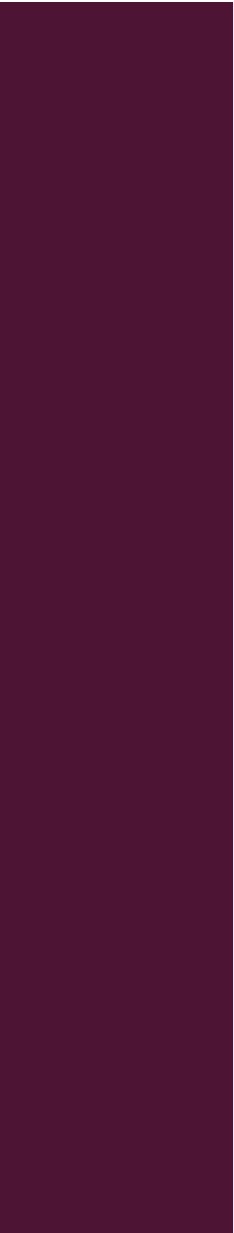


$$W(n)(d_1, \alpha_1; \dots d_n, \alpha_n) = -RT \ln g^{(n)}(d_1, \alpha_1; \dots d_n, \alpha_n),$$

$g^{(n)}$ - функция корреляции для n частиц, показывает экспериментально наблюдаемую частоту контактов катиона c со смежной парой атомов a, b ;

d - расстояние между ионом и атомом b ;

α - угол (a, c, b) .



РАЗРАБОТКА СОБСТВЕННОЙ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

ПОСТАНОВКА ЗАДАЧИ

- **Источник данных** – банк PDB с аннотациями от URS Database.
- **2 набора данных** – структуры с наилучшим разрешением / наибольшим числом ионов.
- **Элемент выборки** – нуклеотид / фрагмент нуклеотида / атом
- **Целевой признак** – наличие магния в радиусе 3 / 5 / 7 / 3-7 Å.

Всего 24 выборки.



Всего 361 - 383 признаков.

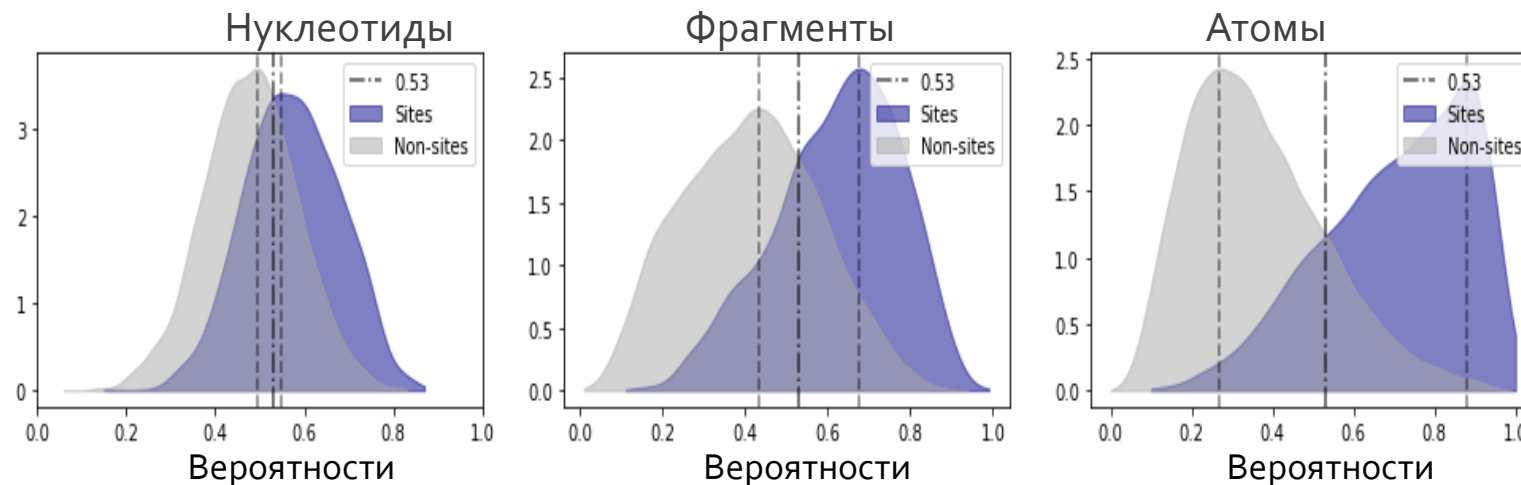
ВЫБОР НАБОРА ДАННЫХ

Классификатор:
Random Forest

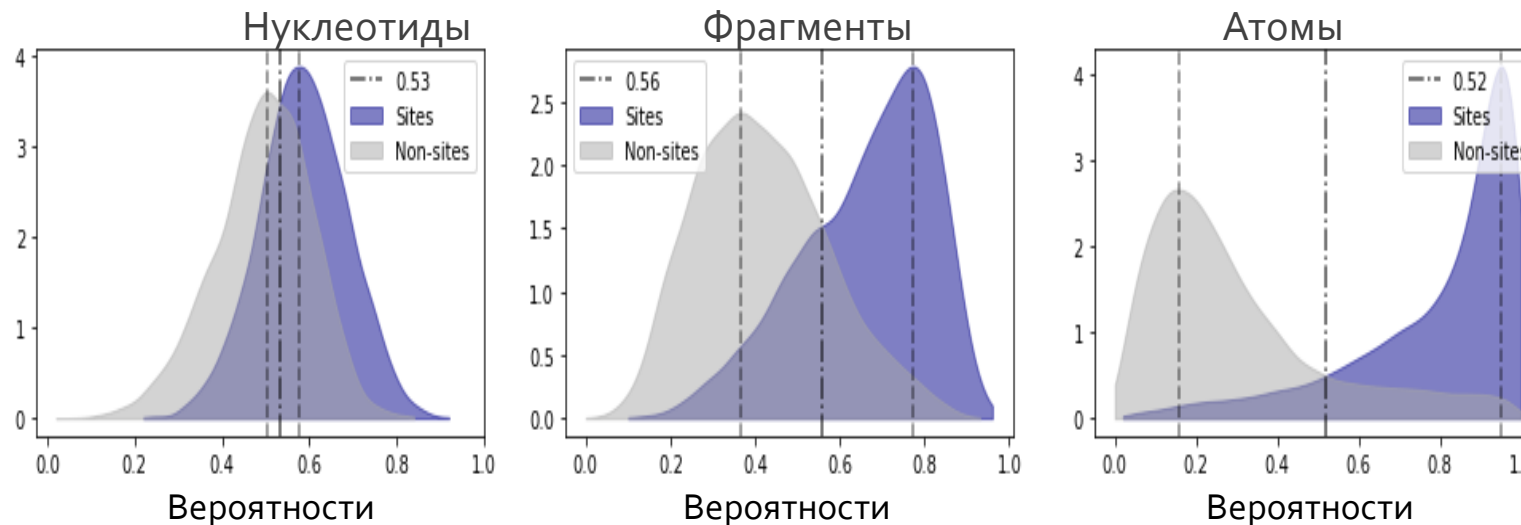
- + не склонен к переобучению;
- + работает с разными типами признаков;
- + быстро работает.

Распределения вероятностей предсказаний Random Forest

Лучшее разрешение, до 5 Å



Лучшее разрешение, до 7 Å



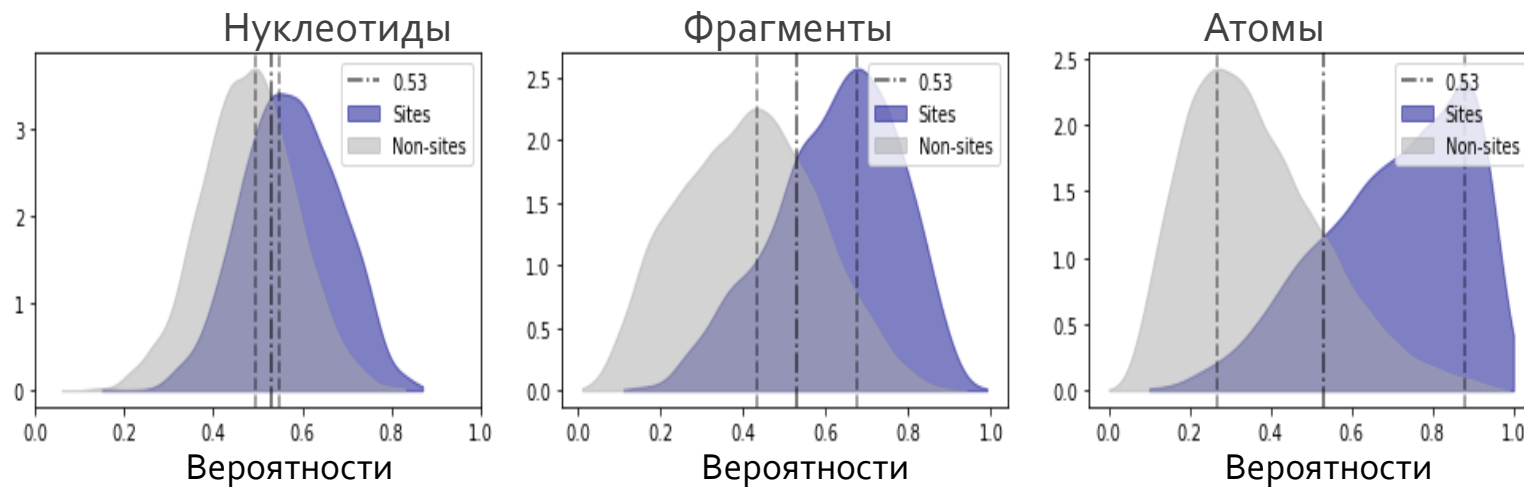
ВЫБОР НАБОРА ДАННЫХ

Классификатор:
Random Forest

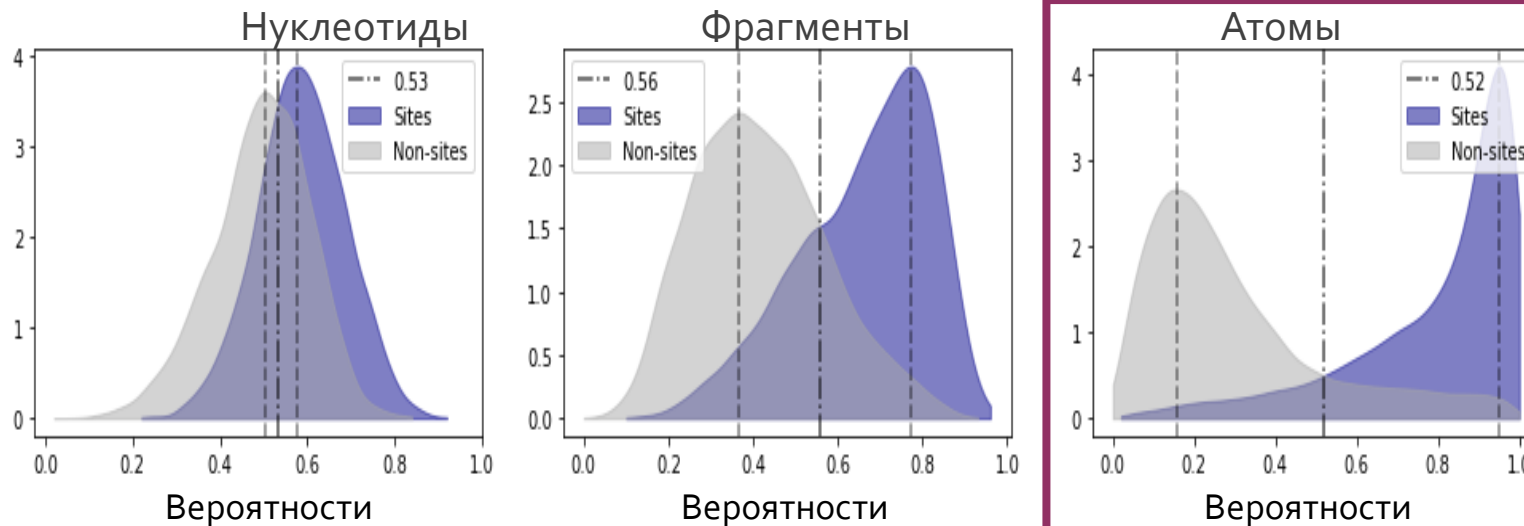
- + не склонен к переобучению;
- + работает с разными типами признаков;
- + быстро работает.

Распределения вероятностей предсказаний Random Forest

Лучшее разрешение, до 5 Å

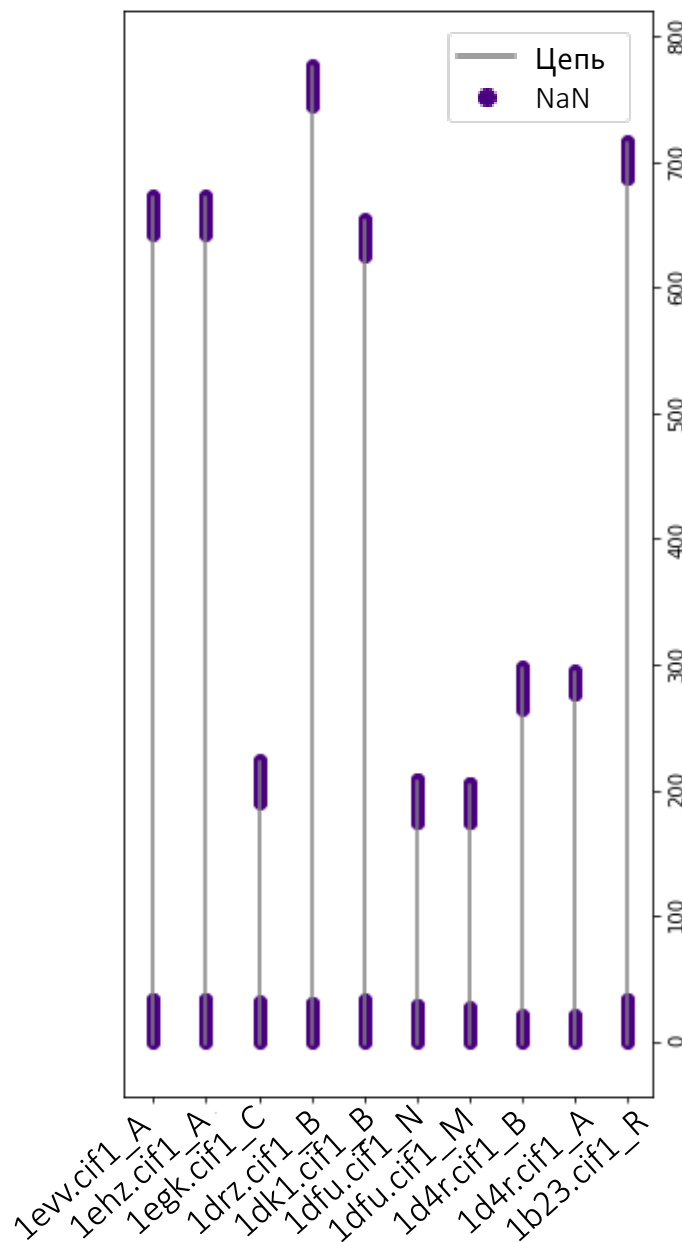


Лучшее разрешение, до 7 Å



АНАЛИЗ ДАННЫХ

- Пропущенные значения.
- Разреженные признаки
- Скоррелированные признаки.



Пропущенные значения встречаются в:



- Углах;
- Спариваниях;
- Информации о нуклеотидах.

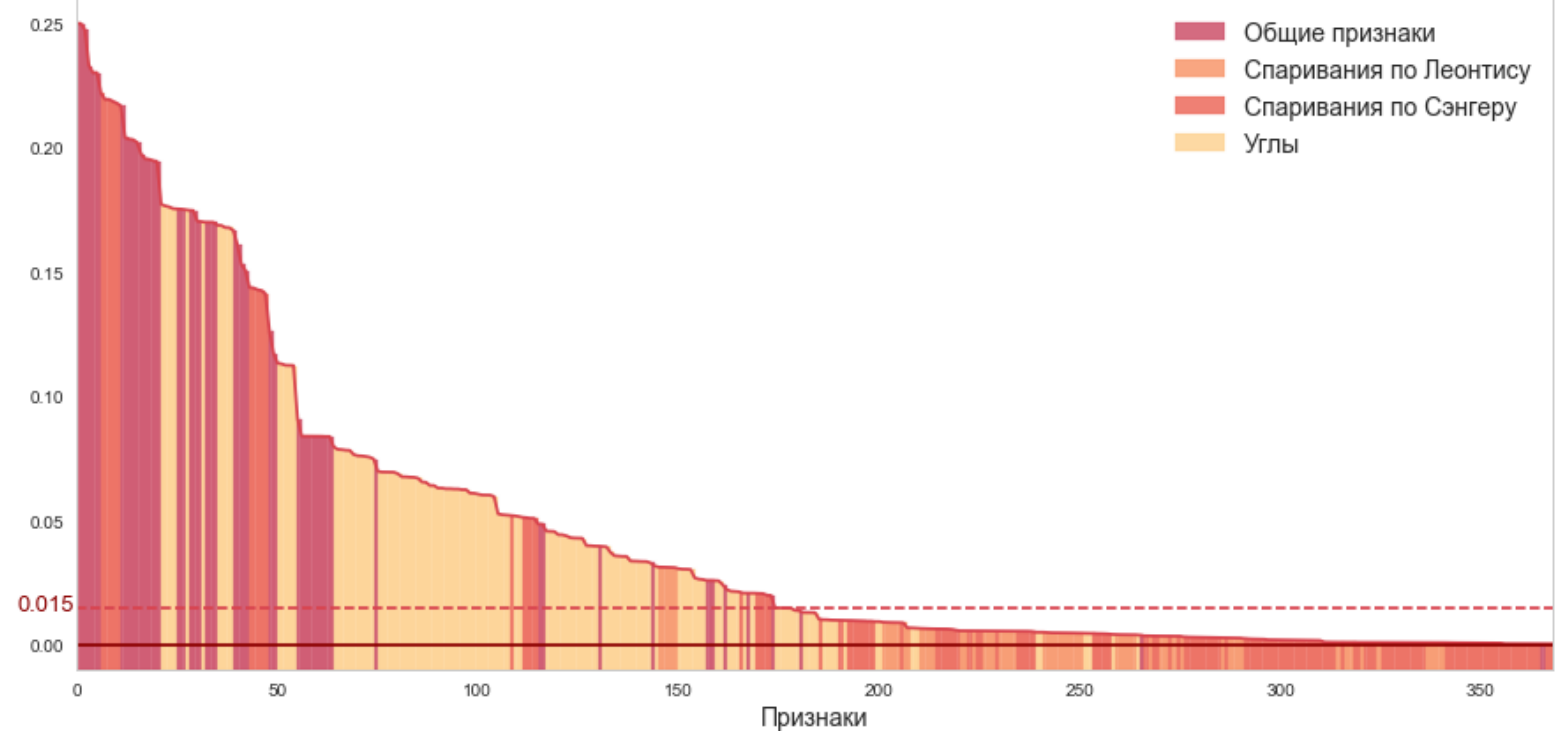
Заполнение нулями не нарушает логики интерпретации признаков.



АНАЛИЗ ДАННЫХ

- Пропущенные значения.
- Разреженные признаки
- Скоррелированные признаки.

Дисперсии признаков



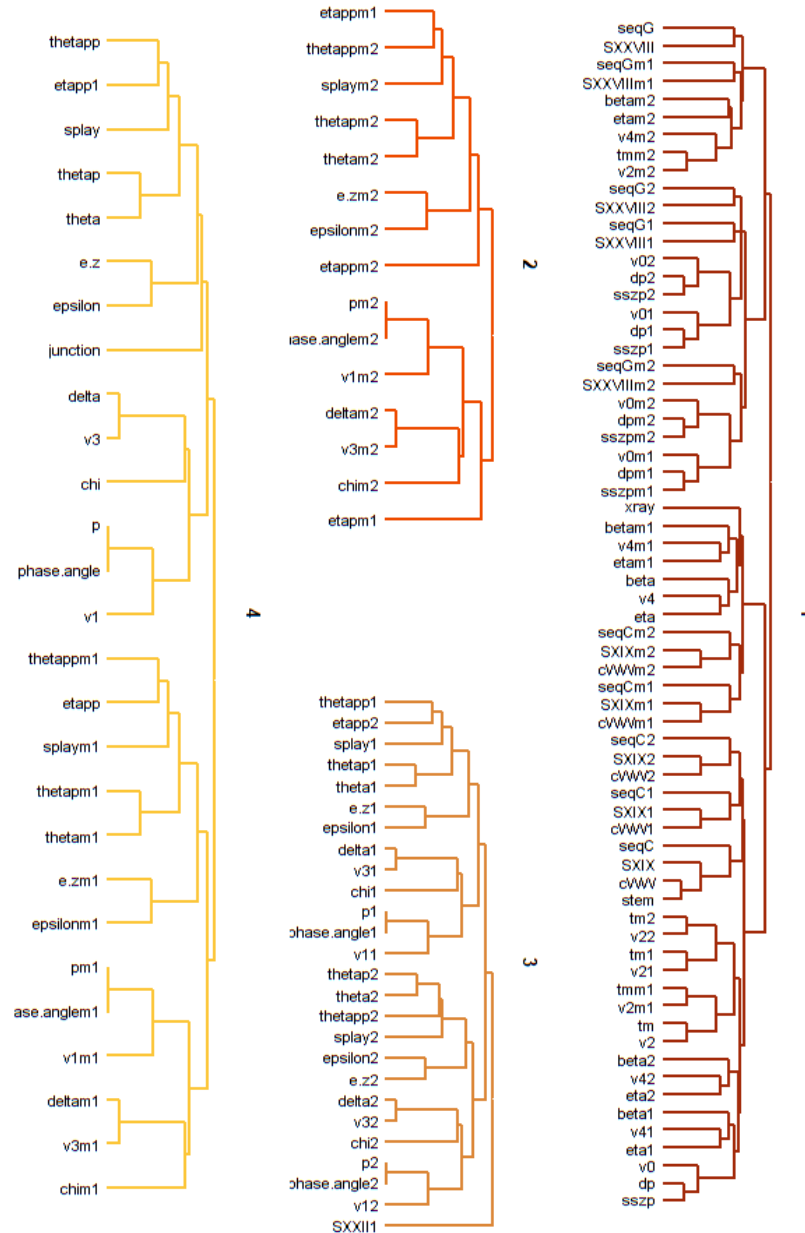
Большая часть спариваний являются неинформативными признаками.



Остается 170 признаков, дисперсия которых выше выбранного порога (0.015).

АНАЛИЗ ДАННЫХ

- Пропущенные значения.
- Разреженные признаки
- Скоррелированные признаки.



Наиболее скоррелированные группы признаков:

1. thetaapp, etapp, splay, thetap, theta;
2. e.z, epsilon;
3. v0, dp, sszp;
4. v1, p, phase.angle;
5. v2, tm;
6. v3, delta.



ОБЩИЙ ВИД МОДЕЛИ

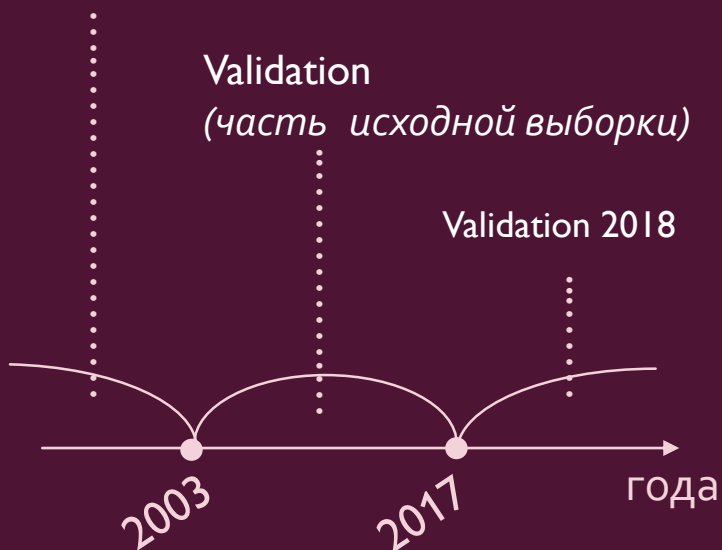
- **Преоброцессинг данных (тренировочной и тестовой выборки).**
 - Заполнение пропущенных значений нулями.
 - Удаление признаков с маленькой дисперсией.
 - Удаление скоррелированных признаков.
- **Тренировка модели.**
 - Балансирование выборки: элементы – не сайты связывания отбираются случайным образом в количестве, равном числу сайтов связывания.
 - Тренировка RandomForest с параметрами:
 - `Max_depth = 26`
 - `Min_samples_leaf = 20`
 - `Max_features = 0.7`
- **Предсказание натренированной модели на тестовой выборке.**

ВАЛИДАЦИЯ РЕЗУЛЬТАТОВ

Validation < 2003

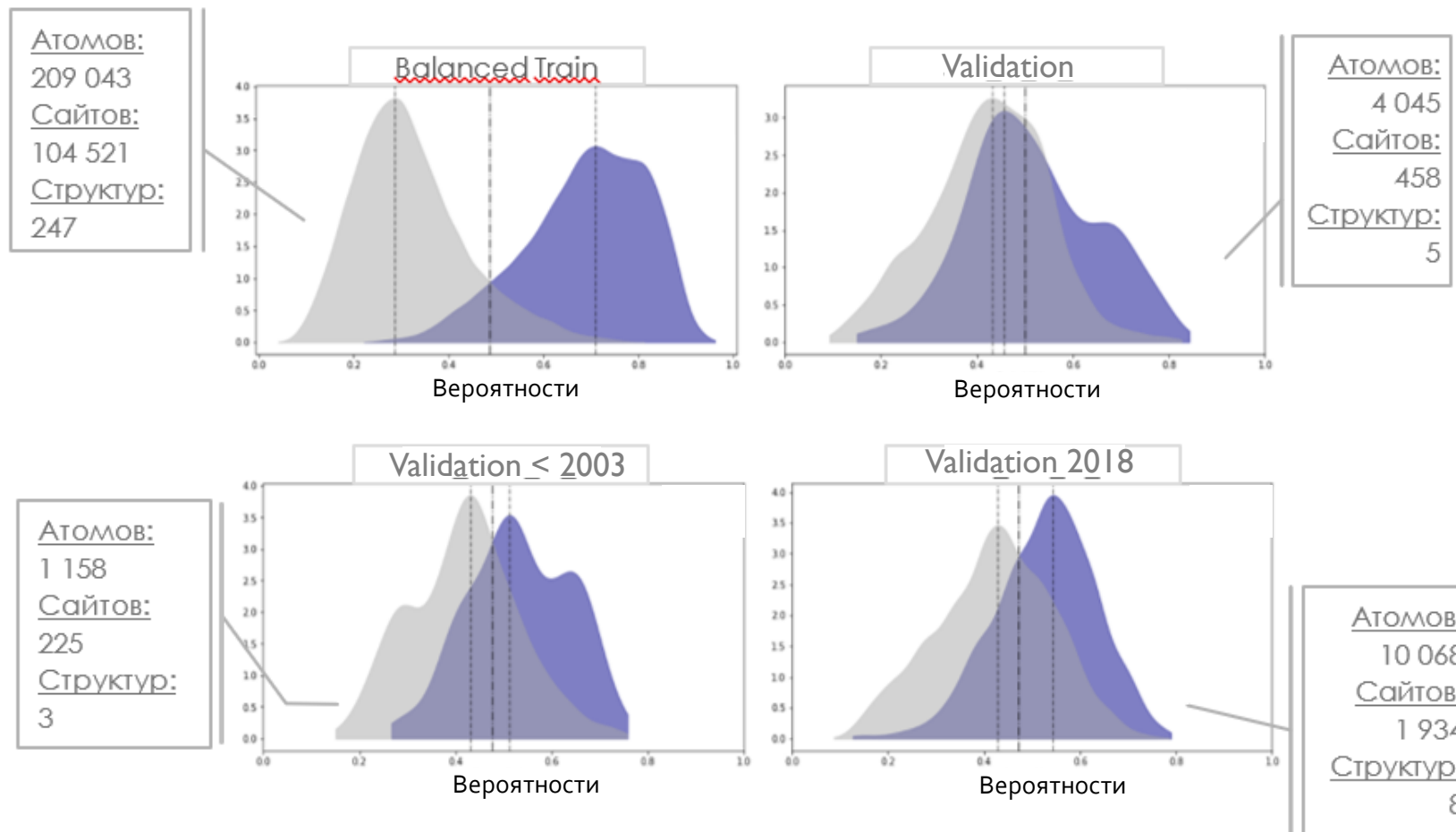
Validation
(часть исходной выборки)

Validation 2018

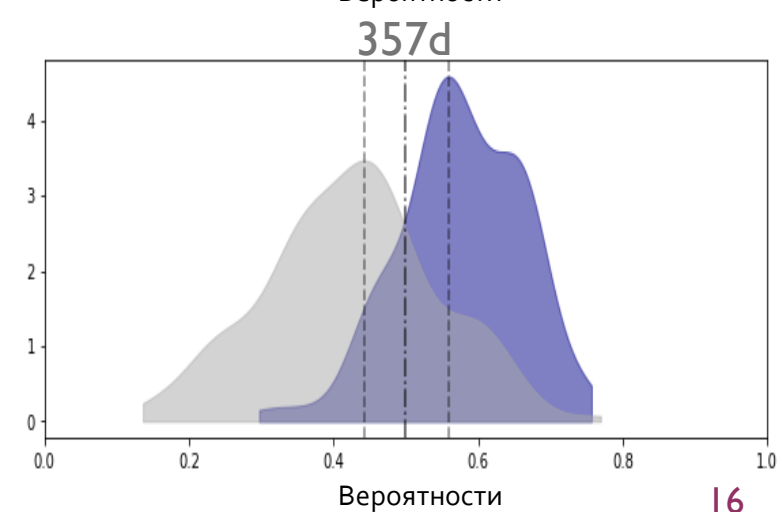
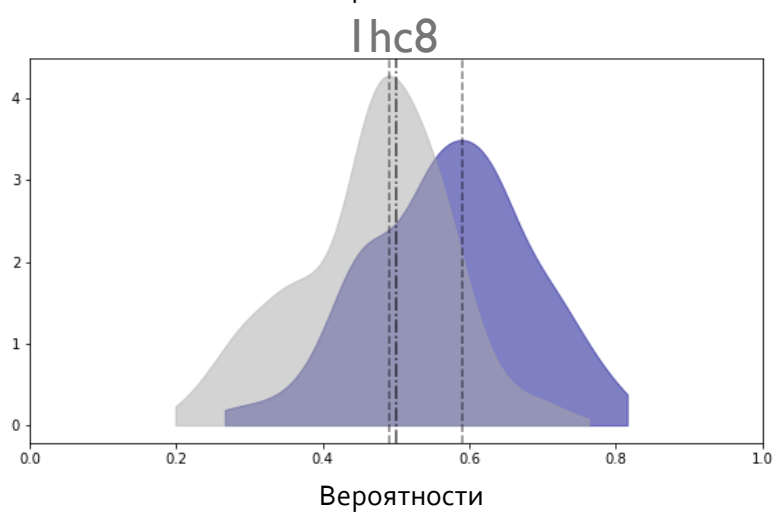
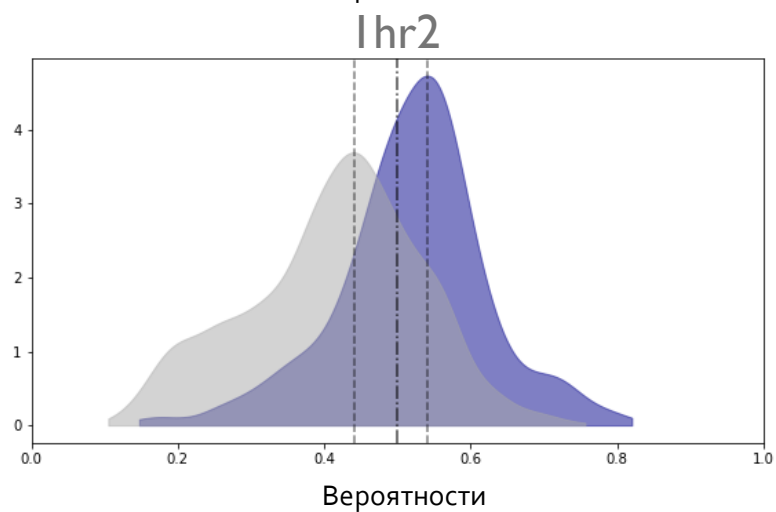
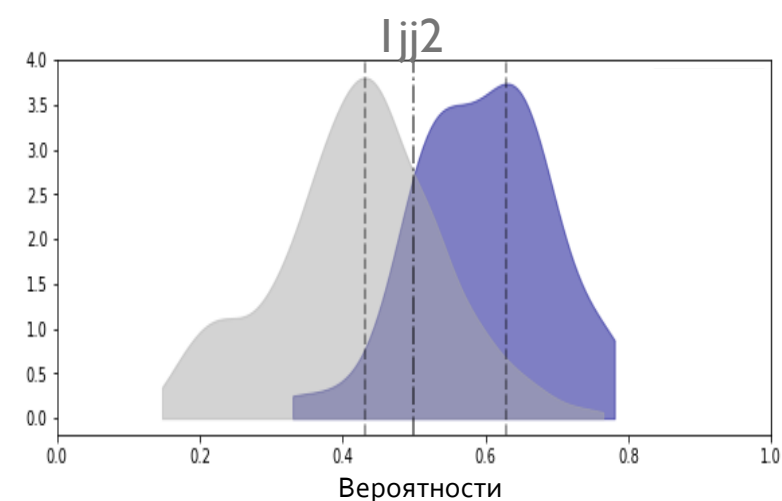
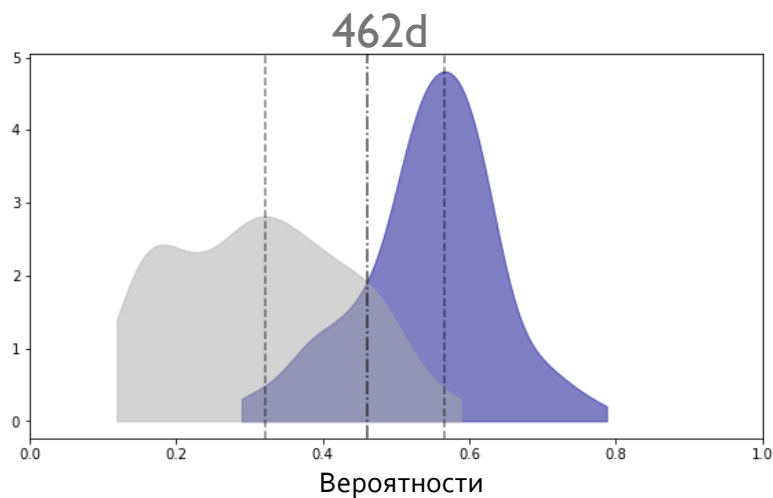
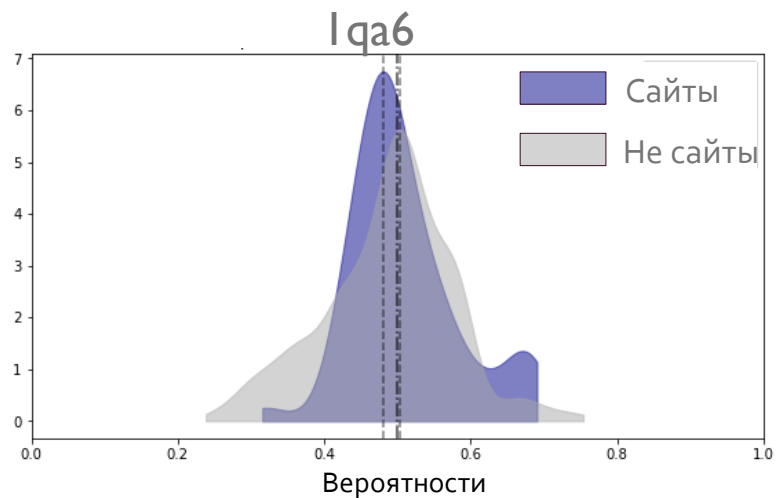


Распределения вероятностей

Сайты
Не сайты



ВАЛИДАЦИЯ ОТДЕЛЬНЫХ ЦЕПОЧЕК

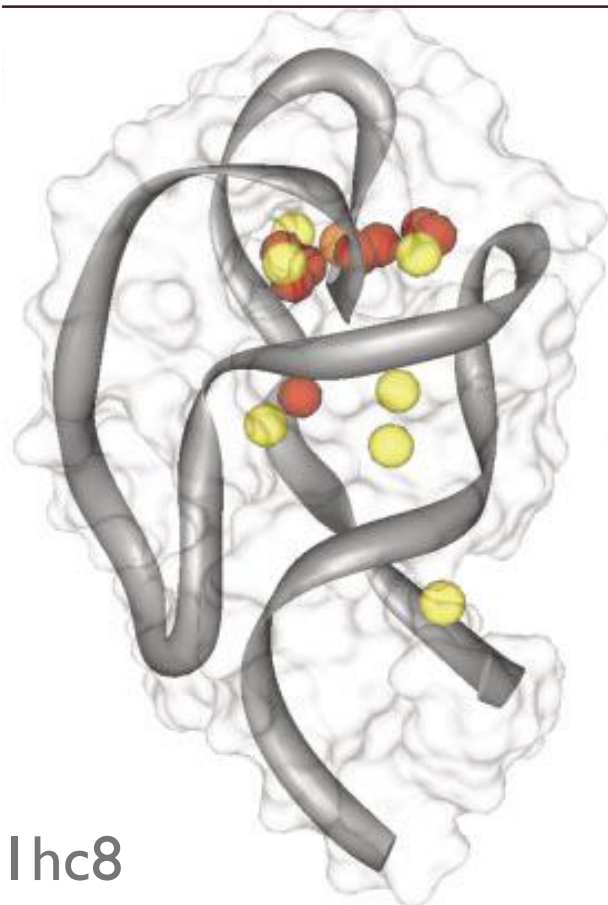


СРАВНЕНИЕ С СУЩЕСТВУЮЩИМИ СЕРВИСАМИ

WebFeature

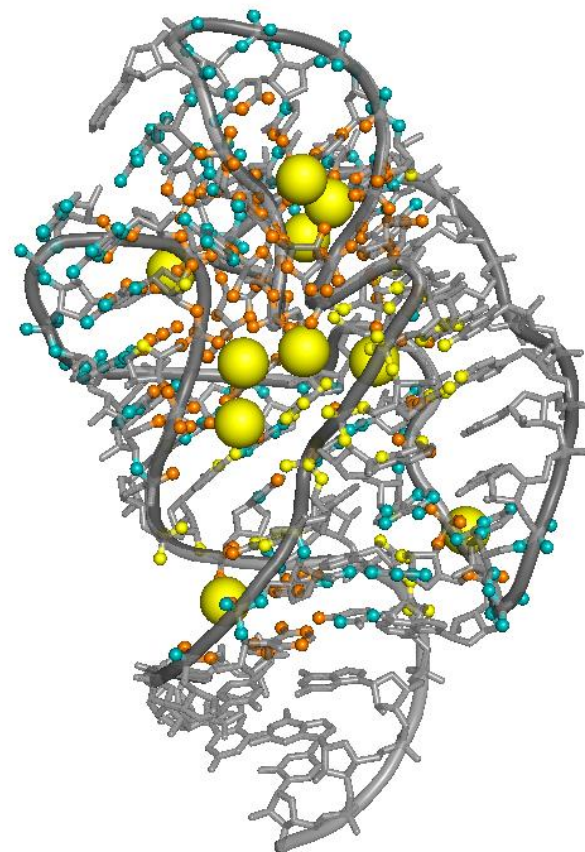
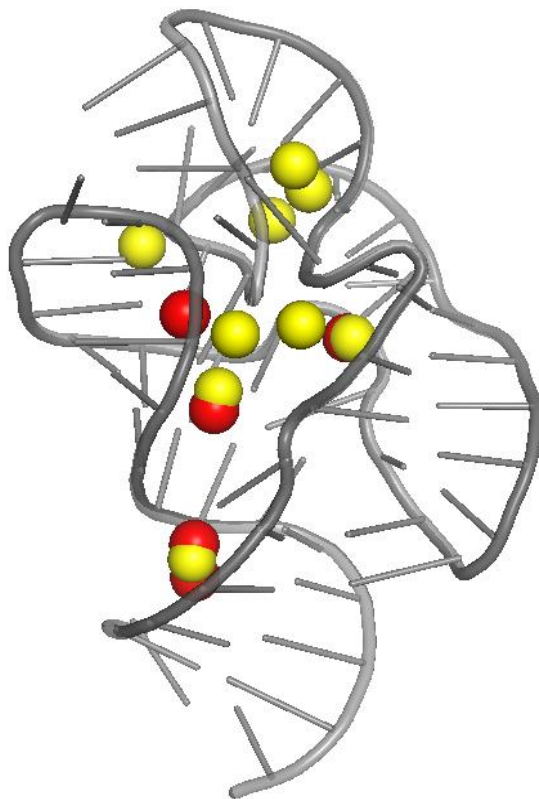
Metallon

RandomForest



Ihc8

(58nt sequence 23s rRNA Bacillus stearothermophilus)




Ион магния

 Реальный

 Предполагаемый

Атом, близкий к иону

 Неверно отмечен

 Верно отмечен

 Не отмечен

АППРОКСИМАЦИЯ КООРДИНАТ ИОНОВ С ПОМОЩЬЮ K-MEANS

k-means разделяет атомы, предсказанные RandomForest, на k групп.

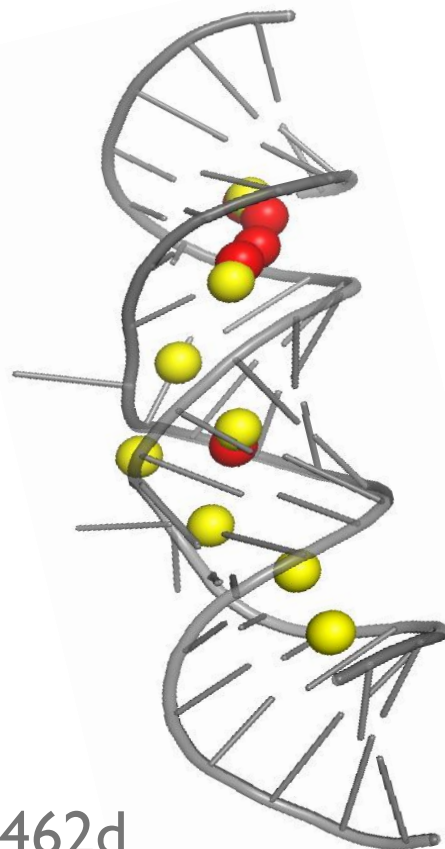
Оптимальное k определяется по качеству покрытия:

$$\frac{|\{a \in A | \min(d(a, c)) \leq 7, c \in C\}|}{|A|},$$

где A – множество предсказанных атомов,
C – множество центров кластеров.

Оптимальный порог покрытия = 0.85

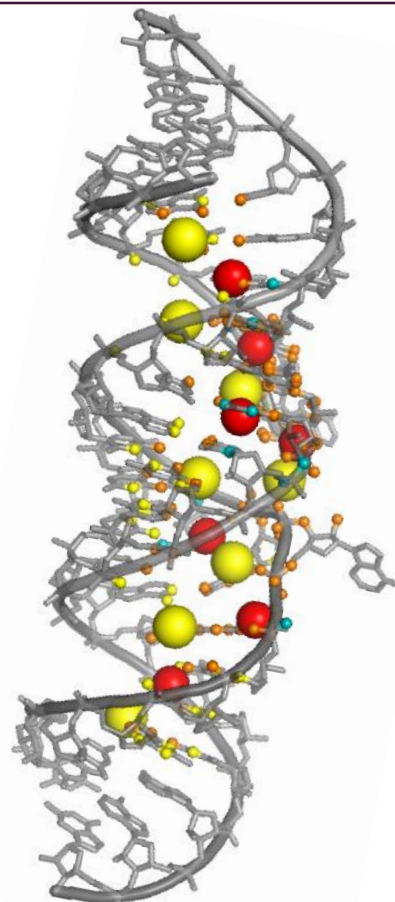
Metallon



462d

(HIV-I genomic RNA dimerization initiation site)

RandomForest



Ион магния

● Реальный

● Предполагаемый

Атом, близкий к иону

● Верно отмечен

● Неверно отмечен

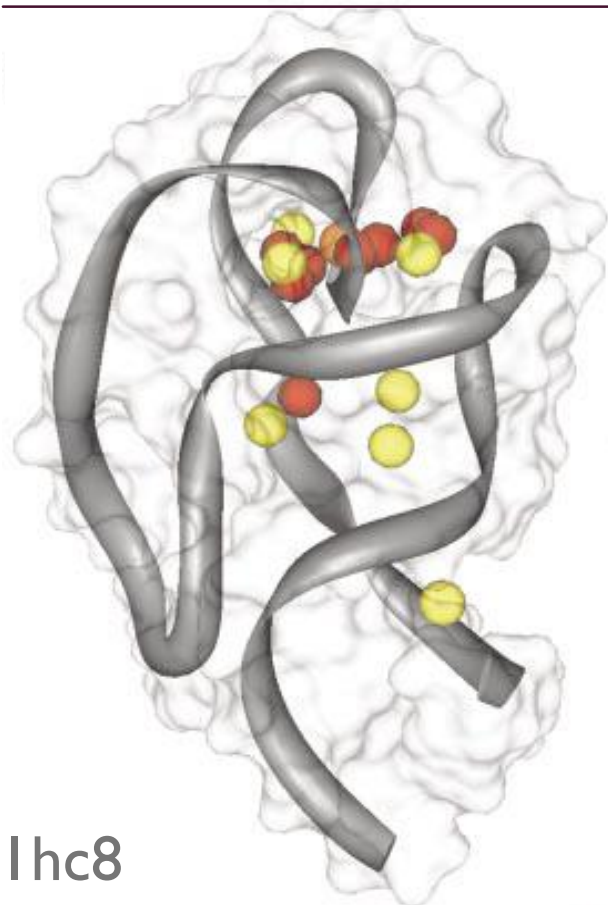
● Не отмечен

АППРОКСИМАЦИЯ КООРДИНАТ ИОНОВ С ПОМОЩЬЮ K-MEANS

WebFeature

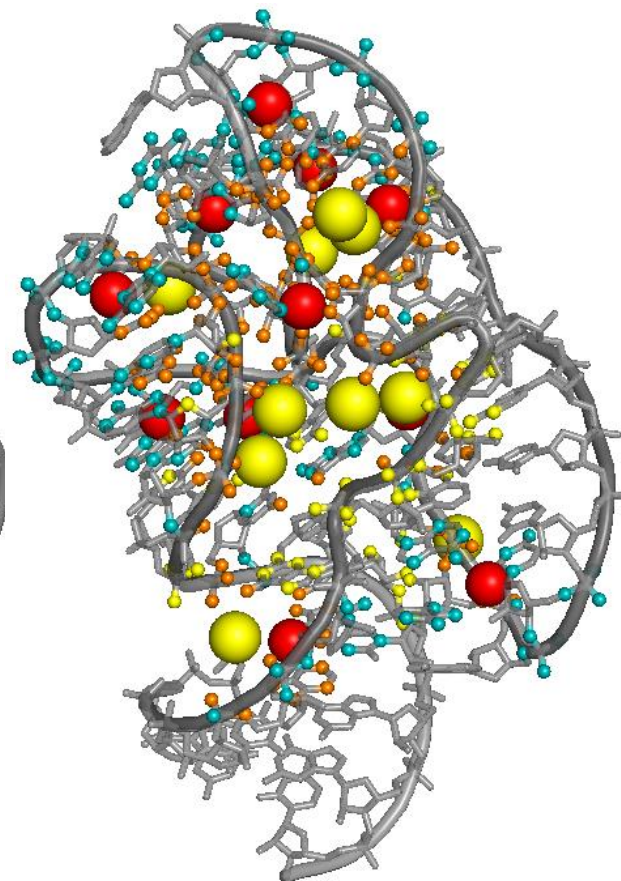
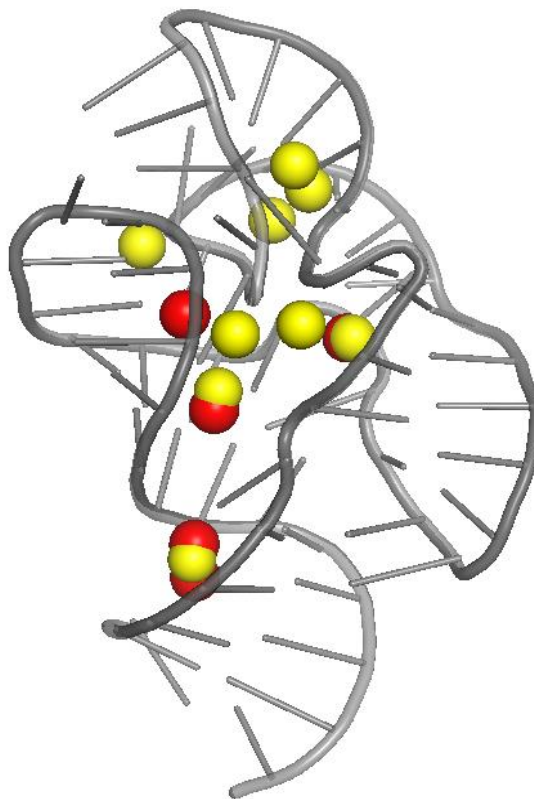
Metallon

RandomForest



Ihc8

(58nt sequence 23s rRNA Bacillus stearothermophilus)



Ион магния

● Реальный

● Предполагаемый

Атом, близкий к иону

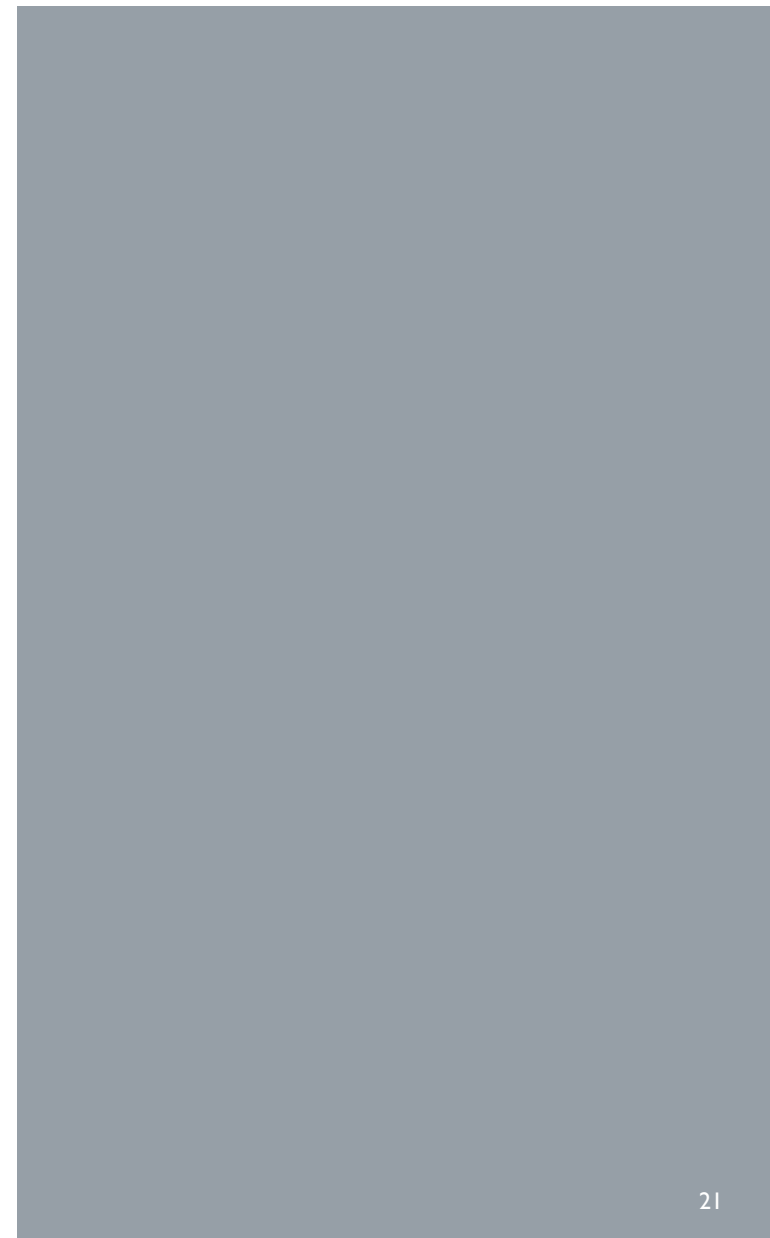
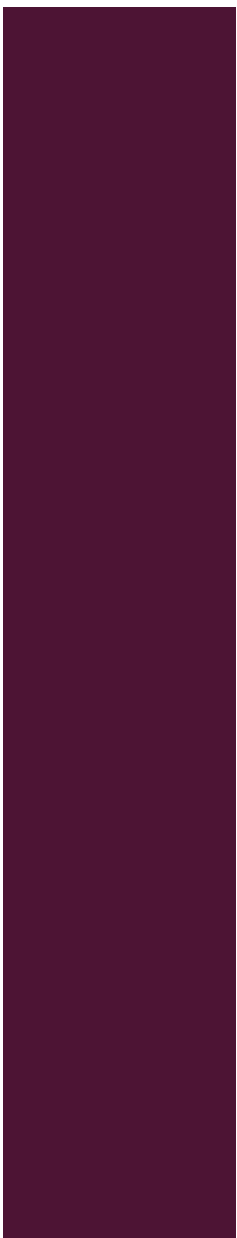
● Верно отмечен

● Неверно отмечен

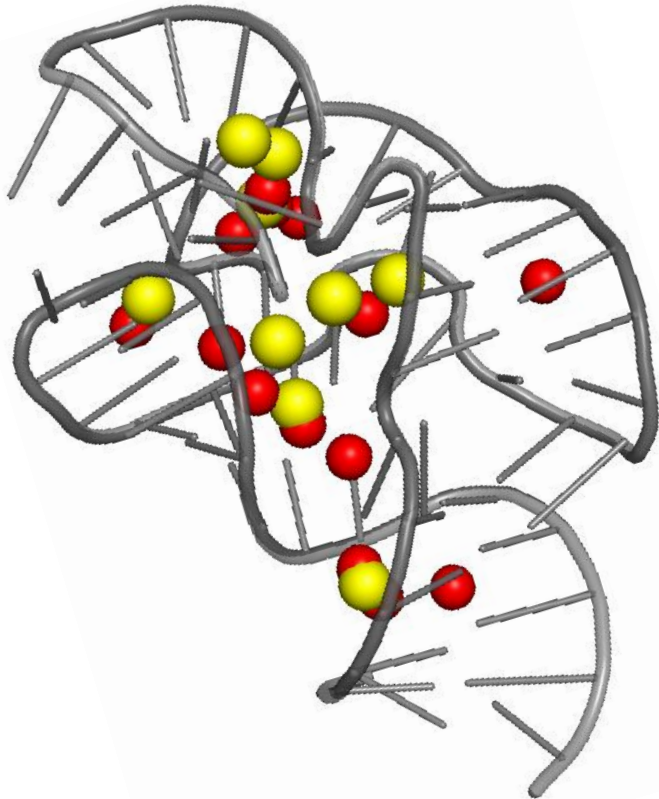
● Не отмечен

ЗАКЛЮЧЕНИЕ

- Таким образом, в ходе этой работы был разработан алгоритм, который для некоторых структур способен достаточно точно определять количество ионов магния, а также вычислять их приближенные координаты.
- На данный момент не существует универсального алгоритма, точно определяющего координаты ионов магния для всех существующих структур.
- В дальнейшем стоит выделить группы структур, для которых координаты ионов магния хорошо распознаются, и тренировать алгоритмы для этих групп.



МЕТАЛИОН RNA С ЗАДАННЫМ ЧИСЛОМ ИОНОВ



1hc8

(58nt sequence 23s rRNA *Bacillus stearothermophilus*)



462d

(HIV-1 genomic RNA dimerization initiation site)

РАССТОЯНИЯ ОТ ИОНОВ ДО ФРАГМЕНТОВ НУКЛЕОТИДОВ

