

Sequence generation

DataCon 3.0. Design a Peptide Vector for Drug Delivery



Agenda

Data

Подготовка датасета 04

База данных 05

Analyze

Part 2

Выравнивание 07

Поиск консервативных последовательностей 08

Models

Part 3

Предсказывание взаимодействия с клеточной мембраной 10

ESM3 11

Validation 13

Q&A

Part 4

Data

Подготовка данных

POSEIDON dataset

1. Конвертация массы в микромоли
2. Перевод времени в минуты
3. Также остальные колонки были приведены в нужные типы
4. Удаление дублей данных

Balanced_dataset.txt

1. Установка правильного парсинга данных
2. Формирование дескрипторов для последовательностей

```
Crot (27-39),NIH-3T3 cells,FITC,21319732,3176,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KMDCRWRWKCCCC  
Crot (27-39) derivative 1,NIH-3T3 cells,FITC,21319732,3344,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,MDCRWRWKCCCC  
Crot (27-39) derivative 2,NIH-3T3 cells,FITC,21319732,1909,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,DCRWRWKCCCC  
CyLoP-1,NIH-3T3 cells,FITC,21319732,4050,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,CRWRWKCCCC  
Crot (27-39) derivative 3,NIH-3T3 cells,FITC,21319732,1511,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,RWRWKCCCC  
Crot (27-39) derivative 4,NIH-3T3 cells,FITC,21319732,1516,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KMDCRWRWKCCCC  
Crot (27-39) derivative 5,NIH-3T3 cells,FITC,21319732,1202,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KMDCRWRWKKKK  
Crot (27-39) derivative 6,NIH-3T3 cells,FITC,21319732,100,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KMDRWRKKK  
Crot (27-39) derivative 7,NIH-3T3 cells,FITC,21319732,1043,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KDCRWRWKCCCC  
Crot (27-39) derivative 8,NIH-3T3 cells,FITC,21319732,1583,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KCRWRWKCCCC  
Crot (27-39) derivative 9,NIH-3T3 cells,FITC,21319732,1226,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,KRWRWKCCCC  
Crot (27-39) derivative 10,NIH-3T3 cells,FITC,21319732,1937,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,MDCRWRWKXCKK  
Crot (27-39) derivative 11,NIH-3T3 cells,FITC,21319732,1741,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,DCRWRWKXCKK  
Crot (27-39) derivative 12,NIH-3T3 cells,FITC,21319732,943,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,CRWRWKXCKK  
Crot (27-39) derivative 13,NIH-3T3 cells,FITC,21319732,2044,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,CRWRWKXCKK  
Crot (27-39) derivative 14,NIH-3T3 cells,FITC,21319732,1347,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,CRWRWKXCKK  
Crot (27-39) derivative 15,NIH-3T3 cells,FITC,21319732,1211,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,RWRWKXCKK  
Crot (27-39) derivative 16,NIH-3T3 cells,FITC,21319732,1256,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,MDCRWRWKXXXX  
Crot (27-39) derivative 17,NIH-3T3 cells,FITC,21319732,872,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,DCRWRWKXXXX  
Crot (27-39) derivative 18,NIH-3T3 cells,FITC,21319732,1390,Mean Fluorescence intensity,2.5 uM,18h,37°C,Fluorescence spectroscopy,Cellular uptake,CRWRWKXXXX
```

```
>14|0  
HHHHHHESGGGGSPGRRRRRRRRRR  
>15|0  
ARRRAARAARRRAARAARRRAARAARRRAARA  
>16|0  
ARRARAARRARAARRARAARRARAARRARA  
>17|0  
AKKAKAAKKAKAAKKAKAAKKAKAAKKAKA  
>18|0  
RARARARARARARARARARARARARARARAR  
>19|0  
YTFGLKTSFNVQYTFGLKTSFNVQ  
>20|0  
RQIKIWFQNRRMKWKKRQIKIWFQNRRMKWK  
>21|0  
LSTAADMQGVVTDGMASGLDKDYLKPDD  
>22|0  
GYLLGHINLHHLAHLHHILC
```

Структура хранения

Peptides

Cpp and non-cpp

Uptake (decimal)

Conc (decimal)

Time (ms, int)

Sequence (varchar)

Type (varchar)

Method (varchar)

Generates

Generate model

human_input (varchar)

sequence (varchar)

coordinates (jsonb)

sequence_prompt (varchar)

sasa_prompt (jsonb)

explanation (decimal)

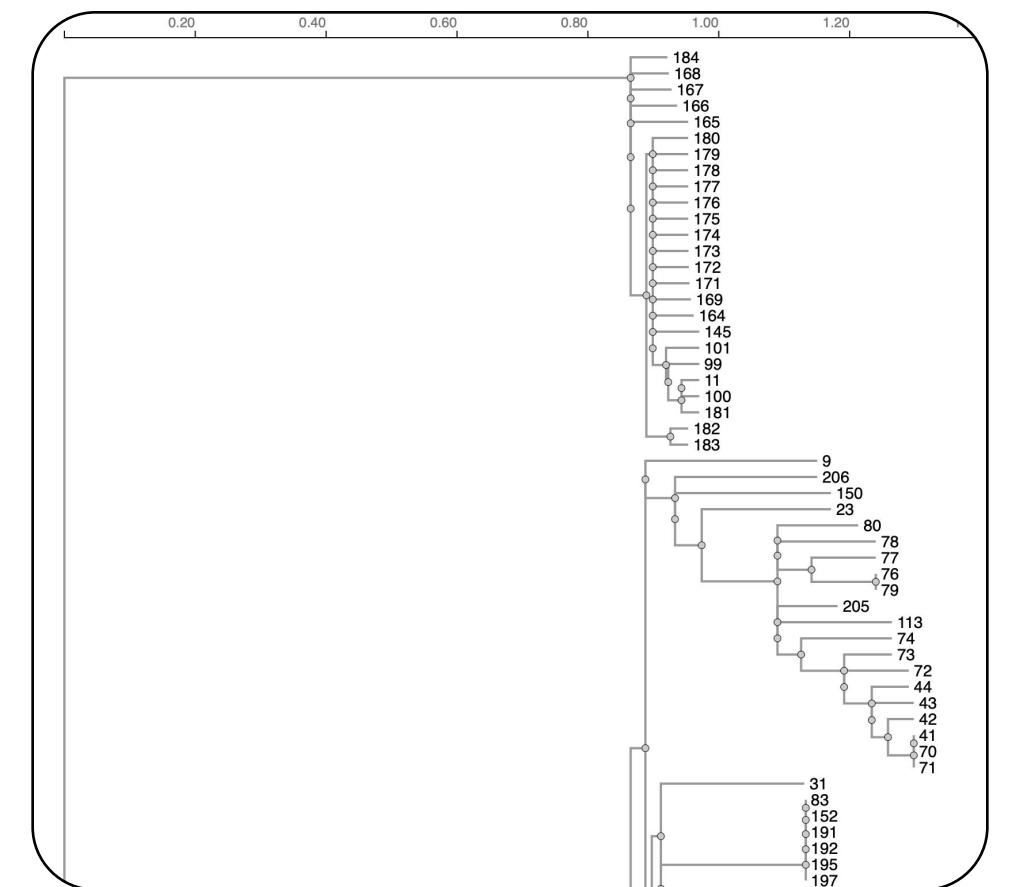
validation (jsonb)

created_at (timezonetz)

Analyze

Выравнивание

Для выравнивания аминокислотных последовательностей был применен алгоритм множественного выравнивания CLUSTAL



172	-	-RLWMRAYSPTTR-RYG-	15
171	-	-RLWMAWYSPTTR-RYG-	15
169	-	-RLWARWYSPTTR-RYG-	15
164	-	-RLAMRWYSPTTR-RYG-	15
145	-	-RAWMRWYSPTTR-RYG-	15
101	-	-KLWMRWYSPWTR-RYG-	15
99	-	-KLWMRWYSATTR-RYG-	15
11	-	-ALWMRWYSPTTR-RYG-	15
100	-	-KLWMRWYSPTTR-RYG-	15
181	-	-RLWMRWYSPTTR-RYG-	15
182	-	-RLWMRWYSPWTR-RWG-	15
183	-	-RLWMRWYSPWTR-RYG-	15
9	-	-AGSHRRL-----	7
206	-	-RRRRWW-----	7
150	-	-RGDGPRRRPRKRRGR-----	15
23	-	-CHHRRRRHHC-----	10
80	-	-HHH-----HHHHRRRRRRRRR-----	17
78	-	-HHHHHH-----HHHHRRRRRRRRR-----	21
77	-	-HHHHHHHHHH-----HHHHRRRRRRRRR-----	25
76	-	-HHHHHHHHHHHH-----HHHHRRRRRRRRR-----	29
79	-	-HHH-----HHHHRRRRRRRR-----	16
205	-	-RRRRRRR-----	8
113	-	-LILIGRRR-RRRRRGC-----	15
74	-	-GSVSRRRRRRGG-----	12
73	-	-GRRR-RRRRRRR-----	11
72	-	-GRRR-RRRRRRR-----	11
44	-	-FF-----LIPKGRRR-RRRRRR-----	16
43	-	-FF-----LIPKGRRR-RRRRRGC-----	17
42	-	-FFGRR--RRRRRG-----	12
41	-	-FFFFGRRR-RRRRRGC-----	15
70	-	-GRRR-R-----	5
71	-	-GRRR-RRRRR-----	9
31	-	-CRLRLH-----LRHHYRRRWHR-----	17

Поиск консервативных последовательностей

При сравнении самых часто встречающихся k-mer у генеральной выборки и у выборки, которая содержит топ 10% последовательностей по uptake, можно увидеть разницу. Например для 3-mer это LKK

	3gram_top	count_3gram	3gram_all	count_3gram	4gram_top	count_4gram	4gram_all	count_4gram	5gram_top	count_5gram	5gram_all	count_5gram
0	RRR	7.0	RRR	68.0	RRRR	4.0	RRRR	34.0	RRRRR	4.0	RRRRR	29.0
1	KKR	5.0	KKR	39.0	PKKK	3.0	RKKR	32.0	LKKLR	2.0	KRRQR	22.0
2	LKK	4.0	KRR	36.0	KLKK	3.0	RRQR	26.0	RRKLK	2.0	KKRRQ	21.0
3	RRK	4.0	RKK	35.0	RRKL	3.0	KKRR	24.0	KLKKL	2.0	RRQRR	21.0
4	KKK	4.0	RRQ	31.0	KKRR	3.0	RQRR	24.0	PKKKR	2.0	RKKRR	21.0
5	KRK	4.0	RQR	30.0	KWRR	2.0	KRRQ	23.0	KKRKV	2.0	RQRRR	20.0
6	GRK	4.0	QRR	28.0	LKKL	2.0	QRRR	22.0	KKKRK	2.0	TTRYG	19.0
7	RKK	4.0	GRK	25.0	KRKV	2.0	TTRY	20.0	WRRKL	2.0	GRKKR	18.0
8	KRR	4.0	YSP	24.0	RKLK	2.0	RRYG	20.0	RKLKK	2.0	YSPTT	17.0
9	PKK	3.0	TRR	23.0	KKLR	2.0	PTTR	18.0	KWRRK	2.0	PTTRR	17.0

Просматривая результаты выравнивания можно заметить закономерности, а именно консервативные последовательности

184	--RLYMRYYSPPTTR-RYG-	15
168	--RLVMRVYSPPTTR-RYG-	15
167	--RLLMRLYSPPTTR-RYG-	15
166	--RLIMRIYSPPTTR-RYG-	15
165	--RLFMRFYSPPTTR-RYG-	15
180	--RLWMRWYSPPTTR-RYA-	15
179	--RLWMRWYSPPTTR-RAG-	15
178	--RLWMRWYSPPTTR-AYG-	15
177	--RLWMRWYSPPTTA-RYG-	15
176	--RLWMRWYSPATAR-RYG-	15
175	--RLWMRWYSPATR-RYG-	15
174	--RLWMRWYSPATTR-RYG-	15
173	--RLWMRWASPTTR-RYG-	15
172	--RLWMRWASPTTR-RYG-	15
171	--RLWMAWYSPPTTR-RYG-	15
169	--RLWARWYSPPTTR-RYG-	15
164	--RLAMRWYSPPTTR-RYG-	15
145	--RAWMRWYSPPTTR-RYG-	15
101	--KLWMRWYSPWTR-RYG-	15
99	--KLWMRWYSATTR-RYG-	15
11	--ALWMRWYSPPTTR-RYG-	15
100	--KLWMRWYSPPTTR-RYG-	15
181	--RLWMRWYSPPTTR-RYG-	15
182	--RLWMRWYSPWTR-RWG-	15
183	--RLWMRWYSPWTR-RYG-	15

Models

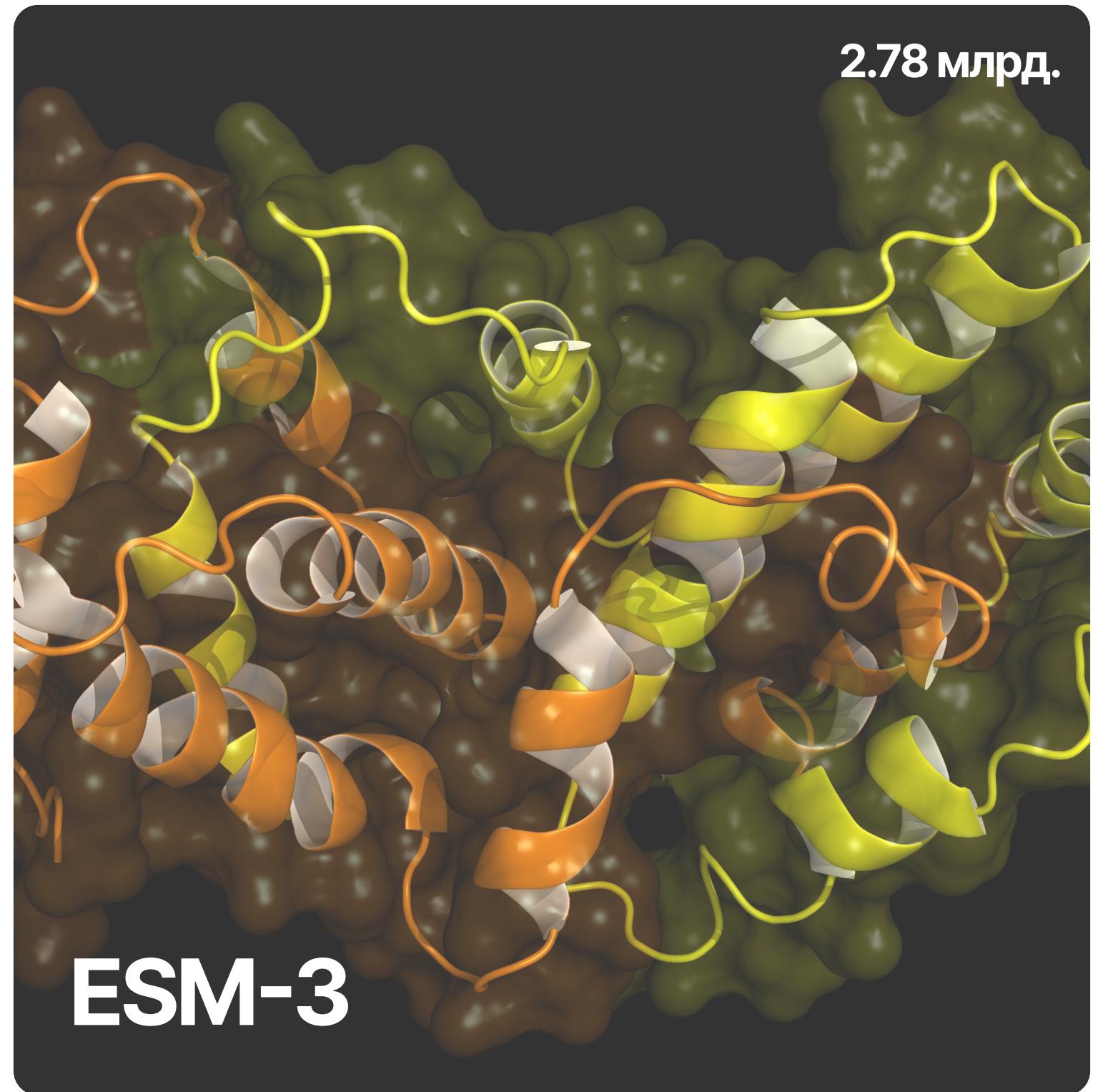
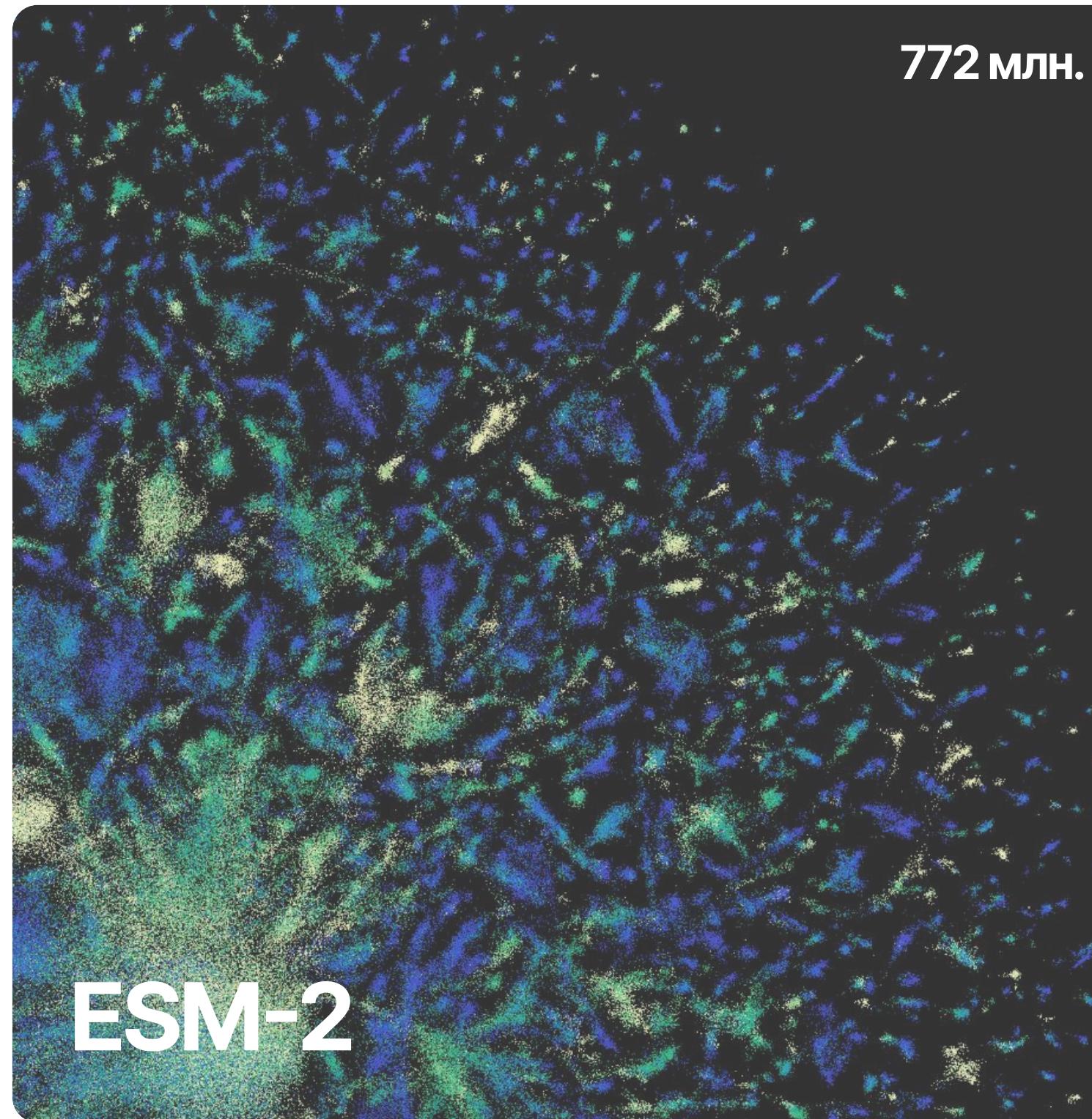
Предсказывание взаимодействия с клеточной мембраной

На датасете Balanced_dataset.txt была обучена модель, которая определяет, будет ли связывание с клеточной мембраной у последовательности аминокислот или нет

Classification Report:					
	precision	recall	f1-score	support	
0	0.92	0.91	0.92	93	
1	0.91	0.92	0.92	91	
accuracy			0.92	184	
macro avg	0.92	0.92	0.92	184	
weighted avg	0.92	0.92	0.92	184	

Поиск моделей

Сравнение



DAG



Validation

Валидация условной генерации CPP: параметры и инструменты

Молекулярная масса

Библиотека: Biopython
Значимость: Влияет на способность проникновения через мембрану
Целевой диапазон: обычно < 5 кДа

Изоэлектрическая точка (pI)

Библиотека: Biopython
Значимость: Определяет заряд пептида при физиологическом pH
Целевой диапазон: обычно > 7

Заряд

Библиотека: modIAMP
Значимость: Необходим для электростатического взаимодействия с мембраной
Целевое значение: обычно $\geq +2$ при pH 7.4

Гидрофобность

Библиотека: modIAMP
Значимость: Влияет на взаимодействие с липидным бислоем
Целевой диапазон: зависит от шкалы (например, ≤ 0.5 по Kyte & Doolittle)

Длина последовательности

Библиотека: встроенные функции Python
Значимость: Критична для эффективности и синтеза
Целевой диапазон: обычно 5-30 аминокислот

Предсказанная активность проникновения

Библиотека: scikit-learn (или другие ML библиотеки)
Значимость: Оценка потенциальной CPP активности
Целевое значение: зависит от модели (например, вероятность > 0.7)

Сходство с известными CPP

Библиотека: Biopython
Значимость: Баланс между новизной и сохранением CPP-подобных свойств
Целевой диапазон: обычно < 80% идентичности

Специфичность к типу клеток

Библиотека: Biopython
Значимость: Оценка целевой специфичности
Целевой диапазон: высокое сходство с CPP для конкретного типа клеток

Стабильность вторичной структуры

Библиотека: Biopython
Значимость: Влияет на эффективность проникновения
Целевой диапазон: обычно > 30% спиральных структур

Взаимодействие с мембраной

Библиотека: MDAnalysis
Значимость: Оценка эффективности взаимодействия с целевой мембраной
Целевой диапазон: значительное время контакта в симуляциях (например, > 50%)

ManuL Team



Андрей
Тиников

Data Analytic



Данил
Кочелаков

Chemist



Сергей
Волчков

Data Scientist



Илья
Морозов

Python developer

Q&A