GENEHACK 2020

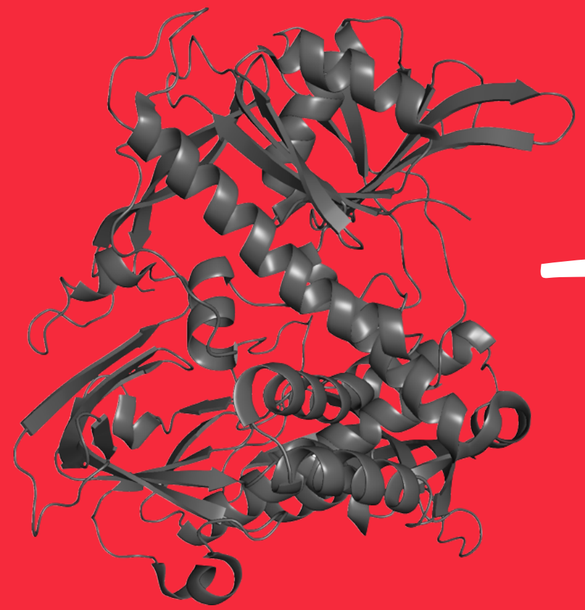
PROPHILE29

OMG
Мыларщиков Д., Камышева А., Азбукина Н., Зинкевич А.
Факультет биоинженерии и биоинформатики МГУ

ИДЕЯ:

ПРЕДСКАЗАНИЕ Ф29-ПОДОБНОЙ ПОЛИМЕРАЗНОЙ АКТИВНОСТИ НА ОСНОВЕ ЗНАЧИМЫХ ДЛЯ ФУНКЦИОНИРОВАНИЯ АМИНОКИСЛОТНЫХ ОСТАТКОВ

Структуро-ориентированное выравнивание

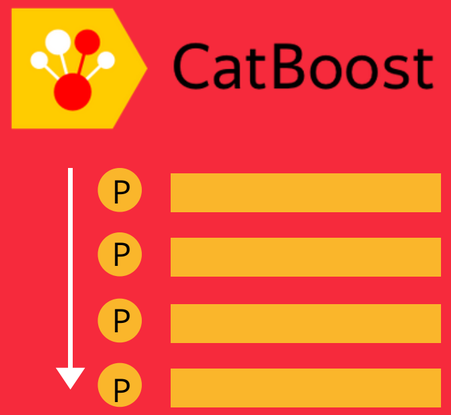


Структуры полимеразы из фагов Ф29 и RB69

Важные для функционирования остатки

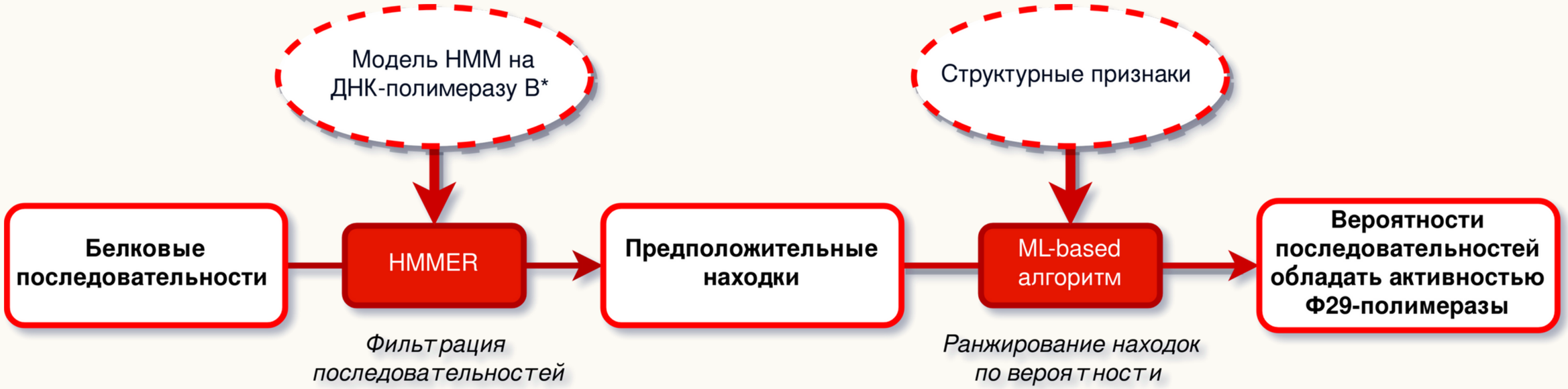
Последовательности полимераз из более широкого семейства

Предсказание активности Ф29 по последовательности



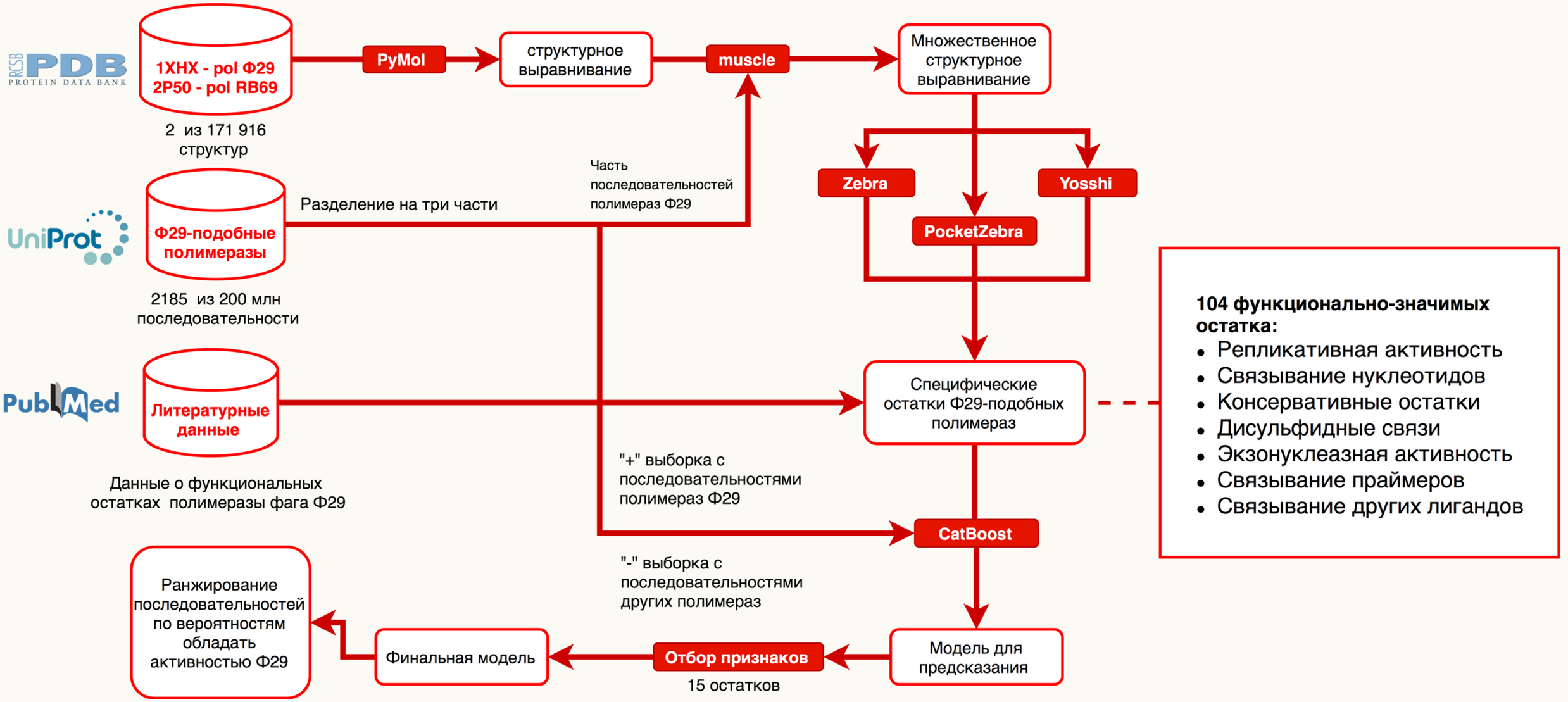
Ранжированный по вероятности список последовательностей белков с предполагаемой активностью Ф29

АЛГОРИТМ ПРЕДСКАЗАНИЯ



* Полимераза Ф29 входит в класс ДНК-полимераз B

МОДЕЛЬ ПРЕДСКАЗАНИЯ Ф29-ПОДОБНОЙ АКТИВНОСТИ С УЧЕТОМ ФУНКЦИОНАЛЬНОЙ АННОТАЦИИ ОСТАТКОВ

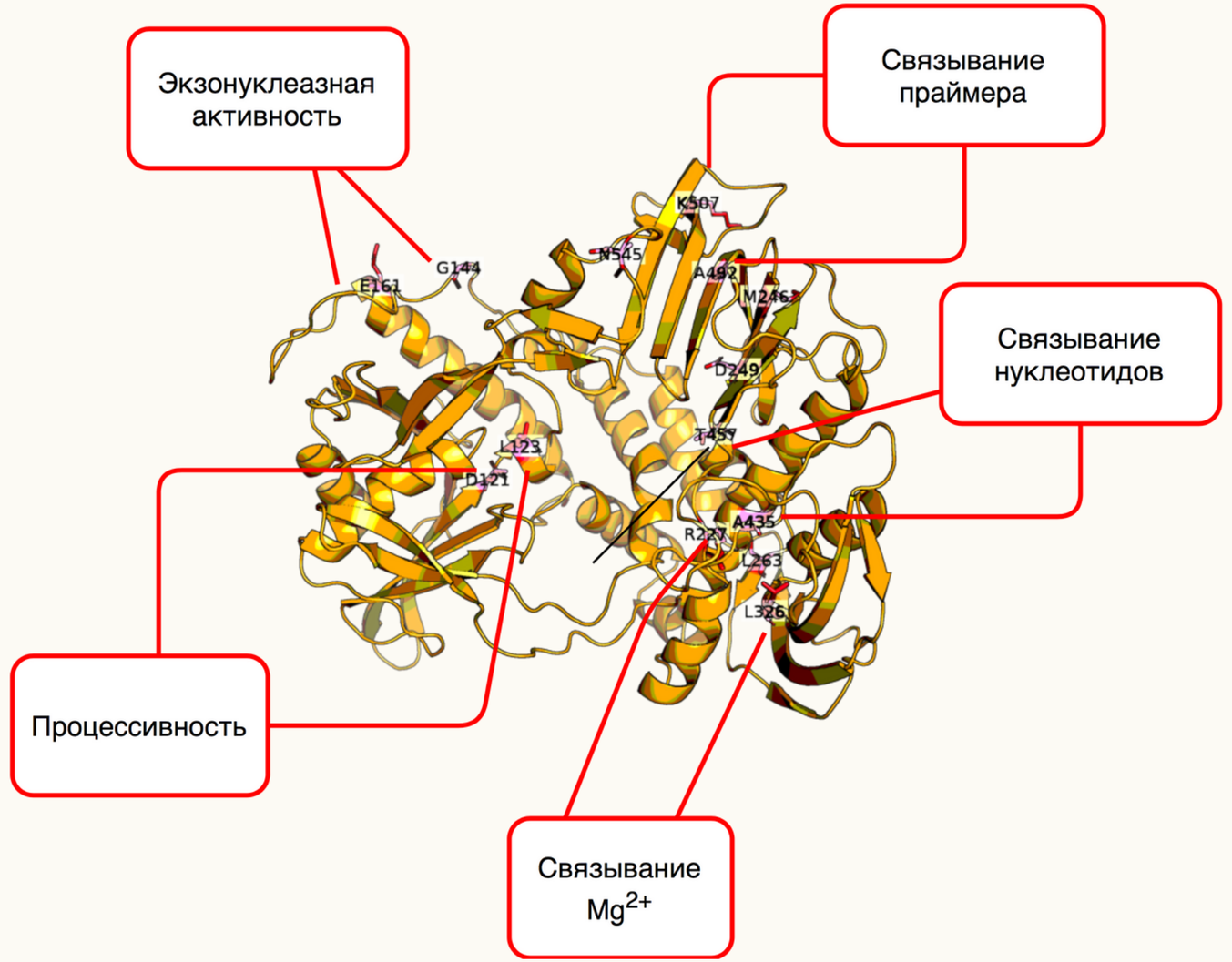


КАЧЕСТВО МОДЕЛИ

Точность предсказания на тестовом датасете - **0.98**

- Из **141 млн бактериальных** белковых последовательностей в **Uniprot** у **95** была предсказана Ф29-подобная активность
- Из **107 780 вирусных белковых** последовательностей в **Uniprot** за этот год у **76** была предсказана Ф29-подобная активность

ОСТАТКИ, СПЕЦИФИЧНЫЕ ДЛЯ Ф29-ПОДОБНОЙ ПОЛИМЕРАЗНОЙ АКТИВНОСТИ



Создан пайплайн PROPHILE29

200 млн

ЗАПИСЕЙ UNIPROT

Было рассмотрено при поиске Ф29-подобных полимераз

2185

БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Отобраны для построения референсного выравнивания и обучения модели **CatBoost**

104

ФУНКЦИОНАЛЬНО-ЗНАЧИМЫХ ОСТАТКА В СТРУКТУРАХ ПОЛИМЕРАЗ

Отобраны для модели по литературе и при аннотации по структурному выравниванию

15

АМИНОКИСЛОТНЫХ ОСТАТКОВ

Вносят наибольший вклад в предсказание активности полимеразы **Ф29**. Они обладают разными функциями и расположением в белке.

76

ПОТЕНЦИАЛЬНЫХ БЕЛКОВ С Ф29-ПОДОБНОЙ АКТИВНОСТЬЮ

Были отобраны из всех вирусных белков в **NCBI**, загруженных в этом году

ПАЙПЛАЙН PROPHILE29

Выложен в открытый доступ по адресу **https://github.com/dmitrymyl/PROphiLE29**, задокументирован и готов к скачиванию