

GENEHACK 2020

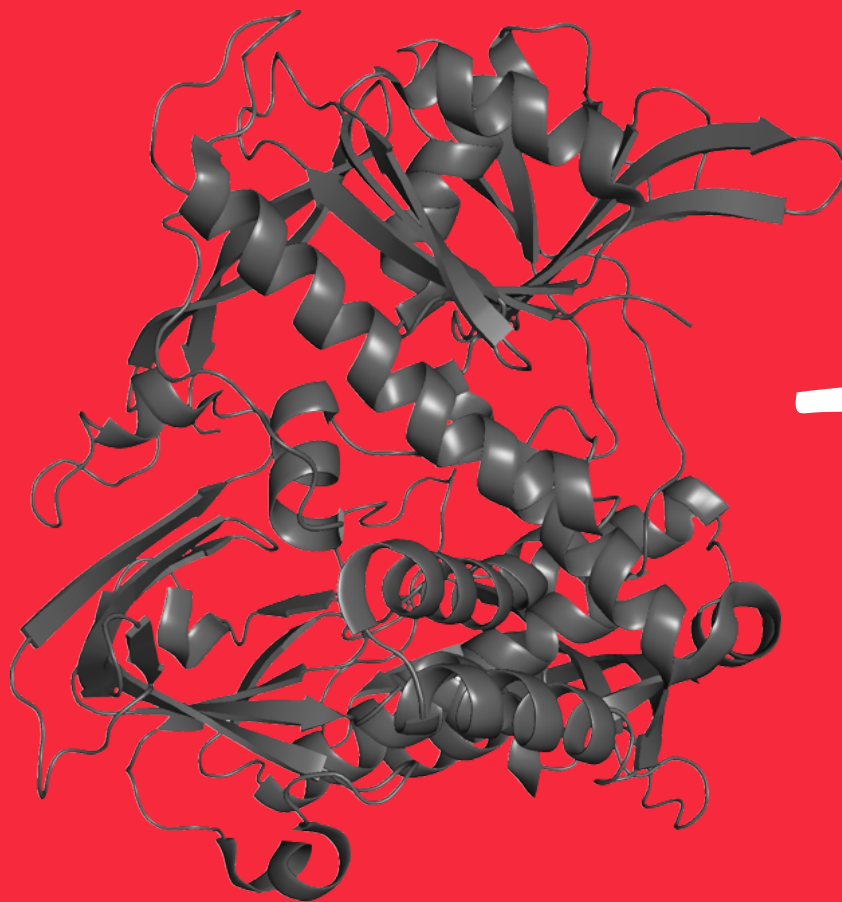
PROPHILE29



OMG

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

Структуро-ориентированное выравнивание



Важные для функционирования остатки

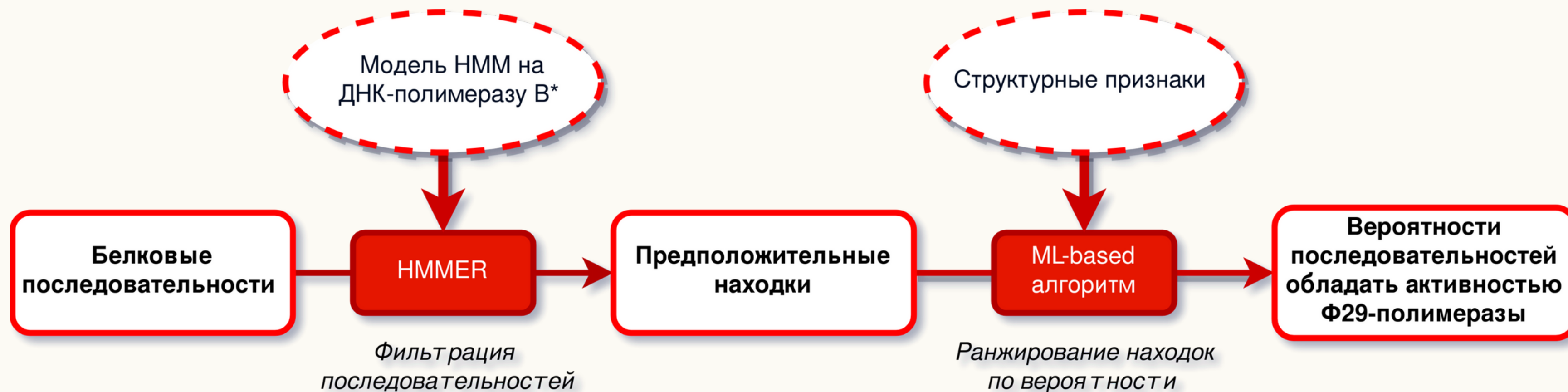
ASVG YFI KKKKYALMI I DDEGT RRD I
GTRGLFL KKKRYG ILKYWEDGFRLDT
GSRGLFL KKKRYAILKYWEDGFRLDV
GSRGLFL KKKRYAILKYWEDGFRLDV
GSRGLFL KKKRYAILKYWEDGFRLDE
GSRGLFL KKKRYAILKYWEDGFRLDE
GSRGLFL KKKRYAILKYWEDGFRLDV
GSRGLFL KKKRYAILKYWEDGFRLDV
GSRGLFL KKKRYAILKYWEDGFRLDI
GSRGLFL KKKRYAILKYWEDGFRLDI
GSRGLFL KKKRYAILKYWEDGFRLDI
GSRGLFL KKKRYAILKYWEDGFRLDI
GSRGLFL KKKRYAILKYWEDGFRLDI
GSRGLFL KKKRYAILKYWEDGFRLDI

Предсказание активности Ф29 по последовательности



Ранжированный по вероятности список последовательностей белков с предполагаемой активностью Ф29

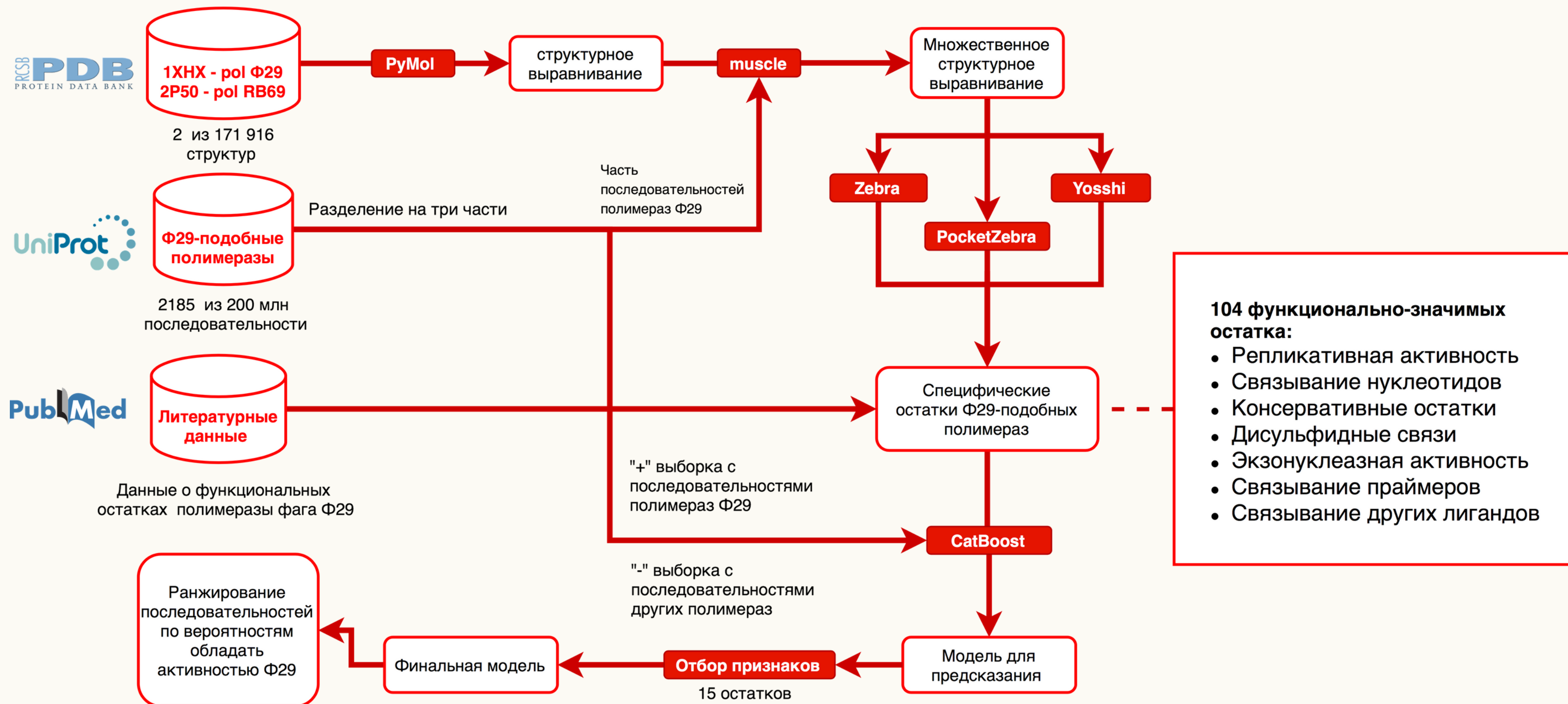
АЛГОРИТМ ПРЕДСКАЗАНИЯ



* Полимераза Φ29 входит в класс ДНК-полимераз B

Q	V	D	R	C	T	A	D	M	K	E	A	K	L	Y	H	R	R	H	K	V	C	E	V	H	A	K	A	S	S	V	F	L	S	G	L	N	Q	R	F	C	Q	Q	C	S	R	F	H	D	L	Q	E	F	D	E	A	K	R	S	C	R	R	R	L	A	G	H	N	E	R	R	R	K	S	S
Q	V	Y	G	C	S	K	D	L	S	S	S	K	D	Y	H	K	R	H	R	V	C	E	A	H	S	K	T	S	V	V	I	V	N	G	L	E	Q	R	F	C	Q	Q	C	S	R	F	H	F	L	S	E	F	D	D	G	K	R	S	C	R	R	R	L	A	G	H	N	E	R	R	R	K	P	A
Q	V	D	N	C	K	E	D	L	S	I	A	K	D	Y	H	R	R	H	K	V	C	E	V	H	S	K	A	T	K	A	L	V	G	K	Q	M	Q	R	F	C	Q	Q	C	S	R	F	H	L	L	S	E	F	D	E	G	K	R	S	C	R	R	R	L	D	G	H	N	R	R	R	R	K	T	Q
Q	V	E	S	C	T	A	D	M	S	K	A	K	Q	Y	H	K	R	H	K	V	C	Q	F	H	A	K	A	P	H	V	R	I	S	G	L	H	Q	R	F	C	Q	Q	C	S	R	F	H	A	L	S	E	F	D	E	A	K	R	S	C	R	R	R	L	A	G	H	N	E	R	R	R	K	S	T
Q	V	E	G	C	G	M	D	L	T	N	A	K	G	Y	Y	S	R	H	R	V	C	G	V	H	S	K	T	P	K	V	T	V	A	G	I	E	Q	R	F	C	Q	Q	C	S	R	F	H	Q	L	P	E	F	D	L	E	K	R	S	C	R	R	R	L	A	G	H	N	E	R	R	R	K	P	Q
Q	V	E	N	C	E	A	D	L	S	K	V	K	D	Y	H	R	R	H	K	V	C	E	M	H	S	K	A	T	S	A	T	V	G	G	I	L	Q	R	F	C	Q	Q	C	S	R	F	H	L	L	Q	E	F	D	E	G	K	R	S	C	R	R	R	L	A	G	H	N	K	R	R	R	K	T	N

МОДЕЛЬ ПРЕДСКАЗАНИЯ Ф29-ПОДОБНОЙ АКТИВНОСТИ С УЧЕТОМ ФУНКЦИОНАЛЬНОЙ АННОТАЦИИ ОСТАТКОВ

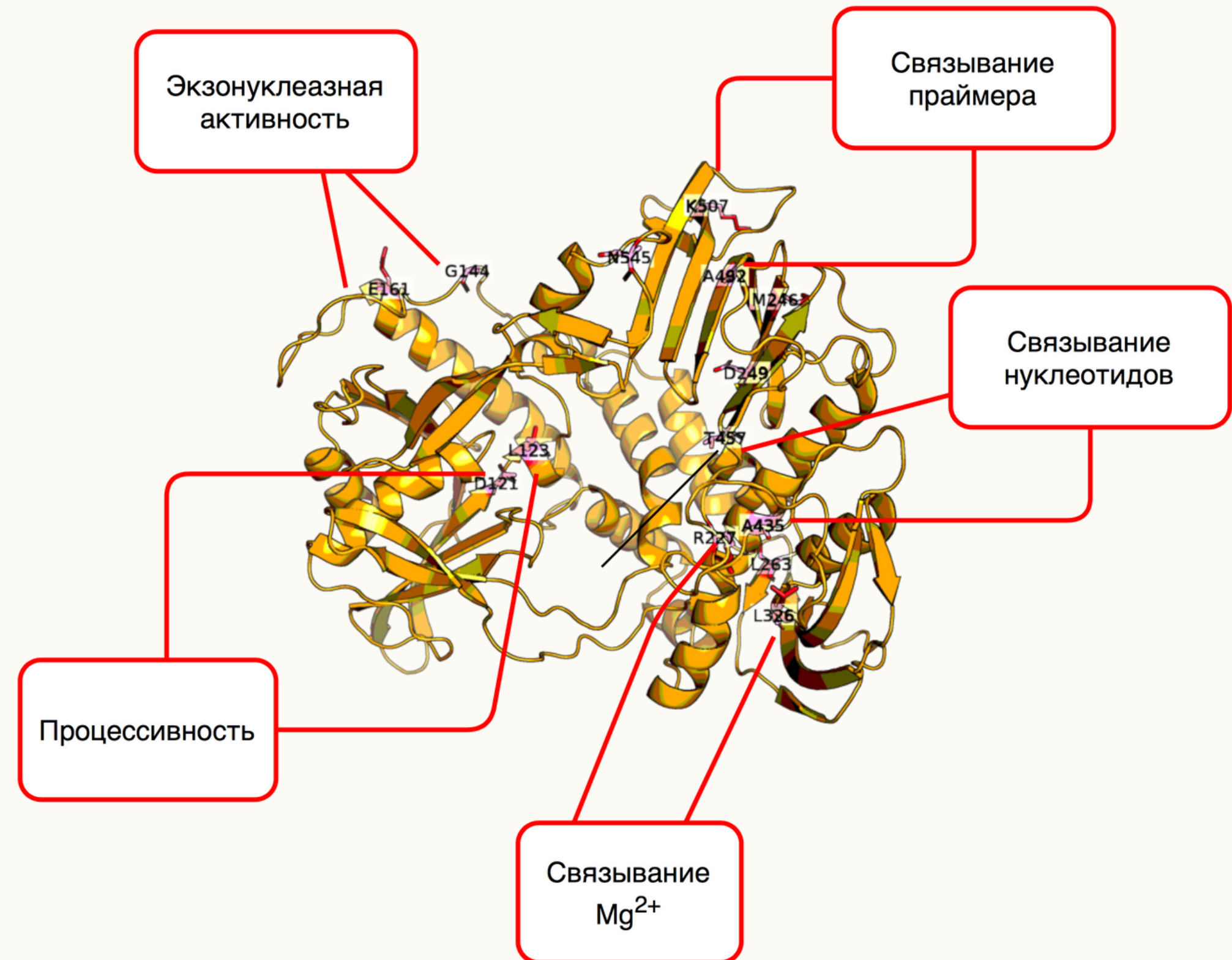


КАЧЕСТВО МОДЕЛИ

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



ЗАПИСЕЙ UNIPROT

Было рассмотрено при поиске Ф29-подобных полимераз



БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Отобраны для построения референсного выравнивания и обучения модели CatBoost



ФУНКЦИОНАЛЬНО-ЗНАЧИМЫХ ОСТАТКА В СТРУКТУРАХ ПОЛИМЕРАЗ

Отобраны для модели по литературе и при аннотации по структурному выравниванию



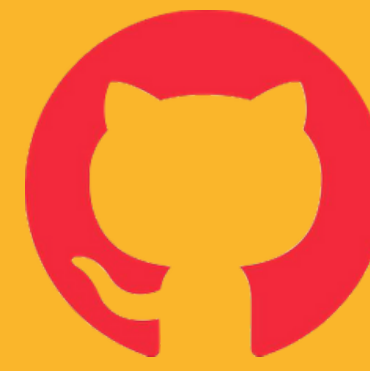
АМИНОКИСЛОТНЫХ ОСТАТКОВ

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



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