LLM модели в биологии

DNABERT, DNABERT-2, GENA-LM, ESM-1v

DNABERT

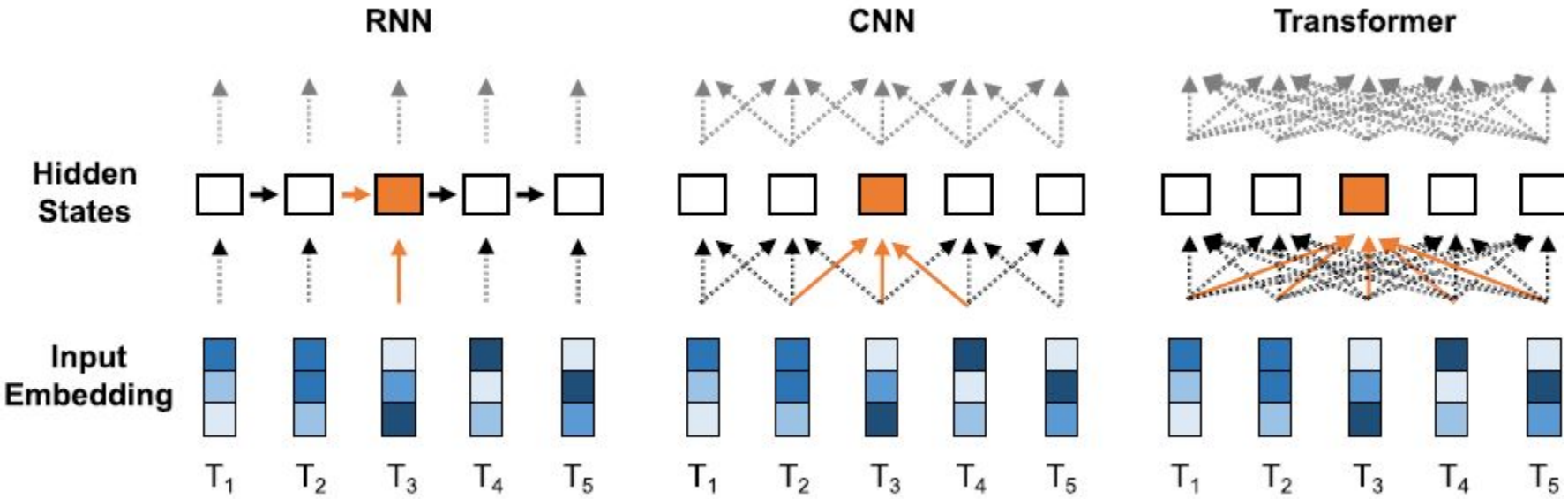
Bioinformatics, 37(15), 2021, 2112–2120
doi: 10.1093/bioinformatics/btab083
Advance Access Publication Date: 4 February 2021
Original Paper

Genome analysis

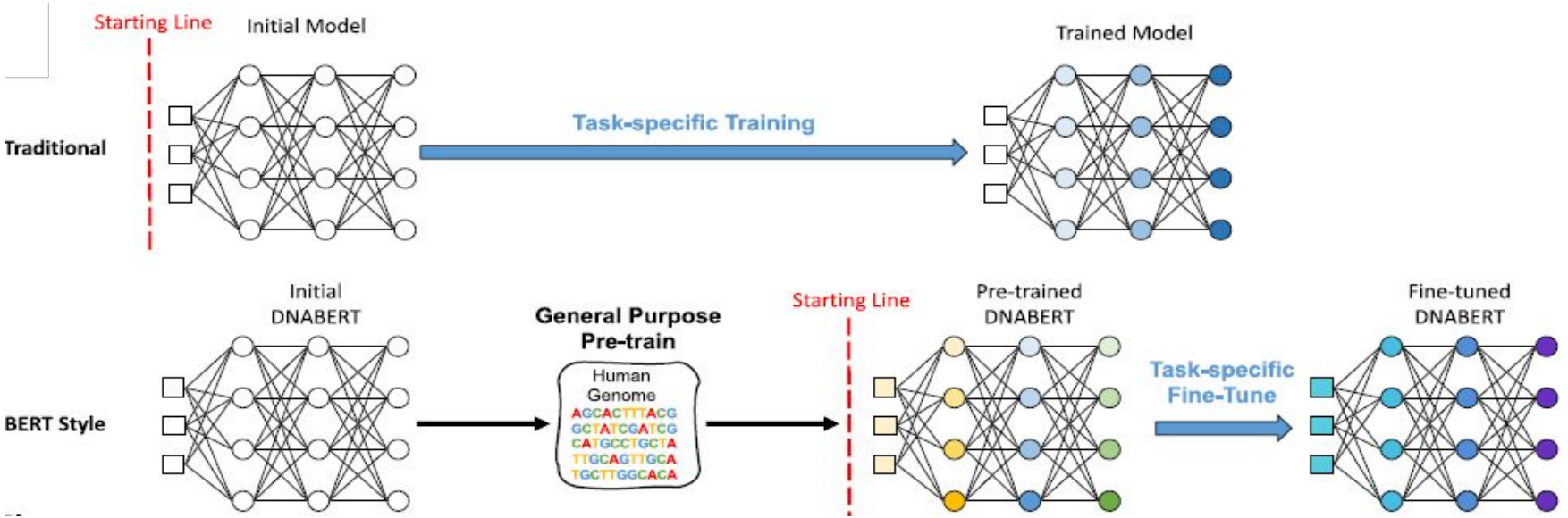
DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

Yanrong Ji[1,†], Zhihan Zhou[2,†], Han Liu[2,*] and Ramana V. Davuluri [3,*]

DNABERT


RNN
CNN
Transformer
Hidden States
Input Embedding
T1
T2
T3
T4
T5

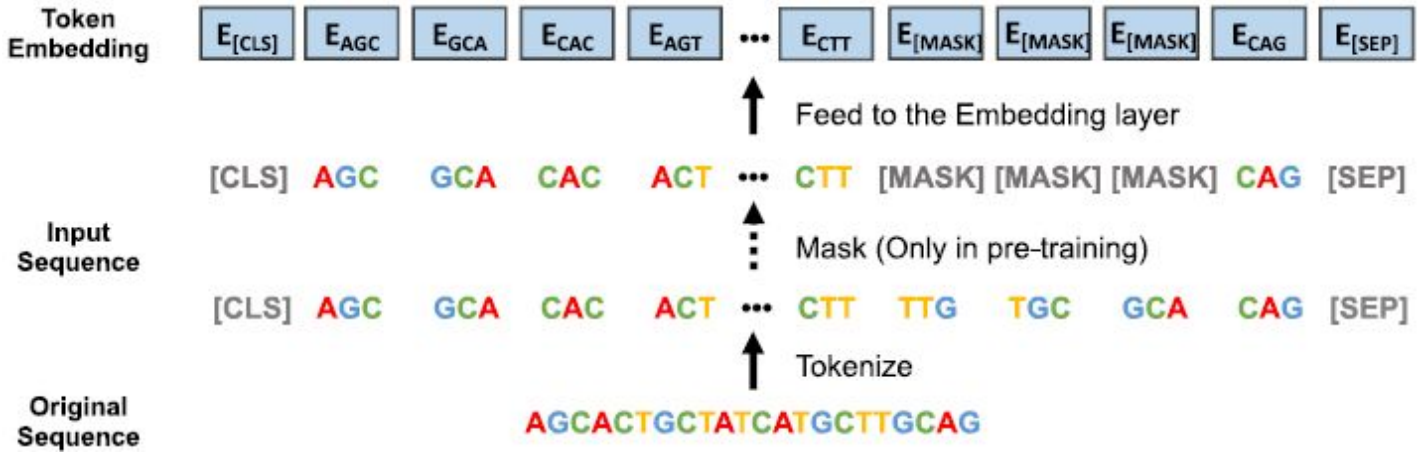
DNABERT


Starting Line
Initial Model
Trained Model
Traditional
Task-specific Training
Initial
DNABERT
General Purpose
Pre-train
Human
Genome
AGCACTTTACG
GCTATCGATCG
CATGCCTGCTA
TTGCAGTTGCA
TGCTTGGCACA
Starting Line
Pre-trained
DNABERT
Fine-tuned
DNABERT
BERT Style
Task-specific
Fine-Tune

DNABERT

Модель обучалась на геноме человека с длиной последовательностей от 5 до 510.

Использовалась k-mer токенизация:



DNABERT

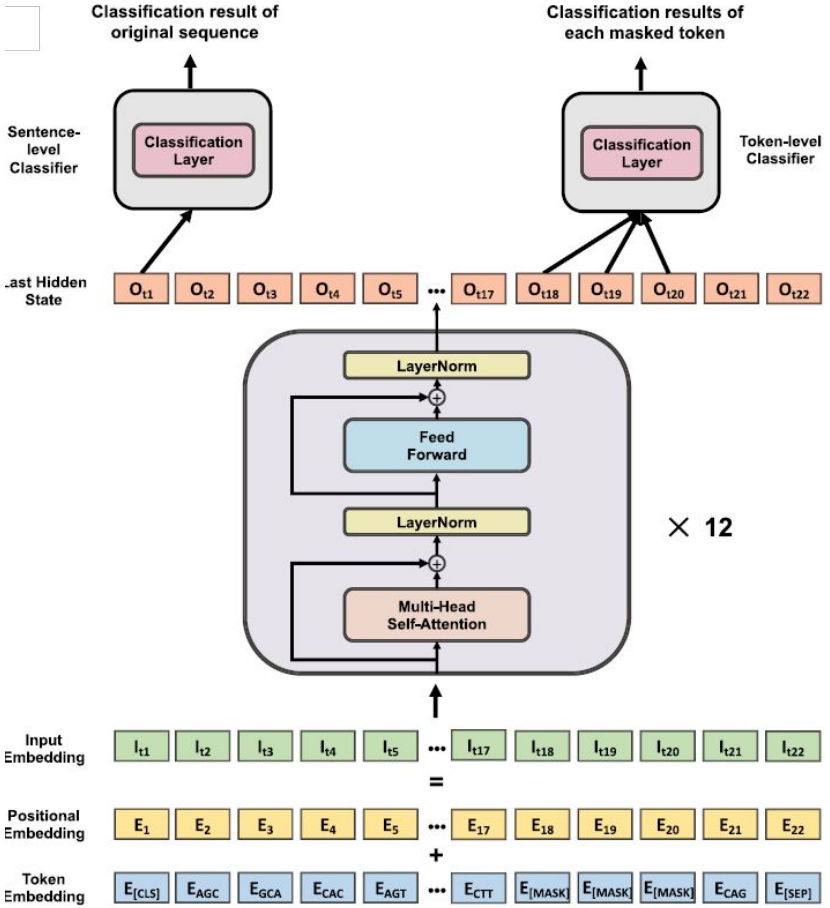
DNABERT captures contextual information by performing the multi-head self-attention mechanism on M:

$$MultiHead(M) = Concat(head_1, \ldots, head_h)W^O$$

where

$$head_i = softmax\left(\frac{MW_i^Q MW_i^K T}{\sqrt{d_k}}\right) \cdot MW_i^V$$

W^O and $W_i^Q, W_i^K, W_i^V \left\{W_i^Q, W_i^K, W_i^V\right\}_{i=0}^{h}$ are learned parameters

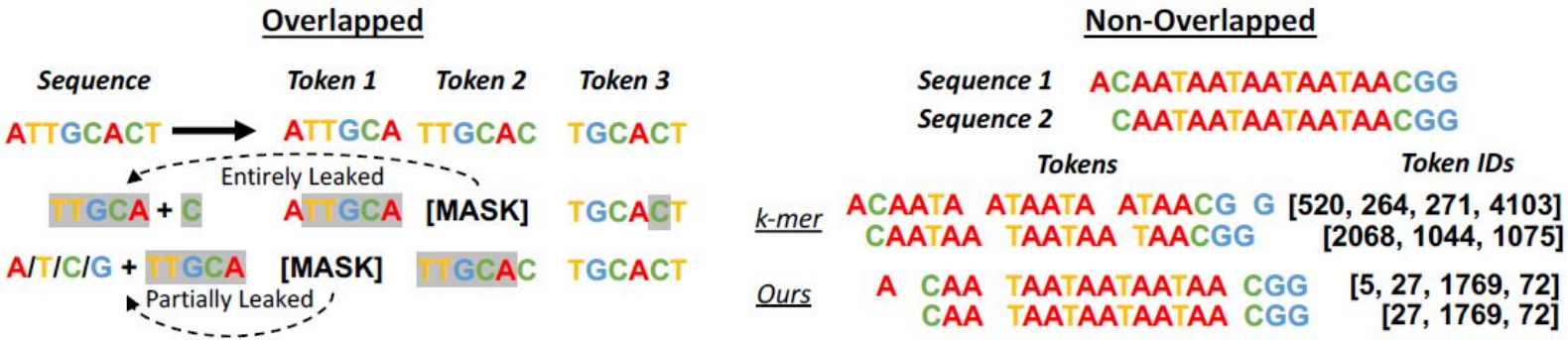


DNABERT-2

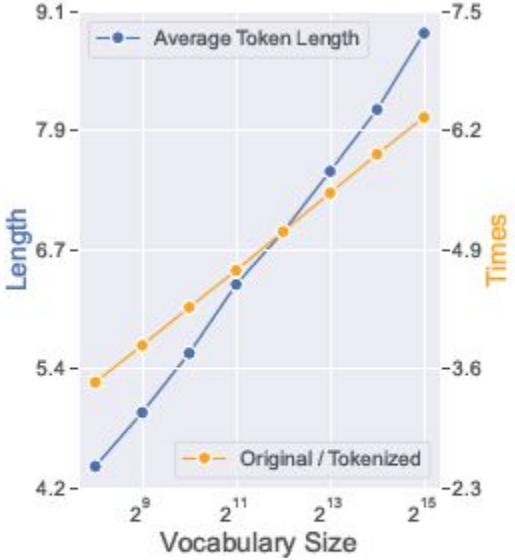
Модель обучалась на геноме 4-х видов с длиной последовательностей от 70 до 10000.

Для токенизации последовательностей ДНК использовался алгоритм Byte Pair Encoding (BPE):

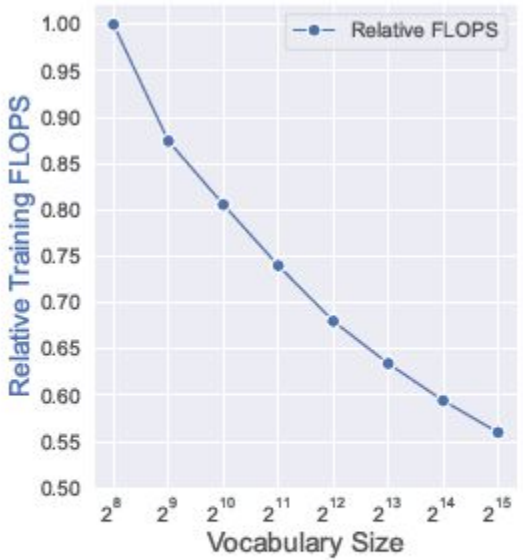
DNABERT-2


Overlapped
Sequence
Token 1
Token 2
Token 3
ATTGCACT
ATTGCA
TTGCAC
TGCACT
Entirely Leaked
TTGCA + C
ATTGCA
[MASK]
TGCACT
A/T/C/G + TTGCA
[MASK]
TTGCAC
TGCACT
Partially Leaked
Non-Overlapped
Sequence 1
ACAATAATAATAATAACGG
Sequence 2
CAATAATAATAATAACGG
Tokens
Token IDs
k-mer
ACAATA ATAATA ATAACG G
[520, 264, 271, 4103]
CAATAA TAATAA TAACGG
[2068, 1044, 1075]
Ours
A CAA TAATAATAATAA CGG
[5, 27, 1769, 72]
CAA TAATAATAATAA CGG
[27, 1769, 72]

DNABERT-2



(a) Average token length and the length ratio of original sequence v.s. tokenized sequence.



(b) Training FLOPs on 500-length sequences compared to model with 2^8 vocabulary.

(c) Model performance averaged over each tasks (macro) and individual dataset (micro).

DNABERT-2, низкоранговая адаптация (LoRA)

Пусть $W_0, W_1 \in \mathbb{R}^{m \times n}$ определяют одну и ту же весовую матрицу до и после тонкой настройки для конкретной задачи

Представим W_1 как $W_1 = W_0 + \Delta W$ где $\Delta W \in \mathbb{R}^{m \times n}$

А ΔW как перемножение 2-х низкоранговых матриц:

$\Delta W = BA$, где $B \in \mathbb{R}^{m \times r}, A \in \mathbb{R}^{r \times n}$ и $r \ll m, r \ll n$

Низкоранговая декомпозиция уменьшает количество обучаемых параметров с **m × n** до **r × (m + n)**

DNABERT-2 vs DNABERT/Nucleotide Transformer

Статистика и производительность моделей:

Model	Params. ↓	FLOPs ↓	Trn. Tokens	Num. Top-2 ↑	Ave. Scores ↑
DNABERT (3-mer)	86M	3.27	122B	2 || 0	61.62
DNABERT (4-mer)	86M	3.26	122B	0 || 1	61.14
DNABERT (5-mer)	87M	3.26	122B	0 || 1	60.05
DNABERT (6-mer)	89M	3.25	122B	0 || 1	60.51
NT-500M-human	480M	3.19	50B	0 || 0	55.43
NT-500M-1000g	480M	3.19	50B	0 || 1	58.23
NT-2500M-1000g	2537M	19.44	300B	0 || 1	61.41
NT-2500M-multi	2537M	19.44	300B	7 || 9	66.93
DNABERT-2	117M	1.00	262B	8 || 4	66.80
DNABERT-2♦	117M	1.00	263B	**11** || **10**	**67.77**

В пяти столбцах представлено количество параметров модели, относительные FLOP по сравнению с DNABERT-2, количество токенов, использованных при предварительном обучении, количество попаданий в топ-2 среди всех моделей (1-й // 2-й), а также средние оценки на 28 наборах данных теста GUE.
♦ предварительное обучение выполнено на обучающих наборах эталонного теста GUE.

NT-500M-human, NT-500M-1000g, NT-2500M-1000g и NT-2500M-multi, где human, 1000g и multi соответственно относятся к эталонному геному человека GRCh38/hg38, набор из 3202 геномов человека с высоким охватом из проекта «1000 геномов» (Byrska-Bishop et al., 2021) и геномы 850 различных видов.

GENA-LM

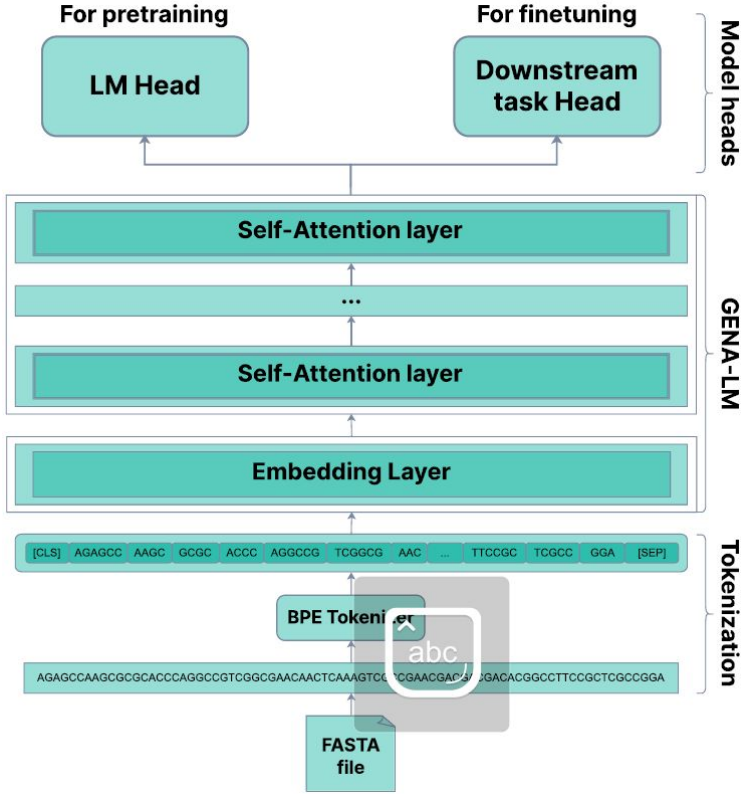
GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences

Veniamin Fishman[1,2*†], Yuri Kuratov[1,3†], Maxim Petrov[1], Aleksei Shmelev[1,4], Denis Shepelin[1], Nikolay Chekanov[1], Olga Kardymon[1,4*], Mikhail Burtsev[5*]

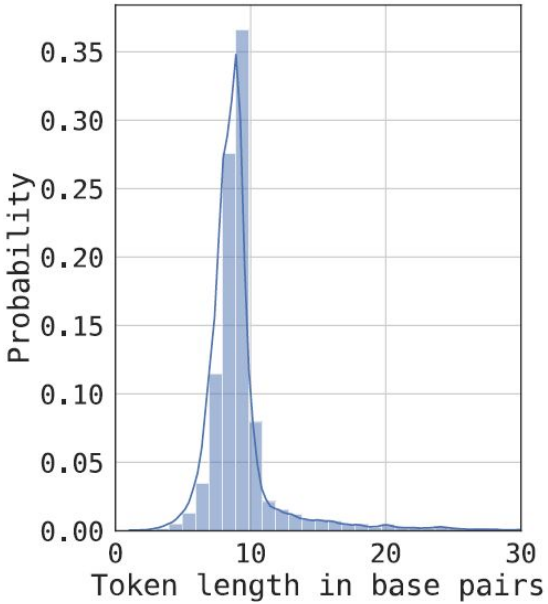
[1]AIRI, Moscow, Russia.
[2]Institute of Cytology and Genetics, Novosibirsk, Russia.
[3]Moscow Institute of Physics and Technology, Dolgoprudny, Russia.
[4]HSE University, Moscow, Russia.
[5]London Institute for Mathematical Sciences, London, UK.

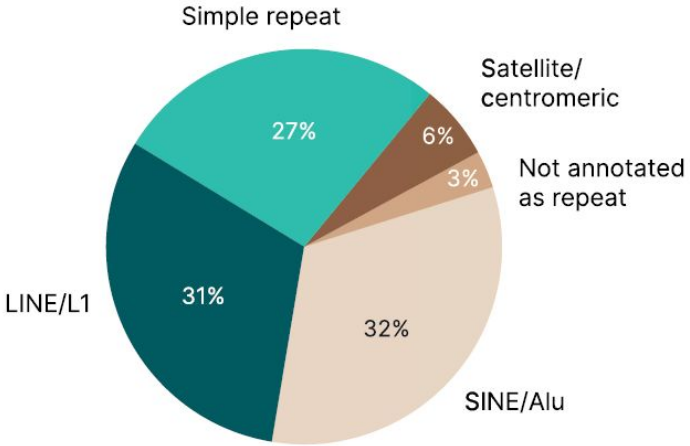
GENA-LM

“В этой работе мы демонстрируем успешное применение передовых нейронных сетей на основе трансформеров для прогнозного анализа различных функциональных геномных элементов в последовательностях ДНК, включая активность промоутера, сплайсинг, сайты полиаденилирования, аннотации энхансеров и профили хроматина. Мы вносим свой вклад в исследовательское сообщество, представляя GENA-LM, семейство моделей с открытым исходным кодом, доступных на GitHub, и предварительно обученные модели (с префиксом gena-lm-) на https://huggingface.co/AIRI-Institute. Мы эмпирически показываем, что точная настройка наших моделей превосходит результаты, полученные с помощью современных архитектур. Более того, мы доказываем, что дополнение GENA-LM архитектурой трансформера рекуррентной памяти (RMT) способствует увеличению длины входной последовательности, увеличивая качество решения сложных биологических задач.”

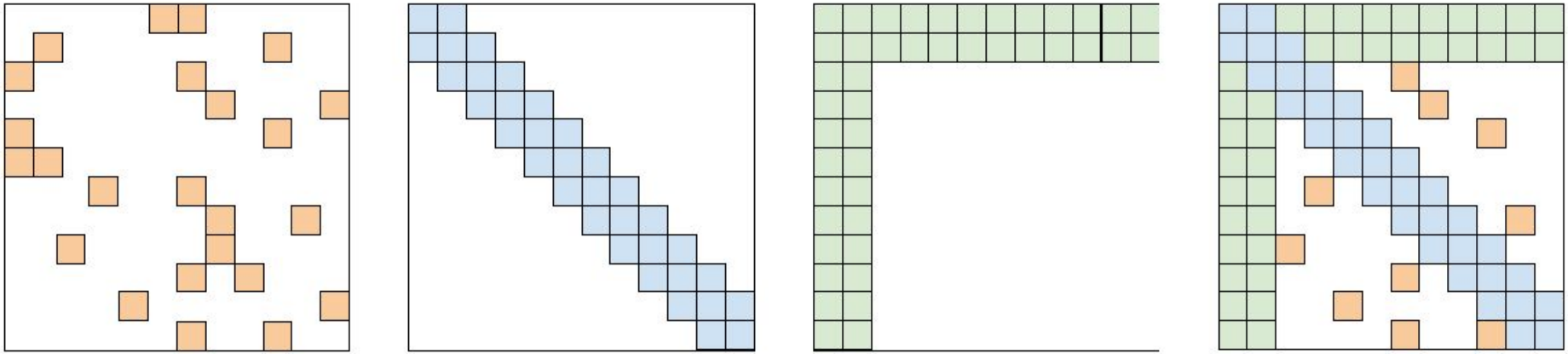


GENA-LM


Probability
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
Token length in base pairs


Bases annotated as repeats among top 100 longest tokens
Simple repeat
27%
Satellite/
centromeric
6%
3%
Not annotated
as repeat
LINE/L1
31%
32%
SINE/Alu

GENA-LM



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

Figure 1: Building blocks of the attention mechanism used in BIGBIRD. White color indicates absence of attention. (a) random attention with $r = 2$, (b) sliding window attention with $w = 3$ (c) global attention with $g = 2$. (d) the combined BIGBIRD model.

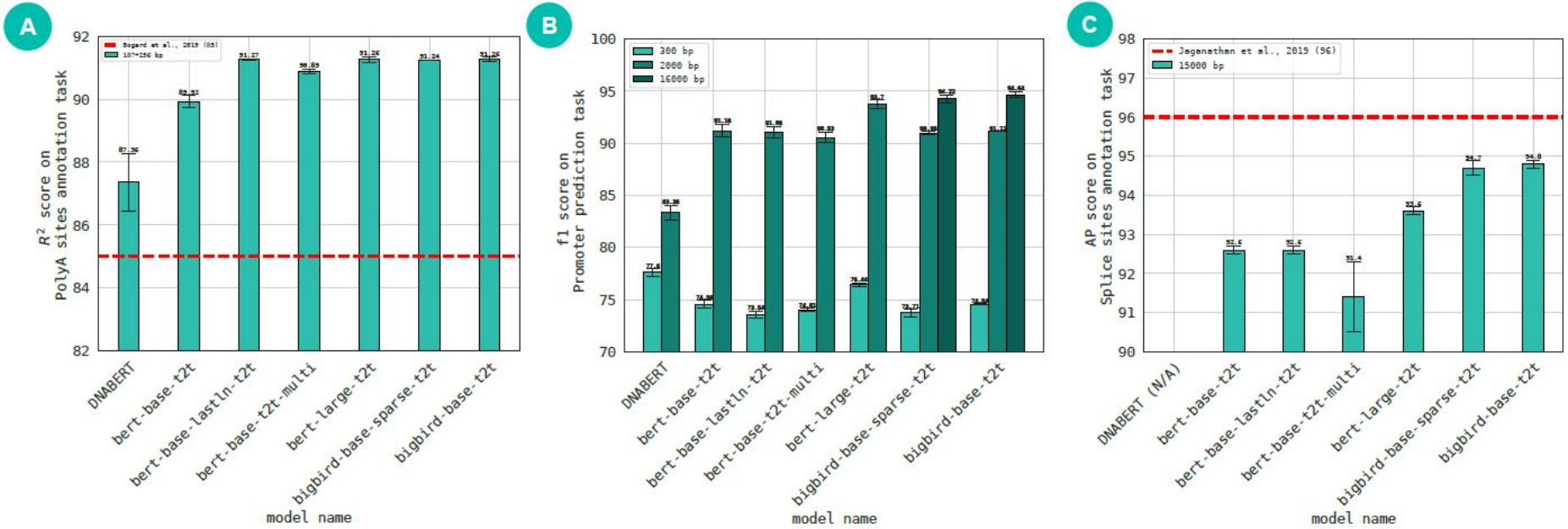
GENA-LM

Model	Architecture	Maximum seq len, tokens (≈ bp)	Tokenizer data	Training data
DNABERT	BERT-12L	512 (512)	3,4,5,6-mer	GRCh38.p13
GENA-LM models:				
bert-base	BERT-12L	512 (4,500)	T2T split v1	T2T split v1
bert-base-t2t	BERT-12L	512 (4,500)	T2T+1KG+M	T2T+1KG
bert-base-lastln-t2t	BERT-12L	512 (4,500)	T2T+1KG+M	T2T+1KG
bert-base-t2t-multi	BERT-12L	512 (4,500)	T2T+1KG+M	T2T+1KG+M
bert-large-t2t	BERT-24L	512 (4,500)	T2T+1KG+M	T2T+1KG
bigbird-base-sparse	BERT-12L, RoPE DS Sparse Att	4,096 (36,000)	T2T split v1	T2T split v1
bigbird-base-sparse-t2t	BERT-12L, RoPE DS Sparse Att	4,096 (36,000)	T2T+1KG+M	T2T+1KG
bigbird-base-t2t	BERT-12L HF Sparse Attention	4,096 (36,000)	T2T+1KG+M	T2T+1KG

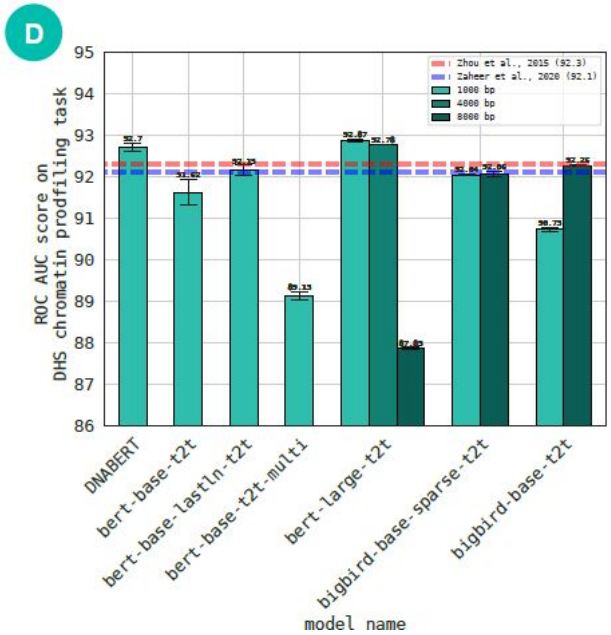
Характеристики моделей GENA-LM. Выделены различия в данных предварительного обучения, количестве слоев, типе attention и длине последовательности. «T2T-сплит v1» - словарь токенов построен на разбиении сборки генома человека T2T, «1KG» является сокращением 1000G, «M» - включение данных нескольких видов.

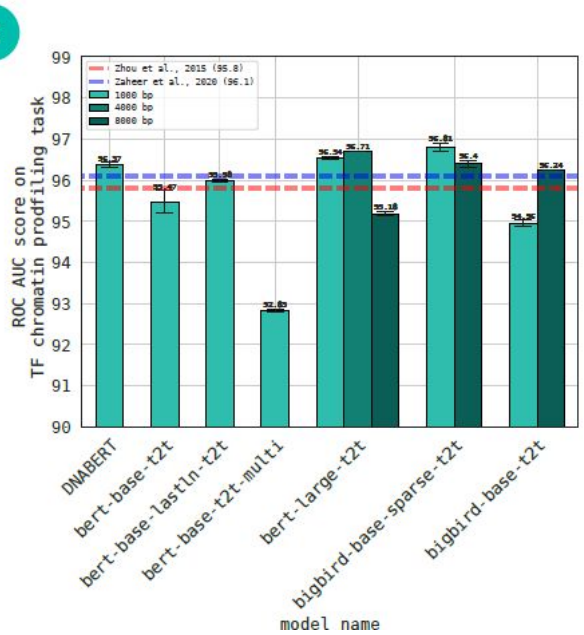
Обозначения «DS Sparse» и «HF Sparse» относятся к реализациям DeepSpeed разреженного внимания и HuggingFace BigBird соответственно. Аббревиатура «RoPE» означает использование rotary position embeddings в качестве альтернативы абсолютного PE в BERT. Модели были структурированы либо из 12 (обозначаемых BERT-12L), либо из 24 (обозначаемых BERT-24L) слоев, содержащих 110M и 336M параметров соответственно.

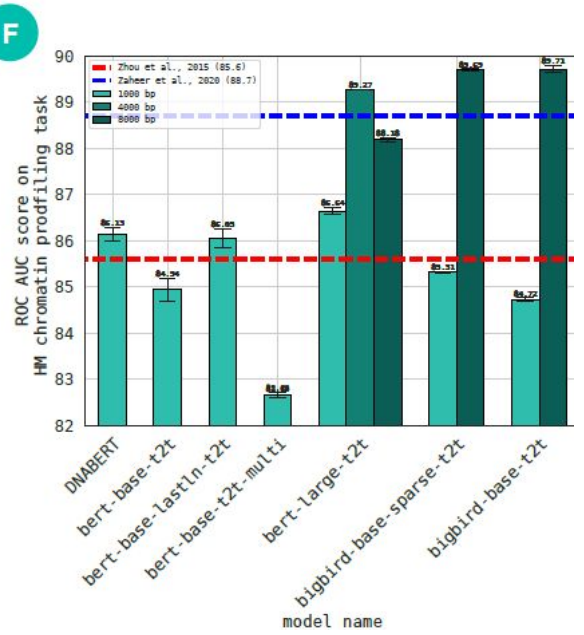
GENA-LM


A
B
C
R^2 score on PolyA sites annotation task
f1 score on Promoter prediction task
AP score on Splice sites annotation task
model name
DNABERT
bert-base-t2t
bert-base-lastln-t2t
bert-base-t2t-multi
bert-large-t2t
bigbird-base-sparse-t2t
bigbird-base-t2t
300 bp
2000 bp
16000 bp
Jaganathan et al., 2019 (96)
15000 bp
DNABERT (N/A)

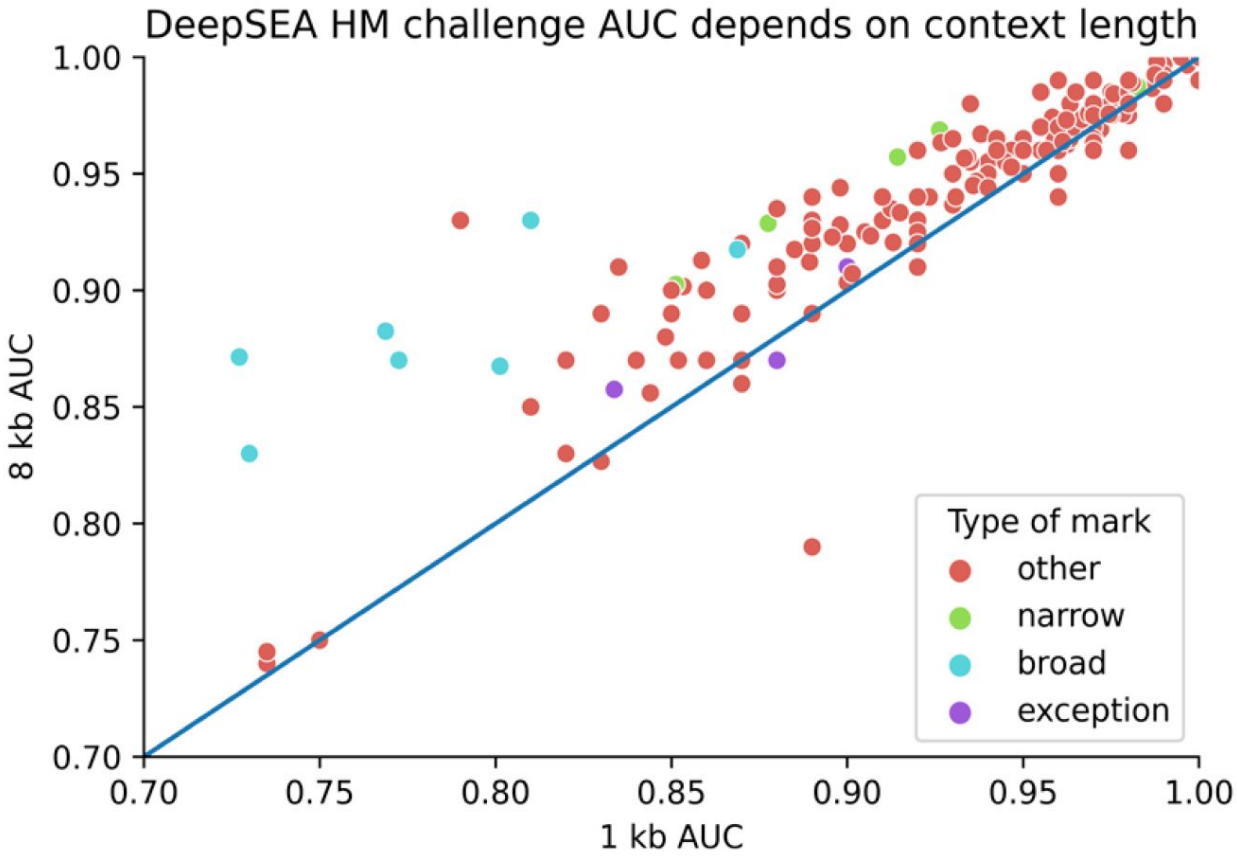
GENA-LM


D
ROC AUC score on DHS chromatin prodfiling task
Zhou et al., 2015 (92.3)
Zaheer et al., 2020 (92.1)
1000 bp
4000 bp
8000 bp
DNABERT
bert-base-t2t
bert-base-lastln-t2t
bert-base-t2t-multi
bert-large-t2t
bigbird-base-sparse-t2t
bigbird-base-t2t
model name

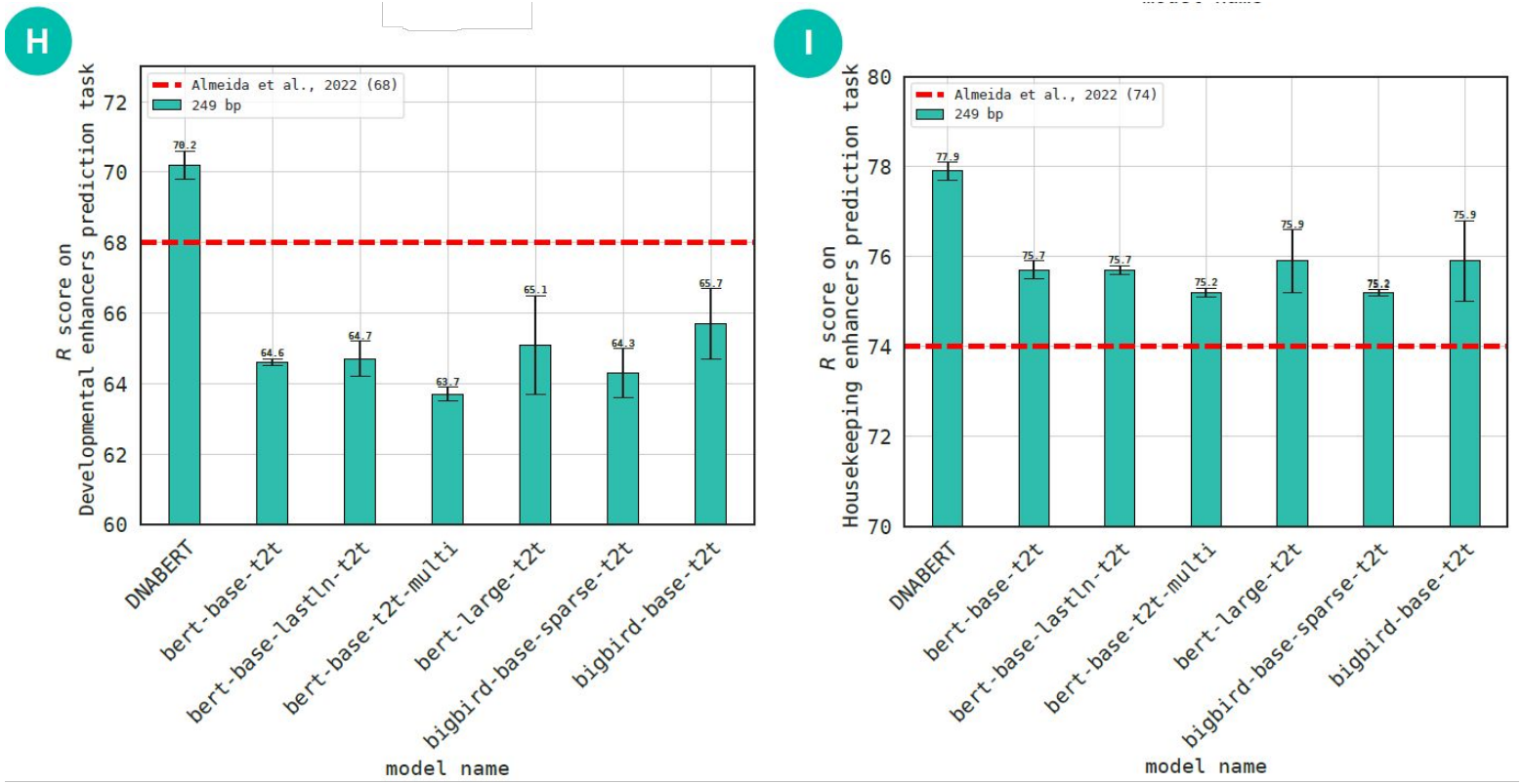

E
ROC AUC score on TF chromatin prodfiling task
Zhou et al., 2015 (95.8)
Zaheer et al., 2020 (96.1)
1000 bp
4000 bp
8000 bp
DNABERT
bert-base-t2t
bert-base-lastln-t2t
bert-base-t2t-multi
bert-large-t2t
bigbird-base-sparse-t2t
bigbird-base-t2t
model name


F
ROC AUC score on HM chromatin prodfiling task
Zhou et al., 2015 (85.6)
Zaheer et al., 2020 (88.7)
1000 bp
4000 bp
8000 bp
DNABERT
bert-base-t2t
bert-base-lastln-t2t
bert-base-t2t-multi
bert-large-t2t
bigbird-base-sparse-t2t
bigbird-base-t2t
model name

GENA-LM

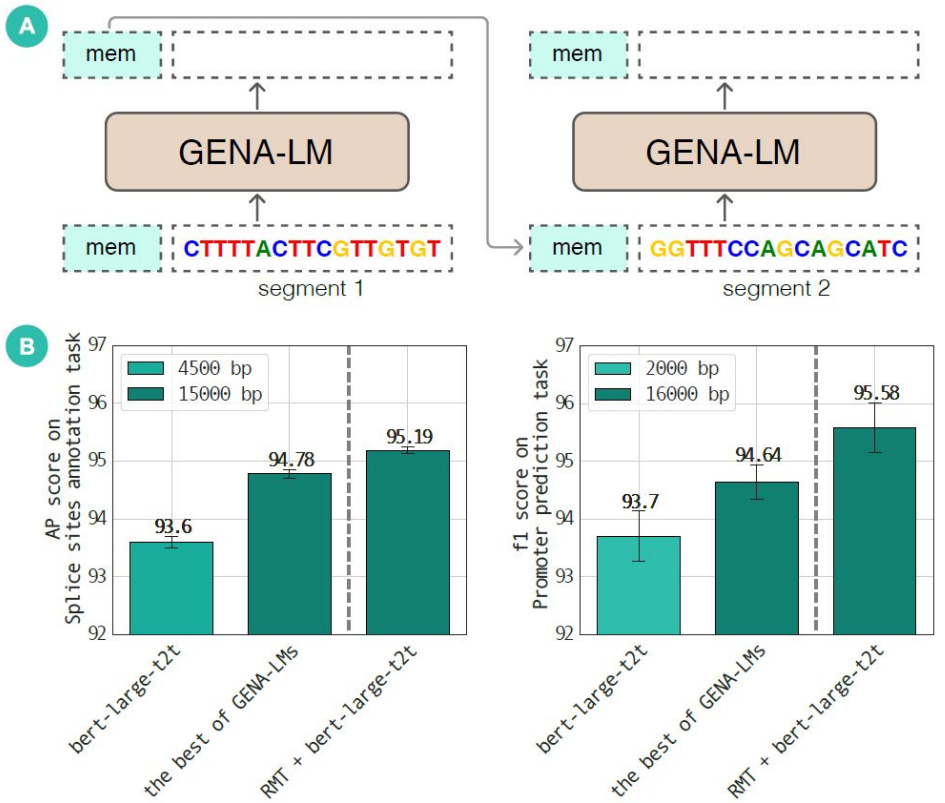

DeepSEA HM challenge AUC depends on context length
8 kb AUC
1 kb AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Type of mark
other
narrow
broad
exception

GENA-LM


H
Almeida et al., 2022 (68)
249 bp
R score on Developmental enhancers prediction task
70.2
64.6
64.7
63.7
65.1
64.3
65.7
DNABERT
bert-base-t2t
bert-base-lastln-t2t
bert-base-t2t-multi
bert-large-t2t
bigbird-base-sparse-t2t
bigbird-base-t2t
model name
I
Almeida et al., 2022 (74)
249 bp
R score on Housekeeping enhancers prediction task
77.9
75.7
75.7
75.2
75.9
75.2
75.9
DNABERT
bert-base-t2t
bert-base-lastln-t2t
bert-base-t2t-multi
bert-large-t2t
bigbird-base-sparse-t2t
bigbird-base-t2t
model name

GENA-LM

Токены памяти добавляются в каждый сегмент для передачи информации между последовательными сегментами, что позволяет им использовать информацию из всех предыдущих сегментов. Таким образом, весь предварительно обученный Трансформер эффективно функционирует как единая рекуррентная единица.

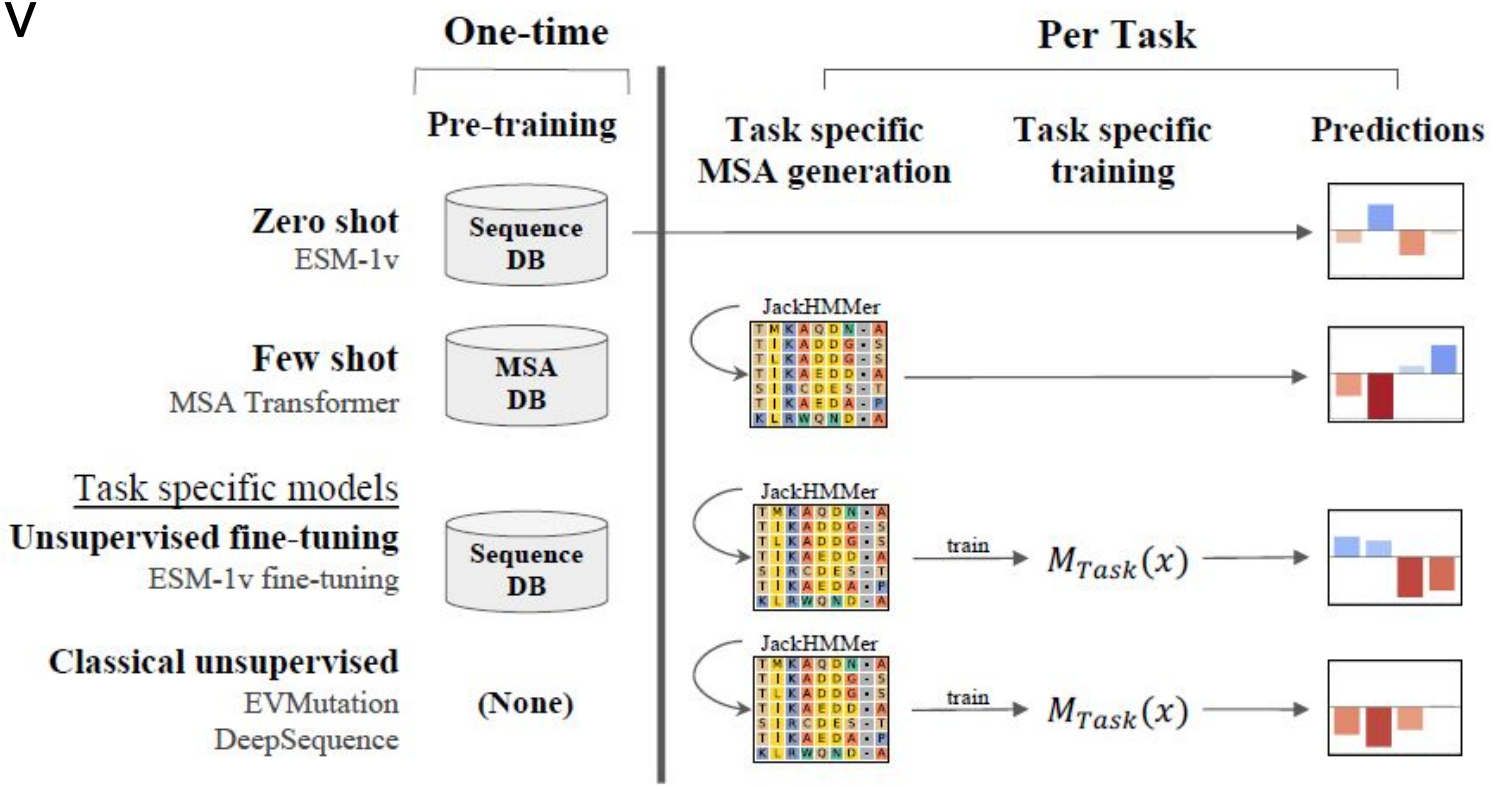


ESM-1v

Language models enable zero-shot prediction of the effects of mutations on protein function

Joshua Meier [1 2] Roshan Rao [3] Robert Verkuil [1] Jason Liu [1]
Tom Sercu [1] Alexander Rives [1 2]

ESM-1v


One-time
Per Task
Pre-training
Task specific MSA generation
Task specific training
Predictions
Zero shot
ESM-1v
Sequence DB
Few shot
MSA Transformer
MSA DB
Task specific models
Unsupervised fine-tuning
ESM-1v fine-tuning
Sequence DB
Classical unsupervised
EVMutation
DeepSequence
(None)
JackHMMer
JackHMMer
train
$M_{Task}(x)$
JackHMMer
train
$M_{Task}(x)$

ESM-1v

Мы оцениваем мутации, используя логарифмическое отношение вероятностей в мутированной позиции, предполагая аддитивную модель, когда несколько мутаций T существуют в одной и той же последовательности:

$$\sum_{t \in T} \log p(x_t = x_t^{mt} | x_{\setminus T}) - \log p(x_t = x_t^{wt} | x_{\setminus T})$$